Finding Causal Relations in Dynamical Data with Locally Linear Models

Alexander Mey – PostDoc at Eindhoven University of Technology

Machines produced by ASML are highly complex, have many components, and are vulnerable to slight deviations in functionality of those components. For that reason many of the components are monitored and produce dynamical data, i.e., data with a temporal component. In this project we are building methods to find causal relationships between those components for root cause analysis, fault detection and prevention and general understanding of the dynamics within the machine beyond their conceptual design.

Introduction

In the context of finding causal relationships in complex dynamical systems we consider in this work the particular goal of finding the direct causal parents of a certain target component Y amongst a set of possible parents $X = (X_1, \cdots, X_D)$. Here Y could be, for example, the power output of a certain component, while $X = (X_1, \cdots, X_D)$ could be measurements of temperature, angle, etc. of other monitored components. While one may of course use statistical methods to determine statistical relationships between Y and X, we want to go in some cases beyond that to find causal relationships. Generally, however, this is only possible if we can intervene on the system, while in this work we only assume that we observe the system. For that reason we use a concept that may be described as `incidental' interventions. The underlying idea is that if we observe the system in heterogeneous environments, those environments can work in some cases similarly to actual interventions. An environment may be the data from a similar machine, but in a different location or setting, or even different time intervals of the same machine, assuming that the machine behaves sufficiently different across some time intervals. The environments have to be sufficiently heterogeneous in the sense that the distributional behavior of X has to be different across (at least some) environments. See Figure 1 for a summary and further explanation of the setting.

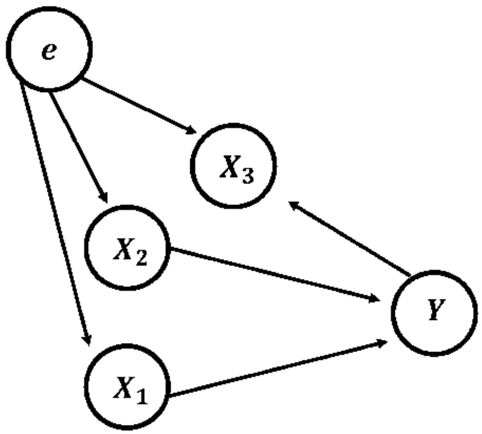

Figure 1: An example causal graph to explain the type of causal discovery task we consider. We want to find the causal parents of Y, which are represented by the nodes X_1, X_2 pointing to Y. Other variables, such as X_3, may be for example causes of Y. The goal is to recover the structure of the arrows by only observing the behavior of the variables. The important assumption in our work is that the non-target variables X_1, X_2, X_3 are affected by an environment variable e, as explained in the main text.

Methods and Approach

The basic modelling idea in this setting follows the work of [1], who propose to model the relationship between X and Y with *one* global linear model, where global means that there is one model for all environments. Modelling our complex dynamical data with one linear model is overly restrictive. Thus, our contribution is to relax this assumption, to use instead locally linear models, and analyze the resulting consequences. For dynamical data this entails that we can model several (short) time intervals with each a linear model. Those linear models are then build for all subsets of potential causal parents and we use a hypothesis test to determine which subsets indeed constitute plausible direct parents of Y. To be a bit more precise we test for the following: Assuming that the causal structure remains the same for different environments, while the distribution X sufficiently change for different environments, the distribution of the residuals of those linear models are the same across all environments if and only if we used all causal parents to build the linear model. Following that, the causal parents are the minimal subset such the above property holds. Alternatively, one may view this procedure as a conditional independence test, where the residuals are independent of the environment if and only if we condition on all causal parents.

Results and Observations

To test our proposed method in a controlled environment, we generated data with a *drifting vector autoregressive model*. This means that the current timestep of the system X, Y is generated with a linear model from the past of X, Y and an additive noise component. The drift means the linear model slowly changes over time, while we changed the variance of the additive noise after every 40 time steps. In Figure 2 we plot the success rate of the method against the chosen sample size, so the interval length, for our locally linear models. We observe that for a range of chosen sample sizes we achieve good to optimal success rate. For a too small sample size we lose success rate as the estimation for the linear models is too difficult. For a large sample size we lose success rate because we introduce a stronger model mismatch as the ground truth linear model drifts over time.

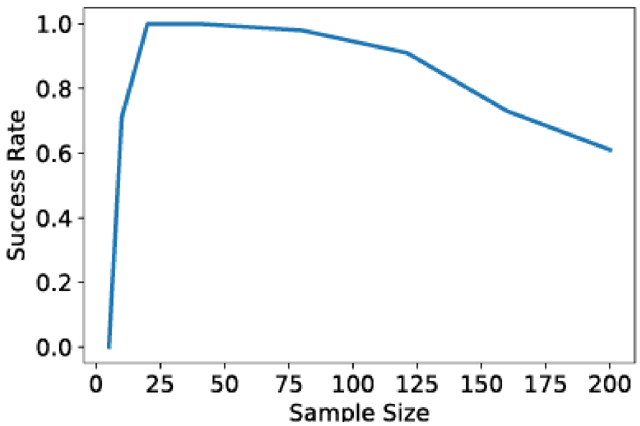

Figure 2: Results with the drifting VAR model. We count a success only if we found all parents.

Limitations and Ongoing Work

While the proposed approach is a promising method to find causal relationships in dynamical systems with observational data, the work is still in progress and has certain limitations. One of the crucial assumptions in this work is that the environment may not directly affect the target Y. This means in practice that ones needs to include all variables that mediate an effect from the environment to Y. Another major drawback of this method is that one has to iterate over all possible subsets of possible parents. While this does not scale well, we are investigating methods that are similar to the proposed approach, but instead work trough the optimization of a single objective function. Besides that we are investigating the behavior of the approach for small sample sizes, and in that relation testing novel ways of hypothesis testing in this context.

References

[1] J. Peters, P. Bühlmann, N. Meinshausen, "Causal Inference by using invariant prediction: identification and confidence intervals." Journal of the Royal Statistical Society. Series B (Statistical Methodology), 78(5):947-1012, 2016.

Facts

Researchers	Alexander Mey * Rui Castro Alessandro Di Bucchianico
Academic partners	Eindhoven University of Technology
Industrial partners	ASML
	* a.mey@tue.nl

https://openreview.net/pdf?id=wuhLhLV3ck-