

Causal models for dynamical systems

Jonas Peters
University of Copenhagen, Denmark
jonas.peters@math.ku.dk

Stefan Bauer
MPI Tübingen, Germany
stefan.bauer@tuebingen.mpg.de

Niklas Pfister
University of Copenhagen, Denmark
np@math.ku.dk

January 20, 2020

Abstract

A probabilistic model describes a system in its observational state. In many situations, however, we are interested in the system's response under interventions. The class of structural causal models provides a language that allows us to model the behaviour under interventions. It can be taken as a starting point to answer a plethora of causal questions, including the identification of causal effects or causal structure learning. In this chapter, we provide a natural and straight-forward extension of this concept to dynamical systems, focusing on continuous time models. In particular, we introduce two types of causal kinetic models that differ in how the randomness enters into the model: it may either be considered as observational noise or as systematic driving noise. In both cases, we define interventions and therefore provide a possible starting point for causal inference. In this sense, the book chapter provides more questions than answers. The focus of the proposed causal kinetic models lies on the dynamics themselves rather than corresponding stationary distributions, for example. We believe that this is beneficial when the aim is to model the full time evolution of the system and data are measured at different time points. Under this focus, it is natural to consider interventions in the differential equations themselves.

We wholeheartedly congratulate Judea Pearl on winning the Turing Award. His groundbreaking work has inspired much of our work, with this book chapter being only one of several examples.

1 Introduction

In causality, we aim to understand how a system reacts under interventions, e.g., in gene knock-out experiments. There are different interventions we can perform (including none at all), and we therefore require different descriptions of the data generating process. Some systems may be adequately described by deterministic equations, but if the system possesses observational noise, unobserved factors or intrinsic randomness, data generating processes are more appropriately modeled using the language of probability. In data-driven sciences, we are used to model the data generating process with a single probability distribution, e.g., using a multivariate Gaussian with a certain covariance matrix. As argued above, however, causal models come with a plethora of distributions: one distribution for each type of modeled intervention.

In general, the intervention distributions are not arbitrarily different as it would be meaningless to talk about a single underlying system otherwise. It is a key challenge to describe which parts of the distribution change and which parts remain invariant when considering different interventions. Many researchers from various disciplines engaged in this question and developed the fundamental assumptions that are often referred to as invariance, autonomy or modularity [Wright, 1921, Haavelmo, 1944, Aldrich, 1989, Hoover, 1990, Imbens and Rubin, 2015, Richardson and Robins, 2013]. The concept of invariance relies heavily on what it means to intervene on a system, making a precise formulation of interventions

crucial for causal modeling. Arguably one of the clearest formulation of interventions is Judea Pearl’s *do*-formalism [Pearl, 2009, Chapter 1]. One starts with a fixed reference distribution called the observational distribution; one may think of it as describing the system in its natural state with no intervention being performed. The system and its corresponding distribution is assumed to have a modular structure, and performing a *do*-intervention means changing some of the modules. This process yields an intervention distribution, often denoted by a *do*(.) subscript. While this description can be made formal in various ways, we focus on one that is based on structural causal models [Wright, 1921, Bollen, 1989, Pearl, 2009]. Usually, the formulation of structural causal models, or SCMs, include random variables. We believe, however, that the descriptive power of SCMs lies in their modular structure, which can be separated from randomness. We therefore introduce two different versions of SCMs: a deterministic version with measurement noise and a version containing random variables.

1.1 Structural causal models with measurement noise

A *deterministic structural causal model* (SCM) over d variables x_1, \dots, x_d is a collection of d assignments

$$x^k := f^k(x^{\mathbf{PA}_k}), \quad k = 1, \dots, d, \quad (1)$$

where for any $k \in \{1, \dots, d\}$, $\mathbf{PA}_k \subseteq \{1, \dots, d\} \setminus \{k\}$ is called the set of direct parents of x^k , and f^k is a real-valued function. If $\mathbf{PA}_k = \emptyset$, then $f^k(x^{\mathbf{PA}_k})$ should be interpreted as a constant. For each SCM, we obtain a corresponding graphical representation of the causal structure over the vertices¹ $(1, \dots, d)$ by drawing directed edges from \mathbf{PA}_k to k for all $k \in \{1, \dots, d\}$. We further assume that the system (1) is uniquely solvable, which may be the case, even if the graph contains directed cycles, such as $3 \rightarrow 1 \rightarrow 4 \rightarrow 3$. Each SCM then induces a state of the system characterized by a single point in \mathbb{R}^d . We will see in Section 1.3 that the modular structure of (1) is key to the ability to serve as a causal model. The assignments in (1) can be thought of as lines in a computer program that generate a specific state of the system. Interventions will be modeled as replacements of some of these lines.

We may now assume to obtain noisy observations of the system, e.g., for each $k \in \{1, \dots, d\}$, we may have

$$X^k := x^k + \varepsilon^k, \quad (2)$$

where $\varepsilon^1, \dots, \varepsilon^d$ are jointly independent random variables. Instead of a single point, this model now induces a joint distribution over the observed random variables X^1, \dots, X^d .

1.2 Structural causal models with driving noise

More common than the above approach is the assumption that the randomness enters inside the structural assignments. Formally, a *stochastic structural causal model* over d random variables X_1, \dots, X_d is a collection of d assignments

$$X^k := f^k(X^{\mathbf{PA}_k}, \varepsilon^k), \quad k = 1, \dots, d, \quad (3)$$

together with a distribution over the noise variables $\varepsilon^1, \dots, \varepsilon^d$. As above, we obtain a corresponding graphical representation of the causal structure over the vertices $(1, \dots, d)$ by drawing directed edges from \mathbf{PA}_k to k for all $k \in \{1, \dots, d\}$. We further assume that the joint noise distribution is absolutely continuous with respect to a product measure and that it factorizes, i.e., the noise components are assumed to be jointly independent. As before, we require the system (3) to be uniquely solvable, which is always satisfied if the graph is acyclic, for example. An SCM induces a unique joint distribution over the variables X_1, \dots, X_d [e.g., Bongers et al., 2016], and an observed data set may be modeled as a collection of i.i.d. realizations from that distribution.

The two approaches described above serve different purposes. The model described in (2) might be helpful when the underlying system is assumed to be deterministic and all randomness can be thought of as measurement noise, for example. While this might be a realistic assumption in many applications, the approach comes with various statistical difficulties, including the famous errors-in-variables problem [Carroll et al., 2006] and an increased difficulty when identifying parameters or causal structure from data [Zhang et al., 2018]. We speculate that this is one of the reasons, why less work seems to be

¹By slight abuse of notation, we identify (x^1, \dots, x^d) with its indices $(1, \dots, d)$.

devoted to the first approach. The second approach is better understood but assumes that the noise is not purely measurement noise, but enters into the causal mechanism. It depends on the application at hand, whether this assumption is reasonable.

1.3 Interventions

SCMs allow us to define *interventions*. For any $j \in \{1, \dots, d\}$, we can replace one of the assignments in (1) or (3). In the former case, we could replace the assignment with $x^j := \tilde{f}^j(x^{\overline{\mathbf{PA}}_j})$ and in the latter case with $X^j := \tilde{f}^j(X^{\overline{\mathbf{PA}}_j}, \tilde{\varepsilon}^j)$, for example. Usually, we restrict ourselves to interventions that yield a new SCM, so the interventions must respect unique solvability. If that is the case, the intervention induces a new state of the system, that we denote by $do(x^j := \tilde{f}^j(x^{\overline{\mathbf{PA}}_j}))$ or $do(X^j := \tilde{f}^j(X^{\overline{\mathbf{PA}}_j}, \tilde{\varepsilon}^j))$, respectively. An intervention on one of the variables propagates through the system, possibly affecting many other variables that are graphical descendants of the targeted node. For the stochastic SCMs from Section 1.2, one may think about randomized experiments as a *do*-intervention and the well-known hard (or point) interventions $do(X^j := 4)$ appear as a special case. Pearl [2009] provides many insightful examples of SCMs and interventions throughout his book. Bongers et al. [2016] give measure theoretic details underlying the construction of SCMs. Below, we extend the concept of SCMs to dynamical systems and give a concrete example of an SCM, a graph, and interventions in that context, see Figure 2.

The above definition clarifies which parts of the distribution remain invariant under interventions. In the case of Section 1.2, each conditional distribution X^k , given $X^{\mathbf{PA}^k} = x$, is determined by the structural assignment for X^k . Thus, two distributions induce the same conditionals $X^k | X^{\mathbf{PA}^k} = x$ if one of the distributions is induced by an SCM and the other one is induced by the same SCM after intervening on a fixed $j \neq k$, for example.

It may further be instructive to think about equivalence of two causal models. They may be called observationally equivalent if they induce the same observational distribution and interventionally equivalent if they induce the same observational distribution as well as the same intervention distributions [e.g., Peters et al., 2017, Section 6.8]. One of the fundamental problems when learning causal structures from data is that two causal models may be observationally equivalent, but not interventionally equivalent.

1.4 Time dependent data

In many practical applications, an i.i.d. data set does not provide an adequate description for the data sample at hand. In particular, the concepts above are lacking the notion of time. Different causal methodology and several extensions of SCMs have been proposed [Wiener, 1956, Granger, 1969, Schreiber, 2000, White and Lu, 2010, Hyttinen et al., 2013, Peters et al., 2013, Pfister et al., 2018b], mostly considering discrete time models such as vector autoregressive models [Lütkepohl, 2007], for example. Peters et al. [2009], Bauer et al. [2016] discuss the relation between causality and the arrow of time. Causal inference for longitudinal studies has been studied extensively, too [e.g. Vanderweele, 2015, Aalen et al., 2008, Robins, 1997], where the results are often formulated in the language of potential outcomes [Imbens and Rubin, 2015], rather than SCMs. In this article, we focus on continuous time systems that are governed by ordinary differential equations. In particular, we propose a natural and straight-forward extension of the notion of SCM to dynamical systems. The construction closely follows the existing ideas of SCMs and interventions. Similar constructions have been suggested elsewhere, and we try our best to provide the relevant references and point out existing differences. Parts of this book chapter are taken from Pfister et al. [2018a, 2019], where we focus on model selection and parameter inference.

2 Chemical reaction networks and ordinary differential equations

In many natural sciences and even some social sciences, there are processes that can be modeled by a set of governing differential equations. Examples are found in diverse areas such as bioprocessing [e.g. Ogunnaike and Ray, 1994], economics [e.g. Zhang, 2005], genetics [e.g. Chen et al., 1999], neuroscience [e.g. Friston et al., 2003] or robotics [e.g. Murray, 2017]. Below, we provide two examples that come from a subclass of dynamical models, namely those that are driven by chemical reactions and connect

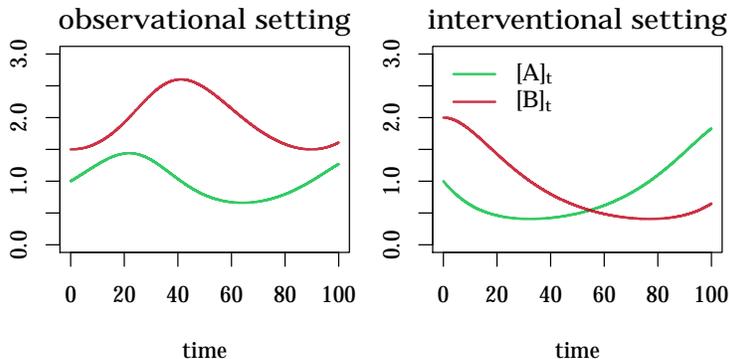
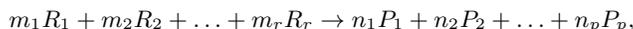


Figure 1. Example trajectories for the basic Lotka-Volterra model given in (7) and (8). Left plot correspond to observation setting with rates $k_1 = 0.1$, $k_2 = 0.05$ and $k_3 = 0.05$ and initial values $[A]_0 = 1$ and $[B]_0 = 1.5$ and right plot to an intervention where we set $k_1 = 0.05$ and $[B]_0 = 2$.

to ordinary differential equation (ODE) based models by mass action kinetics. The general principles, however, can readily be extended to more complex model classes. Formally, a general reaction [e.g. Wilkinson, 2006] takes the form



where r is the number of reactants and p the number of products. Both R_i and P_j can be thought of as molecules and are often called species. The coefficients m_i and n_j are positive integers, called stoichiometries. We now provide two examples: (1) a famous and often used model that describes the abundance of predators and prey, illustrating the law of mass-action kinetics and (2) Michaelis-Menten kinetics which results in nonlinear ODEs.

Lotka-Volterra model The Lotka-Volterra model [Lotka, 1909] takes the form



where A and B describe abundance of prey and predators, respectively. In this model, the prey reproduce by themselves, but the predators require abundance of prey for reproduction. The coefficients k_1 , k_2 , and k_3 indicate the rates, with which the reactions happen.

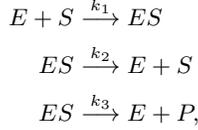
In mass-action kinetics [Waage and Guldberg, 1864], one usually considers the concentration $[X]$ of a species X , the square parentheses indicating that one refers to the concentration rather than to the integer number of abundant species or molecules. The law of mass-action states that the instantaneous rate of each reaction is proportional to the product of each of its reactants raised to the power of its stoichiometry. For the Lotka-Volterra model this yields

$$\frac{d}{dt}[A] = k_1[A] - k_2[A][B] \tag{7}$$

$$\frac{d}{dt}[B] = k_2[A][B] - k_3[B]. \tag{8}$$

Figure 1 shows solutions for these differential equations for both an observational setting (left plot) with rates $k_1 = 0.1$, $k_2 = 0.05$ and $k_3 = 0.05$ and initial values $[A]_0 = 1$ and $[B]_0 = 1.5$ as well as an interventional setting (right plot) where we set $k_1 = 0.05$ and $[B]_0 = 2$. Even though equations (7) and (8) contain interaction terms of the concentration of the different species, they are linear in the model parameters, a property that is exploited by many practical methods.

Michaelis-Menten kinetics In Michaelis-Menten kinetics [Michaelis and Menten, 1913], the starting point is a specific enzyme reaction given by the equations



where the enzyme E binds to a substrate S and finally releases a product P . Under some simplifying assumptions regarding the relation of rates of the reactions, this yields the equation

$$\frac{d}{dt}[P] = c_1 \frac{[S]}{c_2 + [S]}, \quad (9)$$

where c_1, c_2 are constants. There are many reactions that can be described by this model; Michaelis and Menten [1913] used it to describe how the enzyme invertase catalyzes the hydrolysis of sucrose into glucose and fructose.

3 Causal kinetic models

We now define a causal model class for dynamical systems. The reader may think about the example of a Lotka-Volterra model as described in (7) and (8) or Michaelis-Menten kinetics (9), both of which fit into the general framework described below. In analogy to Sections 1.1 and 1.2, we first consider a deterministic version with measurement noise and secondly a version where the randomness enters inside the structural equations.

3.1 Causal kinetic models with measurement noise

We regard the following definition as a natural and straight-forward extension of SCMs, even though we have not seen it in this form before. A *deterministic causal kinetic model* over processes $\mathbf{x} := (\mathbf{x}_t)_t := (x_t^1, \dots, x_t^d)_t$ is a collection of d ODEs and initial value assignments

$$\begin{aligned} \frac{d}{dt}x_t^1 &:= f^1(x_t^{\mathbf{PA}_1}), & x_0^1 &:= \xi_0^1, \\ \frac{d}{dt}x_t^2 &:= f^2(x_t^{\mathbf{PA}_2}), & x_0^2 &:= \xi_0^2, \\ &\vdots & & \\ \frac{d}{dt}x_t^d &:= f^d(x_t^{\mathbf{PA}_d}), & x_0^d &:= \xi_0^d. \end{aligned}$$

Here, for any $k \in \{1, \dots, d\}$, $\frac{d}{dt}x_t^k$ denotes the time derivative of the component x^k at time t and $\mathbf{PA}_k \subseteq \{1, \dots, d\}$ is called the set of direct parents of x^k (and may include x^k itself). We require that the system of initial value problems is uniquely solvable. For each causal kinetic model, we can obtain a corresponding graph over the vertices $(1, \dots, d)$ by drawing edges from \mathbf{PA}_k to k , for $k \in \{1, \dots, d\}$ (see Figure 2). If we consider the initial values as random variables, this induces a distribution over $\mathbf{x} = (\mathbf{x}_t)_t$.

Similarly as in the case of deterministic SCMs, causal kinetic models are deterministic models describing an underlying causal structure. The observed data can then be modeled as noisy observations of the system, i.e.,²

$$\mathbf{X}_t = \mathbf{x}_t + \boldsymbol{\varepsilon}_t, \quad (10)$$

where one may assume for simplicity that each noise component of $\boldsymbol{\varepsilon}_t$ is i.i.d., for example. This induces a distribution over $\mathbf{X} = (\mathbf{X}_t)_t$.

²Alternatively, one may add the noise variables only at observed time points.

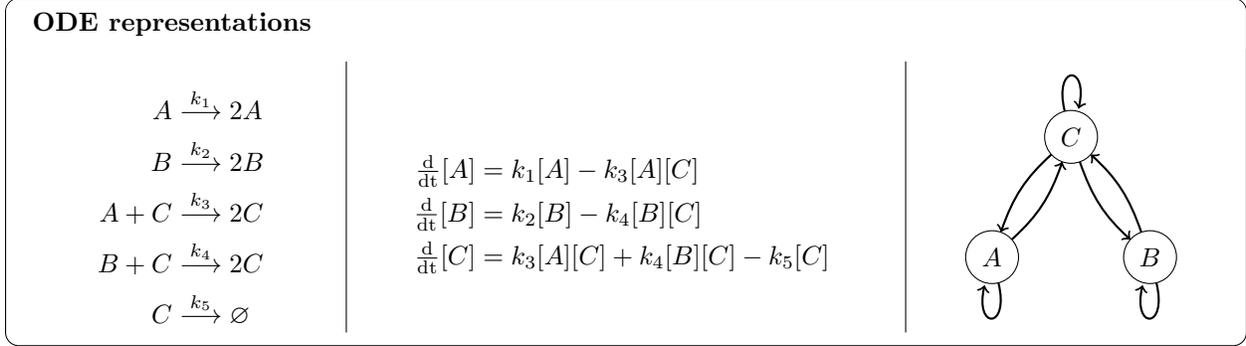


Figure 2. Illustration of different ODE representations: (chemical) reactions (left), ODE system derived by mass-action kinetics (middle) and corresponding graph (right).

3.2 Causal kinetic models with driving noise

As for SCMs, the randomness might also be added directly into the structural assignments. This yields a more involved mathematical formulation though, since the objects of interest are continuous time processes. We define a *stochastic causal kinetic model* over processes $\mathbf{X} := (\mathbf{X}_t)_t := (X_t^1, \dots, X_t^d)_t$ as a collection of d stochastic differential equations (SDEs) and initial value assignments

$$dX_t^k := f^k(X_t^{\mathbf{PA}^k})dt + h^k(X_t^{\mathbf{PA}^k})dW_t^k, \quad X_0^k := \xi_0^k, \quad (11)$$

where dW_t^k are independent white noise processes, i.e., $W_t = \int_0^t dW_s$ is a Brownian motion.³ Again, we require that the SDEs in (11) are uniquely solvable, which in the setting of SDEs becomes substantially harder to verify. The functions f^k are called drift coefficients and the functions h^k are called diffusion coefficients. Intuitively, it can be helpful to think about the change $X_{t+\Delta}^k - X_t^k$ as being normally distributed with expectation $f^k(X_t^{\mathbf{PA}^k}) \cdot \Delta$ and variance $h^k(X_t^{\mathbf{PA}^k})^2 \cdot \Delta$. In the most basic setting, h^k can be assumed to be constant which results in an integrated equations of the form

$$X_t^k := \int_0^t f^k(X_s^{\mathbf{PA}^k})ds + W_t^k.$$

In general, solving SDEs is a difficult problem and numerical procedures often have slower rates compared with their deterministic counterparts [Han and Kloeden, 2017]. We believe that despite these difficulties, SDE-based causal models may potentially prove useful in several areas of applications. There is some work that has made first attempts to circumvent the difficulties of models using SDEs by looking at random differential equations [Bauer et al., 2017, Bongers and Mooij, 2018, Abbati et al., 2019], which still allow including randomness directly into the causal structure. As for SCMs, it depends on the application whether a causal model with measurement noise or the full stochastic setting is the more appropriate choice.

3.3 Interventions

An intervention on the system replaces some of the structural assignments. Interventions can change the dynamics of the process x^k , the initial values or both at the same time. This definition allows for several ways of manipulating the system, which may prove useful when modeling complex dynamical systems and their perturbations; some of the possibilities are discussed below. Formally, for a deterministic causal kinetic model over a process $(\mathbf{x}_t)_t$, an *intervention* on the process x^k for $k \in \{1, \dots, d\}$, corresponds to replacing the k -th initial condition or the k -th ODE with

$$x_0^k := \xi \quad \text{or} \quad \frac{d}{dt}x_t^k := g(x_t^{\mathbf{PA}^k}),$$

³Readers who are not familiar with the formal definition of SDEs may think about them as a notational abbreviation for the integrated form, i.e., $X_t^k := \int_0^t f^k(X_s^{\mathbf{PA}^k})ds + \int_0^t h^k(X_s^{\mathbf{PA}^k})dW_s^k$.

respectively, where $\mathbf{PA} \subseteq \{1, \dots, d\}$ is the set of new parent components. In both cases, we still require that the system of initial value problems is uniquely solvable. The interventions are denoted by

$$do\left(x_0^k := \xi\right) \quad \text{and} \quad do\left(\frac{d}{dt}x_t^k := g(x_t^{\mathbf{PA}})\right),$$

respectively. The same definitions apply in the presence of observational noise ε_t , see (10), where the noise is added after the system has been perturbed. For a stochastic causal kinetic model, we analogously define the interventions

$$do\left(X_0^k := \xi\right) \quad \text{and} \quad do\left(dX_t^k := g(X_t^{\mathbf{PA}}) + j(X_t^{\mathbf{PA}})dW_t^k\right).$$

While we regard both deterministic and stochastic causal kinetic models as potentially relevant for practical applications, we will, in the remainder of this chapter, focus on deterministic causal models.

If the ODE system is induced by a set of reactions, a natural class of interventions is described by replacing one (or some) of the reactions. In the Lotka-Volterra model from Section 2, changing the rate of the first reaction (4), i.e., changing k_1 to \tilde{k}_1 , say, yields a change of the assignment (7). Changing the rate of the second reaction (5), however, yields a change of both assignments (7) and (8). In general, changing one of the reactions induces a change in differential equations for all variables that appear in the reaction (either on the left or on the right). The proposed framework additionally allows us to set a variable x^k to a constant value c by performing the interventions $do(x_0^k := c)$ and $do(\frac{d}{dt}x_t^k := 0)$. To obtain a softer version of this effect, we may also introduce a forcing term that “pulls” the variable x^k to a certain value c . Alternatively, one can keep the dependence of $\frac{d}{dt}x^k$ on x^ℓ intact, but change the strength of this dependence, or even completely change the parent set.

We believe that in a system that is described well by a system of differential equations, it is most natural to formulate the interventions as differential equations, too. Nevertheless, for a differentiable ζ , interventions of the form $x_t^k := \zeta(x_t^A)$ with $A \subseteq \{1, \dots, d\} \setminus \{k\}$ [e.g. Hansen and Sokol, 2014], and $x_t^k := \zeta(t)$ [e.g. Rubenstein et al., 2018] are included in the above formalism as well. The intervention $do(x_t^k := \zeta(x_t^A))$ can be obtained by $do(\frac{d}{dt}x_t^k := \frac{d}{dt}\zeta(x_t^A))$ and $do(x_0^k := \zeta(x_0^A))$. Similarly, $do(x_t^k := \zeta(t))$ is realized by $do(\frac{d}{dt}x_t^k := \frac{d}{dt}\zeta(t))$ and $do(x_0^k := \zeta(0))$.

3.4 Other causal models for dynamical systems and related work

We introduced the formal framework of causal kinetic models that allows us to model dynamical systems with a set of differential equations and specify what we mean by intervening in the system. Several other useful proposals have been made that connect differential equations with causality. Here, we briefly review some of these suggestions and point out a few of the differences. In general, the attempts are tailored towards different goals.

Mooij et al. [2013], Blom and Mooij [2018], Bongers and Mooij [2018] and Rubenstein et al. [2018] consider (deterministic and random) ordinary differential equations. Their goal is to describe the asymptotic solution of such a system as a causal model. The authors consider interventions that fix the full time trajectory of a variable to a pre-defined solution, e.g., to a constant. Mooij et al. [2013] consider interventions on the ODE system itself. In that work, the authors are primarily interested in the equilibrium of the ODE system (assuming that it exists) and its relation to standard structural causal models (SCMs); they explicitly do not distinguish between interventions that yield the same equilibrium. These approaches may be particularly useful when the focus lies on the stationary solution, rather than the full dynamics. Hansen and Sokol [2014] consider stochastic differential equations, which contain ODEs as a special case, and introduce interventions, for which at any point in time the intervened variable can be written as a deterministic function of other variables.

In practice, the application at hand determines which of the models and interventions are most appropriate for describing the real world experiment. The structure of causal kinetic models closely follows the spirit of the SCMs described above. In particular, its modular structure once more highlights which parts remain invariant under interventions.

4 Challenges in causal inference for ODE-based systems

Formalizing a causal model for dynamical systems can be taken as a starting point to conduct causal inference. Similarly to the i.i.d. case, we might be interested in adjustment results, do-calculus, the effect of hidden variables, or causal discovery [see, e.g., Pearl, 2009]. To the best of our knowledge, for dynamical systems, most of such questions are still open. Possible reasons are the difficulties that arise when working with dynamical systems, some of which we highlight below. (1) In the deterministic settings, solving a standard algebraic equation is easier than solving an algebraic equation involving differentials. (2) When adding observational noise, the induced distributions on the left hand side of the structural assignments are more complicated in causal kinetic systems than in SCMs. (3) Suppose that in the i.i.d. case (3), the noise variables are additive. If the parents of each variable (and therefore the structure of the whole system) is known, the causal mechanisms, i.e., the functions f^k , can be estimated by nonlinear regression techniques. In contrast, in the case of dynamical systems, the fitting process is much more involved, and many different methods have been suggested. This includes various versions of goodness-of-fit of the integrated system, nonlinear least squares methods or gradient matching [Bard, 1974, Benson, 1979, Calderhead et al., 2009, Varah, 1982, Ramsay et al., 2007, Macdonald and Husmeier, 2015, Dattner and Klaassen, 2015, Raue et al., 2015, Oates et al., 2014, Wenk et al., 2019]. (4) In the i.i.d. setting, Markov conditions connect properties of the graph, such as d -separation [Pearl, 2009], with properties of the joint distribution, such as conditional independence [Lauritzen, 1996]. For dynamical models, however, it is not apparent that conditional independence is the right notion. For specific model classes, there is interesting work exploiting the concept of local independence [Schweder, 1970, Mogensen et al., 2018, Mogensen and Hansen, 2019, Didelez, 2000, 2008], with several questions still being open. Finally, (5), in most real world systems, not all relevant variables are observed, which means that they need to be modeled as hidden variables. While in the i.i.d. case there is some understanding of the effects of hidden variables on observed distributions, on the identification of causal effects and on causal discovery [e.g. Spirtes et al., 2000, Pearl, 2009, Shpitser, 2018, Richardson and Spirtes, 2002, Hernán and Robins, 2006, Evans, 2015, Richardson et al., 2017, Hyttinen et al., 2012, Verma and Pearl, 1991, Silva et al., 2006], more work is needed in the case of dynamical systems.

5 From invariance to causality and generalizability

In many real world systems, the underlying structure is unknown and needs to be inferred from data. That is, for any k , we do not know which variables are contained in \mathbf{PA}_k . This setting is often referred to as structure learning or causal discovery [Spirtes et al., 2000, Pearl, 2009]. To state the problem let us assume that the observed data consist of n repetitions of discrete time observations of each of the d variables \mathbf{x} , or its noisy version $\tilde{\mathbf{X}}$, on the time grid $\mathbf{t} = (t_1, \dots, t_L)$. By concatenating the time series for the d variables, one may represent the data by an $n \times (d \cdot L)$ matrix. Several methods have been suggested to solve this task [e.g., Oates et al., 2014, Mikkelsen and Hansen, 2017, Raue et al., 2015], most of which combine structure learning, i.e., model selection, with a parameter inference step. Some methods [Oates et al., 2014] explicitly consider the causal nature of this problem. We briefly describe below, in a simplified setting, how it is possible to exploit the invariances induced by the underlying causal kinetic model for causal discovery. Assume there is a target process $y := x^1$, for which the parents are unknown and of particular interest. In short, we assume that each of the n repetitions has been generated by a model of the form

$$\frac{d}{dt}y_t = f^y(\mathbf{x}_t^{\mathbf{PA}_y}), \quad (12)$$

for a fixed function f^y , possibly with additional measurement noise $\tilde{Y}_t = y_t + \varepsilon$. This assumption holds, for example, if the measurements stem from an underlying causal kinetic model under different interventional settings, in none of which the variable y has been intervened on. In practice, the right-hand side of (12) is unknown, and the goal is thus to identify the causal predictors among the \mathbf{x} , i.e., to infer both the parents \mathbf{PA}_y of y as well as the function f^y . In [Pfister et al., 2018a], we propose a procedure that specifically exploits the invariance of (12) to tackle the problem of structure learning. Each of the repetitions is assumed to be part of an environment or experimental condition. We suppose this assignment is known, e.g., repetitions $1, \dots, 6$ are known to belong to experimental condition one, repetitions $7, \dots, 19$ to condition two, and all remaining repetitions to condition three. The method then outputs a ranking of models (or variables) by trading off predictability and invariance of such models. In

the i.i.d. case, trade-offs in a similar spirit have been suggested by Rojas-Carulla et al. [2018], Magliacane et al. [2018] and Rothenhäusler et al. [2018], for example.

The model (12) is valid independently of interventions on variables other than y and can thus be used for prediction in a new experimental setup, even if there are large perturbations on the predictors \mathbf{x} . As a consequence, the method proposed in [Pfister et al., 2018a] outputs models that generalize better to unseen experiments, even when considering real data from large metabolic network experiments. This finding adds to a recent debate suggesting to add invariance as a fitting criteria to data science methodology [Schölkopf et al., 2012, Yu, 2013, Peters et al., 2016, Bareinboim and Pearl, 2016, Meinshausen et al., 2016, Yu and Kumbier, 2019]. At its core lies the modularity of the structure of the causal model and its implied relation between causality and invariance.

6 Conclusions

We have discussed an extension of structural causal models to systems that are governed by differential equations. As in the i.i.d. case, the models may be equipped with either measurement noise or driving noise, where the latter case yields the concept of stochastic differential equations. These two model classes, called causal kinetic models, may serve as a starting point for answering questions commonly asked in the field of causal inference. Many of such questions are neither fully understood nor answered, and more work is needed to gather as much understanding as we have for i.i.d. data.

The mathematical complexity of the models poses a challenge when working with kinetic models. Some aspects of causal inference, however, may become easier. The concept of faithfulness suggests, for example, that in the i.i.d. setting, a child of a random variable is predictive for its parent. This assumption seems less justified in the case of dynamical processes. Also, considering local independence and assuming causal sufficiency, Markov equivalence classes contain only a single directed acyclic graph [Mogensen and Hansen, 2019]. Both of these points may prove to be useful for causal discovery. Furthermore, intervening on a set of differential equations usually affects the whole time trajectory. Relatively mild interventions may thus carry a lot of information about the causal structure. This may be particularly relevant when the available data are not yet sufficient to identify causal mechanisms and additional data have to be collected. There is a close connection between experimentation and causal inference [Imai et al., 2013, Peters et al., 2017]; the selection of measurement readouts, time points or intervention strategies could guide experimentation and has the potential to significantly reduce the number of complicated and expensive experiments.

While there are several differences to the i.i.d. case, causal kinetic models exhibit the same modularity as structural causal models. As a consequence, invariance ideas can be exploited in a similar way as it is done in the i.i.d. case. This includes methods that trade off invariance and predictability to select models that may generalize better to unseen experiments.

Acknowledgements

The authors thank Søren Wengel Mogensen and Niels Richard Hansen for helpful discussions. JP was supported by a research grant (18968) from VILLUM FONDEN.

References

- O. Aalen, O. Borgan, and H. Gjessing. *Survival and Event History Analysis: A Process Point of View*. Springer, New York, NY, USA, 2008.
- G. Abbati, P. Wenk, S. Bauer, M. A. Osborne, A. Krause, and B. Schölkopf. ARS and MaRS – adversarial and MMD-minimizing regression for SDEs. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2019.
- J. Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.
- Y. Bard. *Nonlinear Parameter Estimation*. Academic Press, New York, NY, 1974.

- E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.
- S. Bauer, N. Gorbach, D. Miladinovic, and J. M. Buhmann. Efficient and flexible inference for stochastic systems. In *Advances in neural information processing systems (NIPS)*, pages 6988–6998, 2017.
- Stefan Bauer, Bernhard Schölkopf, and Jonas Peters. The arrow of time in multivariate time series. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 2043–2051, 2016.
- M. Benson. Parameter fitting in dynamic models. *Ecological Modelling*, 6:97–115, 1979.
- T. Blom and J. M. Mooij. Generalized structural causal models. *ArXiv e-prints (1805.06539)*, 2018.
- K. A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, New York, 1989.
- S. Bongers and J. M. Mooij. From random differential equations to structural causal models: the stochastic case. *ArXiv e-prints (1803.08784)*, 2018.
- S. Bongers, J. Peters, B. Schölkopf, and J. M. Mooij. Theoretical aspects of cyclic structural causal models. *ArXiv e-prints (1611.06221v2)*, 2016.
- B. Calderhead, M. Girolami, and N. D. Lawrence. Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. In *Advances in neural information processing systems (NIPS)*, pages 217–224, 2009.
- R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, Boca Raton, FL, USA, 2nd edition, 2006.
- T. Chen, H. He, and G. Church. Modeling gene expression with differential equations. In *Biocomputing'99*, pages 29–40. World Scientific, 1999.
- I. Dattner and C. A. J. Klaassen. Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electronic Journal of Statistics*, 9(2):1939–1973, 2015.
- V. Didelez. *Graphical Models for Event History Analysis based on Local Independence*. PhD thesis, Universität Dortmund, 2000.
- V. Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society, Series B*, 70(1):245–264, 2008.
- R. J. Evans. Margins of discrete Bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656, 2015.
- K. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.
- C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, 1969.
- T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.
- X. Han and P. E. Kloeden. *Random ordinary differential equations and their numerical solution*. Springer, 2017.
- N. R. Hansen and A. Sokol. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19(100):1–24, 2014.
- M. A. Hernán and J. M. Robins. Instruments for causal inference: An epidemiologists dream? *Epidemiology*, 17:360–372, 2006.
- K. D. Hoover. The logic of causal inference. *Economics and Philosophy*, 6:207–234, 1990.
- A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(1):3387–3439, 2012.

- A. Hyttinen, P. O. Hoyer, F. Eberhardt, and M. Järvisalo. Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 301–310, 2013.
- K. Imai, D. Tingley, and T. Yamamoto. Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1):5–51, 2013.
- G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, 2015.
- S. Lauritzen. *Graphical Models*. Oxford University Press, New York, NY, USA, 1996.
- A. J. Lotka. Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry*, 14(3): 271–274, 1909.
- H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, Germany, 2007.
- B. Macdonald and D. Husmeier. Gradient matching methods for computational inference in mechanistic models for systems biology: a review and comparative analysis. *Frontiers in bioengineering and biotechnology*, 3:180, 2015.
- S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems 31*, pages 10846–10856, 2018.
- N. Meinshausen, A. Hauser, J. Mooij, J. Peters, P. Versteeg, and P. Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.
- L. Michaelis and M. L. Menten. Die Kinetik der Invertinwirkung. *Biochem Zeitschrift*, 49:333–369, 1913. translation available at <https://pubs.acs.org/doi/suppl/10.1021/bi201284u>.
- F. V. Mikkelsen and N. R. Hansen. Learning large scale ordinary differential equation systems. *ArXiv e-prints (1710.09308)*, 2017.
- S. W. Mogensen and N. R. Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, to appear, 2019.
- S. W. Mogensen, D. Malinsky, and N. R. Hansen. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 350–360, 2018.
- J. M. Mooij, D. Janzing, and B. Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. In *Proceedings of the 29th Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 440–448, Corvallis, Oregon, USA, 2013. AUAI Press.
- R. Murray. *A mathematical introduction to robotic manipulation*. CRC press, 2017.
- C. J. Oates, F. Dondelinger, N. Bayani, J. Korkola, J. W. Gray, and S. Mukherjee. Causal network inference using biochemical kinetics. *Bioinformatics*, 30(17):i468–i474, 2014.
- B. Ogunnaike and W. Ray. *Process dynamics, modeling, and control*, volume 1. Oxford University Press New York, 1994.
- J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, USA, 2nd edition, 2009.
- J. Peters, D. Janzing, A. Gretton, and B. Schölkopf. Detecting the direction of causal time series. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 801–808. ACM Press, 2009.

- J. Peters, D. Janzing, and B. Schölkopf. Causal inference on time series using structural equation models. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 585–592. Curran Associates, Inc., 2013.
- J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B (with discussion)*, 78(5): 947–1012, 2016.
- J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.
- N. Pfister, S. Bauer, and J. Peters. Identifying causal structure in large-scale kinetic systems: Benefits of a causal approach. *ArXiv e-prints (1810.11776v1)*, 2018a.
- N. Pfister, P. Bühlmann, and J. Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2018b.
- N. Pfister, S. Bauer, and J. Peters. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019.
- J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007.
- A. Raue, B. Steiert, M. Schelker, C. Kreutz, T. Maiwald, H. Hass, J. Vanlier, C. Tönsing, L. Adlung, R. Engesser, W. Mader, T. Heinemann, J. Hasenauer, M. Schilling, T. Höfer, E. Klipp, F. Theis, U. Klingmüller, B. Schöberl, and J. Timmer. Data2dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, 31(21):3558–3560, 2015.
- T. Richardson and J. M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128, 30 April 2013*, 2013.
- T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.
- T. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. *ArXiv e-prints (1701.06686)*, 2017.
- J. M. Robins. Causal inference from complex longitudinal data. In M. Berkane, editor, *Latent Variable Modeling and Applications to Causality*, pages 69–117, New York, NY, USA, 1997. Springer.
- M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Causal transfer in machine learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.
- D. Rothenhäusler, P. Bühlmann, N. Meinshausen, and J. Peters. Anchor regression: heterogeneous data meets causality. *ArXiv e-prints (1801.06229)*, 2018.
- P. Rubenstein, S. Bongers, J. M. Mooij, and B. Schölkopf. From deterministic ODEs to dynamic structural causal models. In *Proceedings of the 34th Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.
- B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.
- T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85:461–464, 2000.
- T. Schweder. Composable Markov processes. *Journal of Applied Probability*, 7:400–410, 1970.
- I. Shpitser. Identification in graphical causal models. In M. Maathuis, M. Drton, S. L. Lauritzen, and M. Wainwright, editors, *Handbook of Graphical Models*. CRC Press, 2018.

- R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.
- P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.
- T. J. Vanderweele. *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press, New York, NY, USA, 2015.
- J. M. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1):28–46, 1982.
- T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 255–270, 1991.
- P. Waage and C. M. Guldberg. Studier over affiniteten (in Danish). *Forhandlinger i Videnskabs-selskabet i Christiania*, pages 35–45, 1864.
- P. Wenk, A. Gotovos, S. Bauer, N. Gorbach, A. Krause, and J. M. Buhmann. Fast Gaussian process based gradient matching for parameter identification in systems of nonlinear ODEs. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.
- H. White and X. Lu. Granger causality and dynamic structural systems. *Journal of Financial Econometrics*, 8(2):193–243, 2010.
- N. Wiener. The theory of prediction. In E. Beckenbach, editor, *Modern Mathematics for Engineers*. McGraw-Hill, New York, NY, 1956.
- D. J. Wilkinson. *Stochastic modelling for systems biology*. Chapman and Hall/CRC mathematical and computational biology series. Chapman & Hall/CRC, 2006.
- S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.
- B. Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.
- B. Yu and K. Kumbier. Three principles of data science: predictability, computability, and stability (pcs). *ArXiv e-prints (1901.08152)*, 2019.
- K. Zhang, M. Gong, J. Ramsey, K. Batmanghelich, P. Spirtes, and C. Glymour. Causal discovery with linear non-Gaussian models under measurement error: Structural identifiability results. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2018.
- W-B. Zhang. *Differential equations, bifurcations, and chaos in economics*, volume 68. World Scientific Publishing Company, 2005.