Causal Discovery in Time Series Data Using Causally Invariant Locally Linear Models

Alexander Mey
Department of Mathematics and Computer Science
Technical University Eindhoven
Eindhoven, 5612 AZ
a.mey@tue.nl

Abstract

Identifying causal relationship is an often desired, but difficult, task, and generally only possible under specific assumptions. In this paper we are considering the task of identifying causal relationships between entities that have a temporal axis, as for example continuous measurements of different components within a complex machine. We introduce a locally linear model class that allows us to recover causal relationships, assuming that the process is locally linear, that we have access to observations in diverse environments and that the causal structure is invariant across the different environments. We validate the model in an idealized theoretical and two experimental settings.

1 Introduction

In this paper we consider the task of identifying causal relationships between different time-series. This setting occurs for example in machine diagnostics, where we are continuously monitoring different components of the machine. From this data we then want to infer which components interact with each other, for example to find the root cause if one component malfunctions. More specifically we consider a setting where we are monitoring the different measurements across different, heterogeneous, environments. Measurements in heterogeneous environments are in practice the observation of similar data under different conditions, for example a machine that has different working modes, or climate measurements in different countries. More generally, heterogeneous environments can also arise as different observational intervals[1] within one single process. Heterogeneity of measurements across different time intervals may occur due to certain distributional shifts within the process, for example induced by degrading components. Under so called invariance conditions, different environments can be useful to identify causal relationships. The underlying principle of the invariance concept is that if the causal structure is invariant, i.e. remains the same, across different environments, those environments can sometimes be seen as 'incidental interventions'. In that sense, having different environments of observational data may contain strictly more causal information than observational data from only one environment. The main idea of our analysis follows the work of Peters et al. [2016], while the main difference is that Peters et al. [2016] encode the invariance idea by assuming one global linear model, while we assume that each environment may have a different linear model, but the support entries of those linear models should be the same across environments. While this makes our model more flexible, it might make the estimation process of the local linear models extremely difficult. For this reason we investigate the small sample per environment case, hoping that we can approximate a potentially complicated process with locally linear models. While our formal theorems are so far only in the large sample case, we present experimental evidence that

[1] We will be using the words environments and intervals interchangeably.

A causal view on dynamical systems workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

the proposed approach still works even in the extreme case that per interval the dimensionality of the problem exceeds the sample size, when using regularized linear regression for the local models.

With this we can summarize our contributions to this field as follows. First, we propose a locally linear model class that can identify causal structures under a causal invariance assumption. We then propose a practical test that can identify the causal structure from finite data. And finally we analyze the model in different settings, in particular when the data has a temporal component, and in the small sample per environment setting.

2 Related Work

The two themes of this paper are invariances in causal discovery and causal discovery in time series. Regarding the invariance part, our ideas resemble mostly the setup of Peters et al. [2016], as elaborated in the introduction. The idea was extended to sequential data in Pfister et al. [2019], while Heinze-Deml et al. [2018] investigates non-linear models. The differences of our paper to Pfister et al. [2019] are the modelling choices. Most notably they test, similarly to Peters et al. [2016], one global linear regression model, while we consider only locally linear models. Christiansen and Peters [2020] also introduce a model class that allows for switching regression parameters as the effect of a hidden unobserved variable. While this model class is includes our setting it is a much harder inference problem, which leads to very specific and vastly different modelling choices compared to our setting. The invariance concept was also carried to the context of out-of-domain generalization for machine learning models Arjovsky et al. [2019], Oberst et al. [2021], Wald et al. [2021], the main idea being that causal features, in contrast to only correlated features, should be useful in all domains. Differences in the approaches in that literature often stem from which parts of the model are assumed to be invariant across environments. While the goal of out-of-domain generalization is different than our goal, which is causal discovery, many of the concepts and models resemble each other. In a way, our proposed locally linear models, are of course very simple machine learning models.

One of the predominant causal discovery ideas in the time-series setting is the concept of Granger causality Granger [1969]. The underlying principle is that we call a time-series X_t a Granger cause of a time series Y_t if the past of X_t has unique, statistically significant, information about the future of Y_t. Although a Granger cause is generally not a cause, for example in the causal sense of a structural causal model (SCM) Pearl [2016], it can be shown to coincide under certain assumptions Peters et al. [2017]. One of those assumptions is that there no instantaneous causal effects, which do occur in practice if the sampling frequency is slower than the underlying causal dynamics. Our framework can deal with this setting.

3 Setting

In this work we assume that we have two observable quantities, a target Y and a set of covariates $(X_1, \cdots, X_D)$ for $D \in \mathbb{R}$. The general goal is to identify the covariates X_d which causally affect the target Y. What we precisely mean with a causal effect is detailed in the next subsection. We furthermore assume that there exist $E \in \mathbb{N}$ different environments, such that in each environment $e \in [E]$ we have the following linear relationship between target and covariate

$$Y^e = X^e \beta^e + \varepsilon^e. \tag{1}$$

Here $X^e = (X_1^e, \cdots, X_D^e)$ is a row vector of random variables, stochastic processes [2] or deterministic processes. Next, $\beta^e \in \mathbb{R}^D$ is a column vector of regression coefficients, fixed for each $e \in [E]$, and $\varepsilon^e \in \mathbb{R}$ is a random noise variable, which is assumed to be generated identically and independently from a zero mean distribution (IID). The different environments could be for example different windows within one long (stochastic) process. The different β^e then allow us to model each window linearly, and are thus the locally linear part in our setting.

For notational convenience we set $[E] \equiv \{1, \ldots, E\}$ and similarly $[D] \equiv \{1, \ldots, D\}$. With X_d^e or β_d^e we indicate the d-th entry of respectively X^e or β^e and similarly X_S^e for $S \subseteq [D]$ denotes the sub-vector with indices given by S. In our notation the superscript e always indicates the corresponding environment. Also note that if we write $\mathbf{E}[P]$ we mean, depending on P, the mean of an IID random

[2] Note that, to ease notation, we forego a time index t for X^e unless needed.

variable or the limiting temporal mean of a stochastic or deterministic series, if they exist. Finally we denote by $N(0, 1)$ the standard normal distribution and by $U(l, h)$ the uniform distribution on $[l, h]$.

3.1 Invariance Assumption

To make use of information we collect in different environments, we connect them with the following invariant support assumption: We assume there is a set $S^* \subseteq [D]$, called the support indices, with cardinality $|S^*| = s \leq D$ such that for $d \notin S^*$ and $e \in [E]$ we have $\beta_d^e = 0$. The non-zero values of β^e are otherwise arbitrary. Note that the value s is generally unknown. The inference goal is to determine S^* from the observed data. From a causal perspective we can see S^* as the set of causes for the target Y^e, so we try to retrieve the causes of Y^e.

While the above already outlines most of our assumptions, the formal assumption we make is the following.

Assumption 1. *There exists a subset $S^* \subseteq [D]$, a zero mean distribution F and a collection of coefficients $\beta^e \in \mathbb{R}^D$ for $e \in [E]$, such that $\beta_d^e = 0$ for all $d \notin S^*$ and $e \in [E]$. Furthermore: For all $e \in [E]$ we have that*

$$Y^e = X^e \beta^e + \varepsilon^e, \quad \varepsilon^e \overset{IID}{\sim} F, \varepsilon^e \perp\!\!\!\perp X_{S^*}^e. \tag{2}$$

Note that the specific meaning of $\varepsilon^e \perp\!\!\!\perp X_{S^*}^e$ depends on how $X_{S^*}^e$ is generated. In the IID case the independence is meant sample-wise. If $X_{S^*}^e$ is a stochastic process, ε^e hast to be independent of the complete process. In the deterministic case this independence is automatically given. Importantly also note that the noise variable ε^e *is* allowed to correlate with the non-support covariates, so the covariates X_d^e for $d \notin S^*$. In a causal sense this can for example happen if one covariate is the effect of our target Y^e.

3.2 Heterogenity of the Environments

So far we talked about the different environments being heterogeneous, i.e. they should be in some aspect different from each other. It is clear that if all of the environments are exactly the same in every aspect, we cannot expect to have any benefit from them, besides having a potentially larger sample size. In the view of Assumption 1 heterogeneity may come into play by having differently behaving covariates X^e or regression parameters β^e for different environments $e \in [E]$. What we precisely require will be formalized and discussed in Section 4.

3.3 Examples

While in Section 6 we provide specific data generation models, we motivate here two scenarios that follow the model idea from above.

As the first example, assume we build a machine that produces watches and we measure the accuracy of the watches, which corresponds to our target Y. We also record other measurements, corresponding to the covariates X, that we consider relevant for the resulting accuracy Y. Those measurements could be some settings of the machine, heat or power output of certain components etc. We now want to understand which of the covariates X directly causally influence on the accuracy Y. To analyse this behavior we collect data from several batches of watches that are produced in different locations. The locations correspond to the different environments. Similarly one may think of the scenario where the same machine produces different types of watches, and one can think of each type as a different environment. In both cases the main assumption of our work is that the causal factors remain the same across environments, and each environment may be well approximated with a linear model, while any additional observational noise in Y has to have the same distribution everywhere.

The other setting we want to consider is the setting of a temporal process, where different intervals of this process can naturally be considered as different environments. For example, consider we drive a car, and we want to understand which components of the car directly influence the engine. While driving the car we may be driving on the highway, in the city and also in different gears. While our measurements are a continuous process, the different intervals of this process give rise to, potentially, heterogeneous data. Similarly heterogeneous data can arise if we have a continuous drift in some measurements, induced for example by degrading components.

4 General Results

In this section we build the theoretical foundations for our proposed methods in a population setting. The next section will then look at specific algorithms for the finite sample case. Based on the previous modelling assumption we want to build a method that identifies, based on observational data, a set $\hat{S} \subseteq [D]$ that is supposed to approximate S^* as good as possible, while ideally we have $\hat{S} = S^*$. To achieve that we first find conditions to control the false positives, i.e. we want that with high probability $\hat{S}$ fulfills $\hat{S} \subseteq S^*$. Subsequently we find conditions that allow us to identify false negatives which will ensure that $\hat{S} = S^*$.

4.1 Controlling False Positives

Similar to Peters et al. [2016] we use a hypothesis test to identify S^*. While adapted to our setting, the essential idea remains the same. With that in mind we define for a subset $S \subseteq [D]$ the following null-hypothesis on the data generating process:

$$H_{0,S} : \begin{cases} \text{exists a distribution } F \text{, and parameters } \beta^e \text{, s.t. } \forall e \in E \\ \beta^e_k = 0 \text{ if } k \notin S \text{ and } Y^e = X^e \beta^e + \varepsilon^e \text{ with} \\ \varepsilon^e \overset{IID}{\sim} F \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_S \end{cases} \tag{3}$$

Contrasting the hypothesis (3) with our main Assumption of Equation 2, we already notice the important property that H_{0,S^*} is true. Assuming we can test the null hypothesis we define an estimator $\hat{S}$ of S^* as

$$\hat{S} := \bigcap_{S: H_{0,S} \text{ is not rejected}} S. \tag{4}$$

The important property of (3) is that it is fulfilled for $S = S^*$, because of Assumption 1. In other words, H_{0,S^*} is true, which implies that $\hat{S} \subseteq S^*$ with high probability, given that we have a valid test for $H_{0,S}$, see Theorem 1 further below. The definition of the null hypothesis also already foreshadows the relevance of the different environments. If we only had one environment e_1, then even $S = \emptyset$ fulfills $H_{0,S}$ with F being equal to the distribution of Y^{e_1}.

The first formal result on controlling false positives if we assume we can test $H_{0,S}$ at a certain confidence level is due to Peters et al. [2016].

Theorem 1. *[Peters et al., 2016, Theorem 1] Assume that we constructed $\hat{S}$ according to Equation* (4) *with a valid test for $H_{0,S}$ in the sense that $P[H_{0,S}$ is rejected $\mid H_{0,S}$ is true$] \leq \alpha$. If Assumption* (1) *holds we have that $\hat{S} \subseteq S^*$ with probability of at least $1 - \alpha$.*

The null hypothesis defined in (3) is not very practical to test, as it is not clear how one might even find the required β^e and ε^e. To move to a more practical version we make the following change. Let

$$\hat{\beta}^e_S = \arg \min_{\beta \,\in\, R^D \,:\, \beta_d \,=\, 0 \text{ if } d \,\notin\, S} \mathbf{E}[(Y^e - X^e \beta)^2],^3 \tag{5}$$

then we can formulate a more practical null hypothesis as

$$H_{0,S} : \begin{cases} \text{exists distribution } F \text{ such that for all } e \in [E] \\ \varepsilon^e \overset{IID}{\sim} F, Y^e = X^e \hat{\beta}^e_S + \varepsilon^e \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_S \end{cases}. \tag{6}$$

Although the hypotheses defined in Equations (3) and (6) are generally not equivalent, H_{0,S^*} remains true if Assumption 1 holds, since $\hat{\beta}^e_{S^*} = \beta^e$. With that also Theorem 1 remains true. Note that the residuals of the regression task, defined as $R^e := Y^e - X^e \hat{\beta}^e_S$, will model the unexplained noise ε^e, and it will be those residuals that we test for equality of distribution, see also the proof of Theorem 2.

[3] If there are multiple solutions we may pick any. The important property we implicitly use to make use of Equation (6) is that $\hat{\beta}^e_{S^*} = \beta^e$ up to an element in the null-space of $X^e_{S^*}$

4.2 Controlling False Negatives

While under the assumptions of the previous section we can guarantee that $\hat{S} \subseteq S^*$, we ideally find conditions that allow us to identify the full support S^*, so that $\hat{S} = S^*$. For that, the heterogeneity of the different environments will come into play, and we make the following assumption regarding that.

Assumption 2. *For every $S \subsetneq S^*$ exists at least two environments $v, w \in [E]$ such that*

$$(X^v_{S^*} - X^v_S \mathbf{E}[{X^v_S}^T X^v_S]^{-1} \mathbf{E}[{X^v_S}^T X^v_{S^*}])\beta^v_{S^*} \overset{d}{\neq} (X^w_{S^*} - X^w_S \mathbf{E}[{X^w_S}^T X^w_S]^{-1} \mathbf{E}[{X^v_S}^T X^w_{S^*}])\beta^w_{S^*} \quad (7)$$

or equivalently

$$X^v_{S^*}\beta^v_{S^*} - X^v_S \hat{\beta}^v_S \overset{d}{\neq} X^w_{S^*}\beta^w_{S^*} - X^w_S \hat{\beta}^w_S. \quad (8)$$

Here the notation $\overset{d}{\neq}$ means that the two quantities are not allowed to be distributionally equivalent. In view of our definition of the expectation operator $\mathbf{E}$ in Section 3, this includes the possible case that either side of the equality is not well defined or does not have a distribution at all. While this assumption is very generic, and indeed just the assumption that will make the proof work, we anticipate one does not require a pair of heterogeneous environments for each subset, but, under possibly further assumptions, it is sufficient to have a pair of environments for each support covariate as in Peters et al. [2016]. Although we believe that under mild conditions Assumption 2 is true, we are aiming at making this precise in future work. In any case, this assumption allows us to draw the following conclusion.

Theorem 2. *Let $S =\subsetneq S^*$. Under Assumptions* (1) *and* (2) *we find that $H_{0,S}$ is not true.*

Proof. For the S as in the theorem let v, w be two environments as given by Assumption 2. Consider the residuals of our model in both environments, which are given by

$$\mathbf{R}^v = X^v_{S^*}\beta^v_{S^*} + \varepsilon^v - X^v \hat{\beta}^v_S \text{ and } \mathbf{R}^w = X^w_{S^*}\beta^w_{S^*} + \varepsilon^w - X^w \hat{\beta}^w_S. \quad (9)$$

By definition $H_{0,S}$ is only true if $\mathbf{R}^v \overset{d}{=} \mathbf{R}^w$. As ε^e is assumed to have the same distribution for all $e \in [E]$, and plugging in the solution for $\hat{\beta}^v_S, \hat{\beta}^w_S$, we find that $\mathbf{R}^v \overset{d}{=} \mathbf{R}^w$ can only hold if

$$(X^v_{S^*} - X^v_S \mathbf{E}[{X^v_S}^T X^v_S]^{-1} \mathbf{E}[{X^v_S}^T X^v_{S^*}])\beta^v_{S^*} \overset{d}{=} (X^w_{S^*} - X^w_S \mathbf{E}[{X^w_S}^T X^w_S]^{-1} \mathbf{E}[{X^w_S}^T X^w_{S^*}])\beta^w_{S^*}$$

This equality, however, is excluded by Assumption (2). □

Two sources of heterogeneity: At this point we want to make a very interesting observation following the previous analysis. Our heterogeneity Assumption 2 can be fulfilled by two means, either through heterogeneity in the covariates X^e themselves, but also through the locally linear regression parameters β^e. Setting $U = S^* \setminus S$, consider that the distributions of the unaccounted variables X^e_U can be affected by a vector of environment dependent noise terms η^e_U. Equation (8) shows that the distributions on the lhs and rhs then will differ at least by the differences of the corresponding terms $\eta^v_U \beta^v_U$ and $\eta^w \beta^w_U$. Equation (7) on the other hand shows that the regression parameters $\beta^v_{S^*}$ themselves can introduce differences in distribution, as long as X^v_S is not perfectly collinear to $X^v_{S^*}$ in all environments. This is in contrast to the work of Peters et al. [2016], where heterogeneity always comes through the covariates. In that sense, allowing for local linearity not only makes our model more flexible, but also introduces a more general concept of heterogeneity.

5 Finite Sample Methods

In this section we propose two finite sample methods that use the ideas of the previous sections to recover S^* with finite observational data. For simplicity we assume that in each environment $e \in E$ we have n observations. Importantly, here E and n denote the ground truth environments and their corresponding ground truth length n. Generally, however, we do *not* assume that our algorithms have access to the correct environments. Thus our algorithms will be provided with environments indexed by $\hat{E}$ which give rise to a model sample size $\hat{n}$, which can be different from n. From here on $\mathbf{Y}^e$ denotes a vector of $\hat{n}$ or n observations from an interval $e \in \hat{E}$ or $e \in [E]$ and $\mathbf{X}^e$ is the data matrix of the appropriate size.

```
Input: $\lambda > 0, (\mathbf{Y}^e, \mathbf{X}^e)$ for $e \in \hat{E}, \alpha$ (In order: Regularization parameter, observations, confidence level)
Output: $\hat{S}$
For all $S \subseteq [D]$
    • Set $\hat{\beta}^e_S := \arg \min_{\beta \in R^D \,:\, \beta_d = 0 \text{ if } d \notin S} \|\mathbf{Y}^e - \mathbf{X}^e\beta\|_2^2 + \lambda\|\beta\|_2$
    • Set $R^e := \mathbf{Y}^e - \mathbf{X}^e\hat{\beta}^e_S$
    • Let $\hat{H}_{0,S}$ a test of the hypothesis that the samples $(R^e)_{e\in\hat{E}}$ have equal distributions, tested
      with the Anderson-Darling test at confidence level $\alpha$
Set $\hat{S} := \bigcap_{S:\hat{H}_{0,S} \text{ is not rejected}} S$
```

Algorithm 1: Our proposed method based on the Anderson-Darling test.

In Algorithm 1 we propose our first method following the ideas and results from Section 4. There are a couple of important things to note. First, for the specific hypothesis test we decided to use a multiple sample version of the Anderson-Darling test Scholz and Stephens [1987].[4] Second, we add a regularization term to the squared loss function to deal with the low sample case, when $\hat{n} \leq D$. One may formally justify that by adding a bounded norm condition of β^e to Assumption 1. Finally, as mentioned earlier, our methods do generally not have access to the true interval length n, which in turn means that as input for the algorithm we do *not* provide the correct indexing of the intervals E, but instead a possibly different indexing $\hat{E}$.

While Algorithm 1 makes use of a formal hypothesis test and enjoys the guarantees and interpretability that come with it, we propose an ad-hoc version in Algorithm 2. Instead of using a hypothesis test, this algorithm rejects a subset $S \subseteq [D]$ if the variance of the residuals R^e varies too much across different intervals $e \in \hat{E}$. There are two reasons to introduce this variation of our approach. First of all, it gives a pathway to algorithms that do not suffer from having to iterate over all possible subsets of $[D]$, see also the discussion Section 7. The second reason is the small sample setting. For the case that $\hat{n} < D$ we observe a sudden degradation of the Anderson-Darling test in terms of power. While in Algorithm 2 we have to set an ad-hoc rejection threshold κ, we want to highlight with it that this sudden degradation for the small sample case may not be inherent to the problem, and different methods may retrieve a good success rate even in that extreme case. If that can be done in a principled way is part of our future research.

```
Input: $\lambda > 0, (\mathbf{Y}^e, \mathbf{X}^e)$ for $e \in \hat{E}, \kappa$ (In order: Regularization parameter, observations, acceptance threshold)
Output: $\hat{S}$
For all $S \subseteq [D]$
    • Set $\hat{\beta}^e_S := \arg \min_{\beta \in R^D \,:\, \beta_d = 0 \text{ if } d \notin S} \|\mathbf{Y}^e - \mathbf{X}^e\beta\|_2^2 + \lambda\|\beta\|_2$
    • Set $R^e := \mathbf{Y}^e - \mathbf{X}^e\hat{\beta}^e_S$
    • Set $\mathbf{V}_S := \mathbf{V}_e\left[\mathbf{V}[R^e]\right]$. Here $\mathbf{V}[R^e]$ is the sample variance of the residuals in interval $e$
      and $\mathbf{V}_e\left[\mathbf{V}[R^e]\right]$ is the sample variance of $\mathbf{V}[R^e]$ over all intervals
Set $\hat{S} := \bigcap_{S:\mathbf{V}_S<\kappa} S$
```

Algorithm 2: Our proposed method based on a simple variance test.

[4] We use an implementation from the python scipy toolbox, licensed under the BSD 3-Clause "New" or "Revised" License.

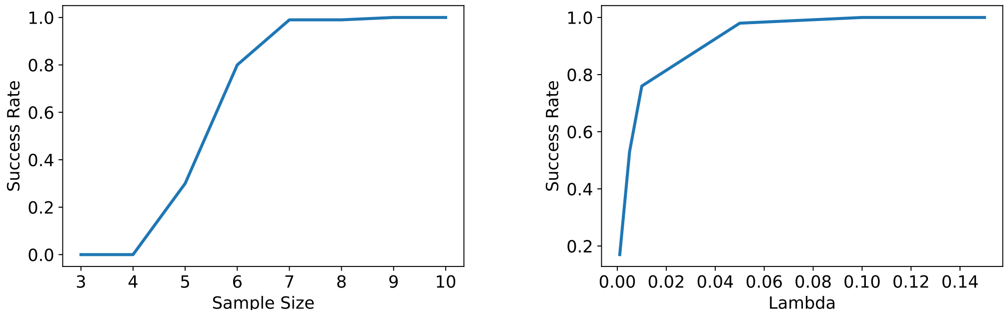
6 Experimental Section[5]

We now test the approaches described in Section 5 on two differently generated data sets. Although our focus is on data with a temporal aspect, we first try the method on a simple setting, where the covariates are created in an IID fashion. Although there is no temporal component, it is still relevant for time-series data, as it simulates possible instantaneous links. Afterwards we test the approach on data generated by a vector autoregressive model. In both cases we are in particular interested in the small sample per interval case. While some of the specific data generation and parameter choices are made to highlight the potential of our approaches, we discuss in Section 7, and further in the appendix, the impact of the various choices.

6.1 The IID Case

Our first experiment is in a simple setting where the data is generated in an IID fashion for $D = 5, |S^*| = 2, E = 100$ and varying n, see Algorithm 3 in the appendix for details. Furthermore we assume here that our algorithm has access to the correct environments, so we set $\hat{n} = n$. With the given data we then use Algorithm 1 with $\lambda = 0.01$ and the ground truth environments to try to recover S^*. For each sample size n we run this experiment 100-times and record a success if the output $\hat{S}$ of Algorithm 1 exactly matches S^*. We plot the resulting success rate against the sample size n in Figure 1a. While we observe that already a moderate sample size of $n = 7 > 5 = D$ suffices to ensure a perfect success rate, the performance quickly degrades in the case where n becomes small and we drop to a rate of 0 for $n < D$. While choosing $\lambda > 0.01$ increased the performance up to 0.25, a higher success rate seemed unachievable for Algorithm 1. Since we believe, however, that the low success rate is due to the specifics of the hypothesis test of Algorithm 1, we continued the next experiment with Algorithm 2.

In this second experiment we tested the success rate of Algorithm 2 with the same data generation process but $n = 4$ fixed. We then use Algorithm 2 with varying λ and $\kappa = 1$, the results are shown in Figure 1b. As we hoped, we indeed see that with a bit of added regularization we can recover a high success rate.



(a) The performance of Algorithm 1 for different sample sizes. Note, here we assume $n = \hat{n}$.

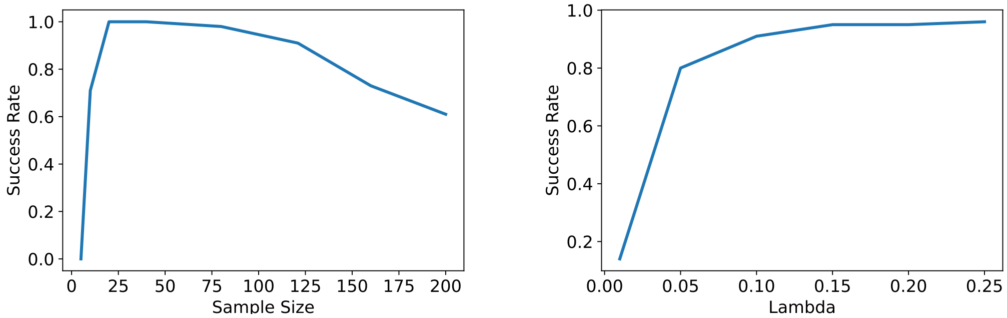
(b) The performance of Algorithm 2 for different λ and $n = \hat{n} = 4$.

Figure 1: Evaluation of our Algorithms in the IID case

6.2 The VAR Setting

This experiment is in the time-series setting and the data is generated with a vector autoregressive model (VAR). In this section we drop the e notation for different environments, instead our quantities will have a time index t. Different environments will now correspond to intervals of a certain length of this process. The VAR model implies that the covariates follow a discrete process $X^{t+1} = X^t W + \gamma$ where $X^t \in \mathbb{R}^D$ is our observation at time t, $W \in \mathbb{R}^{D \times D}$ is the matrix of regression coefficients and $\gamma \in \mathbb{R}^D$ is a driving noise.

[5] The experiments can be found as jupyter notebooks on https://github.com/AlexanderMey/causal-local-linear.



(a) The success rate of Algorithm 1 for varying model sample sizes $\hat{n}$ and $\lambda = 0.01$.

(b) The success rate of Algorithm 2 for $\hat{n} = 4$ and varying λ.

Figure 2: The results of the experiments in the VAR setting.

In our specific setting the regression matrix W and the driving noise γ are allowed to change over time, thus our model can be written as $X^{t+1} = X^t W^t + \gamma^t$. The change of W^t and γ^t corresponds to the heterogeneity of the resulting covariates X^t, and the success of our approach will depend on how W^t and γ^t changes over time. Also in this setting we assume that there are support indices $\mathbb{S}^* \subseteq [D] \times [D]$ with the property that $W_s^t = 0$ if $s \notin \mathbb{S}^*$ for all t.

While our covariates follow a VAR model, the rest of the setup and assumptions remain the same. In particular we set one covariate, w.l.o.g. X_1^t, as the target. By setting our target as $Y^t := X_1^{t+1}$ the model for that target becomes

$$Y^t = X^t \beta^t + \gamma_1^t + \varepsilon^t. \tag{10}$$

Here β^t is the first column of W^t, ε^t is as before an independent noise term and γ_1^t the first entry of γ^t. This already reveals an important assumption we have to make: Contrasting Equation (10) with our model from Assumption (1), we note that the term $\gamma_1^t + \varepsilon^t$ takes the role of the independent noise ε^e. Thus, following Assumption (1), this implies that $\gamma_1^t + \varepsilon^t$ has to have the same distribution for all intervals, i.e. does distributionally not change over time. This is in direct contrast to the driving noise γ_d^t for $d \neq 1$, as this is ideally sufficiently diverse to ensure heterogeneity.

Nevertheless, this model is fairly general and flexible and there are many possible choices on how γ^t and W^t evolve. For our initial experiment we decided that the distribution of γ^t remains fixed for a certain period and then switches. The intervals in which γ^t remains fixed correspond to the ground truth intervals E, and the length of those intervals will be denoted as before with n.

In comparison to the previous IID experiments we make this task more challenging. First we do not assume anymore that we have access to E or n, instead we run our methods for intervals of varying $\hat{n}$, while n remains fixed. Furthermore we allow W^t to drift in *every* time-step and the support entries of W^t follow a biased random walk. Note that this introduces some model misspecification since our model assumes that W^t remains fixed within the intervals. A summarization and some more specific choices are detailed in Algorithm 4 in the appendix. With that, we first generated data with Algorithm 4, setting $n = 40, E = 100, D = 5$ and W^0 detailed in the appendix. We then used Algorithm 1 again with $\lambda = 0.01$ and varying model samples $\hat{n}$ per interval to recover the support indices in S^*, with $|S^*| = 3$, that correspond to our target $Y^t = X_1^{t+1}$, thus the non-zero entries of β^t. As before we report the success rate over 100 independent runs, where a success is only counted if the output $\hat{S}$ of Algorithm 1 exactly matches S^*. In Figure 2a we see that the method still achieves a very high success rate for samples sizes $\hat{n}$ different from $n = 40$. This shows that even if we do not know the ground truth sample size n, we may achieve good performance for a range of $\hat{n}$. It is clear that for model sample sizes $\hat{n}$ chosen too large, the method will lose power and validity, as we have fewer environments to compare to each other, and the model-misspecifiaction increases due to the drift of the regression variables W^t. As before we do also observe a steep drop in performance of Algorithm 1 when the sample size is getting relatively small. However, in principle the case of an extremely small size of $\hat{n}$ does not exclude success since no assumption is violated. To see if even in a small sample size setting we can recover a decent success rate we used the same data with Algorithm 2 with varying λ fixing $\kappa = 1$ and $|\hat{E}| = 1000$, which leaves only $\hat{n} = 4$ samples per interval for parameter estimation. We plot the results in Figure 2b. Once again we see that using a moderate λ significantly helps in recovering the correct support.

7 Discussion and Ongoing Work

While the previous experiments show that the proposed methods are promising, the effectiveness of them depends on a couple of choices of data and the methods them-self. We want to discuss the most interesting observations we made, while further experiments and discussions can be found in the appendix.

One of our main assumptions is the heterogeneity Assumption 2, which, for example in the IID case, is introduced through the randomization in step 1. and 3. of the data generating Algorithm 3. As suggested by the theory, we observed in some initial experiments little to no drop in success rate if we reduced the ranges of the uniform distributions of the covariates or the regression parameters (step 1. and 3. respectively of Algorithm 3), as long as not both were reduced at the same time. As also explained in Section 4.2, however, to make use of the heterogeneity introduced by the regression parameters we rely on the additional assumptions that some relevant covariates are not always perfectly collinear. In our IID setting this is clearly fulfilled as we generate the covariates independently from each other. In the VAR setting the story becomes more complicated and understanding this case is part of our ongoing research.

A problem of our experiment in the time-series setting is that the driving noises for all covariates, except the target, have to be heterogeneous across intervals, while the driving noise for the target needs a fixed distribution. This may pose an additional difficulty if we want causal explanations for all covariates. Regarding this, one may need additional expert knowledge to identify a sub-series of the complete data stream where the driving noise for the chosen target follows roughly the same distribution. On the other hand, one may wonder if heterogeneity introduced by the changing regression parameters W^t is sufficient if all driving noises are fixed. In how far this is indeed the case, is part of ongoing work.

One reason to propose a second algorithm, which is not based on a formal hypothesis test, was that the Anderson-Darling test lost a lot of power in the small sample setting. Motivated by the results of Algorithm 2, this loss in power may not be an inherent identifiability problem of our setting. Following that we are looking for ways to adapt the Anderson-Darling test, or find different, but still principled, tests altogether that are suited for the small sample case.

A major issue of our proposed methods in practice is that it scales exponentially with the dimension of the problem. We are currently investigating a possible solution based on a group-lasso approach, with a group corresponding to all regression parameters of one covariate d, i.e. $(\beta_d^{e_1}, \cdots, \beta_d^{e_I})$. This in principle would allow us to look at all covariates at once as the method finds the most important groups, while other grouped parameters are jointly set to 0. In that formulation the variance term in Algorithm 2 could be used as an additional penalty to look for invariant solutions.

Acknowledgments and Disclosure of Funding

This research has been funded by NWO under the grant PrimaVera (https://primavera-project.com) number NWA.1160.18.238 and was done in collaboration with our project partner ASML.

References

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019.

Rune Christiansen and Jonas Peters. Switching regression models and causal inference in the presence of discrete latent variables. *J. Mach. Learn. Res.*, 21:41:1–41:46, 2020.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Michael Oberst, Nikolaj Thams, Jonas Peters, and David A. Sontag. Regularizing towards causal invariance: Linear models with proxies. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8260–8270. PMLR, 2021.

Judea Pearl. *Causal inference in statistics : a primer*. Wiley, Chichester, West Sussex, 2016. ISBN 9781119186847.

Jonas Peters, Peter Bühlmann, and Nicolai 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.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

F.-W. Scholz and Michael A. Stephens. K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82:918–924, 1987.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 2215–2227, 2021.

A Appendix

A.1 Data Generation Algorithms

```
Input: $D, S^*, n, E$. (In order: Global dimension, support indices, samples per interval, intervals)
Output: $(\bigcup_{e \in [E]} \mathbf{Y}^e, \bigcup_{e \in [E]} \mathbf{X}^e) \in \mathbb{R}^{E \cdot n} \times \mathbb{R}^{E \cdot n, D}$
For all $e \in [E]$
    1. For all $d \in [D]$ sample $(a_d, b_d)$ from $U(-30, 0) \times U(0, 30)$
    2. Draw $\mathbf{X}_{i,d} \in \mathbb{R}$ as an IID $n$-sample $\sim U(a_d, b_d)$ for all $d \in [D]$
    3. Sample $\beta^e_d \sim U(-1, 4)$ for $d \in S^*$, otherwise $\beta^e_d = 0$
    4. $\varepsilon^e$ is an IID $n$-sample $\sim N(0, 1)$
    5. Set $\mathbf{Y}^e = \mathbf{X}^e \beta^e + \varepsilon^e$
```

Algorithm 3: Data generation in the IID case.

```
Input: $D, \mathbb{S}^*, n, E, W^0$. (In order: Global dimension, support indices, samples per interval, intervals, initial regression matrix)
Output:$(\bigcup_{e\in[E]} \mathbf{Y}^e, \bigcup_{e\in[E]} \mathbf{X}^e) \in \mathbb{R}^{E\cdot n} \times \mathbb{R}^{E\cdot n,D}$
Set $X_0 \in \mathbb{R}^D$ as initial state
For all $t \in [n \cdot E]$
    1. If $t$ mod $n = 0$: For all $d \in [D] - \{1\}$ sample $a_d \sim U(-30, 0)$ and $b_d \sim U(0, 30)$
    2. Sample $\gamma_d \in \mathbb{R}$ from $U(a_d, b_d)$ for all $d \in [D] - \{1\}$
    3. Sample $\gamma_1$ from $U(-1, 1)$
    4. Set $\mathbf{X}^{t+1} = \mathbf{X}^t W^t + \gamma$
    5. Set $\mathbf{Y}^{t+1} = \mathbf{X}_1^{t+1} + \varepsilon$ for $\varepsilon \sim N(0, 1)$
    6. Set $W_s^{t+1} = W_s^t + \eta$ for $\eta \sim N(-10^{-4}, 10^{-3})$ for all $s \in \mathbb{S}^*$. If $W_s^{t+1} > 0.9$, set $W_s^{t+1} = 0.8$, if $W_s^{t+1} < 0.9$ set $W_s^{t+1} = -0.8$
```

Algorithm 4: Data Generation in the VAR Case.

A.2 Further Experiments

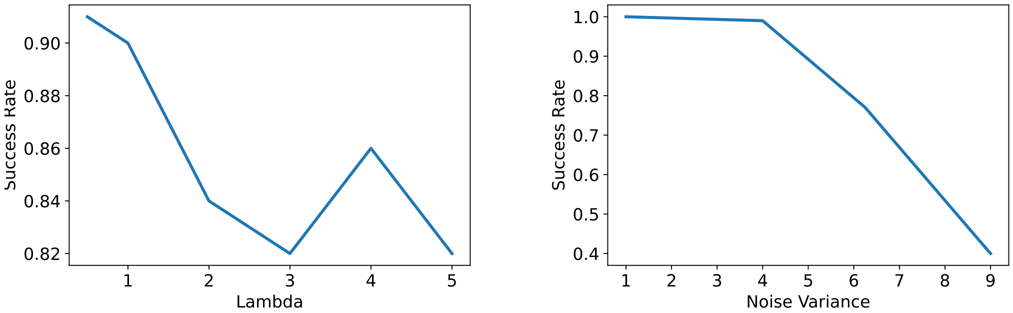
Here we discuss a few further choices of our experiments, and how we anticipate or observed them to change the outcome of the experiments.

The initial W^0 in the VAR experiment. We set the initial W^0 as

$$\begin{pmatrix} 0.5 & -0.7 & 0 & 0 & 0.9 \\ 0 & 0.4 & 1 & 0 & 0.5 \\ 0 & 0 & -0.9 & 0 & 0 \\ 0 & 0.5 & 0 & 0.8 & -1 \\ 0 & 0 & 0 & 0 & 0.4 \end{pmatrix} \tag{11}$$

Our target was the second covariate, so that for $t = 0$ we have $\beta^t = (-0.7, 0.4, 0, 0.5, 0)^T$.

The κ threshold. The choice for $\kappa = 1$ worked well in our experiments, but certainly needs to be addressed in future work. In our IID experiments, the choice was almost irrelevant, a choice of κ in $[1, 1000]$ led to the same results. In our VAR setting, however, changing κ had bigger effects. We believe, however, that the choice is still not crucial as we observed that for other values of κ we often still obtained the same results, just for a different range of λ. In future work want to investigate a replacement of κ for a well chosen hypothesis test on the equality of variances.



(a) The results of Algorithm 2 in the VAR setting, with a larger range of λ.

(b) The success rate of Algorithm 1 in the VAR setting with changing variance of the observational noise.

Figure 3: Additional experiments.

The regularization parameter λ. While a regularized regression approach allows us to perform analysis in the low sample per interval case, it forces us to set a hyperparameter λ. First note that we used in our experiments the same λ across all environments. This was fine in our experiments because the environments were comparable in sample size, but more generally one might have to chose a λ^e for each $e \in [E]$. In future work we are investigating a principled way of choosing λ, either motivated by theoretical results, or by data driven methods as leave-one-out cross validation. The other question a reader might have is, how sensitive the results are to the choice of λ. To investigate that we repeated the experiment that led to Figure 2b with a larger range of λ. The results are shown in 3a and show that increasing λ further certainly degrades the performance, although we would argue in a rather graceful than abrupt manner.

The independent noise ε^e. Looking at the proof of Theorem 1 one expects that for larger noise the methods degrade in terms of controlling the false negatives. The proof works by testing if the residuals in different environments have the same distribution. If, however, the independent noise ε^e strongly strongly dominates those residuals R^e, we expect that we cannot detect differences in those residuals. In Figure 3b we confirm this by using Algorithm 1 on the VAR data, where we provide the ground truth sample size n to the algorithm, but vary the variance of the independent noise ε^t. The results are shown in Figure 3b and we indeed observe a drop in performance for increasing noise variance.