

Identifying Causal Structure in Large-Scale Kinetic Systems

Niklas Pfister

ETH Zürich, Switzerland

`niklas.pfister@stat.math.ethz.ch`

Stefan Bauer

ETH Zürich, Switzerland

MPI Tübingen, Germany

`stefan.bauer@tuebingen.mpg.de`

Jonas Peters

University of Copenhagen, Denmark

`jonas.peters@math.ku.dk`

October 30, 2018

In the natural sciences, differential equations are widely used to describe dynamical systems. The discovery and verification of such models from data has become a fundamental challenge of science today. From a statistical point of view, we distinguish two problems: *parameter estimation* and *structure search*. In parameter estimation, we start from a given differential equation and estimate the parameters from noisy data that are observed at discrete time points. The estimate depends nonlinearly on the parameters. This poses both statistical and computational challenges and makes the task of structure search even more ambitious. Existing methods use either standard model selection techniques or various types of sparsity enforcing regularization, hence focusing on predictive performance. In this work, we develop novel methodology for structure search in ordinary differential equation models. Exploiting ideas from causal inference, we propose to rank models not only by their predictive performance, but also by taking into account stability, i.e., their ability to predict well in different experimental settings. Based on this model ranking we also construct a ranking of individual variables reflecting causal importance. It provides researchers with a list of promising candidate variables that may be investigated further in interventional experiments. Our ranking methodology (both for models and variables) comes with theoretical asymptotic guarantees and is shown to outperform current state-of-the art methods based on extensive experimental evaluation on simulated data. Practical applicability of the procedure is illustrated on a not yet published biological data set. Our methodology is fully implemented. Code will be provided online and will also be made available as an R package.

1. Introduction

Quantitative models of dynamical systems have become a cornerstone of the modern natural sciences. Their development began with the formalization of mathematical analysis by Leibniz [1684] and Newton [1736] introducing differential equations as a tool to model the behavior of complicated dynamical systems over time, e.g., in Newton’s famous laws of motion. Nowadays, differential equations are universally used in scientific fields as diverse as physics, neuroscience [e.g. Friston et al., 2003], genetics [e.g. Chen et al., 1999], bioprocessing [e.g. Ogunnaike and Ray, 1994], robotics [e.g. Murray, 2017], or economics [e.g. Zhang, 2005]. The popularity across science together with new data acquisition technologies [Ren et al., 2003, Regev et al., 2017, Rozman et al., 2018] has shifted the focus in many disciplines to data-driven inference of these models. Assume that a set X^1, \dots, X^d of variables (in biological applications this might be a set of protein concentrations) is governed by the differential equation

$$\dot{\mathbf{X}}_t = f_\theta(\mathbf{X}_t), \quad \mathbf{X}_0 = \mathbf{x}_0. \quad (1)$$

In practice, this set of equations is often unknown, but one is able to constrain the function class to contain only mass-action kinetics or Michaelis-Menten dynamics, for example. If we have access to noisy observations $\tilde{\mathbf{X}}_t$ of \mathbf{X}_t that are measured at a finite set of time points we may then try to estimate the differential equation from the observed data. From a statistical point of view, the challenge of learning underlying equations from time series data can be split into two sub-problems: *parameter estimation* and *structure search*.

The first problem of estimating parameters in known dynamical models is of interest in many applications, for example, if the parameters correspond to rate constants in biological or chemical reactions. There is a rich literature on statistical methods that solve this problem for various settings. We discuss some of the most prominent approaches in Section 1.1. Since the estimate depends nonlinearly on the parameters, the problem becomes both statistically and computationally hard. This is the reason why the second problem, the estimation of the model structure itself, also known as model learning, network inference or system identification, is even more challenging. It corresponds to inferring the underlying natural laws that govern a system, and becomes important when it is possible to constrain the type of relationships between variables but it is unclear which variables enter (1) on the right hand side. Existing approaches, some of which we revise in Section 1.1, are mostly procedures based on predictability and have difficulties in capturing the underlying causal mechanism. As a result, they may not predict well the outcome of experiments that are different from the ones used for fitting the model.

This paper introduces a novel methodological framework, called *Causal KinetiX*, for structure search on dynamical models described by ordinary differential equations (ODEs). By drawing on tools from causality we aim to infer models that represent the underlying causal mechanisms better and therefore show improved ability to generalize to unseen experiments. More specifically, we exploit the idea of invariance whose relation to causality is well studied and has been of interest to researchers across various disciplines for many decades, see Section 2.4. Assume that, as it is usually the case in regression and classification settings, we are mainly interested in a *target variable* Y and that we measure several predictor variables X^j . A causal model for Y is invariant in the sense that it is capable of describing a system well across different experimental settings, where some of the predictor variables X^j have been intervened on. Motivated by this well-known observation, we construct a procedure that scores ODE-based models for the dynamics of the target variable according to their invariance under different observed experimental settings. This, in particular, yields a stability ranking of models. By analyzing the appearance of single variables in the top invariant models, we are then able to draw conclusions about the

causal nature of such variables, too. (This, again, can be expressed as a ranking.) In practice, the stability ranking of models and variables can be used to generate novel causal hypotheses that can then be verified in targeted follow-up experiments. Figure 1 shows an overview of the proposed method, the details can be found in Section 3.

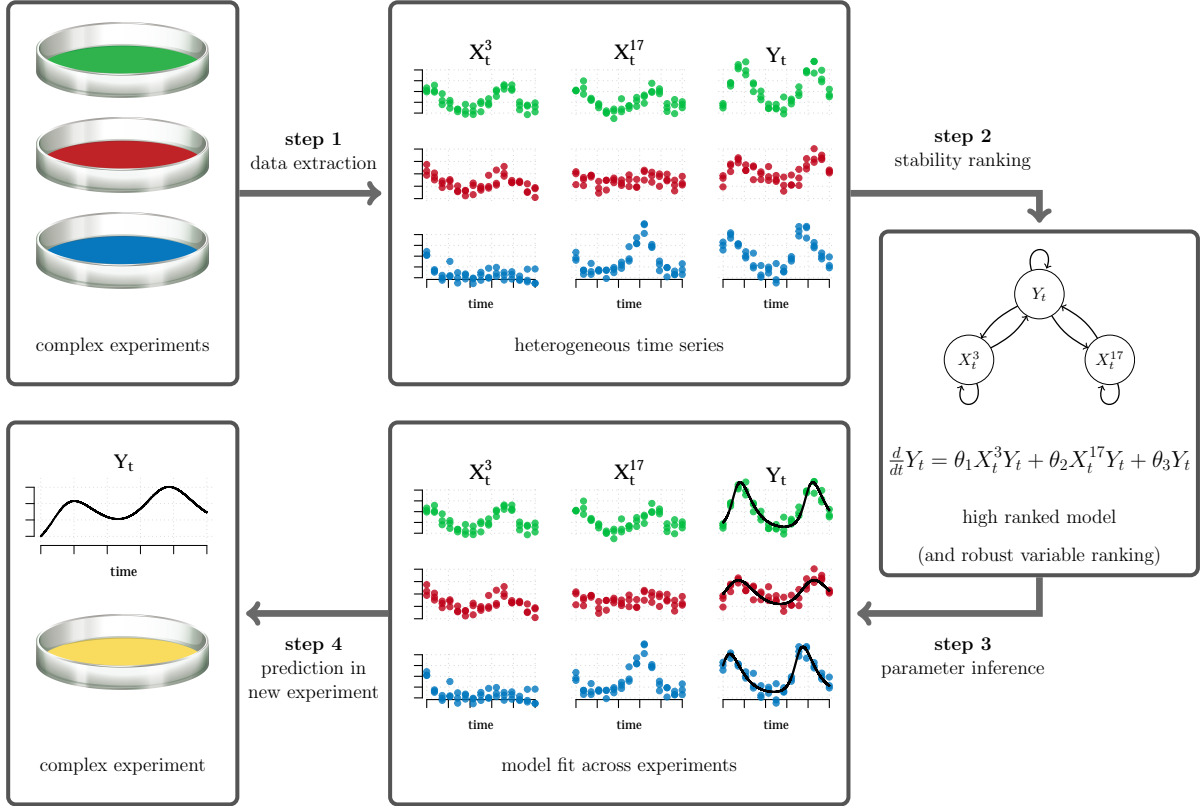


Figure 1. The framework of Causal KinetiX: the data for a target variables Y and predictors X come from different experiments; we rank models according to their ability to fit the target well in all experiments; the top ranked model is then fit to the data; it allows to predict the target in an unseen experiment.

To illustrate the applicability of our procedure, we focus on examples from the field of systems biology, where the need for automated modeling has steadily grown in the last decade [Bongard and Lipson, 2007, Natale et al., 2017]. Due to the modularity of our procedure, it adapts particularly well to problems in systems biology, since researchers there are frequently confronted with the paradoxical situation that the system to be modeled is only partially observed and at the same time models are needed to better understand the system in order to formulate testable hypotheses [Engelhardt et al., 2016].

1.1. Related work

Modeling dynamical systems has a wide variety of applications and there is extensive literature on the topic. We summarize some existing work in terms of three categories: parameter estimation (Section 1.1.1), structure search (Section 1.1.2), and finally a more recent development, causal models for dynamical systems (Section 1.1.3). Our main contribution falls into the latter two topics.

1.1.1. Parameter estimation

In this paper, we are mainly interested in structure search, but the methods used for parameter estimation are often important ingredients for structure search. Assume it is known that the dynamics of a system is described by $\dot{\mathbf{X}}_t = f_\theta(\mathbf{X}_t)$, see (1). In this paper, we will be focusing on a target variable $Y_t = \mathbf{X}_t^1$, say, but we will describe the method in its full generality here. Here, the function f_θ is known up to a parameter vector θ and $\dot{\mathbf{X}}_t$ refers to the time derivative of \mathbf{X}_t . The goal in parameter estimation is to use noisy observations $\tilde{\mathbf{X}}_t$ of \mathbf{X}_t , observed at the time instances t_1, \dots, t_L to infer the parameters θ . This problem is often formulated as a nonlinear least squares problem

$$\operatorname{argmin}_{\theta} \sum_{\ell=1}^L \left(\tilde{\mathbf{X}}_{t_\ell} - \mathbf{X}_{t_\ell} \right)^2, \quad (2)$$

whose exact solution is often considered as a “gold standard” [e.g., Chen et al., 2017]. Here, the solution trajectories \mathbf{X}_t depends on the parameter θ . In practice, solving this optimization is computationally very difficult. For a given candidate value θ , one has to solve the initial value problem (1). The theorem by Picard-Lindelöf guarantees the existence of a unique solution if f_θ is uniformly Lipschitz [e.g., Sideris, 2014, Theorem 3.9]. But except for some special problems, this solution cannot be found by symbolic computation, and numerical integration methods have to be used instead. Many of the instances we consider in this paper are so-called stiff problems whose solutions are particularly difficult to obtain. Standard explicit methods such as RK4 may fail and one must use implicit methods instead, with the consequence of additional computing time. Therefore, even single evaluations of the objective function are computationally expensive and require carefully chosen numerical integration methods. Optimizing over θ comes with the additional challenge that in practice, the number of time points L is very small (we consider an example with $L = 11$), and the objective function is non-convex. Nevertheless, implementations of this procedure exist. There are various versions, and here we concentrate on the highly optimized Matlab implementation `data2dynamics` by Raue et al. [2015]. It is often used in systems biology and can be considered as a state-of-the-art implementation for solving (2). Their numerical integration is based on CVODES of the SUNDIALS suite [Hindmarsh et al., 2005]. It is a variable order, variable step size method, combining Adams-Moulton and the backward differentiation method. To solve nonlinear least squares, they implemented different algorithms, and suggest to use a deterministic trust region approach, a Newton type algorithm implemented in Matlab as LSQNONLIN [Coleman and Li, 1993, 1994]. Finally, they combine their algorithm with a multi-start strategy to avoid local minima. For more details, see Raue et al. [2015, supplementary information].

Due to the computational cost of (2), several alternative approaches have been proposed. In gradient matching, one avoids solving the ODE explicitly by first smoothing the trajectories and then fitting the gradients to the data, e.g., by ordinary least squares. Naturally, the solution depends strongly on the quality of the initial smoother. Varah [1982] suggest to use splines, Ramsay et al. [2007] suggest to regularize the smoother using the ODE itself, and Calderhead et al. [2009], Dondelinger et al. [2013], Wenk et al. [2018] use Gaussian Processes. Macdonald and Husmeier [2015] provide an overview of the different approaches based on gradient matching with a focus on parameter estimation in biological systems. Instead of matching gradients, it has been suggested to perform integral matching [e.g., Dattner and Klaassen, 2015], see also Sections 3.4.4 and 5.1.

1.1.2. Structure search

The task of model selection is arguably even more difficult than parameter estimation and less work has been dedicated to this problem. Existing methods are most often constrained to low-dimensional examples.

The most prominent area of current research views structure search as a model selection problem and considers sparsity enforcing regularization procedures to simultaneously infer parameters and the structure of the model. Brunton et al. [2016], Wu et al. [2014], for example, propose to apply ℓ^1 -penalization, see the baseline implemented in the Section 5.1. Mikkelsen and Hansen [2017] introduce an extension called AIM, which combines the ℓ^1 -penalization with direct minimization of the nonlinear optimization problem that emerges from (2). A similar idea for polynomials is proposed by Tran and Ward [2017] and for stochastic differential equations by Boninsegna et al. [2018]. These approaches can also be extended and applied to finding PDEs [Rudy et al., 2017, Schaeffer, 2017]. On a critical note, it has been suggested that approaches relying on sparsity regularization to identify a model structure should be combined with additional prior information, at least in systems biology [Szederkényi et al., 2011].

Employing other types of regularization or classical model selection techniques is possible, too. Mangan et al. [2017], for example, combine sparsity regularization with a model selection step based on the information criteria AIC and BIC. Common alternatives are based on approximate Bayesian computation or likelihood ratio [Toni et al., 2009, Vyshemirsky and Girolami, 2007]. Procedures using a Gaussian process prior on the solution of the differential equation include Raissi et al. [2017], Dony et al. [2018], for example, and, in a Bayesian setting, by Calderhead et al. [2009], Gorbach et al. [2017], Wenk et al. [2018]. Other procedures consider additional regularization to the inference problem by restricting the functional family to additive models [e.g. Henderson and Michailidis, 2014, Chen et al., 2017].

Some attempts to describe time series by parametric equations rely on symbolic learning as proposed in Crutchfield and McNamara [1987], Schmidt and Lipson [2009] who use evolutionary algorithms over restricted functional forms. Bongard and Lipson [2007] consider an active learning setting, where one can decide on performing new experiments (by changing initial conditions), whose data are then included in the inference procedure, too. Methods based on inductive logic programming are proposed in Todorovski and Dzeroski [1997], Zembowicz and Zytchow [1992], Washio et al. [1999], while modern approaches even try end-to-end learning [Raissi, 2018, Martius and Lampert, 2016] and combine the identification of underlying differential equations from data with prediction.

Some approaches tackle the problem by using model-free network reconstructions as in Äijö and Lähdesmäki [2009], Guo et al. [2014], Casadiego et al. [2017] and also methods based on general time-series models, which focus on non-parametric kernel methods [Grosse et al., 2012, Duvenaud et al., 2013]. For a review of network inference methods with a focus on systems biology, we refer to Oates and Mukherjee [2012] and Siegenthaler and Gunawan [2014] and for general systems to the classic textbook of Ljung [1998].

All of the above approaches focus on predictive performance in the model selection step. We will argue that incorporating invariance yields more robust models and brings us closer to a causal model. The method that comes closest to a causality idea is arguably the one proposed by Oates et al. [2014]. The authors consider a dynamical system as a chemical reaction graph with associated kinetic parameters, where both the graph and kinetic parameters have to be inferred from data. It employs a full Bayesian approach which comes with a computational cost, especially for a large number of variables. The links can be argued to be robust through model averaging.

In this paper, we restrict ourselves to the task of inferring the model for a single target variable $Y_t := \mathbf{X}_t^1$, for example. Almost all of the above methods that consider the same model class aim at the more challenging task of modeling the full system of variables. Exceptions include the approaches with an ℓ^1 type penalty and the AIM method by [Mikkelsen and Hansen, 2017]. These can be straightforwardly adapted to our setting and are included as baselines in our numerical simulations. Modeling the dynamics of a single target variable rather than the full system is more robust against model misspecification and the existence of hidden variables, see Section 3.1. Moreover, the resulting methods become scalable to large numbers of variables as exhaustive model searches continue to remain feasible.

1.1.3. Causal models for ODE based systems

In Section 2.2, we introduce the formal framework of causal kinetic models that do not only allow us to model dynamical systems with a set of differential equations but also to specify what we mean by intervening in the system. There have been several other proposals to connect differential equations with causal models. We review some of them and argue why they were not sufficiently general for our purpose.

Mooij et al. [2013], Blom and Mooij [2018] and Rubenstein et al. [2018] consider (deterministic) ordinary differential equations. Their goal is to describe the asymptotic solution of such a system as a causal model. They 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. They are, however, interested in the equilibrium of the ODE systems (assuming that they exist) and its relation to standard structural causal models (SCMs). Therefore, they explicitly do not distinguish between interventions that yield the same equilibrium. Another class of causal models was introduced by Hansen and Sokol [2014], who consider stochastic differential equations, which contain ODEs as a special case. They introduce interventions, for which at any time point the intervened variable can be written as a deterministic function of other variables.

As opposed to the above approaches, causal kinetic models, remain on the level of ODEs. We believe that this comes with the following benefits. (1) We aim to model the full time evolution of the system and aim at modeling data measured at different time points. The focus in our work may therefore be slightly different from the one in the approaches mentioned above. (2) Modeling the full time evolution, it is natural to also consider interventions in the differential equations themselves. Our framework allows us to work with a general class of interventions that we believe to be natural in several applications, see Section 2.2.1. Formally, we show at the end of Section 2.2.1 that our proposed causal kinetic models indeed also include the interventions considered in the causal models mentioned above. For example, setting the initial condition of one of the variables to a constant and the corresponding differential equation to zero yields an intervention that effectively puts the trajectory of a variable to a constant. Finally, (3), we will later exploit an invariance in the dynamics of the variables, which is easiest phrased in terms of the original ODE, see Assumption 1.

1.2. Contributions

We propose a general and rigorous modeling framework (Causal KinetiX) for performing causal inference in dynamical systems. It contains the concept of causal kinetic models, offering a language for intervening in dynamical systems. It also comes with two novel methodologies for structure identification from heterogeneous data: (a) model ranking and (b) variable ranking.

While we provide full details and analysis for the case of mass-action kinetics, the concepts and principles are readily applicable to other types of models. In contrast to established approaches, our framework inherently benefits from heterogeneity in the experiments and attempts to learn causal structure from time series data. Experimental evaluation on both simulated and real data examples shows that our framework is robust to (ubiquitous) model misspecification. The methods are flexible in that they can be equipped with any parameter estimation method. Here, we consider a gradient matching type approach, which allows for fast parameter estimation and polynomial runtime in the number of variables, repetitions, environments and number of time points. Opposed to any method that searches over an exponentially growing space of model structures, and performs parameter estimation combined with model selection, our inference framework is scalable to very large systems. It is applicable in cases where only few observations for each state are available. For both proposed ranking procedures we prove theoretical guarantees including an asymptotic consistency result. Finally, we will provide easy to use code, which will be published as an open source R-package. It includes our simulation models (most notably the Maillard reaction) which can be used for testing and benchmarking network inference algorithms.

1.3. Outline

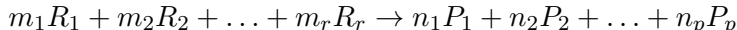
The remainder of this paper is organized as follows. In Section 2, we introduce the framework of causal kinetic models that allows us to define interventions in kinetic systems. We describe the well-studied relation between invariance and causality. In Section 3, we propose the main methodological components. They allow us to score models by stability, i.e., the ability to generalize across experiments. Additionally, we can rank variables according to their importance for stability. Section 4 contains theoretical guarantees. We apply our method to various artificial data sets and a real world biological example in Section 5.

2. Causal models for ordinary differential equations

We begin with a well-known example from biology that mainly serves the purpose of introducing notation. We show how mass-action kinetics connects reactions (as they are common in chemistry, for example) to ODE based models. One may start directly with the ODEs, but often, the underlying reactions give rise to a natural class of interventions (see Definition 2 below).

2.1. An example: the Lotka-Volterra model

A general reaction [e.g. Wilkinson, 2006] takes the form



with r being 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 called *stoichiometries*. A famous example is the Lotka-Volterra model [Lotka, 1909].



where A is the pray and B the predator. 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 indicate that one refers to the concentration rather than to the integer number of abundant species or molecules itself. 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, for example, this yields

$$\frac{d}{dt}[A] = k_1[A] - k_2[A][B] \quad (6)$$

$$\frac{d}{dt}[B] = k_2[A][B] - k_3[B]. \quad (7)$$

This is the type of model that we consider throughout the paper.

2.2. Causal kinetic models

We now define an ODE based model class that includes the set of reactions in (6) and (7), for example. Statistical models are expected to model the observational distribution of a data generating process. In this paper we will introduce a *causal* model class [see Pearl, 2009, Imbens and Rubin, 2015, for i.i.d. data] that additionally models intervention distributions. This allows us to model the system’s behaviour after having perturbed some parts of the system.¹ In this work, we extend the ideas of SCMs (see Section 1.1) to the setting of ordinary differential equations (ODEs). To the best of our knowledge, the following formulation has not been used before.

Definition 1 *A 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*

$$\begin{aligned} \dot{X}_t^1 &:= f^1(X_t^{\mathbf{PA}_1}, X_t^1), & X_0^1 &:= x_0^1 \\ \dot{X}_t^2 &:= f^2(X_t^{\mathbf{PA}_2}, X_t^2), & X_0^2 &:= x_0^2 \\ &\vdots & \\ \dot{X}_t^d &:= f^d(X_t^{\mathbf{PA}_d}, X_t^d), & X_0^d &:= x_0^d. \end{aligned}$$

Here, for any $k \in \{1, \dots, d\}$, \dot{X}_t^k denotes the time derivative of the component X^k at time t and $\mathbf{PA}_k \subseteq \{1, \dots, d\} \setminus \{k\}$ is called the set of direct parents of X^k . We require that the system of ODEs is solvable. For each causal kinetic model we can obtain a corresponding graph over the vertices² $(1, \dots, d)$ by drawing edges from \mathbf{PA}_k to k , $k \in \{1, \dots, d\}$, see Figure 2 for an example. If we consider the initial values as random variables, this induces a distribution over $\mathbf{X} = (\mathbf{X}_t)_t$.

In many practical applications, we might have access to noisy observations only, i.e., we observe

$$\tilde{\mathbf{X}}_t = \mathbf{X}_t + \varepsilon_t, \quad (8)$$

¹A subclass of causal models is furthermore able to model counterfactual distributions, that are not considered in this work. The model formulation we present in Definition 1 can be used for counterfactual statements, too.

²By slight abuse of notation, we identify (X^1, \dots, X^d) with its indices $(1, \dots, d)$.

for example, where we assume for simplicity that each noise component of ε_t is i.i.d. and the components themselves are jointly independent of \mathbf{X}_t . Again, the model induces a distribution over $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_t)_t$.

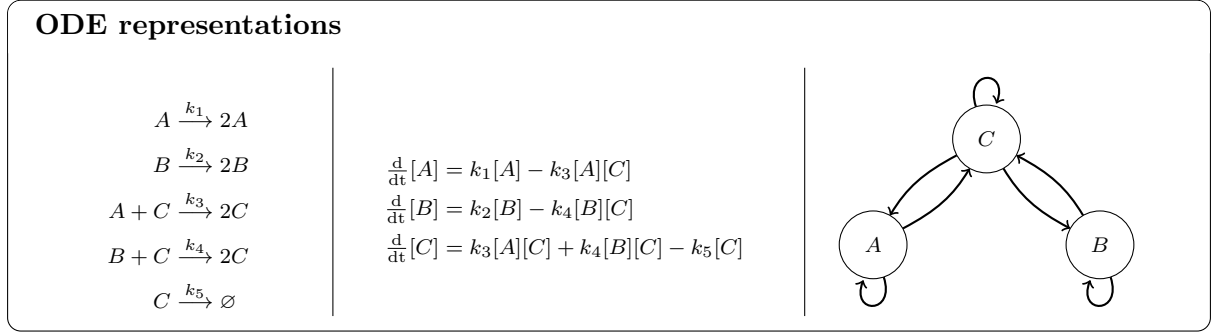


Figure 2. Illustration of different ODE representations: mass-action (chemical) reactions (left), standard ODE system (middle) and corresponding graph (right).

For now, we assume that all variables in the system are observed, an assumption that almost never holds in practice. Our method, however, requires that only one of such assignments holds, see Assumption 1. The modular structure of this model is the key to formalize the concept of interventions.

2.2.1. 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 includes several types of interventions, which may prove useful when modeling complex dynamical systems. We discuss some examples below.

Definition 2 Consider a deterministic causal kinetic model, i.e., a causal kinetic model without noise. An intervention on the process X^k , $1 \leq k \leq d$ corresponds to replacing the k -th initial condition or the k -th ODE with

$$X_0^k := \xi \quad \text{or} \quad \dot{X}_t^k := g(X_t^{\mathbf{PA}}, X_t^k),$$

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 ODEs is solvable. The interventions are denoted by

$$do\left(X_0^k := \xi\right) \quad \text{and} \quad do\left(\dot{X}_t^k := g(X_t^{\mathbf{PA}}, X_t^k)\right),$$

respectively. The same definitions apply in the presence of observational noise ε_t , as in (8).

If the ODE system stems from a set of reactions replacing one of them constitutes a natural intervention; it corresponds to a change of the assignments for all involved variables. In the example from Section 2.1, changing the rate of the first reaction (3), i.e., changing k_1 to \tilde{k}_1 , say, yields a change of assignment (6). Changing the rate of the second reaction (4), however, yields a change of both assignments (6) and (7).

The framework further allows us to set a variable X^k to a constant value c by performing the interventions $do(X_0^k := c)$ and $do(\dot{X}_t^k := 0)$. To obtain a similar effect, we may introduce

a forcing term that “pulls” the variable X^k to a certain value c . Alternatively, one can keep the dependence of \dot{X}^k on X^ℓ , say, but change how strongly \dot{X}^k depends on the value of X^ℓ . In mass-action kinetics, for example, this can be realized by changing the corresponding rates k_j , see Figure 2. It is furthermore possible to change the parent set of X^k .

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, 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 our formalism as well. In our notation, the intervention $do(X_t^k := \zeta(X_t^A))$ can be written as $do(\dot{X}_t^k := \frac{d}{dt}\zeta(X_t^A))$ and $do(X_0^k := \zeta(X_0^A))$. Similarly, $do(X_t^k := \zeta(t))$ is realized by $do(\dot{X}_t^k := \dot{\zeta}(t))$ and $do(X_0^k := \zeta(0))$.

The application at hand determines which of these interventions provides the description that is closest to reality.³

2.3. Structure learning for ODE based systems

Given that the parents \mathbf{PA}_j are unknown, estimation of a causal kinetic model involves both structure learning, i.e., model selection, as well as a parameter inference step. As previously mentioned in Section 1.1, the case where the parents of each variable (and thus the structure of the whole system) is known has received a lot of attention. Our work, however, focuses on settings, in which the system’s underlying structure is unknown and needs to be inferred from data. This setting is often referred to as structure learning, structure identification or causal discovery in the causal community [Spirtes et al., 2000, Pearl, 2009, Peters et al., 2017].

Instead of considering the problem of learning the entire structure, we, here, assume that there is a target process $Y := X^1 \in \mathbf{X}$, for which the parents are unknown and of particular interest. To make this precise, we denote the parents of Y by \mathbf{PA}_Y and assume that each of the n repetitions has been generated by a model of the form

$$\dot{Y}_t = f_Y(\mathbf{X}_t^{\mathbf{PA}_Y}), \quad (9)$$

for a fixed function f_Y . Our goal is to infer both the function f_Y as well as the parents \mathbf{PA}_Y .

The observed data consists of n repetitions of discrete time observations of each of the d variables \mathbf{X} (in general the noise version $\tilde{\mathbf{X}}$) on the time grid $\mathbf{t} = (t_1, \dots, t_L)$. Each of the repetitions is assumed to be part of an environment or experimental condition $\{e_1, \dots, e_m\}$. These experimental conditions should be thought of as different states of the system that can stem from, for example, one of the interventions described in Section 2.2.1. By concatenating the time series for the d variables, we thus represent the data by the following $n \times (d \cdot L)$ matrix.

$$e = e_1 \left\{ \begin{pmatrix} \tilde{X}_{t_1}^{1,(1)} & \dots & \tilde{X}_{t_L}^{1,(1)} & \dots & \tilde{X}_{t_1}^{d,(1)} & \dots & \tilde{X}_{t_L}^{d,(1)} \\ \vdots & & \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots & & \vdots \\ \tilde{X}_{t_1}^{1,(n)} & \dots & \tilde{X}_{t_L}^{1,(n)} & \dots & \tilde{X}_{t_1}^{d,(n)} & \dots & \tilde{X}_{t_L}^{d,(n)} \end{pmatrix} \right. \quad (10)$$

One of the variables, X^1 , say, is considered as the target, i.e., $\tilde{Y}_t^{1,(i)} = \tilde{X}_t^{1,i}$. In Section 3, we propose a two-step method that specifically exploits the difference between experimental setups to tackle the problem of structure identification. The first step is to rank different models by

³One may further be interested in modeling delays in the system, i.e., derivatives depending on the earlier state of a variable. This is possible by adapting Definition 1.

how well they are able to explain the different experimental settings. This leads to a list of candidate models (or functions) f_Y which are then used in the second step to rank variables according to their importance in leading to invariant models and hence whether they belong to \mathbf{PA}_Y . The relation between causality and invariance will be one of the key components.

2.4. Invariance and causality

The notions of invariance and causality are tightly linked to each other. As illustrated in Section 2.2 causal models do not only describe a data generating process in its observational state but they also describe how the system can be intervened on. For many practical applications this is essential as it allows to make statements about the behavior of a system under certain changes. All causal models (for i.i.d. data or dynamical systems) build upon a common assumption: intervening on one part of the system, leaves another part intact. This is commonly referred to as autonomy or modularity [Haavelmo, 1944, Aldrich, 1989, Pearl, 2009, Schölkopf et al., 2012].

This principle can be turned around. Let us therefore assume that the system is observed under a discrete set of experimental (or interventional) settings, none of which affect the target variable Y directly, see (9). In many applications, such an assumption is quite natural, e.g., if the target is a phenotype while the predictors are protein concentrations or gene expressions. This means that the data from all environments can be described by a single model. We therefore propose to evaluate different models by their ability to describe the data from all environments equally well. Only such invariant models have the potential for being causal and restricting inference to the class of invariant models will bring us closer to the underlying causal ground truth. In the i.i.d. settings, such methods have been proposed by Eaton and Murphy [2007], Peters et al. [2016a], Pfister et al. [2018], for example. Note, however, that an extension of the above approaches to dynamical systems is not straightforward due to different scales across experiments, errors in the observation model and the validity of the testing procedure.

In the following section, we propose a method that ranks models by their stability and then uses these scores to find certain variables which appear to be important for the most stable models. By the above reasoning these variables are therefore likely (depending on the number of observed interventional settings) to be part of the true causal model. Moreover, even if they are not part of the true causal model they are good predictors for explaining changes across different interventional settings.

3. Causal KinetiX: Identifying causal predictors

In this section, we introduce a procedure that infers a part of the causal kinetic models described in Section 2. It is based on the assumption that the dynamics of the target variable $Y := X^1$ are described by an invariant (or stable) model, depending only on an unknown subset of the predictors \mathbf{PA}_Y .

Given an assignment $\dot{Y}_t = g(\mathbf{X}_t)$ it will be convenient to speak about the components of \mathbf{X}_t that have an influence on the outcome of g . For any set $S \subseteq \{1, \dots, d\}$, we therefore define

$$\mathcal{F}(S) := \left\{ g : \mathbb{R}^d \rightarrow \mathbb{R} \mid \exists f : \mathbb{R}^{|S|} \rightarrow \mathbb{R} : \forall \mathbf{x} \in \mathbb{R}^d \ g(\mathbf{x}) = f(\mathbf{x}^S) \right\}.$$

A set of functions $\mathcal{G} \subseteq \mathcal{F}(S)$ then contains only functions that do not depend on variables outside S . In the class of mass-action kinetics, for example, we could have

$$\begin{aligned} \mathcal{G}_1 &= \{g \mid g(x) = \theta_1 x_1 + \theta_7 x_7, \text{ where } \theta_1, \theta_7 \in \mathbb{R}\}, \\ \mathcal{G}_2 &= \{g \mid g(x) = \theta_2 x_2 + \theta_3 x_3, \text{ where } \theta_2, \theta_3 \in \mathbb{R}\}, \end{aligned}$$

which implies $\mathcal{G}_1 \subseteq \mathcal{F}(\{1, 7\})$ and $\mathcal{G}_2 \subseteq \mathcal{F}(\{2, 3\})$, see Section 3.1 for more details. In practice, the underlying structure is unknown and many methods therefore include a model selection step. For the remainder of this paper, we assume that we are given a family of target models $\mathcal{M} = \{\mathcal{G}^1, \dots, \mathcal{G}^m\}$, where individual models can depend on various different subsets of variables $S \subseteq \{1, \dots, d\}$. We refer to \mathcal{G} as a *target model* and \mathcal{M} as a *collection of target models*. Based on these definitions, we can make precise what it means for the target trajectories Y to be described by invariant dynamics.

Assumption 1 (invariance) *There exists a set S^* and a function $f^* : \mathbb{R}^{|S^*|} \rightarrow \mathbb{R}$ satisfying for all $i \in \{1, \dots, n\}$ and all $t \in \mathbf{t}$ that*

$$\dot{Y}_t^{(i)} = f^*(\mathbf{X}_t^{S^*,(i)}). \quad (11)$$

Further, S^ is minimal for f^* in the following sense: there is no $S \subsetneq S^*$ such that $f^* \in \mathcal{F}(S)$.*

For causal kinetic models, the pair (f^Y, \mathbf{PA}_Y) satisfies Assumption 1 whenever the environments consist of interventional data, which do not contain interventions on the target Y itself. Even if Assumption 1 is satisfied the pair (f^*, S^*) is not necessarily unique, i.e., there may be one or several pairs satisfying (11). In general, both the set S^* as well as the invariant function f^* are of interest in practice as both are strongly related to the causal mechanisms of the underlying system. In this paper, we propose a method that can rank the target models in \mathcal{M} or the individual predictors in \mathbf{X} according to their importance in achieving the invariance in (11). As discussed in Section 2.4 this means that highly ranked models and variables will be important for understanding the causal mechanisms in the system. Our proposed method consists of two major components: (i) a procedure which assess the stability of each fitted model in \mathcal{M} , which we describe in Section 3.2, and (ii) a principled way of ranking individual variables based on the stability scores of the fitted models depending on them, see Section 3.3. Although the method's underlying principles are more general, our paper focuses on mass-action kinetics, which we introduce formally in the following Section 3.1.

3.1. Parametric models for mass-action kinetics

Many ODE based systems in biology are described by the law of mass-action kinetics, see, e.g., Section 2.1. The resulting ODE models are linear combinations of various orders of interactions between the predictor variables \mathbf{X} . Assuming that the underlying ODE model of our target Y is described by a version of the mass-action kinetic law, the derivative \dot{Y}_t equals

$$\dot{Y}_t = f_\theta(\mathbf{X}_t) = \sum_{k=1}^d \theta_{0,k} X_t^k + \sum_{j=1}^d \sum_{k=j}^d \theta_{j,k} X_t^j X_t^k, \quad (12)$$

where $\theta = (\theta_{0,1}, \dots, \theta_{0,d}, \theta_{1,1}, \theta_{1,2}, \dots, \theta_{d,d}) \in \mathbb{R}^{d(d+1)/2+d}$ is a parameter vector. Correspondingly, the function on the right-hand side of (11) in Assumption 1 has such a parametric form, too. The assumption that the model only depends on the variables in S^* can be expressed by a sparsity on the parameter θ , i.e., $\theta_{j,k} = 0$ for all j, k with $j \notin S^*$ or $k \notin S^*$. Also a target model \mathcal{G} can be constructed by a certain sparsity pattern. Formally, a sparsity pattern is described by a vector $v \in \{0, 1\}^{d(d+1)/2+d}$ which specifies the zero entries in θ . For every such v , we define

$$\mathcal{G}^v := \left\{ f_\theta : \mathbb{R}^d \rightarrow \mathbb{R} \mid \forall \mathbf{x} \in \mathbb{R}^d : f_\theta(\mathbf{x}) = \sum_{k,j} \theta_{k,j} x^k x^j \text{ and } v * \theta = 0 \right\},$$

where $*$ denotes the element-wise product. In principle, one could now search over all sparsity patterns of θ , i.e., define $\mathcal{M} = \{\mathcal{G}^v, v \in \{0, 1\}^{d(d+1)/2+d}\}$, but this becomes computationally infeasible already for small values of d . In this work, we suggest two different collections of target models. Other choices, in particular those motivated by prior knowledge, are possible, too, and can easily be included in our code package.

Exhaustive models. Using only the constraint on the number of terms p leads to the following collection of models

$$\mathcal{M}_p^{\text{Exhaustive}} = \{\mathcal{G}^v \mid v \text{ has at most } p \text{ non-zeros}\}.$$

Every model in $\mathcal{M}_p^{\text{Exhaustive}}$ consists of a linear combination of a fixed number of at most p terms of the form $X^1, \dots, X^d, X^1 X^1, X^1 X^2, \dots, X^{d-1} X^d$ or $X^d X^d$.

Main effect models. Alternatively, one can also add the restriction that the models including interaction terms for variables, include the corresponding main effects, too.

$$\mathcal{M}_p^{\text{MainEffect}} = \{\mathcal{G}^v \mid v \text{ has at most } p \text{ non-zeros and} \\ v_{0,j} \neq 0 \text{ implies } v_{k,j} \neq 0 \forall k < j \text{ and } v_{j,k} \neq 0 \forall k \geq j\}.$$

While the number of main effect models is much smaller it generally requires to fit larger models, which can lead to overfitting. For example, if the true invariant model only depends on the two terms X^1 and $X^4 X^5$ there exists a exhaustive model with two parameters that is invariant, while the smallest main effect model has nine parameters.

Our method is based on Assumption 1, in combination with a more concrete modeling step, as described in this section, for example. Importantly, we only model the dependence of the target on its causal predictors. We do not put any constraints on the distribution of the predictor variables themselves. Apart from computational advantages that we briefly discuss in Section 5.2.4, this comes with the following two modeling benefits: (1) The predictor variables do not have to follow a specific model, such as mass-action kinetics. They can either come from a more complex models or can even be influenced or set by experiments. (2) Additionally, it allows the existence of an arbitrary number of hidden variables: as long as the hidden variables do not influence the target variable directly, they do not affect the validity of Assumption 1. The case, where also the target variable itself is affected by hidden variables is discussed in Section 5.2.6.

3.2. Ranking models by stability

In this section, we introduce a fast, integration-free procedure to compute a stability score for each model in \mathcal{M} . The resulting score can be used to rank each model by how good it preserves the invariance assumption (11). Our proposed stability score is based on a comparison between the fits of an unconstrained smoother of the target trajectories and a constrained smoother in which the constraint enforces a structure based on the considered model. It can be summarized by the following steps. More details are provided in Section 3.4.

(M1) Input

Data as described in Equation (10) and a collection $\mathcal{M} = \{\mathcal{G}^1, \mathcal{G}^2, \dots, \mathcal{G}^m\}$ of models over d variables that is assumed to be rich enough to describe the desired dynamics. In the case of mass-action kinetics, examples include $\mathcal{M} = \mathcal{M}_p^{\text{MainEffect}}$ or $\mathcal{M} = \mathcal{M}_p^{\text{Exhaustive}}$.

(M2) **Screening of predictor terms (optional)**

For large scale systems, one can reduce the method's search space to include only a smaller number of predictor terms.

(M3) **Smooth target trajectories**

For each repetition $i \in \{1, \dots, n\}$, smooth the (noisy) data $\tilde{Y}_{t_1}^{(i)}, \dots, \tilde{Y}_{t_L}^{(i)}$ using a smoothing spline

$$\hat{y}_a^{(i)} := \operatorname{argmin}_{y \in \mathcal{H}_C} \sum_{\ell=1}^L (\tilde{Y}_{t_\ell}^{(i)} - y(t_\ell))^2 + \lambda \int \ddot{y}(s)^2 ds, \quad (13)$$

where λ is a regularization parameter, which in practice is chosen using cross-validation and $\mathcal{H}_C = \{y : [0, T] \rightarrow \mathbb{R} \text{ smooth} \mid \sup_{t \in [0, T]} \max(|y(t)|, |\dot{y}(t)|, |\ddot{y}(t)|) \leq C\}$. We denote the resulting functions by $\hat{y}_a^{(i)} : [0, T] \rightarrow \mathbb{R}$, $i \in \{1, \dots, n\}$. For each of the m candidate target models $\mathcal{G} \in \mathcal{M}$ perform the steps (M4)–(M6).

(M4) **Fit candidate target model**

Fit the target model \mathcal{G} , i.e., find the best fitting function $g \in \mathcal{G}$ such that

$$\dot{Y}_t^{(i)} = g(\mathbf{X}_t^{(i)}), \quad (14)$$

holds for all $i \in \{1, \dots, n\}$ and $t \in \mathbf{t}$. In Section 3.4.1, we describe two procedures for this estimation step resulting in an estimate \hat{g} . For each repetition $i \in \{1, \dots, n\}$, this yields L fitted values $\hat{g}(\tilde{\mathbf{X}}_{t_1}^{(i)}), \dots, \hat{g}(\tilde{\mathbf{X}}_{t_L}^{(i)})$. A slight modification is discussed in Section 3.4.2.

(M5) **Smooth target trajectories with derivative constraint**

Refit the target trajectories for each repetition $i \in \{1, \dots, n\}$ by constraining the smoother to these derivatives, i.e., find the functions $\hat{y}_b^{(i)} : [0, T] \rightarrow \mathbb{R}$ which minimize

$$\hat{y}_b^{(i)} := \operatorname{argmin}_{y \in \mathcal{H}_C} \sum_{\ell=1}^L (\tilde{Y}_{t_\ell}^{(i)} - y(t_\ell))^2 + \lambda \int \ddot{y}(s)^2 ds, \quad (15)$$

such that $\dot{y}(t_\ell) = \hat{g}(\tilde{\mathbf{X}}_{t_\ell}^{(i)})$ for all $\ell = 1, \dots, L$.

(M6) **Compute score**

If the candidate model \mathcal{G} allows for an invariant fit, the fitted values $\hat{g}(\tilde{\mathbf{X}}_1^{(i)}), \dots, \hat{g}(\tilde{\mathbf{X}}_L^{(i)})$ computed in (M4) will be reasonable estimates of the derivatives $\dot{Y}_{t_1}^{(i)}, \dots, \dot{Y}_{t_L}^{(i)}$. This, in particular, means that the constrained fit in (M5) will be good, too. If, conversely, the candidate model \mathcal{G} does not allow for an invariant fit, the estimates produced in (M4) will be poor. We thus score the models by comparing the fitted trajectories $\hat{y}_a^{(i)}$ and $\hat{y}_b^{(i)}$ across repetitions as follows

$$T^{\mathcal{G}} := \frac{1}{n} \sum_{i=1}^n \left[|\text{RSS}_b^{(i)} - \text{RSS}_a^{(i)}| \right] / \left[\text{RSS}_a^{(i)} \right], \quad (16)$$

where $\text{RSS}_*^{(i)} := \frac{1}{L} \sum_{\ell=1}^L (\hat{y}_*^{(i)}(t_\ell) - \tilde{Y}_{t_\ell}^{(i)})^2$.

The scores $T^{\mathcal{G}}$ induce a ranking on the models in \mathcal{M} , where models with a smaller score have more stable fits than models with larger scores. It is further possible to use these scores to rank not only models, but also individual variables. By the relation between stability and causality, this yields variables that are causally related to the target variable.

3.3. Ranking variables by stability

The following idea allows us to rank individual variables according to their importance in stabilizing models. We score all models in the collection \mathcal{M} based on their stability as described in Section 3.2 and then rank the variables according to how many of the top ranked models depend on them. This can be summarized in the following steps.

(V1) Input (same as above)

Data as described in Equation (10) and a collection $\mathcal{M} = \{\mathcal{G}^1, \mathcal{G}^2, \dots, \mathcal{G}^m\}$ of models over d variables that is assumed to be rich enough to describe the desired dynamics. In the case of mass-action kinetics, examples include $\mathcal{M} = \mathcal{M}_p^{\text{MainEffect}}$ or $\mathcal{M} = \mathcal{M}_p^{\text{Exhaustive}}$.

(V2) Compute stabilities

For each model $\mathcal{G} \in \mathcal{M}$ compute the stability score $T^{\mathcal{G}}$ as described in (16). Denote by $S_{(1)}, \dots, S_{(K)}$ the K top ranked models, where $K \in \mathbb{N}$ is chosen to be the number of expected invariant models in \mathcal{M} . See Section 3.4 for how to choose K in practice.

(V3) Score variables

For each variable $j \in \{1, \dots, d\}$ compute the following score

$$s_j := \frac{|\{k \in \{1, \dots, K\} \mid \mathcal{G}_{(k)} \text{ depends on } j\}|}{K}. \quad (17)$$

Here, “ $\mathcal{G}_{(k)}$ depends on j ” means that there is no $S \subseteq \{1, \dots, d\}$ with $j \notin S$ and $\mathcal{G}_{(k)} \subseteq \mathcal{F}(S)$. If there are exactly K invariant models, the above score then represents the fraction of invariant models that depend on variable j . In particular, it equals 1 for a variable j if and only if every invariant model depends on that variable.

Due to the combinatorial nature of these scores, we can construct hypothesis tests of whether a particular score is higher than one would expect given that the models are randomly ranked. The details are given in Section 4.1. Moreover, in Section 4.2, we show that under appropriate assumptions this scoring is consistent in the sense that it is approximately 1 for all variables contained in the true causal model.

3.4. Implementation details

We now provide more details on how to implement the steps described Sections 3.2 and 3.3. Furthermore, we propose to use pre-screening step, when the method is applied to data sets with a large number of predictors, i.e., $d > 30$ or $d > 50$, depending on the considered model classes, see Section 3.4.4.

3.4.1. Fitting target models (M4)

We now propose a method that fits a given target model \mathcal{G} , see (M4) in Section 3.2; that is, we want to find a function $g \in \mathcal{G}$ such that the equation

$$\dot{Y}_t^{(i)} = g(\mathbf{X}_t^{(i)}) \quad (18)$$

is satisfied as well as possible for the observed data $(\tilde{\mathbf{X}}_{t_\ell}^{(i)})_{i \in \{1, \dots, n\}, \ell \in \{1, \dots, L\}}$. This problem is difficult for two reasons. First, the derivative values $\dot{Y}_t^{(i)}$ are not directly observed in the data and, second, even if we had access to noisy unbiased versions of $\dot{Y}_t^{(i)}$, we are dealing with an

error-in-variables problem. Nevertheless, for certain model classes \mathcal{G} it is possible to perform this estimation consistently and since the predictions are only used as constraints, one expects even rough estimates to work as long as they preserve the general dynamics. Here, we first propose a general method which can be adapted to many model classes and then discuss a second method that often performs slightly better but requires the target models to be linear in their parameters.

The first procedure tackles the problem directly by first estimating the derivatives and then performing a regression based on the model class under consideration. More precisely, we first fit the smoother $y_a^{(i)}$ from (M3) and then compute its derivatives. When using the first derivative of a smoothing spline it has been argued that the penalty term in (13) contains the third derivative rather than the second derivative of y [see Ramsay and Silverman, 2005, Section 5.2.8], however, this can lead to numerical instabilities and only works when the problem is sufficiently smooth. Assuming that this results in reasonably unbiased noisy versions of $\dot{Y}_t^{(i)}$, we then perform a regression of the estimated derivatives on the data. The regression procedure depends on the model class \mathcal{G} and can be anything from linear ordinary least squares in the case of a linear model class to random forests if the functions are very nonlinear. As mentioned above most regression procedures have difficulties with errors-in-variables and therefore return biased results. Sometimes it can therefore be helpful to use smoothing or averaging of the predictors to reduce the impact of this problem.

The second method we suggest only works for models which are linear in the parameters, i.e., models that consists of function of the form

$$g(\mathbf{x}) = \sum_{k=1}^p \theta_k g_k(\mathbf{x}),$$

where the functions g_1, \dots, g_p are known transformations. In this case we can transform (18) by integration to

$$Y_{t_\ell}^{(i)} - Y_{t_{\ell-1}}^{(i)} = \sum_{k=1}^p \theta_k \int_{t_{\ell-1}}^{t_\ell} g_k(\tilde{\mathbf{X}}_s^{(i)}) ds.$$

Using this equation in the fit no longer requires estimation of the derivatives of Y but instead the integral of the predictors. It is well-known [e.g., Chen et al., 2017] that integration is numerically more stable than differentiation, which makes the estimation easier. In particular, it is often sufficient to approximate the integrals using the trapezoidal rule, i.e.

$$\int_{t_{\ell-1}}^{t_\ell} g_k(\tilde{\mathbf{X}}_s^{(i)}) ds \approx \frac{g_k(\tilde{\mathbf{X}}_{t_\ell}^{(i)}) + g_k(\tilde{\mathbf{X}}_{t_{\ell-1}}^{(i)})}{2} (t_\ell - t_{\ell-1}).$$

In many applications, it may happen that the noise in the predictors is in fact greater than the error in this approximation. The resulting bias is then negligible.

3.4.2. Leave-one-environment-out (M4)

In Section 3.2, we proposed to fit the model on all n repetitions and then use the predictions of that model to compute the score in (M6). However, since the repetitions are grouped in terms of environments e_1, \dots, e_m (if this is not the case one can also assume each repetition belongs to a different environment) one can incorporate this additional information. Instead of simply fitting the model on all n repetitions, it is often beneficial for every $i \in \{1, \dots, n\}$ to fit the model on all repetitions belonging to all environments other than the one to which i belongs.

Then, using this fitted model, predict the derivatives for the repetition i . Doing this puts more weight on how well a certain model is able to generalize to a different environmental condition, which is exactly what we are trying to measure. It is, therefore, generally preferable to use this leave-one-out procedure whenever sample size and computational time allows it.

3.4.3. Choosing parameter K in variable ranking (V2)

In this section, we discuss the choice of the constant K used in the scores s_j in Section 3.3. As mentioned in (V2), K should ideally be equal to the number of invariant models, as this will be the choice for which we show theoretical consistency guarantees in Section 4.2. We found that the method's results are robust to the choice of K . In doubt, we propose to choose K to be small, to ensure that it is smaller than the number of invariant models. Depending on the collection \mathcal{M} of target models it is sometimes possible to give a heuristic number of invariant models simply based on the structure. For the examples given in Section 3.1 we have the following reasoning. Let us first consider exhaustive models $\mathcal{M}_{p+1}^{\text{Exhaustive}}$. If the smallest invariant model only has p terms (i.e., it corresponds to a $v^* \in V_{p+1}$ with $\sum_j v_j^* = p$) it is clear that any super-model (i.e., any $v \in V_{p+1}$ with $\sum_j |v_j - v_j^*| = 1$) is also an invariant model. The number of super-models, however, is simply given by $2d + \binom{d}{2} - p$. Hence, if we use the model collection $\mathcal{M}_{p+1}^{\text{Exhaustive}}$, where p is assumed to be the expected number of terms contained in the smallest invariant model, a reasonable choice is to set $K = 2d + \binom{d}{2} - p$. A similar reasoning for models in $\mathcal{M}_{p+1}^{\text{MainEffect}}$ leads to the choice $K = d - p$.

3.4.4. Screening of terms (M2)

For large scale, i.e., large d , and even high-dimensional systems, the computational complexity of the method can be significantly reduced by including a prior screening step. The collections of models we propose in Section 3.1 scale as

$$|\mathcal{M}_p^{\text{Exhaustive}}| = \sum_{k=1}^p \binom{\binom{d}{2}}{k} = \mathcal{O}(d^{2p}) \quad \text{and} \quad |\mathcal{M}_p^{\text{MainEffect}}| = \sum_{k=1}^p \binom{d}{k} = \mathcal{O}(d^p).$$

Even though computation of the stability score for a single model is fast, this makes clear that an exhaustive search is infeasible for settings with large d . We therefore propose to reduce the model sizes by first performing a screening step based on how useful certain terms are for prediction in the model estimation in (M4) of Section 3.2 and then continue our procedure with the reduced model class.

Apart from the computational gains such a screening step has the additional advantage of making sure that the terms required for heterogeneity across experimental conditions have predictive power, too. The intuition being that one first reduces to a set of reasonably good predictors by screening and then makes use of the heterogeneity to further assess models according to their importance in creating invariant models. Therefore, even in a low-dimensional, setting combining our method with screening makes sense.

Essentially, any screening or variable selection method based on the model fit in (14) can be used. Here, we give two explicit options based on ℓ^1 -penalized least squares, also known as Lasso [Tibshirani, 1994], for the linear models described Section 3.1. Although we are not aware of a method with this exact formulation, these methods are in spirit very similar to existing methods, see, e.g., [Wu et al., 2014, Brunton et al., 2016, Mikkelsen and Hansen, 2017] and references therein.

DerivLasso This method fits a smoother to each trajectory of the target variable Y and computes the corresponding derivatives. Then, one fits an ℓ^1 -penalized sparse linear model on these estimated derivatives

$$\hat{Y}_{t_\ell}^{(i)} = \sum_{j=1}^d \sum_{k=j}^d \theta_{k,l} X_{t_\ell}^{k,(i)} X_{t_\ell}^{j,(i)} + \varepsilon_{t_\ell}^{(i)},$$

where again $\varepsilon_{t_\ell}^{(i)}$ are assumed independent and identically distributed Gaussian noise variables and the regression coefficient θ is assumed to be sparse. This results in a ranking of terms $X^{k,(i)} X^{j,(i)}$ by when they enter the model for the first time.

IntegratedLasso Similarly, we can proceed as in Section 3.4.1 and integrate the linear model to avoid estimating the often numerically unstable derivatives. In this case, one fits a ℓ^1 -penalized sparse linear model on the difference of the form

$$\begin{aligned} \hat{Y}_{t_\ell}^{(i)} - \hat{Y}_{t_{\ell-1}}^{(i)} &= \sum_{j=1}^d \sum_{k=j}^d \theta_{k,l} \int_{t_{\ell-1}}^{t_\ell} X_s^{k,(i)} X_s^{j,(i)} ds + \varepsilon_{t_\ell}^{(i)} \\ &\approx \sum_{j=1}^d \sum_{k=j}^d \theta_{k,l} \frac{X_{t_\ell}^{k,(i)} X_{t_\ell}^{j,(i)} + X_{t_{\ell-1}}^{k,(i)} X_{t_{\ell-1}}^{j,(i)}}{2} (t_\ell - t_{\ell-1}) + \varepsilon_{t_\ell}^{(i)} \end{aligned}$$

where again $\varepsilon_{t_\ell}^{(i)}$ are assumed independent and identically distributed Gaussian noise and the regression coefficient θ is assumed to be sparse. Again one gets a ranking of the term $X^{k,(i)} X^{j,(i)}$ depending on when they first enter the model. In the numerical simulation section (Section 5.2.1) we show that numerically it appears that IntegratedLasso performs better than DerivLasso (see Figure 5). Intuitively, this is the case whenever the dynamics are hard to detect due to noise as the estimated derivatives will then have strongly time-dependent biases. Similar observations have been made by for example Chen et al. [2017].

The idea of using ℓ^1 -penalized procedures for model inference is not new and has been applied extensively, as also mentioned in Section 1.1. In our numerical experiments in Section 5 we use the above mentioned DerivLasso and IntegratedLasso as two competing methods to assess the performance of our proposed procedure and illustrate that one can indeed profit from enforcing invariance across experiments.

4. Theoretical guarantees

4.1. Significance of the variable ranking

Due to the combinatorial nature of the variable scores introduced in Section 3.3 we can test whether a given score s_j , defined in (17), is significant in the sense that the number of top ranked models depending on variable j is higher than one would expect if the ranking of all models in \mathcal{M} was random. More precisely, consider the null hypothesis

$$H_0 : \text{ the top ranked models } \mathcal{G}_{(1)}, \dots, \mathcal{G}_{(K)} \text{ are drawn uniformly from all models in } \mathcal{M}.$$

It is straightforward to show that under H_0 it holds that $K \cdot s_j$ follows a hypergeometric distribution with parameters $|\mathcal{M}|$ (population size), $|\{\mathcal{G} \in \mathcal{M} \mid \mathcal{G} \text{ depends on } j\}|$ (number success in population) and K (number of draws). For each variable we can hence compute a p -value to assess whether it is significantly important for stability. A small p -value, however, only implies

that a variable is potentially causal but does not directly guarantee causality. It can thus be used to determine how many of the top ranked variables show interesting behavior. In the numerical simulations (see Section 5.2.6) we show that this can often be helpful in applications.

4.2. Consistency of ranking procedure

We now provide some conditions under which the proposed procedure is consistent. To this end, we fix the number of environments (or experiments) to m and assume that there exist R repetitions for each experiment observed on L time points. In total this means we observe $n = m \cdot R$ trajectories, each on a grid of L time points. As asymptotics, consider a growing number of repetitions R_n and simultaneously a growing number of time points L_n . Here, increasing repetitions R and time points L corresponds to collecting more data and obtaining a finer time resolution, respectively. Both of these effects are achieved by novel data collection procedures. To make this more precise, assume that for each $n \in \mathbb{N}$ we are given a time grid

$$\mathbf{t}_n = (t_{n,1}, \dots, t_{n,L_n})$$

on which the data are observed and such that $L_n \rightarrow \infty$ for $n \rightarrow \infty$. For simplicity we will only analyze the case of a uniform grid, i.e., we assume that $\Delta t := t_{n,k+1} - t_{n,k} = \frac{1}{L_n}$ for all $k \in \{1, \dots, L_n\}$. As in Section 2.3 we denote the m environments by $e_1, \dots, e_m \subseteq \{1, \dots, n\}$ and assume that for all $k \in \{1, \dots, m\}$ that $|e_k| = R_n$ which grows as n increases.

To achieve consistency of our ranking procedures (both for models and variables) we require the following three conditions (C1)–(C3) below. These three conditions should be understood as high-level conditions or guidelines. There may be, of course, other sufficient assumptions that yield the desired result and that might cover other settings and models.

- (C1) **Consistency of target smoothing:** The smoothing procedure in (M3), see Section 3.2, satisfies the following consistency. For all $k \in \{1, \dots, m\}$ it holds that

$$\lim_{n \rightarrow \infty} \mathbb{E} \left(\sup_{t \in [0, T]} \left(\hat{y}_a^{(e_k)}(t) - Y_t^{(e_k)} \right)^2 \right) = 0,$$

where, by slight abuse of notation, the superscript (e_k) denotes a fixed repetition from the environment e_k .

- (C2) **Consistency of model estimation:** For every invariant model $\mathcal{G} \in \mathcal{M}$ and for all $k \in \{1, \dots, m\}$ it holds that

$$L_n \max_{\ell \in \{1, \dots, L_n\}} |\hat{g}_n(\tilde{\mathbf{X}}_{t_{n,\ell}}^{(e_k)}) - \dot{Y}_{t_{n,\ell}}^{(e_k)}| \xrightarrow{\mathbb{P}} 0$$

as $n \rightarrow \infty$, i.e., the estimation procedure \hat{g}_n in (M4) is consistent. Furthermore, for all non-invariant models $\mathcal{G} \in \mathcal{M}$ there exists a smooth function $g \in \mathcal{G}$ such that g and its first derivative are bounded and it holds for all $t \in [0, T]$ and for all $k \in \{1, \dots, m\}$ that

$$L_n \max_{\ell \in \{1, \dots, L_n\}} |\hat{g}_n(\tilde{\mathbf{X}}_{t_\ell}^{(e_k)}) - g(\mathbf{X}_{t_\ell}^{(e_k)})| \xrightarrow{\mathbb{P}} 0$$

as $n \rightarrow \infty$, i.e., the estimation convergences to a fixed function.

- (C3) **Uniqueness of invariant model:** There exists a unique function $g^* \in \cup_{\mathcal{G} \in \mathcal{M}} \mathcal{G}$ and a unique set $S^* \subseteq \{1, \dots, d\}$ such that for all $n \in \{m, m+1, \dots\}$ the pair $f^*(\mathbf{x}) := g^*(\mathbf{x}^{S^*})$ and S^* satisfy Assumption 1. This condition is fulfilled if the experiments are sufficiently heterogeneous, e.g., because there are sufficiently many and strong interventions.

Note that (C2) relates to the problem of error-in-variables, see the discussion in Section 3.4.1. Relying on the conditions (C1)–(C3), we are now able to prove consistency results for both the model ranking from Section 3.2 and the variable ranking from Section 3.3. Recalling the definition of $T_n^{\mathcal{G}}$ given in (16) (small values of $T_n^{\mathcal{G}}$ indicate invariance), we define

$$\text{RankAccuracy}_n := 1 - \frac{|\{\mathcal{G} \in \mathcal{M} \mid T_n^{\mathcal{G}} < \max_{\{\tilde{\mathcal{G}} \in \mathcal{M} \mid \tilde{\mathcal{G}} \text{ invariant}\}} T_n^{\tilde{\mathcal{G}}} \text{ and } \mathcal{G} \text{ not invariant}\}|}{|\{\mathcal{G} \in \mathcal{M} \mid \mathcal{G} \text{ not invariant}\}|} \quad (19)$$

as performance measure of our model ranking. RankAccuracy is thus equal to 1 minus “proportion of non-invariant models that are ranked better than the worst invariant model”. In particular, it equals 1 if and only if all invariant models are ranked better than all other models. Given the above conditions the following consistency holds.

Theorem 3 (rank consistency) *Let Assumption 1 and conditions (C1) and (C2) be satisfied. Additionally, assume that for all $k \in \{1, \dots, m\}$ it holds for all $i \in e_k$ and $\ell \in \{1, \dots, L_n\}$ that the noise variables $\varepsilon_{t_\ell}^{(i)}$ are i.i.d., symmetric, sub-Gaussian and satisfy $\mathbb{E}(\varepsilon_{t_\ell}^{(i)}) = 0$ and $\text{var}(\varepsilon_{t_\ell}^{(i)}) = \sigma_k^2$. Let Y_t and its first and second derivative be bounded and assume that for all non-invariant sets $\mathcal{G} \in \mathcal{M}$ the sets $\{t \mapsto g(X_t) \mid g \in \mathcal{G}\}$ are closed with respect to the sup norm. Then, it holds that*

$$\lim_{n \rightarrow \infty} \mathbb{E}(\text{RankAccuracy}_n) = 1.$$

If, in addition, condition (C3) holds, we have the following guarantee for the variable scores $s_j^n = s_j$, defined in (17):

- for all $j \in S^*$ it holds that $\lim_{n \rightarrow \infty} \mathbb{E}(s_j^n) = 1$ and
- for all $j \notin S^*$ it holds that $\lim_{n \rightarrow \infty} \mathbb{E}(s_j^n) \leq \frac{K-1}{K}$,

where $K := |\{\mathcal{G} \in \mathcal{M} \mid \mathcal{G} \text{ is invariant}\}|$.

The result is proved in Appendix B (which also contains the choice of C for (M3)) and it is verified empirically in Section 5.2.3.

5. Numerical experiments

We believe that the main features of our method are as follows: (a) We model only a part of the system, namely the dependence from the target derivative on its causal predictors, see (11) and (12). The predictors themselves do not need to be modeled. This is an advantage not only because of speed but also because the predictors’ dynamics may follow a complex set of differential equations or may depend on variables that are unobserved, see Section 5.2.6, for example. (b) Our method does not solve any numerically expensive steps, e.g., there is no integration step. Together with the first point, this makes our method fast and scalable to large systems, see Section 3.4.4. The experiment on real data includes $d = 411$ predictor variables, see Section 5.3. (c) We take the heterogeneity of the data into account. The method therefore outputs not only predictive variables but also those that yield a prediction that is particularly invariant across different settings. This is a distinctive feature of causal predictors (Section 2), and such a model may still perform well in a new, unseen experiment, see Section 5.3. (d) The output is presented as a ranking that can be seen as a generation of causal hypotheses.

We introduce some competing methods in Section 5.1. Various experiments on synthetic data (partly simulated from real systems) shown in Section 5.2 indicate superior performance of our

proposed method. Finally, Section 5.3 shows an application to a real data set illustrating our method’s potential to practically relevant problems. In the numerical experiments, we consider ODE systems derived from the law of mass-action kinetics, see Section 3.1.

5.1. Competing methods

There exist a large number of methods that aim to perform model selection for ODEs, see also Section 1.1. In essence, these methods combine parameter inference with classical model selection techniques such as information criteria, e.g., AIC or BIC, or ℓ^1 -penalized approaches. All of them have in common that they solely optimize predictive performance of the resulting model and do not make use of any heterogeneity in the data. Here, we compare with two basic ℓ^1 -penalized approaches that we believe are representative for this group of methods. The first method performs the regularization on the level of the derivatives (DerivLasso) and the second on the integrated problem (IntegratedLasso). Both are common in literature and can also be used as screening procedures in our method, see Section 3.4.4. Finally, we also compare with a more involved method called adaptive integral matching (AIM) introduced by Mikkelsen and Hansen [2017]. Rather than only fitting the target equation it fits an entire system of ODEs on all variables, hence utilizing information shared across different variables.

5.2. Simulation experiments

We perform experiments on three ODE models. The first is a biological model of the Maillard reaction [Maillard, 1912], whereas the second and third are artificially constructed ODE models. The relatively small sizes of these systems ($d < 13$) allow for fast data simulation which enables us to compare the performance under various settings and conditions. It further makes the presentation easier to understand. Section 5.3 contains a larger example.

5.2.1. Finding causal predictors in the Maillard reaction

The first simulation study is based on a biological ODE system from *BioModels Database* due to Li et al. [2010]. More specifically, we use the model BIOMD0000000052 due to Brands and van Boekel [2002] which describes reactions in heated monosaccharide-casein systems. This system has the advantages to be relatively small (11 variables), it consists entirely of mass-action type equations, and it remains stable under various random interventions (that we can use to simulate different experimental conditions). The simulation setup is described in Data Set 1.

Data Set 1: Maillard reaction

The ODE structure is given in Appendix C. For the simulations, we randomly select one of the $d = 11$ variables to be the target and generate data from 5 experimental settings and sample 3 repetitions for each experiment. The experimental conditions are as follows.

- **experimental condition 1 (observational data):**
Trajectories are simulated using the parameters given in BIOMD0000000052, see Appendix C.
- **experimental conditions 2 to 5 (interventional data):**
Trajectories are simulated based on the following two types of interventions
 - **initial value intervention:** Initial values are sampled for [Glu] and [Fru] uniform between 0 and $5 \cdot 160$ and for [lys R] uniform between 0 and $5 \cdot 15$, the

remainder of the quantities are kept at zero initially as they are all products of the reactions.

- **blocking reactions:** Random reactions that do not belong to the target equations are set to zero by fixing the corresponding reaction constant $k_i \equiv 0$. The expected number of reactions set to zero is 3.

Based on these experimental conditions each of the true model trajectories are computed using numerical integration. Finally, the observations are given as noisy versions of the values of these trajectories on a quadratic time grid with $L = 11$ time points between 0 and 100. The noise at each observation is independently normal distributed with mean 0 and variance proportional to the total variation norm of the trajectory plus a small positive constant (in case the trajectory is constant). Sample trajectories are given in Figure 3.

As a first assessment of our method, we sample $B = 500$ realizations of the system described in Data Set 1 and apply our method as well as the competing methods to rank the variables according to which variable is most likely to be a parent variable of the target. To remove any effect resulting from ordering of the variables we relabel them in each repetition by randomly permuting the labels. Each ranking is then assessed by computing the area under the operator receiver curve (AUROC) based on the known ground truth (i.e., parents \mathbf{PA}_Y). The results are given in Figure 4. Here, we applied Causal KinetiX using both the exhaustive and the main effect model classes discussed in Section 3.1. For the exhaustive models we considered all possible models consisting of individual variables and interactions (66 potential predictor terms) and restricted the search to models with at most 4 such terms after reducing the number of terms by a prior screening step to 33. For the main effect models we performed no prior screening and considered all models with at most 4 variables. The results show that our method can improve on all competing methods. In particular, we are able to get a median AUROC of 1 implying that in more than half of all repetitions our method ranks the correct models first. Moreover, by comparing with IntegratedLasso one can see that utilizing the heterogeneity (via the stability score) does indeed improve on plain prediction based methods.

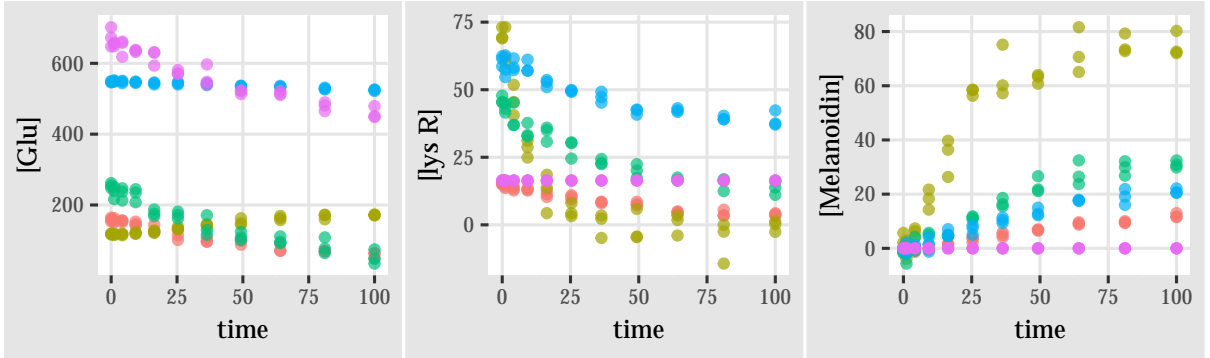


Figure 3. Sample observations (the method’s input) for the variables Glu, lys R and Melanoidin from Data Set 1. Points represent noisy observations with different colors for the 5 different experimental conditions, e.g., red corresponds to experimental condition 1.

5.2.2. Comparison of screening procedures

Using data generated as in Data Set 1, we compare the two screening methods DerivLasso and IntegratedLasso introduced in Section 3.4.4. To this end, we sample $B = 1000$ data sets and

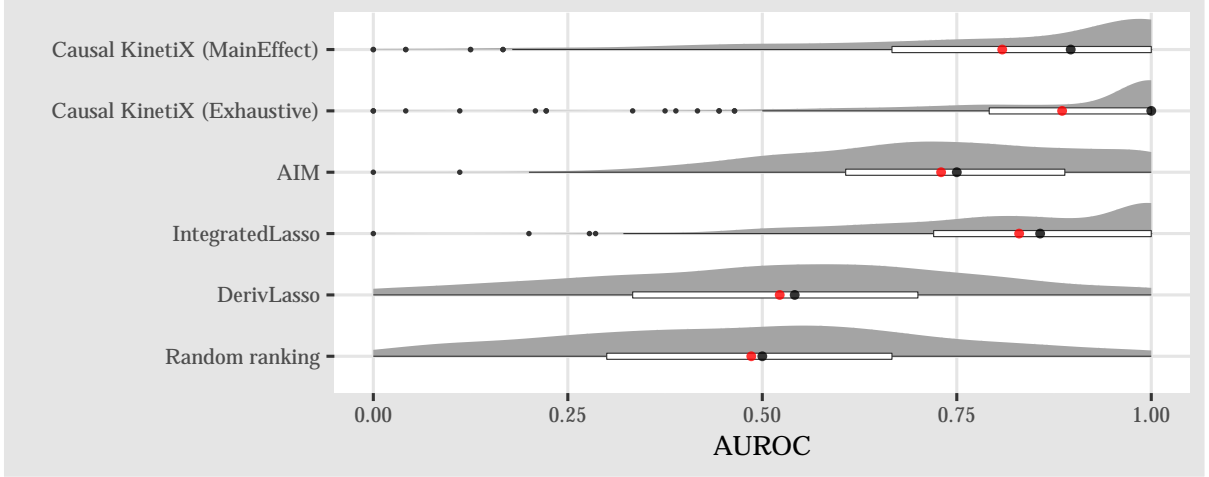


Figure 4. Results for simulation in Section 5.2.1. In each of the 500 simulations, the methods rank predictors for a randomly chosen target. If the AUROC equals one, the correct variables are ranked highest. Red points correspond to mean AUROC, black points to median AUROC.

apply both IntegratedLasso and DerivLasso to rank all $11 \cdot 10 \cdot 0.5 + 22 = 77$ individual terms of the form $X^k X^j$ and X^j ($d = 11$) based on their first entrance into the model. For each data set, we then compute the worst rank of any true term and plot them in Figure 5. For comparison, we also include the results from a random ranking, i.e., a random permutation of the terms. Both methods perform better than the random baseline, and the IntegratedLasso outperforms DerivLasso in this setting. This might be because the integral approximation used in IntegratedLasso is more robust than the estimation of the derivatives required for DerivLasso.

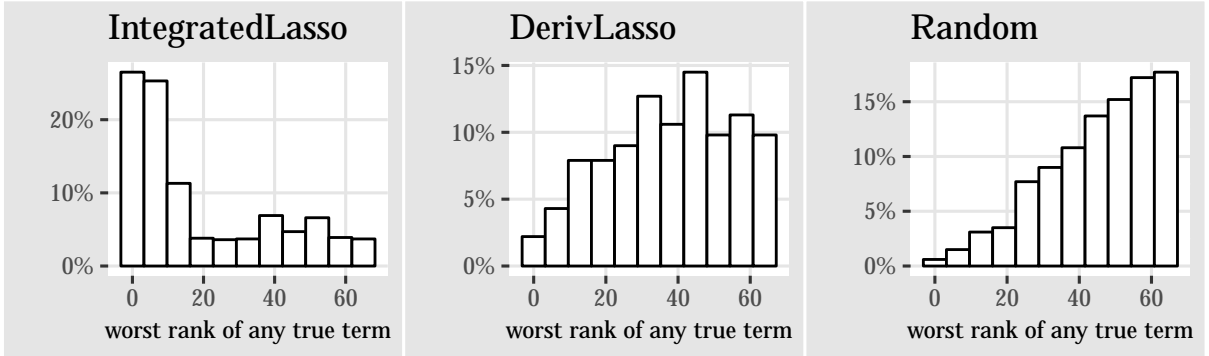


Figure 5. Comparison between different screening methods based on $B = 1000$ simulations from Data Set 1 in Section 5.2.1. All 77 terms of the form $X^k X^j$ and X^j are ranked according to the screening procedure. The x-axis shows the rank of the worst ranked term from the true model. High concentration on the left implies good screening performance. Here, IntegratedLasso outperforms DerivLasso.

5.2.3. Consistency analysis

We now illustrate our theoretical consistency result from Section 4.2. Again, we simulate from Data Set 1, where we consider different values of L and n to analyze the asymptotic behavior. Here, instead of increasing the value of n we decrease the noise variance as this has a similar effect but is computationally faster. Moreover, in light of condition (C3), we now use 10 experimental conditions. The results shown in Figure 6 demonstrate the convergence of the RankAccuracy (19) towards one as the number of time steps L goes to infinity and the noise variance goes to zero.

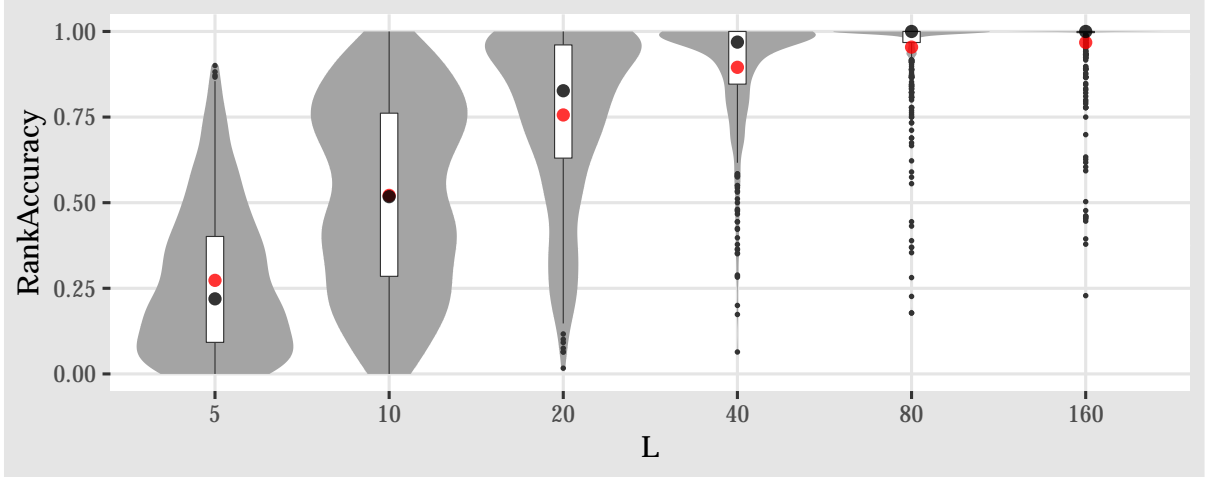


Figure 6. Results for simulation in Section 5.2.3. For different numbers of time points L and noise variance proportional $\frac{10}{L^2}$ we sampled 500 simulations from Data Set 1. For each simulation we compute the RankAccuracy. Red points correspond to mean RankAccuracy, black points to median RankAccuracy.

5.2.4. Scalability

We now analyze how our method scales with the number of variables d , the number of environments m , the number of repetitions in each environment R , and the number of observed time points for each trajectory L . Figure 7 illustrates the run-time of our method when one of these parameters is varied while the others are kept fixed. The data are generated according to Data Set 2. The key steps driving the computational cost of our procedure are the smoothing in steps (M3) and (M5), as well as the estimation step (M4). In our case, the cost of the estimation procedure, fitting a linear model with ordinary least squares, is negligible. Since the number of smoothing operations, we have to perform grows linearly with respect to m and R , we expect a linear increase in run-time. Accordingly, the slopes in Figure 7 (bottom left and top right) are close to one.

We compute the smoothing spline in (M3) using a convex quadratic program, which can be solved in polynomial time – even if the number of constraints is growing linearly, see (M5). The data points in Figure 7 (bottom right) do not lie on a straight line, which may be due to some computational overhead for small values of L or due to the quadratic program itself. The worst case complexity of convex quadratic programming is cubic in sample size, but many instances can be solved more efficiently. Correspondingly, the slope in Figure 7 is not larger than three. When only values $L \geq 64$ are taken into account, the slope is estimated as 2.9, which is close

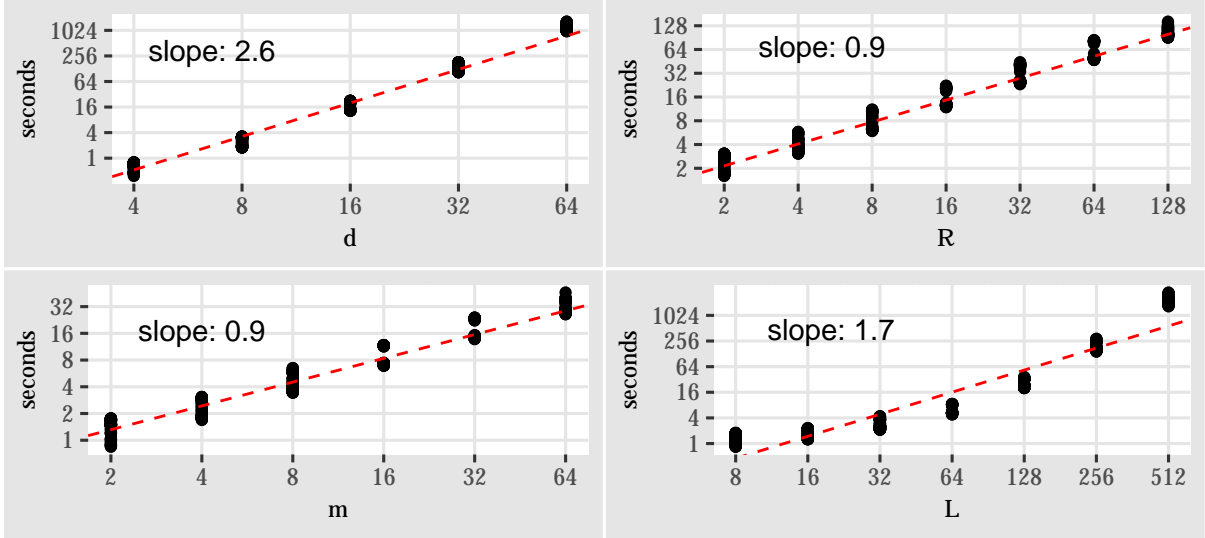


Figure 7. Run-time analysis for the parameters number of variables d , number of repetitions R , number of environments m and number of time points L . In each plot one parameter is varied while the remaining are kept constant and the run-time is computed 100 times for a full application of our method. The dotted red line is a linear fit to the log-log-transformed plots where the slope estimates the polynomial runtime order.

to the worst case guarantee of 3. Finally, varying the number of variables impacts the size of \mathcal{M} , that is, the number of models. In Figure 7, we consider the case of main-effects models of up to three variables (see Section 3.1), which results in $\mathcal{O}(d^3)$ models. If we again assume that run-time of the estimation step can be neglected, we expect a slope of 3 in the log-log plot. In our empirical experiments the slope is estimated as 2.6 (Figure 7 top left).

5.2.5. Allowing for complex predictor models

Our procedure requires that only the dynamics of the target variable are given by an ODE model. We do not model the dynamics of the predictors, which as a consequence may follow any arbitrarily complex model. As an illustration we sample trajectories such that the predictors are completely random and only the target variable satisfies an invariant model, according to Assumption 1. The details of the data generation are shown in Data Set 2.

Data Set 2: Target model based on predictor trajectories

Consider functions of the form

$$f_{c_1, c_2, c_3, c_4}(t) = \frac{c_1}{1 + e^{c_2(t-3)}} + \frac{c_3}{1 + e^{c_4(t-3)}},$$

i.e., these functions are linear combinations of sigmoids which have smooth trajectories that imitate dynamics observed in some real data experiment. For each of the 5 experimental conditions we sample $d = 12$ trajectories $X_t^j = f_{c_1, c_2, c_3, c_4}(t)$ for $t \in [0, 6]$, where c_1, c_2, c_3, c_4 are i.i.d. standard normal. Based on these trajectories and the ODE given by

$$\dot{Y}_t = \theta_1 X^1 + \theta_2 X^2, \quad Y_0 = 0,$$

where $\theta_1 = 0.0001$ and $\theta_2 = 0.0002$, we compute the trajectories of the target variable Y by numerical integration. Finally, the observations are given as noisy versions of the values of these trajectories on an equally spaced time grid with $L = 15$ time points between 0 and 10. The noise at each observation is independently normal distributed with mean 0 and variance proportional to the total variation norm of the trajectory plus a small positive constant (in case the trajectory is constant). Sample trajectories are given in Figure 8.

The results are shown in Figure 9. Here, again we applied Causal KinetiX for both the exhaustive and main effects model class. For the exhaustive model class we again consider individual variables and interactions as possible terms (78 terms) and reduce to half these (39 terms) using screening and then apply our method for all models with at most 3 terms. For the main effect models we again perform no screening and consider all models consisting of at most 3 variables. Even though none of the predictors follows an ODE model our procedure is capable of recovering the true causal parents and again improves on plain prediction (IntegratedLasso and DerivLasso).

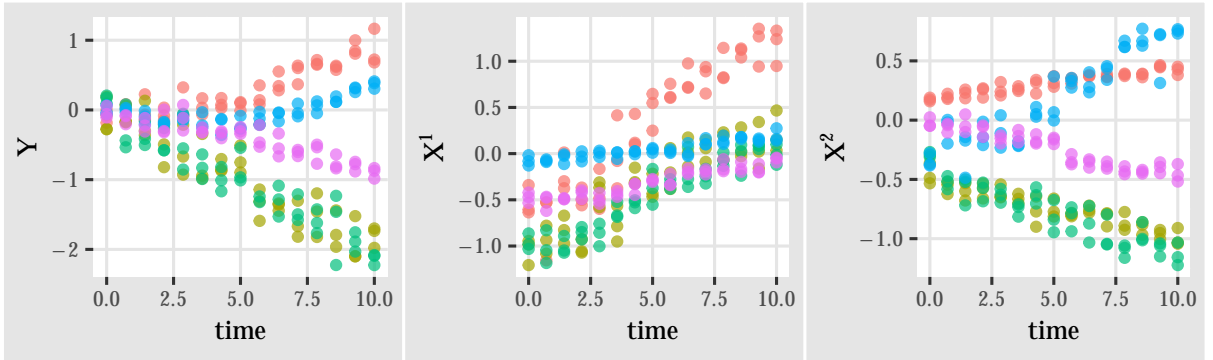


Figure 8. Sample observations for the target variable Y and its two parents X^1 and X^2 Data Set 2. Points represent noisy observations with different colors used for the 5 different experiments, e.g., red corresponds to experimental condition 1.

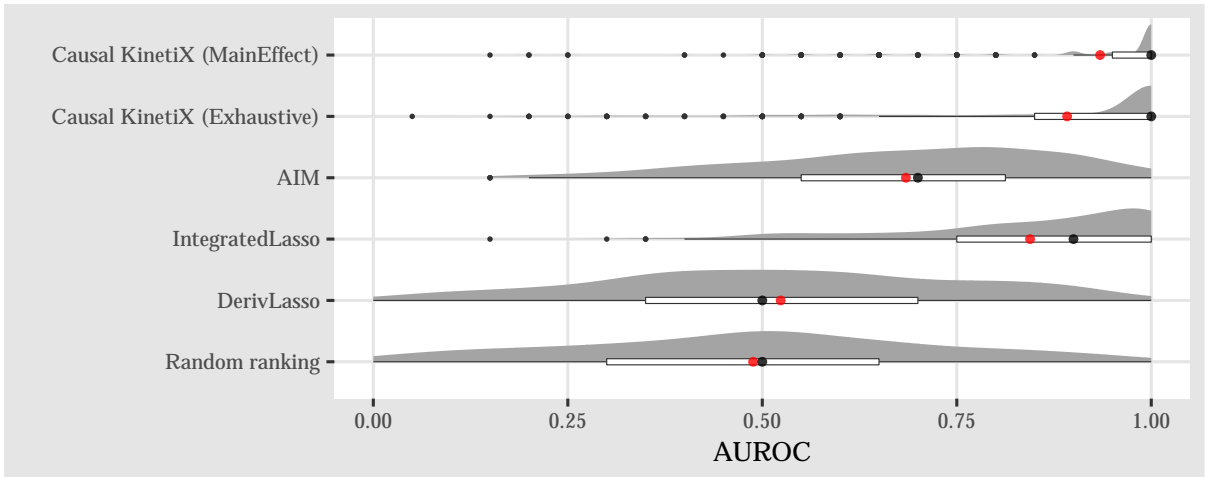


Figure 9. Results for simulation in Section 5.2.5. Red points correspond to mean AUROC, black points to median AUROC.

5.2.6. Robustness in the presence of hidden variables

In many practical applications hidden (unobserved) variables are omnipresent. Since we only model the target equation, see Section 3.1, hidden variables do not affect our methodology if they appear anywhere outside the target variable Y . In this section, we show that even if they enter the target equation our procedure is generally expected to behave well. The data in this section are generated according to Data Set 3, which is based on an artificially constructed ODE system, for which some of the variables are assumed to be hidden. Example trajectories are shown in Figure 10.

Data Set 3: Hidden variable model

The exact ODE structure is given in Appendix D. We generate data from 16 experimental conditions and sample 3 repetitions for each experiment. The experimental conditions are the following.

- **experimental condition 1 (observational data):**

Trajectories are simulated using the parameters given in Appendix D.

- **experimental conditions 2 to 16 (interventional data):**

Trajectories are simulated based on the following two types of interventions

- **initial value intervention:** Initial values are sampled for X^1 , X^2 and X^5 uniform between 0 and 10, the remainder of the quantities are kept at zero initially as they are all products of the reactions.
- **blocking reactions:** Random reactions other than k_4 , k_5 and k_7 are set to zero by fixing the corresponding reaction constant $k_i \equiv 0$. The expected number of reactions set to zero is 2. Additionally, the rate k_7 is randomly perturbed either by sampling it uniform on $[0, 0.2]$ or uniform on $[-0.1, 0.3]$.

Based on these experimental conditions each true trajectory is computed using numerical integration. Finally, the observations are noisy versions of the values of these trajectories on an exponential time grid with $L = 20$ time points between 0 and 100. The noise at each observation is independently normal distributed with mean 0 and variance proportional to the total variation norm of the trajectory plus a small positive constant (in case the trajectory is constant). Example trajectories for the variables X^2 and H^2 depending on the values of k_7 are illustrated in Figure 10.

We conduct three experiments, whose results are shown in Figures 11 and Figure 12. In the first setting (left plots), all variables are observed. The system of equations is built such that X^2 and H^2 obey very similar but not identical trajectories (here k_7 is perturbed less). Most methods are able to correctly identify X^3 and H^2 as the direct causes of Y – those variables are usually ranked highest, see Figures 11 and 12 (left). In the second setting (middle plots), H^2 is unobserved. Because of the similarity between H^2 and X^2 , the methods now infer X^2 as a direct cause. Finally, the third setting (right plots) differ from the second setting in the sense that H^2 and X^2 are significantly different (here k_7 is perturbed more). The latter variable still helps for prediction but does not yield an invariant model. Our method still reliably infers X^3 as a direct cause, which is usually ranked higher than any of the other variables, see Figures 11 and 12 (right).

The results show that our method is relatively robust against the existence of unobserved variables.

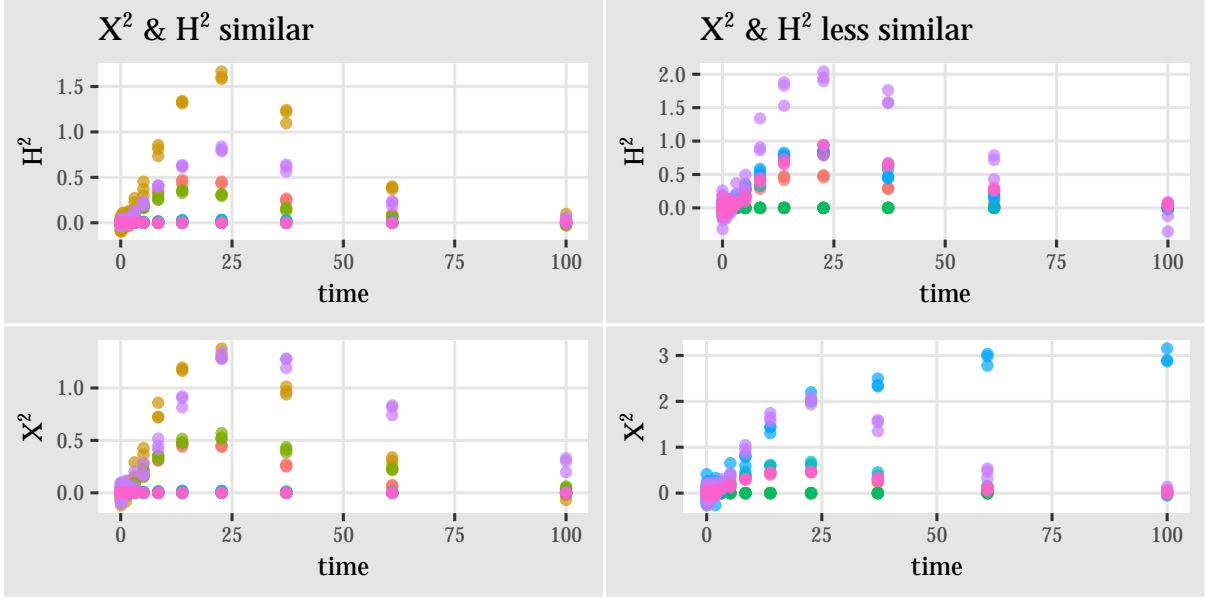


Figure 10. Sample observations of the two predictors X^2 and H^2 from Data Set 3 (only first 8 experiments) for two different choices of perturbations of k_7 . From left to right: For k_7 uniform on $[0, 0.2]$ the dynamics are similar but not identical, for k_7 uniform on $[-0.1, 0.3]$ the dynamics become very different. Points represent noisy observations of the underlying ODE trajectories.

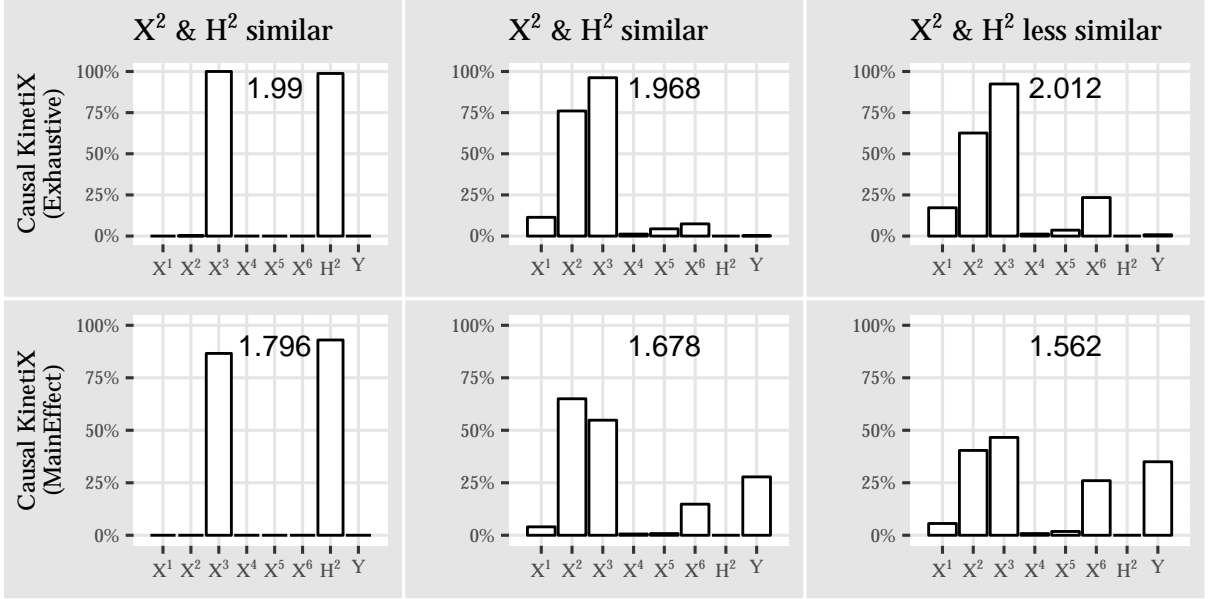


Figure 11. Results for the experiment described in Section 5.2.6 (hidden variables). Plot shows how often each variable gets a p -value smaller than 0.01, see Section 4.1. The number on each histogram is the average number of significant variables at a 1% level.

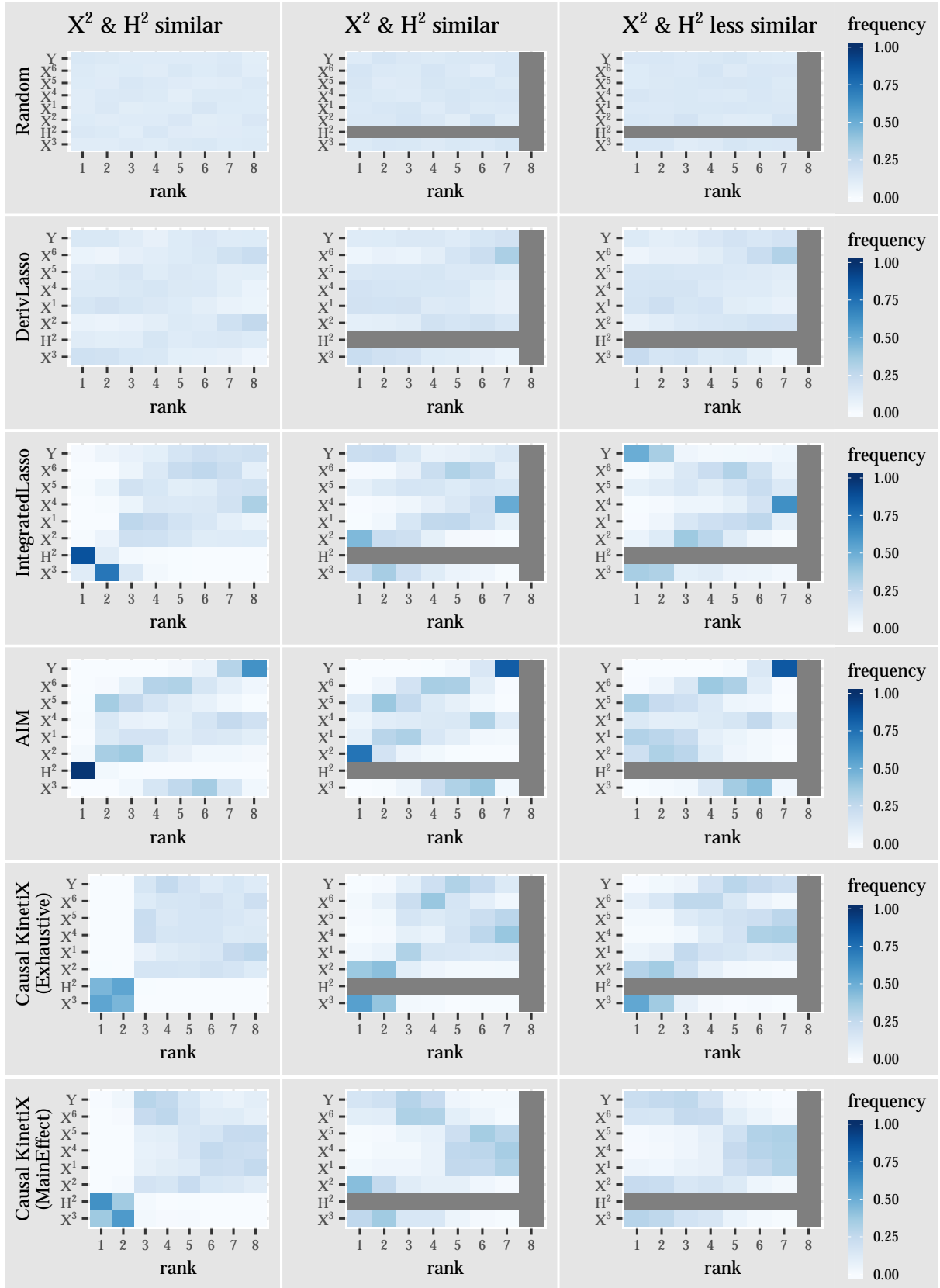


Figure 12. Results for the experiment described in Section 5.2.6 (hidden variables). Left: all variables are observed. From top to bottom: Random, DerivLasso, IntegratedLasso, AIM, Causal KinetiX (Exhaustive), Causal KinetiX (Main Effect)

5.3. Real data example

We apply the proposed ranking method to a real biological data set. At eleven time points, we observe cell concentration or ion count measurements of a target variable and 411 metabolites. In this application one is particularly interested in distinguishing between up- and downshifts, i.e., whether the target trajectory increases or decreases, respectively, compared to its starting value. The system is measured under five different conditions (experiments), each of which contains three biological replicates. Defining the auxiliary variable $Z_t := 2 - Y_t$, we expect that the target species Y_t and Z_t are tightly related: $Y_t \rightleftharpoons Z_t$, i.e., Y_t is formed into Z_t and vice versa. We therefore expect models of the type

$$\begin{aligned}\dot{Y}_t &= \theta_1 Z_t X_t^j X_t^k + \theta_2 Z_t X_t^p X_t^q - \theta_3 Y_t X_t^r X_t^s \\ \dot{Z}_t &= -\theta_1 Z_t X_t^j X_t^k - \theta_2 Z_t X_t^p X_t^q + \theta_3 Y_t X_t^r X_t^s,\end{aligned}$$

where $j, k, p, q, r, s \in \{1, \dots, 411\}$ and $\theta_1, \theta_2, \theta_3 \geq 0$. By the conservation of mass both target equations mirror themselves, which makes it sufficient to only learn the model for Y_t . More precisely, we use the model class consisting of three term models of the form $Z_t X_t^j X_t^k$, $Y_t X_t^j X_t^k$, $Z_t X_t^j$, $Y_t X_t^j$, Z_t , or Y_t , where the sign of the parameter is constrained to being positive or negative depending on whether the term contains Z_t or Y_t , respectively. We constrain ourselves to three terms, as we found this to be the smallest number of terms that results in sufficiently good in-sample fits. Given sufficient computational resources, one may include more terms, too, of course. The sign constraint can be incorporated into our method by performing a constrained least squares fit instead of OLS in step (M4). This constrained regression can then be solved efficiently by a quadratic program with linear constraints.

As the biological data is high-dimensional, our method first screens down to 60 terms and then searches over all models consisting of 3 terms. To get more accurate fits of the dynamics, we pool and smooth over the three biological replicates and only work with the smoothed data. As a baseline, we compare with the results for IntegratedLasso (screened down to 3 terms). DerivLasso performs worse than IntegratedLasso (results not shown).

We now evaluate the applied models according to two criteria: their ability to describe the dynamics in the observed experiments (in-sample performance) and their ability to generalize to similar but unseen experiments (out-of-sample performance). To verify the in-sample accuracy, we consider how well, based on all five experiments, the top ranked model is able to fit the data. The result is illustrated in Figure 13, which shows the target variable and its fit in each of the five experiments. Overall, our method outperforms IntegratedLasso. This is also depicted in the numbers: the averaged RSS between the integrated solution (green) and the smoother (blue) equals 0.013, 0.219 and 0.043 for Causal KinetiX and the two versions of IntegratedLasso, respectively.

We furthermore plot the predicted derivatives (red lines) on the smoother fit. This draws a picture different from the integrated solution: if the latter is only slightly off from the data, it becomes hard to assess the discrepancy between the measured data and the modeled dynamics, see, e.g., Figure 13 (second row, third column). We have not seen such a plot before but regard it as an indispensable tool when analyzing the fit of dynamical models. Note that for both methods, we use the parameters output by the estimation step (see (M4)), in this case an estimator based on ordinary least squares. Additionally, we also fit the parameters for the IntegratedLasso model using the software package Data2Dynamics (d2d) [Raue et al., 2015], which is often used in practice. The large difference between the two estimates is due to the different optimization targets (see Section 1.1.1) and much smaller for better fitting models.

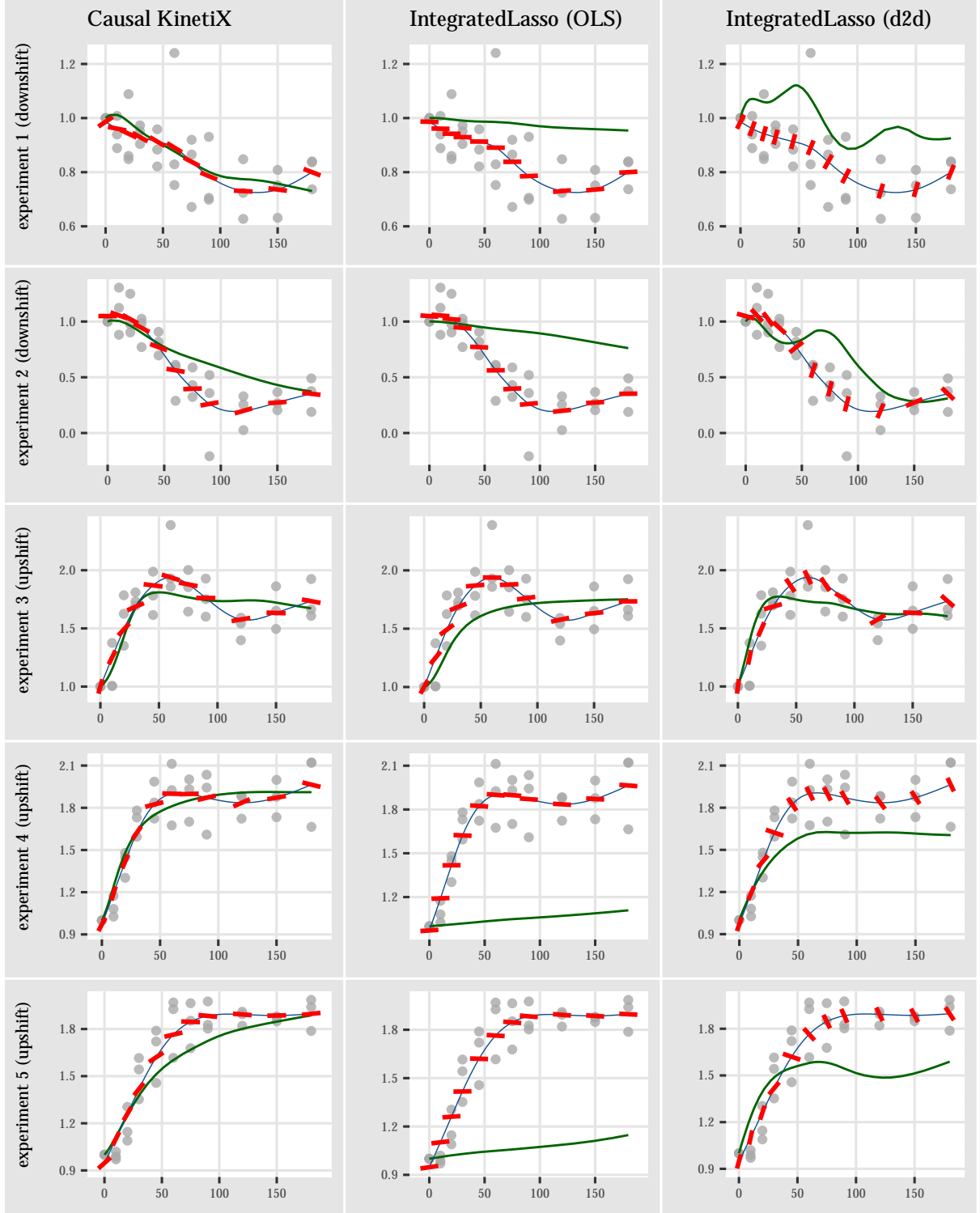


Figure 13. In-sample fit for real world data containing several up- and downshifts. The left model is selected by Causal KinetiX, the center and right model by IntegratedLasso (where the parameters are either estimated with OLS or d2d). The green line shows the integrated solution. Additionally, we fit a smoother to the raw data (blue) and plot at each observed time point the gradient predicted by the model (red). Even though the integrated solution (green) may look reasonable, the dynamics can still be a bad fit, see, e.g., second row, third column. Overall, Causal KinetiX fits the dynamics better than IntegratedLasso.

To assess the ability to generalize, we consider the best ranked model, hold out one experiment, fit the parameters on the remaining four experiments and finally predict the dynamics on the held out experiment. The results (see Figure 14) show that our method is indeed able find models that perform well on experiments that have not been used for parameter estimation.

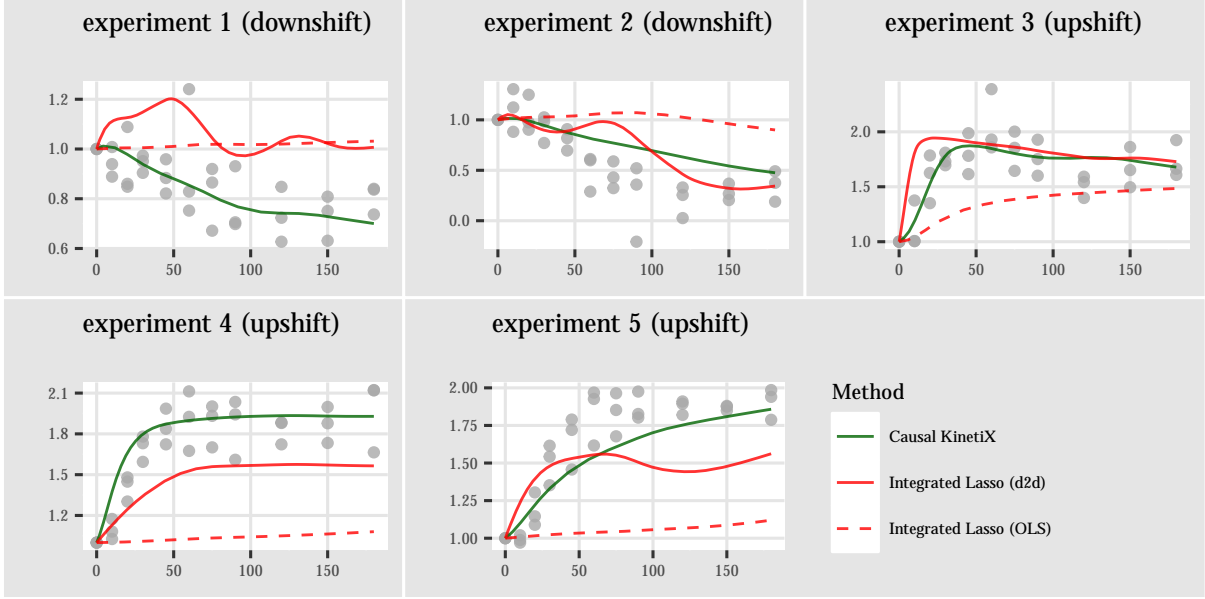


Figure 14. Out-of-sample parameter estimation for the best ranked model from the in-sample experiment. The top ranked model selected by Causal KinetiX is able to generalize well to the previously unseen experiments.

Lastly, we perform an out-of-sample experiment, where during training, the data of one environment is held out. The inferred model is then used to predict the trajectory of the target variable on that held out experiment. This task is particularly difficult since the goal is to predict the outcome of an unseen intervention setting that can deviate arbitrarily from the observed experiments. The results are presented in Figure 15. Our method is capable of predicting two experiments very well (experiment 1 and experiment 5) and one experiment fairly well (experiment 3). For the remaining two experiments the prediction is less accurate, but the predicted direction (up- or downshift) is correct. When holding out experiment 2, the high ranked out-of-sample models performed well on the remaining four in-sample experiments (not shown), but were mostly incapable of accurately generalizing to experiment 2. This indicates that the heterogeneity in the four other experiments was not sufficiently strong to learn a model that generalizes to experiment 2. Model selection by IntegratedLasso performs poorly on all held-out experiments.

Despite the lack of heterogeneity in two of these held-out experiments, the Causal KinetiX variable ranking is very robust. As three model terms were sufficient to find stable models, we applied Causal KinetiX with four terms to compute the variable ranking (see Section 3.4.3). The results are presented in Figure 16. The true causal variables, as well as the true causal model, are unknown. For illustration purposes, we indicate which of the highly ranked variables appear in the model from above which has obtained the best score when based on all five experiments. (This model was able to explain all the variation in the different experiments, see, e.g., Figure 14.)

It can be seen that many of the top ten variables from the fully out-of-sample experiments

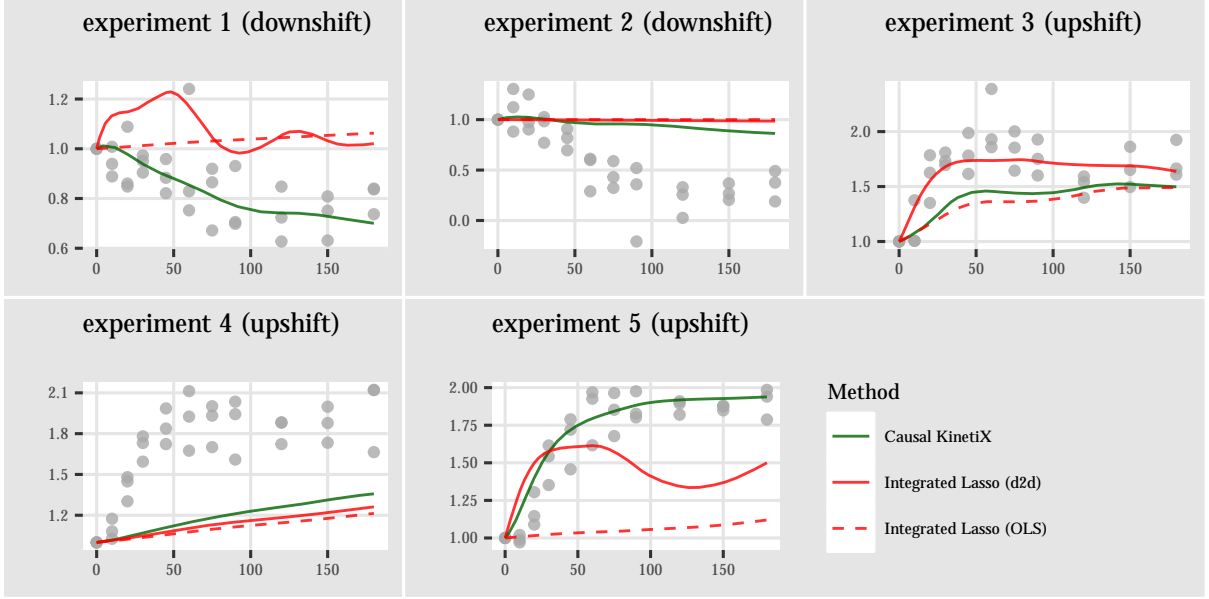


Figure 15. The plots show the ability to generalize to unseen experiments. Here, one of the experiments was entirely held out during training. In many cases, the top ranked model selected by Causal KinetiX is nevertheless able to predict the dynamics in the unseen experiments. For experiments 2 and 4 only the predicted direction (up- or downshift) is correct. See text for details.

rank	held-out-experiment				
	1	2	3	4	5
1	X^{33}	X^{168}	X^{33}	X^{33}	X^{33}
2	X^{56}	X^{231}	X^{56}	X^{168}	X^{56}
3	X^{122}	X^{59}	X^{122}	X^{138}	X^{122}
4	X^{128}	X^{246}	X^{138}	X^{60}	X^{138}
5	X^{168}	X^{33}	X^{168}	X^{61}	X^{168}
6	X^{138}	X^{373}	X^{61}	X^{128}	X^{377}
7	X^{245}	X^{56}	X^{68}	X^{347}	X^{266}
8	X^{355}	X^{122}	X^{132}	X^{73}	X^{60}
9	X^{14}	X^{190}	X^{215}	X^{122}	X^{128}
10	X^{61}	X^{206}	X^{259}	X^{190}	X^{132}

$$\dot{Y}_t = \theta_1 Z_t X_t^{56} X_t^{122} + \theta_2 Z_t X_t^{128} X_t^{168} - \theta_3 Y_t X_t^{33} X_t^{138}$$

Figure 16. Causal KinetiX variable rankings for each held out experiment (left) and best ranked model based on all five experiments (right). As the best ranked in-sample model performs well those variables can be seen as important for modeling the heterogeneity across experiments. To ease visualization, these variables have been colored green.

are variables from the top ranked model. This is a promising result, since these variables seem to play an important role in describing the dynamics of the target under all five experiments.

In summary, we consider our findings on this real data set encouraging. Our method infers a model that provides good fits when trained on the full data and when one environment is held out for parameter estimation. Compared to state-of-the art model selection techniques, our method also generalizes better to experiments that contain unseen and unknown interventions.

Even though the data may not be sufficiently heterogeneous to identify a single invariant model, we found that the variable ranking gave more robust solutions than the model ranking.

Further details including the setup of the biological experiments and a detailed analysis of the biological findings will be published in a biological, non-methodological paper.

6. Summary and conclusion

Learning dynamical systems from data is one of the core challenges in many fields. Existing approaches infer the structure of ordinary differential equations from one experiment, possibly containing data pooled from several experiments, and focus on predictive performance. Causal KinetiX proposes a new framework of causal kinetic models for identifying structure in heterogeneous experiments. The results on both simulated and real-world examples suggests that learning the structure of dynamical systems indeed benefits from taking into account invariance, rather than focusing solely on predictive performance. In situations, where there is not sufficient heterogeneity to guarantee identification of a single invariant model, the proposed variable ranking may still be used to generate causal hypotheses and interesting candidates for further investigation. Code for the proposed implementation will be made available as an open source R-package. For future benchmarking on simulated data, it will include our simulation models (most notably the Maillard reaction).

In this paper, we focus on applications in systems biology. The principle of searching for invariant models, however, may transfer to other areas of applications, too. In robotics, for example, the concept of model learning and structure search is becoming increasingly more prominent [e.g., Nguyen-Tuong and Peters, 2011, Peters et al., 2016b]. Future extensions may also include the application to stochastic, partial and delay differential equations.

Causal KinetiX is a rather general framework that can be combined with a wide range of dynamical models and any parameter inference method. It opens up a promising direction of learning causal time series models from realistic, heterogeneous datasets.

Acknowledgements

We thank Robbie Loewith, Enric Montanana Sayas, Brendan Ryback, Uwe Sauer and Jörg Stelling for providing the real biological data set as well as helpful biological insights. We further thank Niels Richard Hansen and Nicolai Meinshausen for helpful discussions, as well as Antonio Orvieto for his help with data2dynamics. This research was partially supported by the Max Planck ETH Center for Learning Systems and the SystemsX.ch project SignalX. JP was supported by a research grant (18968) from VILLUM FONDEN.

References

- T. Äijö and H. Lähdesmäki. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, 25(22):2937–2944, 2009.
- J. Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.
- T. Blom and J. M. Mooij. Generalized structural causal models. *arXiv preprint arXiv:1805.06539*, 2018.

- J. Bongard and H. Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104(24):9943–9948, 2007.
- L. Boninsegna, F. Nüske, and C. Clementi. Sparse learning of stochastic dynamical equations. *The Journal of Chemical Physics*, 148(24):241723, 2018.
- C. M. J. Brands and M. A. J. S. van Boekel. Kinetic modeling of reactions in heated monosaccharide-casein systems. *Journal of agricultural and food chemistry*, 50(23):6725–6739, 2002.
- S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 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.
- J. Casadiego, M. Nitzan, S. Hallerberg, and M. Timme. Model-free inference of direct network interactions from nonlinear collective dynamics. *Nature communications*, 8(1):2192, 2017.
- S. Chen, A. Shojaie, and D. M. Witten. Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association*, pages 1–11, 2017.
- T. Chen, H. He, and G. Church. Modeling gene expression with differential equations. In *Biocomputing’99*, pages 29–40. World Scientific, 1999.
- T. F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. Technical report, Cornell University, Ithaca, NY, USA, 1993.
- T. F. Coleman and Y. Li. On the convergence of interior-reflective newton methods for nonlinear minimization subject to bounds. *Mathematical Programming*, 67(2):189–224, 1994.
- J. P. Crutchfield and B. S. McNamara. Equation of motion from a data series. *Complex systems*, 1(417-452):121, 1987.
- 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.
- F. Dondelinger, S. Rogers, and D. Husmeier. Ode parameter inference using adaptive gradient matching with gaussian processes. In *16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.
- L. Dony, F. He, and M. Stumpf. Parametric and non-parametric gradient matching for network inference. *bioRxiv*, page 254003, 2018.
- D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 14th International Conference on Machine Learning (ICML)*, pages 1166–1174, 2013.
- D. Eaton and K. P. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 107–114, 2007.

- B. Engelhardt, H. Fröhlich, and M. Kschischo. Learning (from) the errors of a systems biology model. *Scientific reports*, 6:20772, 2016.
- K. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.
- N. S. Gorbach, S. Bauer, and J. M. Buhmann. Scalable variational inference for dynamical systems. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4806–4815, 2017.
- R. Grosse, R. R. Salakhutdinov, W. T. Freeman, and J. B. Tenenbaum. Exploiting compositionality to explore a large space of model structures. *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.
- X. Guo, Y. Zhang, W. Hu, H. Tan, and X. Wang. Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. *PloS one*, 9(2):e87446, 2014.
- T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.
- N. R. Hansen and A. Sokol. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19(100):1–24, 2014.
- J. Henderson and G. Michailidis. Network reconstruction using nonparametric additive ode models. *PloS one*, 9(4):e94003, 2014.
- A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward. SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software*, 31(3):363–396, 2005.
- G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, 2015.
- G. W. Leibniz. Nova methodus pro maximis et minimis, itemque tangentibus, quae nec fractas, nec irrationales quantitates moratur, et singulare pro illis calculi genus. *Acta Eruditorum*, pages 467–473, 1684.
- C. Li, M. Donizelli, N. Rodriguez, H. Dharuri, L. Endler, V. Chelliah, L. Li, E. He, A. Henry, M. I. Stefan, J. L. Snoep, M. Hucka, N. Le Novère, and C. Laibe. Biomodels database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC systems biology*, 4(1):92, 2010.
- L. Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.
- A. J. Lotka. Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry*, 14(3):271–274, 1909.
- 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.
- L. C. Maillard. Action des acides amines sur les sucres; formation des melanoidines par voie methodique. *Comptes rendus de l’Académie des Sciences*, 154:66–68, 1912.

- N. M. Mangan, J. N. Kutz, S. L. Brunton, and J. L. Proctor. Model selection for dynamical systems via sparse regression and information criteria. *Proceedings of the Royal Society A*, 473(2204):20170009, 2017.
- G. Martius and C. H. Lampert. Extrapolation and learning equations. *arXiv preprint arXiv:1610.02995*, 2016.
- P. Meyer. *Probability and potentials*. Blaisdell Publishing Company, 1966.
- F. V. Mikkelsen and N. R. Hansen. Learning large scale ordinary differential equation systems. *arXiv preprint arXiv:1710.09308*, 2017.
- 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.
- J. L. Natale, D. Hofmann, D. G. Hernández, and I. Nemenman. Reverse-engineering biological networks from large data sets. *arXiv preprint arXiv:1705.06370*, 2017.
- I. Newton. *Method of Fluxions*. Henry Woodfall, 1736.
- D. Nguyen-Tuong and J. Peters. Model learning for robot control: a survey. *Cognitive processing*, 12(4):319–340, 2011.
- C. Oates and S. Mukherjee. Network inference and biological dynamics. *The annals of applied statistics*, 6(3):1209, 2012.
- 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, 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, 2016a.
- J. Peters, D. D. Lee, J. Kober, D. Nguyen-Tuong, J. A. Bagnell, and S. Schaal. Robot learning. In *Springer Handbook of Robotics*, pages 357–398. Springer, 2016b.
- J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.
- N. Pfister, P. Bühlmann, and J. Peters. Invariant causal prediction for sequential data. *JASA (accepted)*, *arXiv preprint arXiv:1706.08058*, 2018.
- M. Raissi. Deep hidden physics models: Deep learning of nonlinear partial differential equations. *arXiv preprint arXiv:1801.06637*, 2018.

- M. Raissi, P. Perdikaris, and G. E. Karniadakis. Machine learning of linear differential equations using Gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017.
- J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, NY, 2005.
- 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.
- A. Regev, S. Teichmann, E. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundberg, P. Majumder, J. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe’er, A. Phillipakis, C. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. Shin, O. Stegle, M. Stratton, M. Stubbington, F. Theis, M. Uhlen, A. van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef, and Human Cell Atlas Meeting participants. Science forum: the human cell atlas. *Elife*, 6:e27041, 2017.
- S-X. Ren, G. Fu, X-G. Jiang, R. Zeng, Y-G. Miao, H. Xu, Y-X. Zhang, H. Xiong, G. Lu, L-F. Lu, H-Q. Jiang, J. Jia, Y-F. Tu, J-X. Jiang, W-Y. Gu, Y-Q. Zhang, Z. Cai, H-H. Sheng, H-F. Yin, Y. Zhang, G-F. Zhu, M. Wan, H-L. Huang, Z. Qian, S-Y. Wang, W. Ma, Z-J. Yao, Y. Shen, B-Q. Qiang, Q-C. Xia, X-K. Guo, A. Danchin, S. Girons, R. Somerville, Y-M. Wen, M-H. Shi, Z. Chen, J-G. Xu, and G-P. Zhao. Unique physiological and pathogenic features of leptospira interrogans revealed by whole-genome sequencing. *Nature*, 422(6934):888, 2003.
- J. Rozman, B. Rathkolb, M. Oestereicher, C. Schütt, A. Ravindranath, S. Leuchtenberger, S. Sharma, M. Kistler, M. Willershäuser, R. Brommage, T. Meehan, J. Mason, H. Hase-limashhadi, IMPC Consortium, T. Hough, A-M. Mallon, S. Wells, L. Santos, C. Lelliott, J. White, T. Sorg, M-F. Champy, L. Bower, C. Reynolds, A. Flenniken, S. Murray, L. Nutter, K. Svenson, D. West, G. Tocchini-Valentini, A. Beaudet, F. Bosch, R. Braun, M. Dobbie, X. Gao, Y. Herault, A. Moshiri, B. Moore, K. Lloyd, C. McKerlie, H. Masuya, N. Tanaka, P. Flicek, H. Parkinson, R. Sedlacek, J. Seong, C-K. Wang, M. Moore, S. Brown, M. Tschöp, W. Wurst, M. Klingenspor, E. Wolf, J. Beckers, F. Machicao, A. Peter, H. Staiger, H-U. Häring, H. Grallert, M. Campillos, H. Maier, H. Fuchs, V. Gailus-Durner, T. Werner, and M. Hrabe de Angelis. Identification of genetic elements in metabolism by high-throughput mouse phenotyping. *Nature communications*, 9(1):288, 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)*. AUAI Press, 2018.
- S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

- H. Schaeffer. Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A*, 473(2197):20160446, 2017.
- M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.
- 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. Sideris. *Ordinary Differential Equations and Dynamical Systems*. Atlantis Press, 2014.
- C. Siegenthaler and R. Gunawan. Assessment of network inference methods: how to cope with an underdetermined problem. *PloS one*, 9(3):e90481, 2014.
- P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.
- G. Szederkényi, J. R. Banga, and A. A. Alonso. Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC systems biology*, 5(1):177, 2011.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.
- L. Todorovski and S. Dzeroski. Declarative bias in equation discovery. In *Proceedings of the 14th International Conference on Machine Learning (ICML)*, pages 376–384, 1997.
- T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.
- G. Tran and R. Ward. Exact recovery of chaotic systems from highly corrupted data. *Multiscale Modeling & Simulation*, 15(3):1108–1129, 2017.
- 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.
- V. Vyshemirsky and M. Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2007.
- P. Waage and C. M. Guldberg. Studier over affiniteten (in Danish). *Forhandlinger i Videnskabs-selskabet i Christiania*, pages 35–45, 1864.
- T. Washio, H. Motoda, and Y. Niwa. Discovering admissible model equations from observed data based on scale-types and identity constraints. In *Proceedings of the 16th international joint conference on Artificial intelligence (IJCAI)*, pages 772–779, 1999.
- 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. *arXiv preprint arXiv:1804.04378*, 2018.
- D. J. Wilkinson. *Stochastic modelling for systems biology*. Chapman and Hall/CRC mathematical and computational biology series. Chapman & Hall/CRC, 2006.

H. Wu, T. Lu, H. Xue, and H. Liang. Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109(506): 700–716, 2014.

R. Zembowicz and J. M. Zytkow. Discovery of equations: Experimental evaluation of convergence. In *Proceedings of the 10th National Conference on Artificial intelligence (AAAI)*, 1992.

W-B. Zhang. *Differential equations, bifurcations, and chaos in economics*, volume 68. World Scientific Publishing Company, 2005.

A. Smoothing splines with derivative constraints

In this section, we show how to transform the constrained spline optimization in (15) into a quadratic program with linear constraints. For the classical spline fit in (13) just removes the linear constraint and only considers the quadratic program. Assume the function $y : [0, T] \rightarrow \mathbb{R}$ has the following form

$$y(t) = \sum_{k=1}^K c_k \varphi_k(t),$$

where $(\varphi_k)_{k \in \mathbb{N}}$ is a fixed spline basis. Moreover, define the following three matrices

$$A = \begin{pmatrix} \varphi_1(t_1) & \cdots & \varphi_K(t_1) \\ \vdots & & \vdots \\ \varphi_1(t_L) & \cdots & \varphi_K(t_L) \end{pmatrix}, \quad B = \begin{pmatrix} \dot{\varphi}_1(t_1) & \cdots & \dot{\varphi}_K(t_1) \\ \vdots & & \vdots \\ \dot{\varphi}_1(t_L) & \cdots & \dot{\varphi}_K(t_L) \end{pmatrix}$$

and

$$C = \begin{pmatrix} \int \ddot{\varphi}_1(t) \ddot{\varphi}_1(t) dt & \cdots & \int \ddot{\varphi}_1(t) \ddot{\varphi}_K(t) dt \\ \vdots & & \vdots \\ \int \ddot{\varphi}_K(t) \ddot{\varphi}_1(t) dt & \cdots & \int \ddot{\varphi}_K(t) \ddot{\varphi}_K(t) dt \end{pmatrix},$$

where the second-derivatives in the matrix C can be replaced by the desired order of derivatives used for the smoothness penalty. Then, the solution of the constrained minimization in (15) can be expressed as

$$\hat{y} = \sum_{k=1}^K \hat{c}_k \varphi_k,$$

where

$$\hat{\mathbf{c}} := \underset{\mathbf{c} \in \mathbb{R}^{K \times 1}: B\mathbf{c} = \boldsymbol{\xi}}{\operatorname{argmin}} \left(-2\tilde{\mathbf{Y}}A\mathbf{c} + \mathbf{c}^\top (A^\top A + \lambda C)\mathbf{c} \right). \quad (20)$$

The minimization in (20) can be solved using any standard quadratic program solver which allows for linear constraints.

B. Proof of Theorem 3

To simplify notation, we will whenever it is clear from the context drop the n in the grid time points $t_{n,\ell}$ and simply write t_ℓ .

B.1. Intermediate results

In order to prove Theorem 3 we require the following two auxiliary results.

Lemma 4 *Let $y_1, y_2 : [0, T] \rightarrow \mathbb{R}$ be two smooth functions satisfying that there exists $c_1 > 0$ such that*

$$\exists t^* \in [0, T] \text{ with } |\dot{y}_1(t^*) - \dot{y}_2(t^*)| \geq c_1. \quad (21)$$

Moreover, assume $c_2 := \sup_{t \in [0, T]} (|\ddot{y}_1(t)| + |\ddot{y}_2(t)|) < \infty$. Then, there exists an interval $[l_1, l_2] \subseteq [t^ - \frac{c_1}{4c_2}, t^* + \frac{c_1}{4c_2}]$ satisfying that $l_2 - l_1 = \frac{c_1}{8c_2}$ and*

$$\inf_{t \in [l_1, l_2]} |y_1(t) - y_2(t)| \geq \frac{c_1^2}{16c_2}.$$

Proof To simplify presentation, we will assume that t^* from (21) is not on the boundary of the interval $[0, T]$ and that all the intervals considered in this proof are contained in $(0, T)$. We first show that the bound on the second derivative of the functions implies that the difference in first derivatives is lower bounded on a closed interval. Using a basic derivative inequality it holds for $i \in \{1, 2\}$ and $t \in [t^*, t^* + \frac{c_1}{4c_2}]$ that

$$\dot{y}_i(t) \leq \dot{y}_i(t^*) + \frac{c_1}{4c_2} \cdot \sup_{s \in [0, T]} |\ddot{y}_i(s)| \leq \dot{y}_i(t^*) + \frac{c_1}{4}.$$

Similarly, for $i \in \{1, 2\}$ and $t \in [t^* - \frac{c_1}{4c_2}, t^*]$ it holds that

$$\dot{y}_i(t) \geq \dot{y}_i(t^*) - \frac{c_1}{4c_2} \cdot \sup_{s \in [0, T]} |\ddot{y}_i(s)| \geq \dot{y}_i(t^*) - \frac{c_1}{4}.$$

Combining these inequalities with (21) yields

$$\inf_{t \in [t^* - \frac{c_1}{4c_2}, t^* + \frac{c_1}{4c_2}]} (\dot{y}_1(t) - \dot{y}_2(t)) \cdot \text{sign}(\dot{y}_1(t^*) - \dot{y}_2(t^*)) \geq \frac{c_1}{2}. \quad (22)$$

Next, we show that this lower bound on the difference of the first derivatives implies the statement of the lemma. To this end, consider the two intervals $I_1 = [t^* - \frac{c_1}{4c_2}, t^* - \frac{c_1}{8c_2}]$ and $I_2 = [t^* + \frac{c_1}{8c_2}, t^* + \frac{c_1}{4c_2}]$. We show that at least one of the following two inequalities holds

$$(a) \inf_{t \in I_1} |y_1(t) - y_2(t)| \geq \frac{c_1^2}{16c_2},$$

$$(b) \inf_{t \in I_2} |y_1(t) - y_2(t)| \geq \frac{c_1^2}{16c_2}.$$

Assume that (a) does not hold. Then, there exists $t \in I_1$ such that

$$|y_1(t) - y_2(t)| < \frac{c_1^2}{16c_2}. \quad (23)$$

Let $s \in I_2$, then since the sign of the difference in first derivatives remains constant on the interval $[t^* - \frac{c_1}{4c_2}, t^* + \frac{c_1}{4c_2}]$ (see (22)) it holds by integration that

$$\begin{aligned} \int_t^s |\dot{y}_1(r) - \dot{y}_2(r)| dr &= [(y_1(s) - y_2(s)) - (y_1(t) - y_2(t))] \cdot \text{sign}(\dot{y}_1(t^*) - \dot{y}_2(t^*)) \\ &= |y_1(s) - y_2(s)| - |y_1(t) - y_2(t)|. \end{aligned} \quad (24)$$

By (22), it additionally holds that

$$\int_t^s |\dot{y}_1(r) - \dot{y}_2(r)| dr \geq (s-t) \frac{c_1}{2} \geq \frac{c_1}{4c_2} \frac{c_1}{2} = \frac{c_1^2}{8c_2}. \quad (25)$$

Finally, combining (23), (24) and (25) we get that

$$|y_1(s) - y_2(s)| \geq \frac{c_1^2}{16c_2},$$

which implies that (b) holds since $s \in I_2$ was arbitrary. An analogous argument can be used if we assume that (b) does not hold. Hence, since at least one of (a) and (b) holds, we have proved Lemma 4. \square

Lemma 5 *For any a smooth function $f : \mathbb{R} \rightarrow \mathbb{R}$, with $\max(\sup_t |f(t)|, \sup_t |\dot{f}(t)|, \sup_t |\ddot{f}(t)|) \leq C$, any $a > 1$, and any $r \in \mathbb{R}$, there exists a smooth function g satisfying $\dot{g}(0) = r$, $g(t) = f(t)$ for all $|t| \geq 1/a$, and*

$$\max \left(\sup_t |g(t)|, \sup_t |\dot{g}(t)|, \sup_t |\ddot{g}(t)| \right) \leq C + 16a|r - \dot{f}(0)|.$$

Proof Assume that we are given a smooth function b_a that is supported on $[-1/a, 1/a]$, and that has derivative $\dot{b}_a(0) = 1$. We can then define

$$g(t) := f(t) + (r - \dot{f}(0)) \cdot b_a(t),$$

which is equal to f outside the interval $[-1/a, 1/a]$ and which satisfies $\dot{g}(0) = r$.

Let us first create such a function b_a . To do so, define

$$b(t) := \begin{cases} \sin(t) \exp\left(1 - \frac{1}{1-t^2}\right) & \text{if } |t| < 1 \\ 0 & \text{otherwise.} \end{cases}$$

This function is smooth and satisfies $\sup_t |b(t)| \leq 1$, $\sup_t |\dot{b}(t)| \leq 1$, $\sup_t |\ddot{b}(t)| \leq 16$, and $\dot{b}(0) = 1$. We now define the function

$$b_a(t) := \frac{1}{a} b(at),$$

whose support contained in $[-1/a, 1/a]$. Because of $\dot{b}_a(t) = \dot{b}(at)$, we have $\dot{b}_a(0) = 1$. Finally, we find

$$\begin{aligned} \sup_t |\dot{g}(t)| &\leq C + |c - \dot{f}(0)| \sup_t |\dot{b}_a(t)| \leq C + |c - \dot{f}(0)| \\ \sup_t |\ddot{g}(t)| &\leq C + |c - \dot{f}(0)| 16a, \end{aligned}$$

where the last line follows from $\ddot{b}_a(t) = a\ddot{b}(at)$. This completes the proof of Lemma 5. \square

Lemma 6 *Let $((\varepsilon_{n,k})_{k \in \{1, \dots, n\}})_{n \in \mathbb{N}}$ be a triangular array of i.i.d. sub-Gaussian (with parameter ν) random variables. Moreover, assume $((X_{n,k})_{k \in \{1, \dots, n\}})_{n \in \mathbb{N}}$ is a triangular array of random variables which satisfies that*

$$\max_{k \in \{1, \dots, n\}} X_{n,k} \xrightarrow{\mathbb{P}} 0 \text{ as } n \rightarrow \infty \quad \text{and} \quad \exists K > 0 : \sup_{n \in \mathbb{N}} \max_{k \in \{1, \dots, n\}} |X_{n,k}| \leq K.$$

Then, it holds that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mathbb{E}(|X_{n,k} \varepsilon_{n,k}|) = 0.$$

Proof Fix $\delta, \theta > 0$, then by the convergence in probability it holds that there exists $N \in \mathbb{N}$ such that for all $n \in \{N, N+1, \dots\}$ it holds that

$$\mathbb{P}\left(\max_{k \in \{1, \dots, n\}} |X_{n,k}| > n\right) \leq \mathbb{P}\left(\max_{k \in \{1, \dots, n\}} |X_{n,k}| > 1\right) \leq \frac{\delta}{2}. \quad (26)$$

Furthermore, using independence, sub-Gaussianity and Bernoulli's inequality we get for all $c > 0$ and $n \in \mathbb{N}$ that

$$\begin{aligned} \mathbb{P}\left(\max_{k \in \{1, \dots, n\}} |\varepsilon_{n,k}| > c\right) &= 1 - \mathbb{P}\left(\max_{k \in \{1, \dots, n\}} |\varepsilon_{n,k}| \leq c\right) \\ &= 1 - \mathbb{P}(|\varepsilon_{n,1}| \leq c)^n \\ &= 1 - (1 - \mathbb{P}(|\varepsilon_{n,1}| > c))^n \\ &\leq 1 - \left(1 - Ce^{-\nu c^2}\right)^n \\ &\leq nCe^{-\nu c^2}. \end{aligned} \quad (27)$$

Combining (26) and (27) this proves that for all $n \in \{N, N+1, \dots\}$ it holds that

$$\begin{aligned} \mathbb{P}\left(\max_{k \in \{1, \dots, n\}} |X_{n,k}\varepsilon_{n,k}| > \theta\right) &= \mathbb{P}\left(\max_{k \in \{1, \dots, n\}} |X_{n,k}\varepsilon_{n,k}| > \theta, \max_{k \in \{1, \dots, n\}} |X_{n,k}| \leq n\right) \\ &\quad + \mathbb{P}\left(\max_{k \in \{1, \dots, n\}} |X_{n,k}\varepsilon_{n,k}| > \theta, \max_{k \in \{1, \dots, n\}} |X_{n,k}| > n\right) \\ &\leq \mathbb{P}\left(\max_{k \in \{1, \dots, n\}} |\varepsilon_{n,k}| > \frac{\theta}{n}\right) + \mathbb{P}\left(\max_{k \in \{1, \dots, n\}} |X_{n,k}| > n\right) \\ &\leq nCe^{-\nu(\frac{\theta}{n})^2} + \frac{\delta}{2}. \end{aligned}$$

Since the term $nCe^{-\nu(\frac{\theta}{n})^2}$ converges to zeros as n goes to infinity, there exists $N^* \in \{N, N+1, \dots\}$ such that for all $n \in \{N^*, N^*+1, \dots\}$ it holds that

$$nCe^{-\nu(\frac{\theta}{n})^2} \leq \frac{\delta}{2}.$$

Finally, we combine these results to show that for all $n \in \{N^*, N^*+1, \dots\}$ it holds that

$$\mathbb{P}\left(\max_{k \in \{1, \dots, n\}} |X_{n,k}\varepsilon_{n,k}| > \theta\right) \leq \delta,$$

which implies that $\max_{k \in \{1, \dots, n\}} |X_{n,k}\varepsilon_{n,k}|$ converges to zero in probability as $n \rightarrow \infty$. In particular, $\frac{1}{n} \sum_{k=1}^n |X_{k,n}\varepsilon_{k,n}|$ also converges to zero in probability as it is \mathbb{P} -a.s. dominated by $\max_{k \in \{1, \dots, n\}} |X_{n,k}\varepsilon_{n,k}|$. Furthermore, due to boundedness assumption on $X_{n,k}$ it also holds that

$$\sup_{n \in \mathbb{N}} \mathbb{E}\left(\left|\frac{1}{n} \sum_{k=1}^n X_{k,n}\varepsilon_{k,n}\right|^2\right) < \infty,$$

which by de la Vallée-Poussin's theorem [Meyer, 1966, p.19 Theorem T22] implies uniform integrability. Since uniform integrability and convergence in probability is equivalent to convergence in L^1 , this completes the proof of Lemma 6. \square

The following two lemmas are the key steps used in the proof of the Theorem 3. They prove some essential properties related to the constraint optimization, i.e., the estimation of \hat{y}_b .

Lemma 7 Consider the setting of Theorem 3, that is, let Assumption 1 and conditions (C1) and (C2) be satisfied. Additionally, assume that for all $k \in \{1, \dots, m\}$ it holds for all $i \in e_k$ and $\ell \in \{1, \dots, L_n\}$ that the noise variables $\varepsilon_{t_\ell}^{(i)}$ are i.i.d., symmetric, sub-Gaussian and satisfy $\mathbb{E}(\varepsilon_{t_\ell}^{(i)}) = 0$ and $\mathbf{var}(\varepsilon_{t_\ell}^{(i)}) = \sigma_k^2$. Let Y_t and its first and second derivative be bounded by $c < \infty$ and define $C := c + 16$ for the set \mathcal{H}_C , see (M3). Then, for an invariant model $\mathcal{G} \in \mathcal{M}$ and for all $k \in \{1, \dots, m\}$ it holds that

$$\lim_{n \rightarrow \infty} \mathbb{E} \left(\sup_{t \in [0, T]} \left(\hat{y}_b^{(e_k)}(t) - Y_t^{(e_k)} \right)^2 \right) = 0,$$

i.e., the outcome of step (M5) converges towards the true target trajectory. Furthermore, for $\mathcal{G} \in \mathcal{M}$ non-invariant there exists $k^* \in \{1, \dots, m\}$ and $c_{\min} > 0$ such that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\frac{1}{L_n} \sum_{\ell=1}^{L_n} \left(\hat{y}_b^{(e_{k^*})}(t_\ell) - Y_{t_\ell}^{(e_{k^*})} \right)^2 \geq c_{\min} \right) = 1. \quad (28)$$

Proof First, recall the definition of \mathcal{H}_C (see (M3)) and define the smoother function $\hat{y}_c^{(i)} \in \mathcal{H}_C$ corresponding to the constrained optimization based on the true derivatives, i.e.,

$$\begin{aligned} \hat{y}_c^{(i)} &:= \operatorname{argmin}_{y \in \mathcal{H}_C} \sum_{\ell=1}^L \left(\tilde{Y}_{t_\ell}^{(i)} - y(t_\ell) \right)^2 + \lambda \int \ddot{y}(s)^2 ds, \\ \text{such that } \dot{y}(t_\ell) &= \dot{Y}_{t_\ell}^{(i)} \text{ for all } \ell = 1, \dots, L_n. \end{aligned}$$

Fix $k \in \{1, \dots, m\}$. To simplify notation we will drop the superscript (e_k) in the following. Fix $\delta \in (0, 1)$ and define the sets

$$A_\delta := \left\{ L_n \max_{\ell \in \{1, \dots, L_n\}} |\hat{g}_n(\tilde{\mathbf{X}}_{t_\ell}) - \dot{Y}_{t_\ell}| \leq \delta \right\} \quad \text{and} \quad B_\delta := \left\{ \left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} \varepsilon_{t_\ell} \right| \leq \delta \right\}.$$

Then, by condition (C2) it holds that

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_\delta) = 1, \quad (29)$$

and, by the law of large numbers,

$$\lim_{n \rightarrow \infty} \mathbb{P}(B_\delta) = 1. \quad (30)$$

Note that on the set A_δ , our method is well-defined: for $a = L_n$, Lemma 5 shows us that the function \hat{y}_b exists since the corresponding optimization problem has at least one solution. Then, on the event $A_\delta \cap B_\delta$ it holds that

$$\begin{aligned} \max_{\ell \in \{1, \dots, L_n\}} |\hat{y}_b(t_\ell) - \hat{y}_c(t_\ell)| &\leq \sum_{k=1}^{L_n} \int_{t_{k-1}}^{t_k} |\dot{\hat{y}}_b(s) - \dot{\hat{y}}_c(s)| ds + |\hat{y}_b(t_1) - \hat{y}_c(t_1)| \\ &\leq L_n \max_{\ell \in \{2, \dots, L_n\}} \left(\int_{t_{\ell-1}}^{t_\ell} \frac{2C}{L_n} + |\dot{\hat{y}}_b(t_{\ell-1}) - \dot{\hat{y}}_c(t_{\ell-1})| ds \right) + |\hat{y}_b(t_1) - \hat{y}_c(t_1)| \\ &\leq \frac{2C}{L_n} + \delta + |\hat{y}_b(t_1) - \hat{y}_c(t_1)|, \end{aligned} \quad (31)$$

where the second last inequality follows from the bound on the second derivative. Moreover, define the function $y_{b*} := \hat{y}_b - \hat{y}_b(t_1) + Y_{t_1}$ then similar arguments show that

$$\max_{\ell \in \{1, \dots, L_n\}} |y_{b*}(t_\ell) - Y_{t_\ell}| = \max_{\ell \in \{1, \dots, L_n\}} |(\hat{y}_b(t_\ell) - \hat{y}_b(t_1)) - (Y_{t_\ell} - Y_{t_1})| \leq \frac{2C}{L_n} + \delta. \quad (32)$$

Using that \hat{y}_c has the true derivatives as constraint the same argument implies for $y_{c*} := \hat{y}_c - \hat{y}_c(t_1) + Y_{t_1}$ that

$$\max_{\ell \in \{1, \dots, L_n\}} |y_{c*}(t_\ell) - Y_{t_\ell}| \leq \frac{2C}{L_n}. \quad (33)$$

Next, define the loss function

$$\text{loss}_n(y) := \sum_{\ell=1}^{L_n} \left(\tilde{Y}_{t_\ell} - y(t_\ell) \right)^2 + \lambda_n \int_0^T \ddot{y}(s)^2 ds.$$

Then using (32) and (33) it holds that

$$\begin{aligned} \text{loss}_n(\hat{y}_b) &= \sum_{\ell=1}^{L_n} \left(\tilde{Y}_{t_\ell} - \hat{y}_b(t_\ell) \right)^2 + \lambda_n \int_0^T \ddot{\hat{y}}_b(s)^2 ds \\ &= \text{loss}_n(y_{b*}) + \sum_{\ell=1}^{L_n} (Y_{t_1} - \hat{y}_b(t_1))^2 + 2(Y_{t_1} - \hat{y}_b(t_1)) \sum_{\ell=1}^{L_n} \left(\tilde{Y}_{t_\ell} - y_{b*}(t_\ell) \right) \\ &\geq \text{loss}_n(y_{b*}) + L_n (Y_{t_1} - \hat{y}_b(t_1))^2 + 2|Y_{t_1} - \hat{y}_b(t_1)| L_n \left(\frac{2C}{L_n^2} + \frac{\delta}{L_n} \right) + 2(Y_{t_1} - \hat{y}_b(t_1)) \sum_{\ell=1}^{L_n} \varepsilon_{t_\ell} \end{aligned}$$

Now, y_{b*} has the same derivatives as \hat{y}_b and since \hat{y}_b minimizes loss_n under fixed derivative constraints it holds that $\text{loss}_n(\hat{y}_b) \leq \text{loss}_n(y_{b*})$. This implies

$$L_n (Y_{t_1} - \hat{y}_b(t_1))^2 \leq 2|Y_{t_1} - \hat{y}_b(t_1)| L_n \left(\frac{2C}{L_n} + \delta \right) + 2(Y_{t_1} - \hat{y}_b(t_1)) \sum_{\ell=1}^{L_n} \varepsilon_{t_\ell}, \quad (34)$$

which is equivalent to

$$|Y_{t_1} - \hat{y}_b(t_1)| \leq 2 \cdot \left(\frac{2C}{L_n} + \delta \right) + 2 \left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} \varepsilon_{t_\ell} \right|. \quad (35)$$

Since, we are on the set B_δ this in particular as $n \rightarrow \infty$ implies that

$$\limsup_{n \rightarrow \infty} |Y_{t_1} - \hat{y}_b(t_1)| \leq 4\delta. \quad (36)$$

With the same arguments as in (34) and (35) for the function \hat{y}_c we get that

$$\limsup_{n \rightarrow \infty} |Y_{t_1} - \hat{y}_c(t_1)| \leq 2\delta. \quad (37)$$

Combining (36) and (37) with the triangle inequality it holds that

$$\limsup_{n \rightarrow \infty} |\hat{y}_b(t_1) - \hat{y}_c(t_1)| \leq 6\delta. \quad (38)$$

Hence, we can combine this with (31) to get that

$$\limsup_{n \rightarrow \infty} \max_{\ell \in \{1, \dots, L_n\}} |\hat{y}_b(t_\ell) - \hat{y}_c(t_\ell)| \leq 7\delta,$$

which together with the global bound on the first derivative also implies that

$$\limsup_{n \rightarrow \infty} \sup_{t \in [0, T]} |\hat{y}_b(t) - \hat{y}_c(t)| \leq \limsup_{n \rightarrow \infty} \left(\max_{\ell \in \{1, \dots, L_n\}} |\hat{y}_b(t_\ell) - \hat{y}_c(t_\ell)| + \frac{C}{L_n} \right) \leq 7\delta.$$

Finally, we use this, the global bound and the dominated convergence theorem to show that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \mathbb{E} \left(\sup_{t \in [0, T]} \left(\hat{y}_b^{(e_k)}(t) - Y_t^{(e_k)} \right)^2 \right) \\ &= \lim_{n \rightarrow \infty} \left(\mathbb{E} \left(\sup_{t \in [0, T]} \left(\hat{y}_b^{(e_k)}(t) - Y_t^{(e_k)} \right)^2 \mathbb{1}_{A_\delta \cap B_\delta} \right) + \mathbb{E} \left(\sup_{t \in [0, T]} \left(\hat{y}_b^{(e_k)}(t) - Y_t^{(e_k)} \right)^2 \mathbb{1}_{A_\delta^c \cup B_\delta^c} \right) \right) \\ &\leq 7\delta + \lim_{n \rightarrow \infty} \mathbb{P}(A_\delta^c \cup B_\delta^c) = 7\delta. \end{aligned}$$

Since $\delta > 0$ was arbitrary this proves the first part of the lemma.

Next, we show the second part. To that end, let $\mathcal{G} \in \mathcal{M}$ be non-invariant. Since we assumed that the set $\{t \mapsto g(X_t) \mid g \in \mathcal{G}\}$ is closed with respect to the sup norm there exist $c > 0$, $k^* \in \{1, \dots, m\}$ and $(t_n^*)_{n \in \mathbb{N}} \subseteq [0, T]$ such that for all $n \in \mathbb{N}$ it holds that

$$|\dot{Y}_{t_n^*}^{(e_{k^*})} - \hat{g}_n(\mathbf{X}_{t_n^*}^{(e_{k^*})})| \geq c. \quad (39)$$

Next, define $\ell_n^* := \operatorname{argmin}_{\ell \in \{1, \dots, L_n\}} |t_n^* - t_\ell|$ then by the derivative constraint it in particular holds that $\dot{\hat{y}}_b^{(e_{k^*})}(t_{\ell_n^*}^*) = \hat{g}_n(\tilde{\mathbf{X}}_{t_{\ell_n^*}^*}^{(e_{k^*})})$. Moreover, using the global bound from the function class \mathcal{H}_C it holds that

$$\begin{aligned} & |\hat{g}_n(\mathbf{X}_{t_n^*}^{(e_{k^*})}) - \dot{\hat{y}}_b^{(e_{k^*})}(t_n^*)| \\ &\leq |\hat{g}_n(\mathbf{X}_{t_n^*}^{(e_{k^*})}) - \hat{g}_n(\tilde{\mathbf{X}}_{t_{\ell_n^*}^*}^{(e_{k^*})})| + |\dot{\hat{y}}_b^{(e_{k^*})}(t_n^*) - \dot{\hat{y}}_b^{(e_{k^*})}(t_{\ell_n^*}^*)| \\ &\leq |\hat{g}_n(\mathbf{X}_{t_n^*}^{(e_{k^*})}) - g(\mathbf{X}_{t_n^*}^{(e_{k^*})})| + |\hat{g}_n(\tilde{\mathbf{X}}_{t_{\ell_n^*}^*}^{(e_{k^*})}) - g(\mathbf{X}_{t_{\ell_n^*}^*}^{(e_{k^*})})| + |g(\mathbf{X}_{t_n^*}^{(e_{k^*})}) - g(\mathbf{X}_{t_{\ell_n^*}^*}^{(e_{k^*})})| + \frac{C}{L_n} \\ &\leq |\hat{g}_n(\mathbf{X}_{t_n^*}^{(e_{k^*})}) - g(\mathbf{X}_{t_n^*}^{(e_{k^*})})| + |\hat{g}_n(\tilde{\mathbf{X}}_{t_{\ell_n^*}^*}^{(e_{k^*})}) - g(\mathbf{X}_{t_{\ell_n^*}^*}^{(e_{k^*})})| + \frac{2C}{L_n}. \end{aligned} \quad (40)$$

Combining the bounds in (39) and (40) implies that

$$\begin{aligned} |\dot{Y}_{t_n^*}^{(e_{k^*})} - \dot{\hat{y}}_b^{(e_{k^*})}(t_n^*)| &\geq |\dot{Y}_{t_n^*}^{(e_{k^*})} - \hat{g}_n(\mathbf{X}_{t_n^*}^{(e_{k^*})})| - |\hat{g}_n(\mathbf{X}_{t_n^*}^{(e_{k^*})}) - \dot{\hat{y}}_b^{(e_{k^*})}(t_n^*)| \\ &\geq c - |\hat{g}_n(\mathbf{X}_{t_n^*}^{(e_{k^*})}) - g(\mathbf{X}_{t_n^*}^{(e_{k^*})})| - |\hat{g}_n(\tilde{\mathbf{X}}_{t_{\ell_n^*}^*}^{(e_{k^*})}) - g(\mathbf{X}_{t_{\ell_n^*}^*}^{(e_{k^*})})| - \frac{2C}{L_n}. \end{aligned}$$

Next, assume $n \in \mathbb{N}$ is large enough such that $c - \frac{2C}{L_n} > 0$ and define for $\delta \in (0, c - \frac{2C}{L_n})$ the event $C_\delta := \{|\hat{g}_n(\mathbf{X}_{t_n^*}^{(e_{k^*})}) - g(\mathbf{X}_{t_n^*}^{(e_{k^*})})| + |\hat{g}_n(\tilde{\mathbf{X}}_{t_{\ell_n^*}^*}^{(e_{k^*})}) - g(\mathbf{X}_{t_{\ell_n^*}^*}^{(e_{k^*})})| \leq \delta\}$ (which depends on n). Then on C_δ it holds by Lemma 4 that there exist intervals $[l_{1,n}, l_{2,n}] \subseteq [0, T]$ with length strictly greater than a fixed constant (independent of n) and a constant $\mu > 0$ (also independent of n) satisfying that

$$\inf_{t \in [l_{1,n}, l_{2,n}]} |Y_t^{(e_{k^*})} - \hat{y}_b^{(e_{k^*})}(t)| \geq \mu.$$

Since we assumed an equally spaced grid it is clear that at least $\lfloor \frac{l_{n,2}-l_{1,n}}{T}n \rfloor$ grid points are contained in the interval $[l_{1,n}, l_{2,n}]$. Hence, defining $c_{\min} := \mu^2$ we get

$$\begin{aligned}
& \liminf_{n \rightarrow \infty} \mathbb{P} \left(\frac{1}{L_n} \sum_{\ell=1}^{L_n} \left(Y_{t_\ell}^{(e_{k^*})} - \hat{y}_b^{(e_{k^*})}(t_\ell) \right)^2 \geq c_{\min} \right) \\
& \geq \liminf_{n \rightarrow \infty} \mathbb{P} \left(\left\lfloor \frac{l_{2,n}-l_{1,n}}{T}n \right\rfloor \inf_{t \in [l_{1,n}, l_{2,n}]} |Y_t^{(e_{k^*})} - \hat{y}_b^{(e_{k^*})}(t)|^2 \geq c_{\min} \right) \\
& \geq \liminf_{n \rightarrow \infty} \mathbb{P} \left(\left\{ \left\lfloor \frac{l_{2,n}-l_{1,n}}{T}n \right\rfloor \inf_{t \in [l_{1,n}, l_{2,n}]} |Y_t^{(e_{k^*})} - \hat{y}_b^{(e_{k^*})}(t)|^2 \geq c_{\min} \right\} \cap C_\delta \right) \\
& \geq \liminf_{n \rightarrow \infty} \mathbb{P} \left(\left\{ \left\lfloor \frac{l_{2,n}-l_{1,n}}{T}n \right\rfloor \mu^2 \geq c_{\min} \right\} \cap C_\delta \right) \\
& = \liminf_{n \rightarrow \infty} \mathbb{P}(C_\delta) = 1,
\end{aligned}$$

where in the last step we used the second part of condition (C2). This completes the proof of Lemma 7. \square

Simply stated the following lemma proves that under condition (C2) it holds that for non-invariant $\mathcal{G} \in \mathcal{M}$ the estimates \hat{y}_b corresponding to the constraint optimization converge to a fixed function y_{\lim} . The function $y_{\lim}(\cdot)$ can be explicitly constructed as the integral of the derivative function $g(X_\cdot)$ shifted by a fixed constant that is chosen to minimize the area between $y_{\lim}(\cdot)$ and the true function Y .

Lemma 8 *Let condition (C2) be satisfied. Additionally, assume that for all $k \in \{1, \dots, m\}$ it holds for all $i \in e_k$ and $\ell \in \{1, \dots, L_n\}$ that the noise variables $\varepsilon_{t_\ell}^{(i)}$ are i.i.d., symmetric, sub-Gaussian and satisfy $\mathbb{E}(\varepsilon_{t_\ell}^{(i)}) = 0$ and $\mathbf{var}(\varepsilon_{t_\ell}^{(i)}) = \sigma_k^2$. Let Y_t and its first and second derivative be bounded by $c < \infty$ and define $C := c + 16$ for the set \mathcal{H}_C , see (M3). Then, for any non-invariant $\mathcal{G} \in \mathcal{M}$ with $g \in \mathcal{G}$ the limit function from condition (C2) it holds that for all $k \in \{1, \dots, m\}$ the functions $y_{\lim}^{(e_k)} : [0, T] \rightarrow \mathbb{R}$ defined for all $t \in [0, T]$ by*

$$y_{\lim}^{(e_k)}(t) := \int_0^t g(X_s^{(e_k)}) ds + \frac{1}{T} \int_0^T \left(Y_s^{(e_k)} - \int_0^s g(X_r^{(e_k)}) dr \right) ds,$$

satisfy that

$$\sup_{t \in [0, T]} |\hat{y}_b^{(e_k)}(t) - y_{\lim}^{(e_k)}(t)| \xrightarrow{\mathbb{P}} 0,$$

as $n \rightarrow \infty$.

Proof The proof is very similar in spirit to the proof of the second part of Lemma 7. Let $\mathcal{G} \in \mathcal{M}$ be non-invariant, fix $k \in \{1, \dots, m\}$ and let $g \in \mathcal{G}$ be the function from the second part of condition (C2). To simplify notation we will drop the superscript (e_k) in the remainder of this proof. Next, let $\delta \in (0, 1)$ and define the sets

$$A_\delta := \left\{ L_n \max_{\ell \in \{1, \dots, L_n\}} |\hat{g}_n(\tilde{\mathbf{X}}_{t_\ell}) - g(\mathbf{X}_{t_\ell})| \leq \delta \right\} \quad \text{and} \quad B_\delta := \left\{ \left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} \varepsilon_{t_\ell} \right| \leq \delta \right\}.$$

Then, by condition (C2) it holds that

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_\delta) = 1, \tag{41}$$

and, by the law of large numbers,

$$\lim_{n \rightarrow \infty} \mathbb{P}(B_\delta) = 1. \quad (42)$$

Note that on the set A_δ , our method is well-defined: for $a = L_n$, Lemma 5 shows us that the function \hat{y}_b exists since the corresponding optimization problem has at least one solution. Then, on the event $A_\delta \cap B_\delta$ it holds that

$$\begin{aligned} \max_{\ell \in \{1, \dots, L_n\}} |\hat{y}_b(t_\ell) - y_{\lim}(t_\ell)| &\leq \sum_{k=1}^{L_n} \int_{t_{k-1}}^{t_k} |\dot{\hat{y}}_b(s) - \dot{y}_{\lim}(s)| ds + |\hat{y}_b(t_1) - y_{\lim}(t_1)| \\ &\leq L_n \max_{\ell \in \{2, \dots, L_n\}} \left(\int_{t_{\ell-1}}^{t_\ell} \frac{2C}{L_n} + |\dot{\hat{y}}_b(t_{\ell-1}) - \dot{y}_{\lim}(t_{\ell-1})| ds \right) + |\hat{y}_b(t_1) - y_{\lim}(t_1)| \\ &\leq \frac{2C}{L_n} + \delta + |\hat{y}_b(t_1) - y_{\lim}(t_1)|, \end{aligned} \quad (43)$$

where we used that $\dot{y}_{\lim}(t) = g(X_t)$. Moreover, define the function $y_{b*} := \hat{y}_b - \hat{y}_b(t_1) + y_{\lim}(t_1)$ then similar arguments show that

$$\max_{\ell \in \{1, \dots, L_n\}} |y_{b*}(t_\ell) - y_{\lim}(t_\ell)| = \max_{\ell \in \{1, \dots, L_n\}} |(\hat{y}_b(t_\ell) - \hat{y}_b(t_1)) - (y_{\lim}(t_\ell) - y_{\lim}(t_1))| \leq \frac{2C}{L_n} + \delta. \quad (44)$$

Next, define the loss function

$$\text{loss}_n(y) := \sum_{\ell=1}^{L_n} \left(\tilde{Y}_{t_\ell} - y(t_\ell) \right)^2 + \lambda_n \int_0^T \ddot{y}(s)^2 ds.$$

Moreover, it holds that

$$\begin{aligned} \text{loss}_n(\hat{y}_b) &= \sum_{\ell=1}^{L_n} \left(\tilde{Y}_{t_\ell} - \hat{y}_b(t_\ell) \right)^2 + \lambda_n \int_0^T \ddot{\hat{y}}_b(s)^2 ds \\ &= \text{loss}_n(y_{b*}) + \sum_{\ell=1}^{L_n} (y_{\lim}(t_1) - \hat{y}_b(t_1))^2 + 2(y_{\lim}(t_1) - \hat{y}_b(t_1)) \sum_{\ell=1}^{L_n} (\tilde{Y}_{t_\ell} - y_{b*}(t_\ell)) \\ &= \text{loss}_n(y_{b*}) + L_n (y_{\lim}(t_1) - \hat{y}_b(t_1))^2 + 2(y_{\lim}(t_1) - \hat{y}_b(t_1)) \left[\sum_{\ell=1}^{L_n} \varepsilon_{t_\ell} + \sum_{\ell=1}^{L_n} (Y_{t_\ell} - y_{b*}(t_\ell)) \right] \end{aligned}$$

Now, y_{b*} has the same derivatives as \hat{y}_b and since \hat{y}_b minimizes loss_n under fixed derivative constraints it holds that $\text{loss}_n(\hat{y}_b) \leq \text{loss}_n(y_{b*})$. This implies

$$L_n (y_{\lim}(t_1) - \hat{y}_b(t_1))^2 \leq 2(y_{\lim}(t_1) - \hat{y}_b(t_1)) \left[\sum_{\ell=1}^{L_n} \varepsilon_{t_\ell} + \sum_{\ell=1}^{L_n} (Y_{t_\ell} - y_{b*}(t_\ell)) \right], \quad (45)$$

which further implies

$$|y_{\lim}(t_1) - \hat{y}_b(t_1)| \leq 2 \cdot \left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} \varepsilon_{t_\ell} + \frac{1}{L_n} \sum_{\ell=1}^{L_n} (Y_{t_\ell} - y_{b*}(t_\ell)) \right|. \quad (46)$$

Firstly, since we are on the set B_δ we get that

$$\left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} \varepsilon_{t_\ell} \right| \leq \delta. \quad (47)$$

Secondly, using (44) and the definition of the Riemann integral we get that

$$\begin{aligned}
& \limsup_{n \rightarrow \infty} \left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} (Y_{t_\ell} - y_{b*}(t_\ell)) \right| \\
& \leq \limsup_{n \rightarrow \infty} \left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} (Y_{t_\ell} - y_{\lim}(t_\ell)) \right| + \limsup_{n \rightarrow \infty} \left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} (y_{b*}(t_\ell) - y_{\lim}(t_\ell)) \right| \\
& \leq \left| \int_0^T (Y_s - y_{\lim}(s)) ds \right| + \delta \\
& = \delta,
\end{aligned} \tag{48}$$

where in the last step we used the definition of the function y_{\lim} . Hence, combining (46) with (47) and (48) we get that

$$\limsup_{n \rightarrow \infty} |y_{\lim}(t_1) - \hat{y}_b(t_1)| \leq 4\delta. \tag{49}$$

Furthermore, we can combine this with (43) to get that

$$\limsup_{n \rightarrow \infty} \max_{\ell \in \{1, \dots, L_n\}} |\hat{y}_b(t_\ell) - y_{\lim}(t_\ell)| \leq 5\delta,$$

which together with the global bound on the first derivative also implies that

$$\limsup_{n \rightarrow \infty} \sup_{t \in [0, T]} |\hat{y}_b(t) - y_{\lim}(t)| \leq \limsup_{n \rightarrow \infty} \left(\max_{\ell \in \{1, \dots, L_n\}} |\hat{y}_b(t_\ell) - y_{\lim}(t_\ell)| + \frac{C}{L_n} \right) \leq 5\delta.$$

Since $\delta \in (0, 1)$ was arbitrary this proves that $\sup_{t \in [0, T]} |\hat{y}_b(t) - y_{\lim}(t)|$ converges in probability to zero, which completes the proof of Lemma 8. \square

B.2. Proof of theorem

Proof Assume that Y_t and its first and second derivative be bounded by $c < \infty$ and define $C := c + 16$ for the set \mathcal{H}_C , see (M3). The proof of Theorem 3 consists of two parts. First we assume that the following two claims are true and show that they suffice in proving the result. Afterwards, we prove both claims.

Claim 1: For all invariant $\mathcal{G} \in \mathcal{M}$ it holds that

$$\lim_{n \rightarrow \infty} \mathbb{E}(T_n^{\mathcal{G}}) = 0$$

Claim 2: There exists a $c > 0$ such that for all non-invariant $\mathcal{G} \in \mathcal{M}$ it holds that

$$\liminf_{n \rightarrow \infty} \mathbb{E}(T_n^{\mathcal{G}}) \geq c.$$

Combining both claims and using Markov's inequality we get that

$$\begin{aligned}
& \lim_{n \rightarrow \infty} \mathbb{E} \left(\left| \{ \mathcal{G} \in \mathcal{M} \mid T_n^{\mathcal{G}} < \max_{\{\tilde{\mathcal{G}} \in \mathcal{M} \mid \tilde{\mathcal{G}} \text{ invariant}\}} T_n^{\tilde{\mathcal{G}}} \text{ and } \mathcal{G} \text{ not invariant} \} \right| \right) \\
&= \lim_{n \rightarrow \infty} \sum_{\substack{\mathcal{G} \in \mathcal{M}: \\ \mathcal{G} \text{ not invariant}}} \mathbb{E} \left(\mathbf{1}_{\{T_n^{\mathcal{G}} < \max_{\{\tilde{\mathcal{G}} \in \mathcal{M} \mid \tilde{\mathcal{G}} \text{ invariant}\}} T_n^{\tilde{\mathcal{G}}}\}} \right) \\
&= \sum_{\substack{\mathcal{G} \in \mathcal{M}: \\ \mathcal{G} \text{ not invariant}}} \lim_{n \rightarrow \infty} \mathbb{P} \left(T_n^{\mathcal{G}} < \max_{\{\tilde{\mathcal{G}} \in \mathcal{M} \mid \tilde{\mathcal{G}} \text{ invariant}\}} T_n^{\tilde{\mathcal{G}}} \right) \\
&= \sum_{\substack{\mathcal{G} \in \mathcal{M}: \\ \mathcal{G} \text{ not invariant}}} \lim_{n \rightarrow \infty} \mathbb{P} \left(\mathbb{E}(T_n^{\mathcal{G}}) < \max_{\{\tilde{\mathcal{G}} \in \mathcal{M} \mid \tilde{\mathcal{G}} \text{ invariant}\}} T_n^{\tilde{\mathcal{G}}} - T_n^{\mathcal{G}} + \mathbb{E}(T_n^{\mathcal{G}}) \right) \\
&\leq \sum_{\substack{\mathcal{G} \in \mathcal{M}: \\ \mathcal{G} \text{ not invariant}}} \lim_{n \rightarrow \infty} \mathbb{P} \left(\mathbb{E}(T_n^{\mathcal{G}}) < \left| \max_{\{\tilde{\mathcal{G}} \in \mathcal{M} \mid \tilde{\mathcal{G}} \text{ invariant}\}} T_n^{\tilde{\mathcal{G}}} - T_n^{\mathcal{G}} + \mathbb{E}(T_n^{\mathcal{G}}) \right| \right) \\
&\stackrel{\text{Markov}}{\leq} \sum_{\substack{\mathcal{G} \in \mathcal{M}: \\ \mathcal{G} \text{ not invariant}}} \lim_{n \rightarrow \infty} \frac{\mathbb{E} \left(\left| \max_{\{\tilde{\mathcal{G}} \in \mathcal{M} \mid \tilde{\mathcal{G}} \text{ invariant}\}} T_n^{\tilde{\mathcal{G}}} - T_n^{\mathcal{G}} + \mathbb{E}(T_n^{\mathcal{G}}) \right| \right)}{\mathbb{E}(T_n^{\mathcal{G}})} \\
&\leq \sum_{\substack{\mathcal{G} \in \mathcal{M}: \\ \mathcal{G} \text{ not invariant}}} \lim_{n \rightarrow \infty} \frac{\mathbb{E} \left(\left| \max_{\{\tilde{\mathcal{G}} \in \mathcal{M} \mid \tilde{\mathcal{G}} \text{ invariant}\}} T_n^{\tilde{\mathcal{G}}} \right| \right) + \mathbb{E}(|T_n^{\mathcal{G}} - \mathbb{E}(T_n^{\mathcal{G}})|)}{\mathbb{E}(T_n^{\mathcal{G}})} \\
&\stackrel{\text{claim 2}}{\leq} \sum_{\substack{\mathcal{G} \in \mathcal{M}: \\ \mathcal{G} \text{ not invariant}}} \lim_{n \rightarrow \infty} \frac{\mathbb{E} \left(\left| \max_{\{\tilde{\mathcal{G}} \in \mathcal{M} \mid \tilde{\mathcal{G}} \text{ invariant}\}} T_n^{\tilde{\mathcal{G}}} \right| \right) + \mathbb{E}(|T_n^{\mathcal{G}} - \mathbb{E}(T_n^{\mathcal{G}})|)}{c} \\
&\stackrel{\text{claim 1}}{=} 0,
\end{aligned}$$

which proves that $\lim_{n \rightarrow \infty} \mathbb{E}(\text{RankAccuracy}_n) = 1$. This result also proves the second part of Theorem 3. In the limit of infinitely many data points, any invariant model depends on all variables in S^* (otherwise the set S^* would not be unique, see (C3)). Each variable $j \in S^*$ therefore receives a score of one. On the other hand, any variable $j \notin S^*$ receives a score less or equal to $(K-1)/K$ since there exists at least one invariant model, namely the pair $S^*, g^*(\mathbf{x}^{S^*})$ that does not depend on variable j .

It therefore remains to prove claim 1 and claim 2.

Proof of claim 1: Let $\mathcal{G} \in \mathcal{M}$ be invariant and fix $k \in \{1, \dots, m\}$. In the remainder of this proof, the residual sum of square terms $\text{RSS}_a^{(e_k)}$ and $\text{RSS}_b^{(e_k)}$ depend on n , which will not be

reflected in our notation. First, observe that the triangle inequality implies that

$$\begin{aligned}
& \mathbb{E} \left(|\text{RSS}_b^{(e_k)} - \text{RSS}_a^{(e_k)}| \right) \\
& \leq \frac{1}{L_n} \sum_{\ell=1}^{L_n} \mathbb{E} \left(|(\hat{y}_b^{(e_k)}(t_\ell) - \tilde{Y}_{t_\ell}^{(e_k)})^2 - (\hat{y}_a^{(e_k)}(t_\ell) - \tilde{Y}_{t_\ell}^{(e_k)})^2| \right) \\
& = \frac{1}{L_n} \sum_{\ell=1}^{L_n} \mathbb{E} \left(|(\hat{y}_b^{(e_k)}(t_\ell) - \hat{y}_a^{(e_k)}(t_\ell))(\hat{y}_b^{(e_k)}(t_\ell) + \hat{y}_a^{(e_k)}(t_\ell) - 2\tilde{Y}_{t_\ell}^{(e_k)})| \right) \\
& = \frac{1}{L_n} \sum_{\ell=1}^{L_n} \mathbb{E} \left(|[(\hat{y}_b^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)}) - (\hat{y}_a^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)})][(\hat{y}_b^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)}) + (\hat{y}_a^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)}) - 2\varepsilon_{t_\ell}^{(e_k)}]| \right) \\
& \leq \frac{1}{L_n} \sum_{\ell=1}^{L_n} [A(t_\ell, k) + B(t_\ell, k) + C(t_\ell, k) + D(t_\ell, k) + E(t_\ell, k)], \tag{50}
\end{aligned}$$

where we used the following definitions

$$\begin{aligned}
A(t_\ell, k) &:= \mathbb{E} \left((\hat{y}_b^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)})^2 \right) \\
B(t_\ell, k) &:= \mathbb{E} \left((\hat{y}_a^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)})^2 \right) \\
C(t_\ell, k) &:= 2\mathbb{E} \left(|(\hat{y}_b^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)})(\hat{y}_a^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)})| \right) \\
D(t_\ell, k) &:= 2\mathbb{E} \left(|(\hat{y}_b^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)})\varepsilon_{t_\ell}^{(e_k)}| \right) \\
E(t_\ell, k) &:= 2\mathbb{E} \left(|(\hat{y}_a^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)})\varepsilon_{t_\ell}^{(e_k)}| \right).
\end{aligned}$$

First, it holds that

$$\lim_{n \rightarrow \infty} \frac{1}{L_n} \sum_{\ell=1}^{L_n} A(t_\ell, k) = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{1}{L_n} \sum_{\ell=1}^{L_n} B(t_\ell, k) = 0, \tag{51}$$

where the first statement holds by the first part of Lemma 7 and the second by condition (C1). Together with the fact that the functions $\hat{y}_a^{(e_k)} \in \mathcal{H}_C$, $\hat{y}_b^{(e_k)} \in \mathcal{H}_C$ and $Y^{(e_k)} \in \mathcal{H}_C$ it holds \mathbb{P} -a.s. that

$$\sup_{n \in \mathbb{N}} \sup_{t \in [0, T]} |\hat{y}_a^{(e_k)}(t) - Y_t^{(e_k)}| \leq 2C \quad \text{and} \quad \sup_{n \in \mathbb{N}} \sup_{t \in [0, T]} |\hat{y}_b^{(e_k)}(t) - Y_t^{(e_k)}| \leq 2C. \tag{52}$$

Using the second statement together with condition (C1) we get that

$$\begin{aligned}
\lim_{n \rightarrow \infty} \frac{1}{L_n} \sum_{\ell=1}^{L_n} C(t_\ell, k) &= \lim_{n \rightarrow \infty} \frac{1}{L_n} \sum_{\ell=1}^{L_n} 2\mathbb{E} \left(|(\hat{y}_b^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)})(\hat{y}_a^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)})| \right) \\
&\leq 4C \cdot \lim_{n \rightarrow \infty} \frac{1}{L_n} \sum_{\ell=1}^{L_n} \mathbb{E} \left(|\hat{y}_a^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)}| \right) \\
&= 0.
\end{aligned}$$

Using both bounds in (52), condition (C1) and Lemma 7 we can apply Lemma 6 to get that

$$\lim_{n \rightarrow \infty} \frac{1}{L_n} \sum_{\ell=1}^{L_n} D(t_\ell, k) = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{1}{L_n} \sum_{\ell=1}^{L_n} E(t_\ell, k) = 0. \tag{53}$$

Hence, by taking the limit of (50), we have shown that

$$\lim_{n \rightarrow \infty} \mathbb{E} \left(|\text{RSS}_b^{(e_k)} - \text{RSS}_a^{(e_k)}| \right) = 0. \quad (54)$$

Moreover, we can make the following decomposition.

$$\begin{aligned} \mathbb{E} \left(\text{RSS}_a^{(e_k)} \right) &= \frac{1}{L_n} \sum_{\ell=1}^{L_n} \mathbb{E} \left(\left(\hat{y}_a^{(e_k)}(t_\ell) - \tilde{Y}_{t_\ell}^{(e_k)} \right)^2 \right) \\ &= \frac{1}{L_n} \sum_{\ell=1}^{L_n} \mathbb{E} \left(\left(\hat{y}_a^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)} + Y_{t_\ell}^{(e_k)} - \tilde{Y}_{t_\ell}^{(e_k)} \right)^2 \right) \\ &= \frac{1}{L_n} \sum_{\ell=1}^{L_n} B(t_\ell, k) + \frac{1}{L_n} \sum_{\ell=1}^{L_n} \mathbb{E} \left((\varepsilon_{t_\ell}^{(e_k)})^2 \right) + \frac{2}{L_n} \sum_{\ell=1}^{L_n} \mathbb{E} \left((\hat{y}_a^{(e_k)}(t_\ell) - Y_{t_\ell}^{(e_k)}) \varepsilon_{t_\ell}^{(e_k)} \right). \end{aligned} \quad (55)$$

Using (51) and (53) and taking the limit of (55) it holds that

$$\lim_{n \rightarrow \infty} \mathbb{E} \left(\text{RSS}_a^{(e_k)} \right) = \sigma_k^2. \quad (56)$$

Combining (54) and (56) with Slutsky's theorem this shows that $\frac{|\text{RSS}_b^{(e_k)} - \text{RSS}_a^{(e_k)}|}{\text{RSS}_a^{(e_k)}} \xrightarrow{\mathbb{P}} 0$ as $n \rightarrow \infty$. By (52) and (56) it also holds that

$$\sup_{n \in \mathbb{N}} \mathbb{E} \left(\left(\frac{|\text{RSS}_b^{(e_k)} - \text{RSS}_a^{(e_k)}|}{\text{RSS}_a^{(e_k)}} \right)^2 \right) < \infty,$$

which together with de la Vallée-Poussin's theorem [Meyer, 1966, p.19 Theorem T22] implies uniform integrability and thus L^1 convergence, i.e.,

$$\lim_{n \rightarrow \infty} \mathbb{E} \left(\frac{|\text{RSS}_b^{(e_k)} - \text{RSS}_a^{(e_k)}|}{\text{RSS}_a^{(e_k)}} \right) = 0.$$

Finally, since the number of environments m is fixed and it holds for all $i \in e_k$ that

$$\frac{|\text{RSS}_b^{(i)} - \text{RSS}_a^{(i)}|}{\text{RSS}_a^{(i)}} \stackrel{d}{=} \frac{|\text{RSS}_b^{(e_k)} - \text{RSS}_a^{(e_k)}|}{\text{RSS}_a^{(e_k)}}, \quad (57)$$

it holds that

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E} (T_n^{\mathcal{G}}) &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left(\frac{|\text{RSS}_b^{(i)} - \text{RSS}_a^{(i)}|}{\text{RSS}_a^{(i)}} \right) \\ &= \lim_{n \rightarrow \infty} \frac{1}{m} \sum_{k=1}^m \mathbb{E} \left(\frac{|\text{RSS}_b^{(e_k)} - \text{RSS}_a^{(e_k)}|}{\text{RSS}_a^{(e_k)}} \right) = 0. \end{aligned}$$

This completes the proof of claim 1.

Proof of claim 2: Let $\mathcal{G} \in \mathcal{M}$ be non-invariant. Let $k^* \in \{1, \dots, m\}$ be the index and c_{\min} the constant for which (28) in Lemma 7 is satisfied. For every $\delta > 0$ define the following sets

$$\begin{aligned} A_\delta &:= \left\{ \left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} (\hat{y}_a^{(e_{k^*})}(t_\ell) - Y_{t_\ell}^{(e_{k^*})})^2 \right| + \left| \frac{2}{L_n} \sum_{\ell=1}^{L_n} (\hat{y}_a^{(e_{k^*})}(t_\ell) - Y_{t_\ell}^{(e_{k^*})}) \varepsilon_{t_\ell}^{(e_{k^*})} \right| \leq \delta \right\} \\ B_\delta &:= \left\{ \left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} \left(\varepsilon_{t_\ell}^{(e_{k^*})} \right)^2 - \sigma_{k^*} \right| \leq \delta \right\} \\ C_\delta &:= \left\{ \left| \frac{1}{L_n} \sum_{\ell=1}^{L_n} (\hat{y}_b^{(e_{k^*})}(t_\ell) - Y_{t_\ell}^{(e_{k^*})}) \varepsilon_{t_\ell}^{(e_{k^*})} \right| \leq \delta \right\} \\ D_\delta &:= \left\{ \frac{1}{L_n} \sum_{\ell=1}^{L_n} (\hat{y}_b^{(e_{k^*})}(t_\ell) - Y_{t_\ell}^{(e_{k^*})})^2 \geq c_{\min} - \delta \right\}. \end{aligned}$$

Using that both summands in the definition of A_δ converge in L^1 (this follows in exactly the same way, we obtained (51) and (53)) it holds that the sum convergences in probability. This in particular implies that there exists $N_A \in \mathbb{N}$ such that for all $n \in \{N_A, N_A + 1, \dots\}$ it holds that

$$\mathbb{P}(A_\delta) \geq 1 - \delta. \quad (58)$$

Next, by the law of large numbers it holds that $\frac{1}{L_n} \sum_{\ell=1}^{L_n} \left(\varepsilon_{t_\ell}^{(e_{k^*})} \right)^2$ converges to $\sigma_{k^*}^2$ in probability. This implies that there exists $N_B \in \mathbb{N}$ such that for all $n \in \{N_B, N_B + 1, \dots\}$ it holds that

$$\mathbb{P}(B_\delta) \geq 1 - \delta. \quad (59)$$

Finally, observe that since $\varepsilon_{t_\ell}^{(e_{k^*})}$ has mean zero it holds that

$$\mathbb{E} \left((\hat{y}_b^{(e_{k^*})}(t_\ell) - Y_{t_\ell}^{(e_{k^*})}) \varepsilon_{t_\ell}^{(e_{k^*})} \right) = \mathbb{E} \left((\hat{y}_b^{(e_{k^*})}(t_\ell) - y_{\lim}^{(e_{k^*})}(t_\ell)) \varepsilon_{t_\ell}^{(e_{k^*})} \right),$$

where $y_{\lim}^{(e_{k^*})}$ is the limit function given in Lemma 8. The statement of Lemma 8 together with the boundedness of the functions allows us to apply Lemma 6 to get that

$$\lim_{n \rightarrow \infty} \frac{2}{L_n} \sum_{\ell=1}^{L_n} \mathbb{E} \left| (\hat{y}_b^{(e_{k^*})}(t_\ell) - Y_{t_\ell}^{(e_{k^*})}) \varepsilon_{t_\ell}^{(e_{k^*})} \right| = 0.$$

Hence, this term also converges in probability and thus there exists $N_C \in \mathbb{N}$ such that for all $n \in \{N_C, N_C + 1, \dots\}$ it holds that

$$\mathbb{P}(C_\delta) \geq 1 - \delta. \quad (60)$$

Finally, applying Lemma 7 there exists $N_D \in \mathbb{N}$ such that for all $n \in \{N_D, N_D + 1, \dots\}$ it holds that

$$\mathbb{P}(D_\delta) \geq 1 - \delta. \quad (61)$$

Combining (58), (59), (60) and (61) we get for all $n \in \{N^{\max}, N^{\max} + 1, \dots\}$ with $N^{\max} :=$

$\max\{N_A, N_B, N_C, N_D\}$ that

$$\begin{aligned}
\mathbb{E} \left(\frac{|\text{RSS}_b^{(e_{k^*})} - \text{RSS}_a^{(e_{k^*})}|}{\text{RSS}_a^{(e_{k^*})}} \right) &\geq \mathbb{E} \left(\frac{\text{RSS}_b^{(e_{k^*})}}{\text{RSS}_a^{(e_{k^*})}} \right) - 1 \\
&\geq \mathbb{E} \left(\frac{\text{RSS}_b^{(e_{k^*})}}{\text{RSS}_a^{(e_{k^*})}} \mathbb{1}_{A_\delta} \mathbb{1}_{B_\delta} \mathbb{1}_{C_\delta} \mathbb{1}_{D_\delta} \right) - 1 \\
&\geq \mathbb{E} \left(\frac{c_{\min} - \delta - \delta + \sigma_{k^*}^2 - \delta}{2\delta + \sigma_{k^*}^2 + \delta} \mathbb{1}_{A_\delta} \mathbb{1}_{B_\delta} \mathbb{1}_{C_\delta} \mathbb{1}_{D_\delta} \right) - 1 \\
&= \frac{c_{\min} - 3\delta + \sigma_{k^*}^2}{3\delta + \sigma_{k^*}^2} \mathbb{P}(A_\delta \cap B_\delta \cap C_\delta \cap D_\delta) - 1 \\
&\geq \frac{c_{\min} - 3\delta + \sigma_{k^*}^2}{3\delta + \sigma_{k^*}^2} (1 - 4\delta) - 1,
\end{aligned}$$

where for the third inequality we used the expansion

$$\text{RSS}_*^{(e_{k^*})} = \frac{1}{L_n} \sum_{\ell=1}^{L_n} (\hat{y}_*^{(e_{k^*})} - Y_{t_\ell}^{(e_{k^*})})^2 - \frac{2}{L_n} \sum_{\ell=1}^{L_n} (\hat{y}_*^{(e_{k^*})} - Y_{t_\ell}^{(e_{k^*})}) \varepsilon_{t_\ell}^{(e_{k^*})} + \frac{1}{L_n} \sum_{\ell=1}^{L_n} (\varepsilon_{t_\ell}^{(e_{k^*})})^2$$

together with the normal and reverse triangle inequality and the definitions of the sets A_δ , B_δ , C_δ and D_δ . Since δ was arbitrary we can let δ tend to zero which implies that

$$\liminf_{n \rightarrow \infty} \mathbb{E} \left(\frac{|\text{RSS}_b^{(e_{k^*})} - \text{RSS}_a^{(e_{k^*})}|}{\text{RSS}_a^{(e_{k^*})}} \right) \geq \frac{c_{\min}}{\sigma_{k^*}^2}.$$

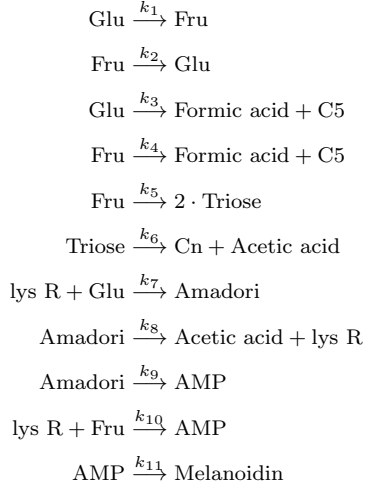
Finally, using this together with (57) we get that

$$\begin{aligned}
\liminf_{n \rightarrow \infty} \mathbb{E} (T_n^{\mathcal{G}}) &= \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left(\frac{|\text{RSS}_b^{(i)} - \text{RSS}_a^{(i)}|}{\text{RSS}_a^{(i)}} \right) \\
&\geq \liminf_{n \rightarrow \infty} \frac{1}{m} \sum_{k=1}^m \mathbb{E} \left(\frac{|\text{RSS}_b^{(e_{k^*})} - \text{RSS}_a^{(e_{k^*})}|}{\text{RSS}_a^{(e_{k^*})}} \right) \\
&\geq \frac{c_{\min}}{\sigma_{k^*}^2} > 0,
\end{aligned}$$

which completes the proof of claim 2 and also completes the proof of Theorem 3. \square

C. Biomodel 52

Reactions equations



ODE equations

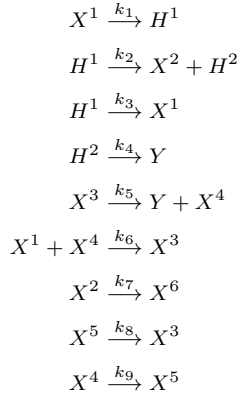
$$\begin{aligned}
 \frac{d}{dt}[\text{Glu}] &= -(k_1 + k_3)[\text{Glu}] + k_2[\text{Fru}] + k_7[\text{Glu}][\text{lys R}] \\
 \frac{d}{dt}[\text{Fru}] &= k_1[\text{Glu}] - (k_2 + k_4 + k_5)[\text{Fru}] - k_{10}[\text{Fru}][\text{lys R}] \\
 \frac{d}{dt}[\text{Formic acid}] &= k_3[\text{Glu}] + k_4[\text{Fru}] \\
 \frac{d}{dt}[\text{Triose}] &= 2k_5[\text{Fru}] - k_6[\text{Triose}] \\
 \frac{d}{dt}[\text{Acetic acid}] &= k_6[\text{Triose}] + k_8[\text{Amadori}] \\
 \frac{d}{dt}[\text{Cn}] &= k_6[\text{Triose}] \\
 \frac{d}{dt}[\text{Amadori}] &= -(k_8 + k_9)[\text{Amadori}] + k_7[\text{Glu}][\text{lys R}] \\
 \frac{d}{dt}[\text{AMP}] &= k_9[\text{Amadori}] - k_{11}[\text{AMP}] + k_{10}[\text{Fru}][\text{lys R}] \\
 \frac{d}{dt}[\text{C5}] &= k_3[\text{Glu}] + k_4[\text{Fru}] \\
 \frac{d}{dt}[\text{lys R}] &= k_8[\text{Amadori}] - k_7[\text{Glu}][\text{lys R}] - k_{10}[\text{Fru}][\text{lys R}] \\
 \frac{d}{dt}[\text{Melanoidin}] &= k_{11}[\text{AMP}]
 \end{aligned}$$

Parameters and initial conditions

$k_1 = 0.01$	$[\text{Glu}] _{t=0} = 160$
$k_2 = 0.00509$	$[\text{Fru}] _{t=0} = 0$
$k_3 = 0.00047$	$[\text{Formic acid}] _{t=0} = 0$
$k_4 = 0.0011$	$[\text{Triose}] _{t=0} = 0$
$k_5 = 0.00712$	$[\text{Acetic acid}] _{t=0} = 0$
$k_6 = 0.00439$	$[\text{Cn}] _{t=0} = 0$
$k_7 = 0.00018$	$[\text{Amadori}] _{t=0} = 0$
$k_8 = 0.11134$	$[\text{AMP}] _{t=0} = 0$
$k_9 = 0.14359$	$[\text{C5}] _{t=0} = 0$
$k_{10} = 0.00015$	$[\text{lys R}] _{t=0} = 15$
$k_{11} = 0.12514$	$[\text{Melanoidin}] _{t=0} = 0$

D. Artificial hidden variable model

Reactions equations



ODE equations

$$\begin{aligned}
 \frac{d}{dt}[X^1] &= -k_1[X^1] + k_3[H^1] - k_6[X^1][X^4] \\
 \frac{d}{dt}[X^2] &= k_2[H^1] - k_7[X^2] \\
 \frac{d}{dt}[X^3] &= -k_5[X^3] + k_6[X^1][X^4] + k_8[X^5] \\
 \frac{d}{dt}[X^4] &= k_5[X^3] - k_6[X^1][X^4] - k_9[X^4] \\
 \frac{d}{dt}[X^5] &= k_9[X^4] - k_8[X^5] \\
 \frac{d}{dt}[X^6] &= k_7[X^2] \\
 \frac{d}{dt}[H^1] &= k_1[X^1] - (k_2 + k_3)[H^1] \\
 \frac{d}{dt}[H^2] &= k_2[H^1] - k_4[H^2] \\
 \frac{d}{dt}[Y] &= k_4[H^2] + k_5[X^3]
 \end{aligned}$$

Parameters and initial conditions

$$\begin{aligned}
 k_1 &= 0.08 \\
 k_2 &= 0.08 \\
 k_3 &= 0.01 \\
 k_4 &= 0.1 \\
 k_5 &= 0.003 \\
 k_6 &= 0.06 \\
 k_7 &= 0.1 \\
 k_8 &= 0.02 \\
 k_9 &= 0.05
 \end{aligned}$$

$$\begin{aligned}
 [X^1] |_{t=0} &= 5 \\
 [X^2] |_{t=0} &= 0 \\
 [X^3] |_{t=0} &= 0 \\
 [X^4] |_{t=0} &= 5 \\
 [X^5] |_{t=0} &= 0 \\
 [H^1] |_{t=0} &= 0 \\
 [H^2] |_{t=0} &= 0 \\
 [Y] |_{t=0} &= 0
 \end{aligned}$$

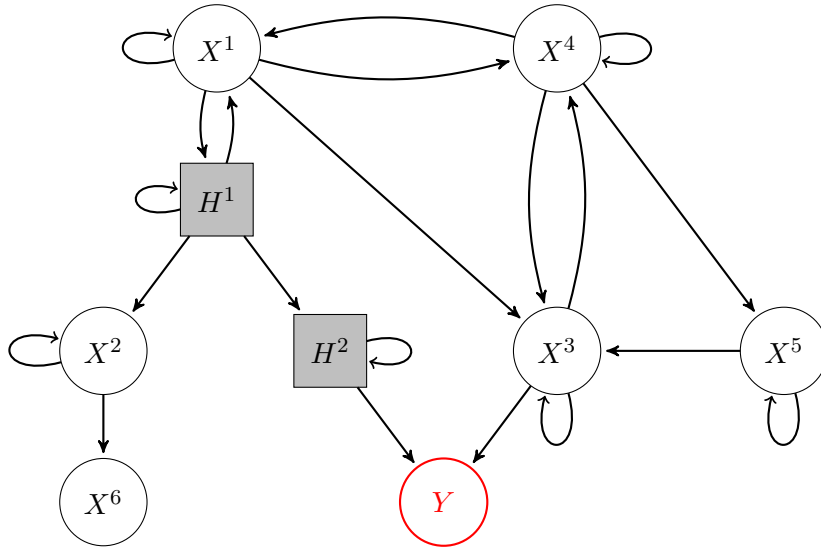


Figure 17. Graph representation of hidden variable ODE model. If the rate k_4 is equal to the rate k_7 the variables X^2 and H^2 will have identical dynamics.