UNIVERSITY OF COPENHAGEN

Switching Regression Models and Causal Inference in the Presence of Discrete Latent Variables

Christiansen, Rune; Peters, Jonas

Published in:
Journal of Machine Learning Research

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Christiansen, R., & Peters, J. (2020). Switching Regression Models and Causal Inference in the Presence of Discrete Latent Variables. *Journal of Machine Learning Research*, *21*, 1-46. [(41)].

Switching Regression Models and Causal Inference in the Presence of Discrete Latent Variables

Rune Christiansen[*]　　　　　　　　　　　　　　　　　　KRUNECHRISTIANSEN@MATH.KU.DK
Jonas Peters　　　　　　　　　　　　　　　　　　　　　　JONAS.PETERS@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark

Abstract

Given a response Y and a vector $X = (X^1, \ldots, X^d)$ of d predictors, we investigate the problem of inferring direct causes of Y among the vector X. Models for Y that use all of its causal covariates as predictors enjoy the property of being invariant across different environments or interventional settings. Given data from such environments, this property has been exploited for causal discovery. Here, we extend this inference principle to situations in which some (discrete-valued) direct causes of Y are unobserved. Such cases naturally give rise to switching regression models. We provide sufficient conditions for the existence, consistency and asymptotic normality of the MLE in linear switching regression models with Gaussian noise, and construct a test for the equality of such models. These results allow us to prove that the proposed causal discovery method obtains asymptotic false discovery control under mild conditions. We provide an algorithm, make available code, and test our method on simulated data. It is robust against model violations and outperforms state-of-the-art approaches. We further apply our method to a real data set, where we show that it does not only output causal predictors, but also a process-based clustering of data points, which could be of additional interest to practitioners.

Keywords: causal discovery, invariance, switching regression models, hidden Markov models, latent variables

1. Introduction

1.1. Causality

In many real world applications, we are often interested in causal rather than purely statistical relations. In the last decades, seminal work by Imbens and Rubin (2015), Spirtes et al. (2000), and Pearl (2009) has provided a solid mathematical basis for formalizing causal questions. They often start from a given causal model in the form of a structural causal model (SCM) or potential outcomes. In practice, we often do not know the underlying causal model, and the field of causal discovery aims at inferring causal models from data. There are several lines of work that are based on different assumptions. Among them are constraint-based methods (Spirtes et al., 2000; Pearl, 2009; Maathuis et al., 2009), score-

---

*. *Address for correspondence*: Rune Christiansen, Department of Mathematical Sciences, Universitetsparken 5, 2100 Copenhagen, Denmark.

based methods (Chickering, 2002; Silander and Myllymak, 2006; Koivisto, 2006; Cussens, 2011), methods based on restricted SCMs (Shimizu et al., 2006; Mooij et al., 2016; Peters et al., 2017), and methods based on the independence of causal mechanisms (Janzing et al., 2012; Steudel et al., 2010). The problem of hidden variables has been addressed in several works (e.g., Spirtes et al., 1995; Silva et al., 2006; Silva and Ghahramani, 2009; Sgouritsa et al., 2013; Claassen et al., 2013; Ogarrio et al., 2016; Silva and Evans, 2016; Richardson et al., 2017; Tsagris et al., 2018). These methods usually consider slightly different setups than our work does; e.g., they concentrate on full causal discovery (rather than estimating causal parents), and consider different model classes.

In this work, instead of aiming to learn all of the data generating structure, we consider the subproblem of inferring the set of causal parents of a target variable Y among a set of variables $X = (X^1, \ldots, X^d)$. We furthermore assume that some of the causal predictors are unobserved. While in general, this is a notoriously hard problem to solve, we will constrain the influence of the hidden variables by assuming that they take only few different values. Such a model is applicable whenever the system may be in one of several unobserved states and was motivated by an example from Earth system science, see Section 5.2. We further assume that the data are not purely observational but come from different environments.

For the case when all causal parents are observed, Peters et al. (2016) recently proposed the method *invariant causal prediction* (ICP). Under the assumption that the causal mechanism generating Y from its causal predictors remains the same in all environments ("invariant prediction"), it is possible to obtain the following guarantee: with large probability, the inferred set is a subset of the true set of causal predictors. A concise description of the method is provided in Section 1.3.

If some of the causal predictors are unobserved, the above guarantee will, in general, not hold anymore. Under the additional assumption of faithfulness, one can still prove that ICP infers a subset of the causal ancestors of the target Y. In many cases, however, the method of ICP infers the empty set, which is not an incorrect, but certainly an uninformative answer. This paper extends the idea of invariant models to situations, in which relevant parts of the system are unobserved. In particular, we suggest a relaxation of the invariance assumption and introduce the formal framework of h-invariance ("hidden invariance"). If the influence of the hidden variable is not too complex, e.g., because it takes only a few discrete values, this property is restrictive enough to be exploited for causal discovery. The assumption of h-invariance gives rise to switching regression models, where each value of the hidden variable corresponds to a different regression coefficient (we provide more details in Section 1.2). For building an invariance-based procedure, we require a test for the equality of switching regression models. In this paper, we provide such a test and show that it satisfies asymptotic level guarantees. This result allows us to prove that our causal discovery procedure is asymptotically correct under mild assumptions. In case of sequential data, we allow for the possibilities that the hidden variables follow an i.i.d. structure or a hidden Markov model (e.g., Zucchini et al., 2016). We suggest efficient algorithms, provide code and test our method on simulated and real data.

1.2. Switching Regression Models

Switching regression models are often used to model statistical dependencies that are subject to unobserved "regime switches", and can be viewed as ordinary regression models that include interactions with a discrete hidden variable. Roughly speaking, each data point (X_i, Y_i) is assumed to follow one of several different regression models; a formal definition is given in Definition 1. Switching regression models have been used in various disciplines, e.g., to model stock returns (Sander, 2018), energy prices (Langrock et al., 2017) or the propagation rate of plant infections (Turner, 2000). Statistical inference in switching regression models is a challenging problem for several reasons: switching regression models are non-identifiable (permuting mixture components does not change the modeled conditional distribution), and their likelihood function is unbounded (one may consider one of the regression models containing a single point with noise variance shrinking toward zero) and non-convex. In this paper, we circumvent the problem of an unbounded likelihood function by imposing parameter constraints on the error variances of the mixture components (e.g., Hathaway, 1985; Goldfeld and Quandt, 1973). We then construct a test for the equality of switching regression models by evaluating the joint overlap of the Fisher confidence regions (based on the maximum likelihood estimator) of the respective parameter vectors of the different models. We establish an asymptotic level guarantee for this test by providing sufficient conditions for (i) the existence, (ii) the consistency and (iii) the asymptotic normality of the maximum likelihood estimator. To the best of our knowledge, each of these three results is novel and may be of interest in itself. We further discuss two ways of numerically optimizing the likelihood function.

Without parameter constraints, the likelihood function is unbounded and global maximum likelihood estimation is an ill-posed problem (e.g., De Veaux, 1989). Some analysis has therefore been done on using local maxima of the likelihood function instead. Kiefer (1978) show that there exists a sequence of roots of the likelihood equations that yield a consistent estimator, but provide no information on which root, in case there is more than one, is consistent. Another popular approach is to impose parameter constraints on the error variances of the mixture components. In the case of ordinary, univariate Gaussian mixture models, Hathaway (1985) formulate such a constrained optimization problem and prove the existence of a global optimum. In this paper, we present a similar result for switching regression models. The proof of Hathaway (1985) uses the fact that the maximum likelihood estimates of all mean parameters are bounded by the smallest and the largest observation. This reasoning cannot be applied to the regression coefficients in switching regression models and therefore requires a modified argument. We also provide sufficient conditions for the consistency and the asymptotic normality (both up to label permutations) of the proposed constrained maximum likelihood estimator. Our proofs are based on the proofs provided by Bickel et al. (1998) and Jensen and Petersen (1999), who show similar results for the maximum likelihood estimator in hidden Markov models with finite state space. Together, (ii) and (iii) prove the asymptotic coverage of Fisher confidence regions and ensure the asymptotic level guarantee of our proposed test.

Readers mainly interested in inference in switching regression models, may want to skip directly to Section 3. Additionally, Sections 2.5 and 2.6 contain our proposed test for the

3

equality of switching regression models that is available as the function `test.equality.sr` in our code package.

1.3. The Principle of Invariant Causal Prediction

This section follows the presentation provided by Pfister et al. (2019b). Suppose that we observe several instances $(Y_1, X_1), \ldots, (Y_n, X_n)$ of a response or target variable $Y \in \mathbb{R}$ and covariates $X \in \mathbb{R}^{1 \times d}$. We assume that the instances stem from different environments $e \subseteq \{1, \ldots, n\}$, and use $\mathcal{E}$ to denote the collection of these, i.e., $\bigcup_{e \in \mathcal{E}} e = \{1, \ldots, n\}$. These environments can, for example, correspond to different physical or geographical settings in which the system is embedded, or controlled experimental designs in which some of the variables have been intervened on. The crucial assumption is then that there exists a subset $S^* \subseteq \{1, \ldots, d\}$ of variables from X that yield a predictive model for Y that is invariant across all environments.

More formally, one assumes the existence of a set $S^* \subseteq \{1, \ldots, d\}$, such that for all x and all $1 \leq s, t \leq n$, we have

$$Y_s \mid (X_s^{S^*} = x) \stackrel{d}{=} Y_t \mid (X_t^{S^*} = x), \tag{1}$$

where $X_t^{S^*}$ denotes the covariates in S^* at instance t. For simplicity, the reader may think about (1) in terms of conditional densities. Also, the reader might benefit from thinking about the set S^* in the context of causality, which is why we will below refer to the set S^* as the set of (observable) direct causes of the target variable. If, for example, data come from a structural causal model (which we formally define in Appendix A), and different interventional settings, a sufficient condition for (1) to hold is that the structural assignment for Y remains the same across all observations, i.e., there are no interventions occurring directly on Y. In Section 2.3, we will discuss the relationship to causality in more detail. Formally, however, this paper does not rely on the definition of the term "direct causes".

Since each instance is only observed once, it is usually hard to test whether Equation (1) holds. We therefore make use of the environments. Given a set $S \subseteq \{1, \ldots, d\}$, we implicitly assume that for every $e \in \mathcal{E}$, the conditional distribution $P_{Y_t | X_t^S}$[1] is the same for all $t \in e$, say $P_{Y|X^S}^e$, and check whether for all $e, f \in \mathcal{E}$, we have that

$$P_{Y|X^S}^e = P_{Y|X^S}^f. \tag{2}$$

In the population case, Equation (2) can be used to recover (parts of) S^* from the conditional distributions $P_{Y|X^S}^e$: for each subset $S \subseteq \{1, \ldots, d\}$ of predictors we check the validity of (2) and output the set

$$\tilde{S} := \bigcap_{S \text{ satisfies (2)}} S \tag{3}$$

of variables that are necessary to obtain predictive stability. Under assumption (1), $\tilde{S}$ only contains variables from S^*. For purely observational data, i.e., $(Y_t, X_t) \stackrel{d}{=} (Y_s, X_s)$ for all s, t, Equation (2) is trivially satisfied for any set $S \subseteq \{1, \ldots, d\}$ and thus $\tilde{S} = \emptyset$. It is

---

1. We use $P_{Y_t | X_t^S}$ as shorthand notation for the family $\left(P_{Y_t | (X_t^S = x)} \right)_x$ of conditional distributions.

the different heterogeneity patterns of the data in different environments that allow for causal discovery. If only a single i.i.d. data set is available, the method's result would not be incorrect, but it would not be informative either. Based on a sample from $(Y_t, X_t)_{t \in e}$ for each environment, Peters et al. (2016) propose an estimator $\hat{S}$ of $\tilde{S}$ that comes with a statistical guarantee: with controllable (large) probability, the estimated set $\hat{S}$ is contained in S^*. In other words, whenever the method outputs a set of predictors, they are indeed causal with high certainty.

In this paper, we consider cases in which the full set of direct causes of Y is not observed. We then aim to infer the set of *observable* causal variables $S^* \subseteq \{1, \ldots, d\}$. Since the invariance assumption (1) cannot be expected to hold in this case, the principle of invariant prediction is inapplicable. We therefore introduce the concept of h-invariance, a relaxed version of assumption (1). If the the latent variables are constrained to take only few values, the h-invariance property can, similarly to (3), be used for the inference of S^*.

1.4. Organization of the Paper

The remainder of the paper is organized as follows. Section 2 explains in which sense the principle of invariant causal prediction breaks down in the presence of hidden variables and proposes an adaptation of the inference principle. It also contains hypothesis tests that are suitable for the setting with hidden variables. In Section 3, we establish asymptotic guarantees for these tests. This section contains all of our theoretical results on the inference in switching regression models, and can be read independently of the problem of causal inference. In Section 4, we combine the results of the preceding sections into our overall causal discovery method (ICPH), provide an algorithm and prove the asymptotic false discovery control of ICPH. The experiments on simulated data in Section 5 support these theoretical findings. They further show that even for sample sizes that are too small for the asymptotic results to be effective, the overall method generally keeps the type I error control. The method is robust against a wide range of model misspecifications and outperforms other approaches. We apply our method to a real world data set on photosynthetic activity and vegetation type. Proofs of our theoretical results are contained in Appendix C. All our code is available as an `R` package at `https://github.com/runesen/icph`, and can be installed by `devtools::install_github("runesen/icph/code")`, for example. Scripts reproducing all simulations can be found at the same url.

2. Invariant Causal Prediction in the Presence of Latent Variables

Consider a collection $(\mathbf{Y}, \mathbf{X}, \mathbf{H}) = (Y_t, X_t, H_t)_{t \in \{1, \ldots, n\}}$ of triples of a target variable $Y_t \in \mathbb{R}$, observed covariates $X_t \in \mathbb{R}^{1 \times d}$ and some latent variables $H_t \in \mathbb{R}^{1 \times k}$. For simplicity, we refer to the index t as time, but we also allow for an i.i.d. setting; see Section 3.1 for details. When referring to properties of the data that hold true for all t, we sometimes omit the index altogether.

In analogy to Section 1.3, we start by assuming the existence of an invariant predictive model for Y, but do not require all relevant variables to be observed. That is, we assume the existence of a set $S^* \subseteq \{1, \ldots, d\}$ and a subvector H^* of H such that the conditional

distribution of $Y_t \,|\, (X_t^{S^*}, H_t^*)$ is the same for all time points t. Based on the observed data $(\mathbf{Y}, \mathbf{X})$, we then aim to infer the set S^*.

Section 2.1 shows why the original version of invariant causal prediction is inapplicable. In Sections 2.2 and 2.4 we introduce the formal concept of h-invariance and present an adapted version of the inference principle discussed in Section 1.3. In Sections 2.5 and 2.6 we then present tests for h-invariance of sets $S \subseteq \{1, \ldots, d\}$, which are needed for the construction of an empirical estimator $\hat{S}$ of S^*. A causal interpretation of the h-invariance property is given in Section 2.3.

2.1. Latent Variables and Violation of Invariance

The inference principle described in Section 1.3 relies on the invariance assumption (1). The following example shows that if some of the invariant predictors of Y are unobserved, we cannot expect this assumption to hold. The principle of ordinary invariant causal prediction is therefore inapplicable.

Example 1 (Violation of invariance assumption due to latent variables) *We consider a linear model for the data $(Y_t, X_t^1, X_t^2, H_t^*)_{t \in \{1, \ldots, n\}} \in \mathbb{R}^{n \times 4}$. Assume there exist i.i.d. zero-mean noise variables $\varepsilon_1, \ldots, \varepsilon_n$ such that for all t, $(X_t^1, H_t^*, \varepsilon_t)$ are jointly independent and*

$$Y_t = X_t^1 + H_t^* + \varepsilon_t.$$

Assume furthermore that the distribution of the latent variable H_t^* changes over time, say $\mathbb{E}[H_r^*] \neq \mathbb{E}[H_s^*]$ for some r, s. Then, with $S^* := \{1\}$, the conditional distribution $P_{Y_t | (X_t^{S^*}, H_t^*)}$ is time-homogeneous, but

$$\mathbb{E}[Y_r | X_r^{S^*} = x] = x + \mathbb{E}[H_r^*] \neq x + \mathbb{E}[H_s^*] = \mathbb{E}[Y_s | X_s^{S^*} = x],$$

which shows that $P_{Y_t | X_t^{S^*}}$ is not time-homogeneous, i.e., S^* does not satisfy (1).

The above example shows that in the presence of hidden variables, assumption (1) may be too strong. The distribution in the above example, however, allows for a different invariance. For all t, s and all x, h we have that[2]

$$Y_t \,|\, (X_t^{S^*} = x, H_t^* = h) \overset{d}{=} Y_s \,|\, (X_s^{S^*} = x, H_s^* = h). \tag{4}$$

Ideally, we would like to directly exploit this property for the inference of S^*. Given a candidate set $S \subseteq \{1, \ldots, d\}$, we need to check if there exist $H_1^*, \ldots, H_n^*$ such that (4) holds true for $S^* = S$. Similarly to (3), the idea is then to output the intersection of all sets for which this is the case. Without further restrictions on the influence of the latent variables, however, the result will always be the empty set.

---

2. In the remainder of this work, we implicitly assume that for every t, (Y_t, X_t, H_t) is abs. continuous w.r.t. a product measure. This ensures the existence of densities $f_t(y, x, h)$ for (Y_t, X_t, H_t). The marginal density $f_t(x, h)$ can be chosen strictly positive on the support of (X_t, H_t) and thereby defines a set of conditional distributions $\{Y_t \,|\, (X_t = x, H_t = h)\}_{(x,h) \in \mathrm{supp}((X_t, H_t))}$ via the conditional densities $f_t(y \,|\, x, h) = f_t(y, x, h) / f_t(x, h)$. Strictly speaking, we therefore assume that the conditional distributions *can be chosen* s.t. (4) holds for all $(x, h) \in \mathrm{supp}((X_t^{S^*}, H_t^*)) \cap \mathrm{supp}((X_s^{S^*}, H_s^*))$.

Proposition 1 (Necessity of constraining the influence of H^*) *Let $S \subseteq \{1,\ldots,d\}$ be an arbitrary subset of the predictors X_t. Then, there exist variables $H_1,\ldots,H_n$ such that (4) is satisfied for $S^* = S$ and $(H_t^*)_{t \in \{1,\ldots,n\}} = (H_t)_{t \in \{1,\ldots,n\}}$.*

The proof is immediate by choosing latent variables with non-overlapping support (e.g., such that for all t, $P(H_t = t) = 1$). Proposition 1 shows that without constraining the influence of H^*, (4) cannot be used to identify S^*. Identifiability improves, however, for univariate, discrete latent variables $H^* \in \{1,\ldots,\ell\}$ with relatively few states $\ell \geq 2$. Equation (4) then translates into the following assumption on the observed conditional distributions $P_{Y_t \mid X_t^{S^*}}$: for all t, x it holds that

$$P_{Y_t \mid (X_t^{S^*} = x)} = \sum_{j=1}^{\ell} \lambda_{xt}^j P_x^j, \tag{5}$$

for some $\lambda_{xt}^1, \ldots, \lambda_{xt}^\ell \in (0,1)$ with $\sum_{j=1}^{\ell} \lambda_{xt}^j = 1$ and distributions $P_x^1, \ldots, P_x^\ell$ that do not depend on t. This fact can be seen by expressing the conditional density of $P_{Y_t \mid (X_t^{S^*} = x)}$ as $f_t(y \mid x) = \int f_t(y \mid x, h) f_t(h \mid x) dh$. By (4), $f_t(y \mid x, h)$ does not depend on t. Property (5) then follows by taking $\lambda_{xt}^j = P(H_t^* = j \mid X_t^{S^*} = x)$ and letting P_x^j denote the distribution of $Y_1 \mid (X_1^{S^*} = x, H_1^* = j)$.

The conditional distributions of $Y_t \mid (X_t^{S^*} = x)$ are thus assumed to follow mixtures of ℓ distributions, each of which remains invariant across time. The mixing proportions λ_{xt} may vary over time. In the following subsection, we translate property (5) into the framework of mixtures of linear regressions with Gaussian noise. The invariance assumption on $P_x^1, \ldots P_x^\ell$ then corresponds to time-homogeneity of the regression parameters of all mixture components.

2.2. Hidden Invariance Property

As motivated by Proposition 1, we will from now on assume that H^* only takes a small number of different values. We now formalize the dependence of Y on (X^{S^*}, H^*) by a parametric function class. We purposely refrain from modeling the dependence between observations of different time points, and come back to that topic in Section 3.1. Since the inference principle described in Section 1.3 requires us to evaluate (5) for different candidate sets S, we state the following definition in terms of a general p-dimensional vector X (which will later play the role of the subvectors X^S, see Definition 2).

Definition 1 (Switching regression) *Let X be a p-dimensional random vector, $\ell \in \mathbb{N}$ and $\lambda \in (0,1)^\ell$ with $\sum_{j=1}^{\ell} \lambda_j = 1$. Let furthermore Θ be a matrix of dimension $(p+2) \times \ell$ with columns $\Theta_{\cdot j} = (\mu_j, \beta_j, \sigma_j^2) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}_{>0}$, for $j \in \{1,\ldots,\ell\}$. The joint distribution P of $(Y,X) \in \mathbb{R}^{(1+p)}$ is said to follow a switching regression of degree ℓ with parameters (Θ,λ), if there exist $H \sim \text{Multinomial}(1,\lambda)$ and $\varepsilon_j \sim \mathcal{N}(0,\sigma_j^2), j \in \{1,\ldots,\ell\}$, with $(\varepsilon_1,\ldots,\varepsilon_\ell) \perp\!\!\!\perp X$, such that*

$$Y = \sum_{j=1}^{\ell} (\mu_j + X\beta_j + \varepsilon_j) \mathbb{1}_{\{H=j\}},$$

where $\mathbb{1}_{\{H=j\}}$ denotes the indicator function for the event $H = j$.

A few remarks are in place. First, we will as of now let $\ell \geq 2$ be fixed. The reader is encouraged to think of $\ell = 2$, which is also the case to be covered in most examples and experiments. (Non-binary latent variables are considered in Appendix E.1.) Second, it will be convenient to parametrize the matrix Θ by a map $\theta \mapsto \Theta(\theta)$, $\theta \in \mathcal{T}$, where $\mathcal{T}$ is a subset of a Euclidean space. This allows for a joint treatment of different types of parameter contraints such as requiring all intercepts or all variances to be equal. We will use $\mathcal{SR}_{\Theta}(\theta, \lambda \,|\, X)$ ("Switching Regression") to denote the distribution P over (Y, X) satisfying Definition 1 with parameters $(\Theta(\theta), \lambda)$, although we will often omit the implicit dependence on Θ and simply write $\mathcal{SR}(\theta, \lambda \,|\, X)$. For now, the reader may think of $(\Theta, \mathcal{T})$ as the unconstrained parametrization, where $\mathcal{T} = (\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}_{>0})^{\ell}$ and where Θ consists of the coordinate projections $\Theta_{ij}(\theta) = \theta_{(j-1)(p+2)+i}$. Finally, we will for the rest of this paper disregard the intercept terms μ_j as they can be added without loss of generality by adding a constant predictor to X.

The following definition and assumption translate (5) into the model class $\mathcal{SR}$.

Definition 2 (h-invariance) *A set $S \subseteq \{1, \ldots, d\}$ is called h-invariant w.r.t. $(\mathbf{Y}, \mathbf{X}) = (Y_t, X_t)_{t \in \{1, \ldots, n\}}$ if there exist θ and $\lambda_1, \ldots, \lambda_n$ such that, for all t, $P_{(Y_t, X_t^S)} = \mathcal{SR}(\theta, \lambda_t \,|\, X_t^S)$.*

Definition 2 describes an invariance in the regression parameters θ and makes no restriction on the mixing proportions $\lambda_1, \ldots, \lambda_n$. This allows the influence of the latent variable to change over time. From now on, we assume the existence of an h-invariant set S^*.

Assumption 1 *There exists a set $S^* \subseteq \{1, \ldots, d\}$ which is h-invariant w.r.t. $(\mathbf{Y}, \mathbf{X})$.*

This assumption is at the very core of the proposed methodology, with the unknown h-invariant set S^* as inferential target. In Section 2.3 we show that if the data $(\mathbf{Y}, \mathbf{X}, \mathbf{H})$ are generated by different interventions in an SCM (see Appendix A), in which the variable $H^* \in \{1, \ldots, \ell\}$ acts on Y, Assumption 1 is satisfied by the set $S^* = \mathrm{PA}^0(Y)$ of observable parents of Y. Here, interventions are allowed to act on the latent variables, and thus indirectly on the target Y. For illustrations of the h-invariance property, see Figures 1 and 2.

2.3. Relation to Causality

Assumption 1 is formulated without the notion of causality. The following proposition shows that if the data $(\mathbf{Y}, \mathbf{X}, \mathbf{H})$ do come from an SCM, the set S^* may be thought of as the set of observable parents of Y.

Proposition 2 (Causal interpretation of S^*) *Consider an SCM over the system of variables $(Y_t, X_t, H_t^*)_{t \in \{1, \ldots, n\}}$, where for every t, $(Y_t, X_t, H_t^*) \in \mathbb{R}^1 \times \mathbb{R}^d \times \{1, \ldots, \ell\}$. Assume that the structural assignment of Y is fixed across time, and for every $t \in \{1, \ldots, n\}$ given by*

$$Y_t := f(X_t^{\mathrm{PA}^0(Y)}, H_t^*, N_t),$$

where $(N_t)_{t \in \{1, \ldots, n\}}$ are i.i.d. noise variables. Here, $\mathrm{PA}^0(Y) \subseteq \{1, \ldots, d\}$ denotes the set of parents of Y_t among $(X_t^1, \ldots, X_t^d)$. The structural assignments for the remaining variables $X^1, \ldots, X^d, H^*$ are allowed to change between different time points. Then, property (4) is satisfied for $S^* = \mathrm{PA}^0(Y)$. If furthermore the assignment $f(\cdot, h, \cdot)$ is linear for all $h \in \{1, \ldots, \ell\}$ and the noise variables N_t are normally distributed, then, Assumption 1
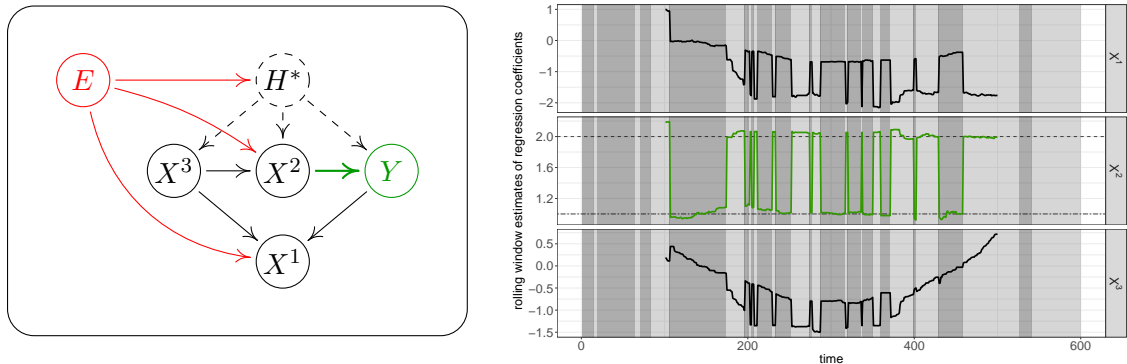
Figure 1: An illustration of the h-invariance property based on simulated data from the SCM in Example 2. The causal graph (left) and rolling window estimates of regression coefficients in the linear interaction model for the conditional distribution of Y given (X^1, H^*), (X^2, H^*) and (X^3, H^*), respectively (right). Within both regimes $H_t^* = 1$ and $H_t^* = 2$ (corresponding to different background colors in the plot), the regression coefficient for X^2 (green) is time-homogeneous, and the set $S^* = \{2\}$ is therefore h-invariant with respect to $(\mathbf{Y}, \mathbf{X})$. Due to heterogeneity in the data ("the variable E acts on X^1, X^2 and H^*"), neither of the sets $\{1\}$ or $\{3\}$ satisfy h-invariance. In practice, we test for h-invariance using environments, rather than rolling windows, see Section 2.5.

is satisfied for $S^* = \mathrm{PA}^0(Y)$. That is, the set of observable parents of Y is h-*invariant with respect to* $(\mathbf{Y}, \mathbf{X}) = (Y_t, X_t)_{t \in \{1, \dots, n\}}$.

From a causal perspective, Proposition 2 informs us about the behavior of $P_{Y|(X^{S^*}=x)}$ under interventions in the data generating process. The set $S^* = \mathrm{PA}^0(Y)$ will be h-invariant under any type of intervention that does not occur directly on the target variable (except through the latent variable H^*). The following example demonstrates the h-invariance property for an SCM in which the assignments of some of the variables change between every time point.

Example 2 *Consider an SCM over the system of variables* $(Y_t, X_t, H_t^*)_{t \in \{1, \dots, n\}}$, *where for every t, the causal graph over* $(Y_t, X_t, H_t^*) \in \mathbb{R}^1 \times \mathbb{R}^3 \times \{1, 2\}$ *is given as in Figure 1. The node E denotes the "environment variable" and the outgoing edges from E to X^1, X^2 and H^* indicate that the structural assignments of these variables change throughout time. The structural assignment of Y is fixed across time, and for every $t \in \{1, \dots, n\}$ given by*

$$Y_t := (1 + X_t^2 + 0.5N_t)1_{\{H_t^*=1\}} + (1 + 2X_t^2 + 0.7N_t)1_{\{H_t^*=2\}},$$

where $(N_t)_{t \in \{1, \dots, n\}}$ *are i.i.d. standard Gaussian noise variables. Then, by Proposition 2, the set $S^* = \{2\}$ of observable parents of Y is h-invariant w.r.t.* $(\mathbf{Y}, \mathbf{X})$, *see Figure 1.*

2.4. Inference of the h-Invariant Set

In general, Definition 2 does not define a unique set of predictors. In analogy to Peters et al. (2016), we thus propose to output the intersection of all h-invariant sets. We define

$$H_{0,S} : S \text{ is } h\text{-invariant with respect to } (\mathbf{Y}, \mathbf{X}), \text{ and} \tag{6}$$

$$\tilde{S} := \bigcap_{S : H_{0,S} \text{ true}} S, \tag{7}$$

where S runs over subsets $S \subseteq \{1, \ldots, d\}$. In (7), we define the intersection over an empty index set as the empty set. In practice, we are given a sample from $(\mathbf{Y}, \mathbf{X})$, and our goal is to estimate $\tilde{S}$. Given a family of tests $(\varphi_S)_{S \subseteq \{1, \ldots, d\}}$ of the hypotheses $(H_{0,S})_{S \subseteq \{1, \ldots, d\}}$, we therefore define an empirical version of (7) by

$$\hat{S} := \bigcap_{S : \varphi_S \text{ accepts } H_{0,S}} S. \tag{8}$$

Using that $\{\varphi_{S^*} \text{ accepts } H_{0,S^*}\} \subseteq \{\hat{S} \subseteq S^*\}$, we immediately obtain the following important coverage property.

Proposition 3 (Coverage property) *Under Assumption 1 and given a family of tests $(\varphi_S)_{S \subseteq \{1, \ldots, d\}}$ of $(H_{0,S})_{S \subseteq \{1, \ldots, d\}}$ that are all valid at level α, we have that $\mathbb{P}(\hat{S} \subseteq S^*) \geq 1 - \alpha$. In words, the (setwise) false discovery rate of (8) is controlled at level α.*

The set S^* in Proposition 3 may not be uniquely determined by the h-invariance property. But since our output is the *intersection* (8) of all h-invariant sets, this ambiguity does no harm—the coverage guarantee for the inclusion $\hat{S} \subseteq S^*$ will be valid for *any* choice of h-invariant set S^*. The key challenge that remains is the construction of the tests $(\varphi_S)_{S \subseteq \{1, \ldots, d\}}$, which we will discuss in Section 2.5.

2.4.1. TESTS FOR NON-CAUSALITY OF INDIVIDUAL PREDICTORS

Proposition 3 proves a level guarantee for the estimator $\hat{S}$. To obtain statements about the significance of individual predictors that could be used for a ranking of all the variables in X, for example, we propose the following construction. Whenever at least one hypothesis $H_{0,S}$ is accepted, we define for every $j \in \{1, \ldots, d\}$ a p-value for the hypothesis $H_0^j : j \notin S^*$ of non-causality of X^j by $p_j := \max\{p\text{-value for } H_{0,S} : j \notin S\}$. When all hypotheses $H_{0,S}$, $S \subseteq \{1, \ldots, d\}$, are rejected (corresponding to rejecting the existence of S^*), we set all of these p-values to 1. The validity of thus defined tests is ensured under the assumptions of Proposition 3, and is a direct consequence of φ_{S^*} achieving correct level α.

2.5. Tests for the Equality of Switching Regression Models

We will now focus on the construction of tests for the hypotheses $H_{0,S}$ that are needed to compute the empirical estimator (8). Let $S \subseteq \{1, \ldots, d\}$ be fixed for the rest of this section. We will make use of the notation $\mathbf{X}^S$ to denote the columns of $\mathbf{X}$ with index in S and
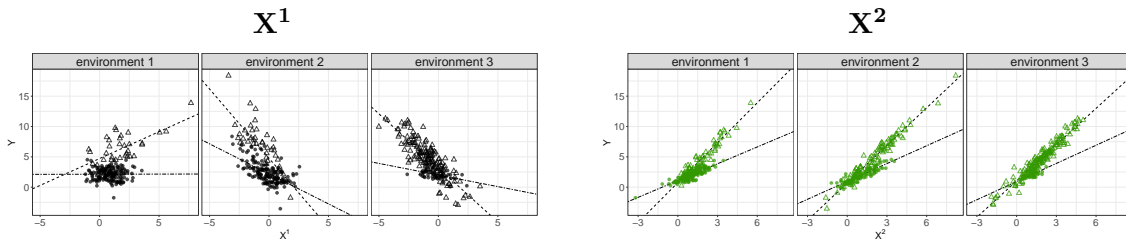
Figure 2: Testing procedure for $H_{0,S}$, here illustrated for the sets $\{1\}$ (black; not h-invariant) and $\{2\}$ (green; h-invariant) using the same data that generated Figure 1. First, we split data up into several environments, here $e_1 = \{1, \ldots, 200\}$, $e_2 = \{201, \ldots, 400\}$ and $e_3 = \{401, \ldots, 600\}$. Then, we fit an $\mathcal{SR}$ model to each data set $(\mathbf{Y}_e, \mathbf{X}_e^S)$, $e \in \mathcal{E}$, separately, and evaluate whether the mixture components remain invariant across all environments. For illustration purposes, we indicate model fits by dashed lines, and assign points to the most likely hidden state ($\bullet : \hat{H}_t^* = 1$, $\triangle : \hat{H}_t^* = 2$). (This explicit classification of points is not part of the proposed testing procedure.)

$\mathbf{Y}_e = (Y_t)_{t \in e}$ and $\mathbf{X}_e^S = (X_t^S)_{t \in e}$ for the restrictions of $\mathbf{Y}$ and $\mathbf{X}^S$ to environment $e \in \mathcal{E}$. For notational convenience, we rewrite $H_{0,S}(\mathcal{E}) := H_{0,S}$ as follows.

$$H_{0,S}(\mathcal{E}) : \begin{cases} \text{There exist } \lambda_1, \ldots, \lambda_n \text{ and } (\theta_e)_{e \in \mathcal{E}}, \text{ such that, for all } e \in \mathcal{E}, \\ P_{(Y_t, X_t^S)} = \mathcal{SR}(\theta_e, \lambda_t \mid X_t^S) \text{ if } t \in e, \text{ and for all } e, f \in \mathcal{E}, \theta_e = \theta_f. \end{cases}$$

Intuitively, a test $\varphi_S = \varphi_S(\mathcal{E})$ of $H_{0,S}(\mathcal{E})$ should reject whenever the parameters θ_e and θ_f differ between at least two environments $e, f \in \mathcal{E}$. This motivates a two-step procedure:

(i) For every $e \in \mathcal{E}$, fit an $\mathcal{SR}$ model to $(\mathbf{Y}_e, \mathbf{X}_e^S)$ to obtain an estimate $\hat{\theta}_e$ with confidence intervals, see Section 3.

(ii) Based on (i), test if $\theta_e = \theta_f$ for all $e, f \in \mathcal{E}$, see Section 2.6.

For (i), we use maximum likelihood estimation and construct individual confidence regions for the estimated parameters $\hat{\theta}_e$ using the asymptotic normality of the MLE. For (ii), we evaluate the joint overlap of these confidence regions. Any other test for the equality of $\mathcal{SR}$ models can be used here, but to the best of our knowledge, we propose the first of such tests. Figure 2 illustrates step (i) for the two canditate sets $\{1\}$ and $\{2\}$. Here, we would expect a test to reject the former set, while accepting the truly h-invariant set $S^* = \{2\}$. A generic approach for comparing ordinary linear regression models across different environments can be based on exact resampling of the residuals (e.g., Pfister et al., 2019b). This procedure, however, is not applicable to mixture models: after fitting the mixture model, the states H_t are unobserved, and thus, there are multiple definitions of the residual $r_t^j = Y_t - X_t^S \hat{\beta}_j$, $j \in \{1, \ldots, \ell\}$.

2.6. Intersecting Confidence Regions

Assume $H_{0,S}(\mathcal{E})$ is true and let θ_0 be the true vector of regression parameters (that is the same for all environments). If for $e \in \mathcal{E}$, $C_e^\alpha = C_e^\alpha(\mathbf{Y}_e, \mathbf{X}_e^S)$ are valid $(1 - \alpha)$–confidence

regions for $\theta_e = \theta_0$, we can obtain a p-value for $H_{0,S}(\mathcal{E})$ by considering their joint overlap. More formally, we construct the test statistic $T_S : \mathbb{R}^{n \times (1+|S|)} \to [0,1]$ by

$$T_S(\mathbf{Y}, \mathbf{X}^S) := \max \left\{ \alpha \in [0,1] : \bigcap_{e \in \mathcal{E}} C_e^{\alpha/|\mathcal{E}|}(\mathbf{Y}_e, \mathbf{X}_e^S) \neq \emptyset \right\}, \qquad (9)$$

and define a test φ_S^α by $\varphi_S^\alpha = 1 :\Leftrightarrow T_S < \alpha$. Due to the Bonferroni correction of the confidence regions, such a test will be conservative. The construction of confidence regions is discussed in the following section.

3. Inference in Switching Regression Models

In this section, we discuss maximum likelihood estimation and the construction of confidence regions for the parameters in $\mathcal{SR}$ models. In Sections 3.1–3.2 we present two different models for time dependencies in the data, introduce the likelihood function for $\mathcal{SR}$ models, and present two types of parameter constraints that ensure the existence of the maximum likelihood estimator. In Section 3.3–3.4 we construct confidence regions based on the maximum likelihood estimator, and in Section 3.5 we show that these confidence regions attain the correct asymptotic coverage. As a corollary, we obtain that the test defined in (9) satisfies asymptotic type I error control.

Let $S \subseteq \{1, \ldots, d\}$ and consider a fixed environment e, say $e = \{1, \ldots, m\}$. Throughout this section, we will omit all indications of S and e and simply write $(Y_t, X_t) \in \mathbb{R}^{1+p}$ for (Y_t, X_t^S) and $(\mathbf{Y}, \mathbf{X})$ for $(\mathbf{Y}_e, \mathbf{X}_e^S)$.

3.1. Time Dependence and Time Independence

Assume there exist parameters θ and $\lambda_1, \ldots, \lambda_m$ such that, for all $t \in \{1, \ldots, m\}$, $(Y_t, X_t) \sim \mathcal{SR}(\theta, \lambda_t \,|\, X_t)$. Let $\mathbf{H} = (H_t)_{t \in \{1,\ldots,m\}} \in \{1, \ldots, \ell\}^m$ be such that for every $t \in \{1, \ldots, m\}$, the distributional statement in Definition 1 holds for (Y_t, X_t, H_t). We will now consider two different models for the dependence between observations of different time points:

- Independent observations ("IID"): All observations (Y_t, X_t, H_t) across different time points $t = 1, \ldots, m$ are jointly independent and the marginal distribution of $\mathbf{H}$ is time-homogeneous. Furthermore, for every $t \in \{1, \ldots, m\}$, the variables X_t and H_t are independent.

- A hidden Markov model ("HMM"): The dependence in the data is governed by a first order Markovian dependence structure on the latent variables $\mathbf{H}$ as described in Figure 3. The Markov chain $\mathbf{H}$ is initiated in its stationary distribution. Furthermore, for every $t \in \{1, \ldots, m\}$, the variables X_t and H_t are independent.

We conveniently assume the independence of X and H, which allows for likelihood inference without explicitly modelling the distribution of X. Our robustness analysis in Section 5.1.5 suggests, however, that violations of this assumption do not negatively affect the performance of our causal discovery method.

For $i, j \in \{1, \ldots, \ell\}$, let $\Gamma_{ij} = P(H_t = j \,|\, H_{t-1} = i)$ denote the transition probabilities of $\mathbf{H}$. By considering different parametrizations $\gamma \mapsto \mathbf{\Gamma}(\gamma)$, $\gamma \in \mathcal{G}$, where $\mathcal{G}$ is a subset of a
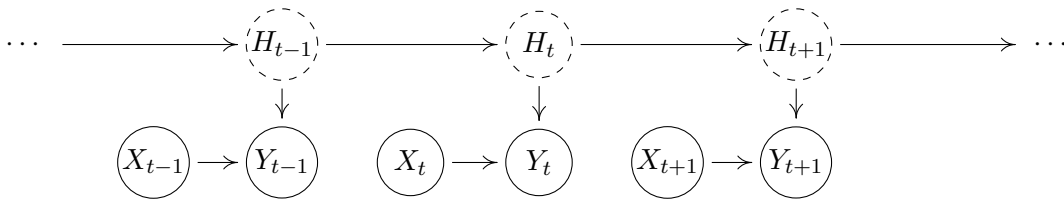
Figure 3: A hidden Markov model for $(\mathbf{Y}, \mathbf{X})$. All observations (across different $t \in \{1, \ldots, m\}$) are conditionally independent given $\mathbf{H}$, and (Y_t, X_t) only depends on $\mathbf{H}$ through the present state H_t. Moreover, the variables in $\mathbf{H}$ resemble a first order Markov chain, that is, $(H_1, \ldots, H_{t-1}) \perp\!\!\!\perp H_{t+1} \mid H_t$ for all $t \in \{2, \ldots, m-1\}$.

Euclidean space, we can encompass both of the above models simultaneously. The model IID then simply corresponds to a map $\mathbf{\Gamma}$ satisfying that, for every $\gamma \in \mathcal{G}$, $\mathbf{\Gamma}(\gamma)$ has constant columns. For details on the parametrizations of the models IID and HMM, see Appendix B.

3.1.1. NOTATION

The characteristics of the model for the joint distribution of $(\mathbf{Y}, \mathbf{X})$ are determined by the parametrizations $(\mathbf{\Theta}, \mathcal{T})$ and $(\mathbf{\Gamma}, \mathcal{G})$ of the regression matrix Θ and the transition matrix Γ, respectively. For every $\gamma \in \mathcal{G}$, let $\lambda(\gamma) = \lambda(\mathbf{\Gamma}(\gamma)) \in \mathbb{R}^{1 \times \ell}$ be the stationary distribution of $\mathbf{\Gamma}(\gamma)$. The stationary distribution $\lambda(\gamma)$ exists (and is unique) if the matrix $\mathbf{\Gamma}(\gamma)$ is irreducible and aperiodic (e.g., Ching and Ng, 2006, Propositions 1.31–1.33). In the remainder of this work, we therefore require the image $\mathbf{\Gamma}(\mathcal{G})$ to be a subset of the space of irreducible and aperiodic matrices of dimension $\ell \times \ell$. We use $\mathcal{SR}_{(\mathbf{\Theta}, \mathbf{\Gamma})}(\theta, \gamma \mid \mathbf{X})$ to denote the joint distribution P over $(\mathbf{Y}, \mathbf{X})$ with marginals $(Y_t, X_t) \sim \mathcal{SR}_{\mathbf{\Theta}}(\theta, \lambda(\gamma) \mid X_t)$ and a dependence structure given by $\mathbf{\Gamma}(\gamma)$. Unless explicit parametrizations are referred to, we will usually omit the dependence on $\mathbf{\Theta}$ and $\mathbf{\Gamma}$ and simply write $\mathcal{SR}(\theta, \gamma \mid \mathbf{X})$. For every $j \in \{1, \ldots, \ell\}$, we use $\beta_j(\cdot)$ and $\sigma_j^2(\cdot)$ to denote the parametrizations of the jth regression coefficient and the jth error variance, respectively, as induced by $(\mathbf{\Theta}, \mathcal{T})$. Finally, ϕ denotes the combined parameter vector (θ, γ) with corresponding parameter space $\mathcal{P} := \mathcal{T} \times \mathcal{G}$.

3.2. Likelihood

Consider a fixed pair of parametrizations $(\mathbf{\Theta}, \mathcal{T})$ and $(\mathbf{\Gamma}, \mathcal{G})$. For $(\theta, \gamma) \in \mathcal{T} \times \mathcal{G}$, the joint density of $(\mathbf{Y}, \mathbf{X}, \mathbf{H})$ induced by the distribution $\mathcal{SR}(\theta, \gamma \mid \mathbf{X})$ is given by

$$p_{(\mathbf{\Theta}, \mathbf{\Gamma})}(\mathbf{y}, \mathbf{x}, \mathbf{h} \mid \theta, \gamma) = p(\mathbf{x})\lambda(\gamma)_{h_1} \prod_{s=2}^{m} \mathbf{\Gamma}_{h_{s-1}h_s}(\gamma) \prod_{t=1}^{m} \mathcal{N}(y_t \mid x_t \beta_{h_t}(\theta), \sigma_{h_t}^2(\theta)),$$

where $p(\mathbf{x})$ is the (unspecified) density of $\mathbf{X}$, and where, for $j \in \{1, \ldots, \ell\}$, $\mathcal{N}(y_t \mid x_t \beta_j, \sigma_j^2)$ is short hand notation for the density of a $\mathcal{N}(x_t \beta_j, \sigma_j^2)$ distribution evaluated at y_t. Given a sample $(\mathbf{y}, \mathbf{x})$ from $(\mathbf{Y}, \mathbf{X})$, the loglikelihood function for the model $\{\mathcal{SR}(\theta, \gamma \mid \mathbf{X}) : (\theta, \gamma) \in \mathcal{T} \times \mathcal{G}\}$ is then given by

$$\ell_{(\mathbf{\Theta}, \mathbf{\Gamma})}(\mathbf{y}, \mathbf{x} \mid \theta, \gamma) = \log \sum_{h_1} \cdots \sum_{h_m} p_{(\mathbf{\Theta}, \mathbf{\Gamma})}(\mathbf{y}, \mathbf{x}, \mathbf{h} \mid \theta, \gamma), \qquad (\theta, \gamma) \in \mathcal{T} \times \mathcal{G}. \tag{10}$$

13

It is well known that, in general, the loglikelihood function (10) is non-concave and may have several local maxima. For unconstrained parametrizations $(\boldsymbol{\Theta}, \mathcal{T})$ and $(\boldsymbol{\Gamma}, \mathcal{G})$, it is even unbounded. To see this, one may, for example, choose $(\theta, \gamma) \in \mathcal{T} \times \mathcal{G}$ such that all entries of $\boldsymbol{\Gamma}(\gamma)$ are strictly positive and such that $x_{t_0} \beta_1(\theta) = y_{t_0}$ for a single fixed t_0. By letting $\sigma_1^2(\theta)$ go to zero while keeping all other regression parameters fixed, $p_{(\boldsymbol{\Theta}, \boldsymbol{\Gamma})}(\mathbf{y}, \mathbf{x}, \mathbf{h} \,|\, \theta, \gamma)$ approaches infinity for all $\mathbf{h}$ with $h_t = 1 \Leftrightarrow t = t_0$.

We consider two kinds of parameter constraints: (i) a lower bound on all error variances, and (ii) equality of all error variances. These constraints can be implemented using the parametrizations $(\boldsymbol{\Theta}^c, \mathcal{T}^c)$ and $(\boldsymbol{\Theta}^=, \mathcal{T}^=)$ given in Appendix B. In the following theorem, we show that either of these parametrizations ensures the existence of the maximum likelihood estimator.

Theorem 1 (Existence of the MLE) *Let $(\mathbf{y}, \mathbf{x})$ be a sample of $(\mathbf{Y}, \mathbf{X}) = (Y_t, X_t)_{t \in \{1, \ldots, m\}}$ and assume that the set $\{(y_t, x_t) \,|\, t \in \{1, \ldots, m\}\}$ is not contained in a union of ℓ hyperplanes of dimension p. Let $\mathcal{G}$ be a compact subset of a Euclidean space and let $\boldsymbol{\Gamma} : \mathcal{G} \to [0, 1]^{\ell \times \ell}$ be a continuous parametrization of the transition matrix Γ. Then, with $(\boldsymbol{\Theta}, \mathcal{T})$ being either of the parametrizations $(\boldsymbol{\Theta}^c, \mathcal{T}^c)$ or $(\boldsymbol{\Theta}^=, \mathcal{T}^=)$ (see Appendix B), the loglikelihood function $\ell_{(\boldsymbol{\Theta}, \boldsymbol{\Gamma})}$ attains its supremum on $\mathcal{T} \times \mathcal{G}$.*

The assumption involving hyperplanes excludes the possibility of a perfect fit. The conditions on $(\boldsymbol{\Gamma}, \mathcal{G})$ ensure that the space of possible transition matrices is a compact set. The continuity of all parametrizations together with the parameter constraints inherent in $(\boldsymbol{\Theta}^c, \mathcal{T}^c)$ and $(\boldsymbol{\Theta}^=, \mathcal{T}^=)$ make for a continuous and bounded likelihood function. We use two different methods for likelihood optimization: a numerical optimization routine[3] and an EM-type algorithm. These methods make use of the R packages `nlm` and `mixreg`, respectively, and will be referred to as "NLM" and "EM"; see Appendix D for details.

3.3. Fisher Confidence Regions

Using the asymptotic normality of maximum likelihood estimators, we can now construct (approximate) confidence regions for θ. Let therefore $\hat{\phi} = (\hat{\theta}, \hat{\gamma})$ be a global maximizer of the likelihood function and let $\mathcal{J}(\hat{\phi})$ be the observed Fisher information (e.g., Lehmann and Casella, 2006, Chapter 2) at $\hat{\phi}$. For $\alpha \in (0, 1)$, we define the region

$$C^\alpha(\hat{\theta}) := \left\{ \hat{\theta} + \mathcal{J}^{-1/2}(\hat{\theta})v \,:\, \|v\|_2^2 \leq q_{\chi^2(\dim(\theta))}(\alpha) \right\}, \tag{11}$$

where $\dim(\theta)$ is the length of the parameter vector θ, $q_{\chi^2(f)}(\alpha)$ is the α-quantile of a $\chi^2(f)$-distribution and $\mathcal{J}^{-1/2}(\hat{\theta})$ is the submatrix of $\mathcal{J}(\hat{\phi})^{-1/2}$ corresponding to $\hat{\theta}$. For these confidence regions to achieve the correct asymptotic coverage, we need to adjust for the label switching problem described in the following subsection.

3.4. Label Permutations

The distribution $\mathcal{SR}(\phi \,|\, \mathbf{X})$ is invariant under certain permutations of the coordinates of the parameter vector ϕ. For example, when $\ell = 2$, the hidden variable has two states. If

---

3. We are grateful to Roland Langrock who who shared parts of his code with us.

we exchange all parameters corresponding to the first state with those corresponding to the second state, the induced mixture distribution is unchanged. In general, the model $\{\mathcal{SR}(\phi \,|\, \mathbf{X}) : \phi \in \mathcal{P}\}$ is therefore not identifiable. More formally, let Π denote the set of all permutations of elements in $\{1, \ldots, \ell\}$. For every permutation $\pi \in \Pi$ with associated permutation matrix M_π, define the induced mappings $\pi_{\mathcal{T}} := \boldsymbol{\Theta}^{-1} \circ (\Theta \mapsto \Theta M_\pi^T) \circ \boldsymbol{\Theta}$, $\pi_{\mathcal{G}} := \boldsymbol{\Gamma}^{-1} \circ (\Gamma \mapsto M_\pi \Gamma M_\pi^T) \circ \boldsymbol{\Gamma}$ and $\pi_{\mathcal{P}} := (\pi_{\mathcal{T}}, \pi_{\mathcal{G}})$ on $\mathcal{T}$, $\mathcal{G}$ and $\mathcal{P}$, respectively. Then, for every $\phi \in \mathcal{P}$ and every $\pi \in \Pi$, the distributions $\mathcal{SR}(\phi \,|\, \mathbf{X})$ and $\mathcal{SR}(\pi_{\mathcal{P}}(\phi) \,|\, \mathbf{X})$ coincide (and thus give rise to the same likelihood). The likelihood function therefore attains its optimum in a set of different parameter vectors, all of which correspond to permutations of one another. Coverage properties of the confidence region (11) depend on which particular permutation of the MLE is output by the optimization routine (even though each of them parametrizes the exact same distribution). To overcome this ambiguity, we introduce the permutation-adjusted confidence regions

$$C_{\text{adjusted}}^{\alpha}(\hat{\theta}) := \bigcup_{\pi \in \Pi} C^{\alpha}(\pi_{\mathcal{T}}(\hat{\theta})). \tag{12}$$

In the following section, we make precise under which conditions these confidence regions achieve the correct asymptotic coverage.

3.5. Asymptotic Coverage of Adjusted Confidence Regions

Assume that the distribution of X_t is stationary across $e = \{1, \ldots, m\}$ and has a density f with respect to the Lebesgue measure on $\mathbb{R}^p$. Consider a fixed pair $(\boldsymbol{\Theta}, \mathcal{T})$ and $(\boldsymbol{\Gamma}, \mathcal{G})$ of parametrizations. Let $\phi^0 = (\theta^0, \gamma^0) \in \mathcal{P} := \mathcal{T} \times \mathcal{G}$ be the true parameters and let $\Theta^0 = \boldsymbol{\Theta}(\theta^0)$ and $\Gamma^0 = \boldsymbol{\Gamma}(\gamma^0)$ be the associated regression matrix and transition matrix, respectively.

Suppose now that the data within environment e accumulates. For every $m \in \mathbb{N}$, write $(\mathbf{Y}_m, \mathbf{X}_m) = (Y_t, X_t)_{t \in \{1, \ldots, m\}}$, let $\mathbb{P}_0^m := \mathcal{SR}(\theta^0, \gamma^0 \,|\, \mathbf{X}_m)$ and use $\mathbb{P}_0$ to denote the (infinite-dimensional) limiting distribution of $\mathbb{P}_0^m$. Similarly, $\mathbb{E}_0$ denotes the expectation with respect to $\mathbb{P}_0$. We require the following assumptions.

(A1) The maximum likelihood estimator exists.

(A2) The true parameter ϕ^0 is contained in the interior of $\mathcal{P}$.

(A3) The transition matrix Γ^0 is irreducible and aperiodic (e.g., Ching and Ng, 2006, Section 1).

(A4) For every $i \in \{1, \ldots, p+1\}$ and $j, k \in \{1, \ldots, \ell\}$, the maps $\theta \mapsto \boldsymbol{\Theta}_{ij}(\theta)$ and $\gamma \mapsto \boldsymbol{\Gamma}_{jk}(\gamma)$ have two continuous derivatives.

(A5) For every $m \in \mathbb{N}$, assume that the joint distribution of $(\mathbf{Y}_m, \mathbf{X}_m)$ has a density with respect to the Lebesgue measure that we denote by f_m. Then, the Fisher information matrix $\mathcal{I}_0$ defined as

$$\mathcal{I}_0 := \mathbb{E}_0[\eta \eta^T], \quad \text{where} \quad \eta = \lim_{m \to \infty} \frac{\partial}{\partial \phi} f_m(Y_m, X_m \,|\, \mathbf{Y}_{m-1}, \mathbf{X}_{m-1}, \phi) \Big|_{\phi = \phi^0},$$

is strictly positive definite.

(A6) All coordinates of X_1 have finite fourth moment.

(A7) $\mathbb{E}[|\log f(X_1)|] < \infty$.

Assumptions (A1) and (A4) are satisfied for the explicit parametrizations of the models IID and HMM given in Appendix B, see Theorem 1. The irreducibility of Γ^0 assumed in (A3) guarantees all latent states to be visited infinitely often, such that information on all parameters keeps accumulating. Assumption (A5) is needed to ensure that, in the limit, the loglikelihood function has, on average, negative curvature and hence a local maximum at ϕ^0. Finally, (A6) and (A7) are mild regularity conditions on the (otherwise unspecified) distribution of X_t.

Essentially, the asymptotic validity of the adjusted confidence regions (12) rests on two results: (1) consistency of the MLE and (2) asymptotic normality of the MLE. For every $\phi \in \mathcal{P}$, let $[\phi] := \{\pi_{\mathcal{P}}(\phi) : \pi \in \Pi\} \subseteq \mathcal{P}$ denote the equivalence class of ϕ, i.e., the set of parameters in $\mathcal{P}$ that are equal to ϕ up to a permutation $\pi_{\mathcal{P}}$ as defined in Section 3.4. Consistency in the quotient topology ("$[\hat{\phi}_m] \to [\phi^0]$") then simply means that any open subset of $\mathcal{P}$ that contains the equivalence class of ϕ^0, must, for large enough m, also contain the equivalence class $\hat{\phi}_m$. With this notation, we can now state an asymptotic coverage result for confidence regions (12). The main work is contained in Theorems 2 and 3. Their proofs make use of results given by Leroux (1992) and Bickel et al. (1998), which discuss consistency and asymptotic normality, respectively, of the MLE in hidden Markov models with finite state space.

Theorem 2 (Consistency of the MLE) *Assume that (A1), (A3), (A4) and (A7) hold true. Then, $\mathbb{P}_0$-almost surely, $[\hat{\phi}_m] \to [\phi^0]$ as $m \to \infty$.*

Theorem 2 says that $(\hat{\phi}_m)_{m \in \mathbb{N}}$ alternates between one or more subsequences, each of which is convergent to a permutation of ϕ^0. The following theorem proves a central limit theorem for these subsequences.

Theorem 3 (Asymptotic normality of the MLE) *Assume that the maximum likelihood estimator is consistent. Then, under (A1)–(A6), it holds that $\mathcal{J}(\hat{\phi}_m)^{1/2}(\hat{\phi}_m - \phi^0) \xrightarrow{d} \mathcal{N}(0, I)$ under $\mathbb{P}_0$.*

Together, Theorems 2 and 3 imply the following asymptotic coverage guarantee.

Corollary 1 (Asymptotic coverage of adjusted confidence regions) *Under Assumptions (A1)–(A7), the adjusted confidence regions (12) achieve the correct asymptotic coverage. That is, for any $\alpha \in (0, 1)$,*

$$\liminf_{m \to \infty} \mathbb{P}_0^m(\theta^0 \in C^\alpha_{\text{adjusted}}(\hat{\theta}_m)) \geq 1 - \alpha. \tag{13}$$

As another corollary, the asymptotic type I error control of the tests defined by (9) follows by applying Corollary 1 to each environment separately.

4. ICPH: Algorithm and False Discovery Control

We can now summarize the above sections into our overall method. In Section 4.1 we provide a pseudo code for this procedure, and Section 4.2 presents our main theoretical result—an asymptotic version of Proposition 3, which states that our procedure is consistent.

4.1. Algorithm

Given data $(\mathbf{Y}, \mathbf{X})$ and a collection $\mathcal{E}$ of environments, we run through all $S \subseteq \{1, \ldots, d\}$, test the hypothesis $H_{0,S}$ with the test defined by (9) using the adjusted confidence regions (12), and output the intersection of all accepted sets. Below, this procedure is formalized in a pseudo code.

Algorithm 1: ICPH ("Invariant Causal Prediction in the presence of Hidden variables")

---

1 Input: response $\mathbf{Y} \in \mathbb{R}^n$, covariates $\mathbf{X} \in \mathbb{R}^{n \times d}$, environment indicator
$\quad$ $\mathbf{E} \in \{1, \ldots, |\mathcal{E}|\}^n$ (i.e., $\mathbf{E}_t = k \Leftrightarrow t \in e_k$);
2 Options: `model` $\in \{$"IID", "HMM"$\}$, `method` $\in \{$"EM", "NLM"$\}$,
$\quad$ `variance.constraint` $\in \{$"lower bound", "equality"$\}$, `number.of.states` $\in \mathbb{N}_{\geq 2}$,
$\quad$ `intercept` $\in \{$TRUE, FALSE$\}$,
$\quad$ `test.parameters` $\subseteq \{$"intercept", "beta", "sigma"$\}$, `alpha` $\in (0,1)$;
3 for $S \subseteq \{1, \ldots, d\}$ **do**
4 $\quad$ **for** $e \in \mathcal{E}$ **do**
5 $\quad\quad$ Fit an $\mathcal{SR}$ model to $(\mathbf{Y}_e, \mathbf{X}_e^S)$, see Section 3.2;
6 $\quad\quad$ Construct the permutation-adjusted confidence region (12);
7 $\quad$ **end**
8 $\quad$ Compute a p-value p_S for $H_{0,S}$ using the test defined by (9);
9 end
10 Output: the empirical estimator $\hat{S} = \bigcap_{S : p_S > \alpha} S$;

Most of the options in Algorithm 1 are self-explanatory. The option `test.parameters` allows the user to specify the "degree of h-invariance" that is required of the sets $S \subseteq \{1, \ldots, d\}$. If, for example, `test.parameters` $= \{$"beta", "sigma"$\}$, a set S will be regarded h-invariant if the mixture components of $P_{Y_t | X_t^S}$ are "invariant in β and σ^2", i.e., time-homogeneous up to changes in the intercept between different environments. Code is available online (see Section 1.4). To make Algorithm 1 scalable to a large number of predictors, it can be combined with a variable screening step, e.g., using Lasso (Tibshirani, 1994); see Section 4.2 for more details.

4.2. Asymptotic False Discovery Control of ICPH

The cornerstone for the false discovery control of ICPH is given in Corollary 1. It proves that if Assumptions (A1)–(A7) are satisfied for the true set S^*, then the test φ_{S^*} achieves the correct asymptotic level, which in turn guarantees an asymptotic version of Proposition 3. We will now summarize this line of reasoning into out main theoretical result.

Assume that we are given data $((\mathbf{Y}_n, \mathbf{X}_n))_{n \in \mathbb{N}} = \left((Y_{n,t}, X_{n,t})_{t \in \{1, \ldots, n\}} \right)_{n \in \mathbb{N}}$ from a triangular array, where, for every n, $(\mathbf{Y}_n, \mathbf{X}_n) \in \mathbb{R}^{n \times (1+d)}$. Consider a fixed number of K environments and let $(\mathcal{E}_n)_{n \in \mathbb{N}}$ be a sequence of collections $\mathcal{E}_n = \{e_{n,1}, \ldots, e_{n,K}\}$, such that, for all n, $e_{n,1}, \ldots, e_{n,K}$ are disjoint with $\cup_k e_{n,k} = \{1, \ldots, n\}$ and such that, for all k, $|e_{n,k}| \to \infty$ as $n \to \infty$. For all n and k, write $(\mathbf{Y}_{n,k}, \mathbf{X}_{n,k}) = (Y_t, X_t)_{t \in e_{n,k}}$. Consider a transition parametrization $(\mathbf{\Gamma}, \mathcal{G})$ and a family of regression parametrizations $\{(\mathbf{\Theta}^S, \mathcal{T}^S)\}_{S \subseteq \{1, \ldots, d\}}$, i.e., for every $S \subseteq \{1, \ldots, d\}$, $\mathbf{\Theta}^S$ maps $\mathcal{T}^S$ into the space of matrices of dimension $(|S| + 1) \times \ell$

with columns in $\mathbb{R}^{|S|} \times \mathbb{R}_{>0}$. For every n and every $S \subseteq \{1, \ldots, d\}$, let $H_{0,S}^n$ denote the hypothesis (6) for the data $(\mathbf{Y}_n, \mathbf{X}_n^S)$ and let φ_S^n be the corresponding test defined by (9) with the confidence regions (12). Finally, define for every n the estimator

$$\hat{S}_n := \bigcap_{S: \varphi_S^n \text{ accepts } H_{0,S}^n} S. \tag{14}$$

We then have the following result.

Theorem 4 (Asymptotic false discovery control) *Assume that Assumption 1 is satisfied. That is, there exists a set $S^* \subseteq \{1, \ldots, d\}$ which, for every n, is h-invariant with respect to $(\mathbf{Y}_n, \mathbf{X}_n)$. Assume furthermore that, for every k, (A1)–(A7) hold true for the data $(\mathbf{Y}_{n,k}, \mathbf{X}_{n,k}^{S^*})$ with parametrizations $(\mathbf{\Theta}^{S^*}, \mathcal{T}^{S^*})$ and $(\mathbf{\Gamma}, \mathcal{G})$. Then, the estimator $\hat{S}_n$ enjoys the following coverage property*

$$\liminf_{n \to \infty} \mathbb{P}_0^n(\hat{S}_n \subseteq S^*) \geq 1 - \alpha, \tag{15}$$

where $\mathbb{P}_0^n$ is the law of $(\mathbf{Y}_n, \mathbf{X}_n)$.

If the number of predictor variables is large, our algorithm can be combined with an upfront variable screening step. Given a family $(\hat{S}_{\text{screening}}^n)_{n \in \mathbb{N}}$ of screening estimators, we can for every $n \in \mathbb{N}$ construct an estimator $\bar{S}_n$ of S^* analogously to (14), but where the intersection is taken only over those S additionally satisfying that $S \subseteq \hat{S}_{\text{screening}}^n$. Given that $\liminf_{n \to \infty} \mathbb{P}_0^n(S^* \subseteq \hat{S}_{\text{screening}}^n) \geq 1 - \alpha$, it then follows from

$$\mathbb{P}_0^n(\bar{S}_n \not\subseteq S^*) = \mathbb{P}_0^n([\bar{S}_n \not\subseteq S^*] \cap [S^* \subseteq \hat{S}_{\text{screening}}^n]) + \mathbb{P}_0^n([\bar{S}_n \not\subseteq S^*] \cap [S^* \not\subseteq \hat{S}_{\text{screening}}^n])$$
$$\leq \mathbb{P}_0^n(\varphi_{S^*}^n \text{ rejects } H_{0,S^*}^n) + \mathbb{P}_0^n(S^* \not\subseteq \hat{S}_{\text{screening}}^n),$$

that the estimator $(\bar{S}_n)_{n \in \mathbb{N}}$ satisfies the asymptotic false discovery control (15) at level $1 - 2\alpha$. In high-dimensional models, assumptions that allow for the screening property have been studied (see, e.g., Bühlmann and van de Geer, 2011).

5. Experiments

In this section, we apply our method to simulated data (Section 5.1) and to a real world data set on photosynthetic activity and sun-induced fluorescence (Section 5.2). We only report results using the NLM optimizer. In all experiments, the results for EM were almost identical to those for NLM, except that the computation time for EM was larger (by approximately a factor of 6). For an experiment that uses the EM-method, see Appendix D.2.

5.1. Simulated Data

We start by testing the sample properties of the adjusted confidence regions, disregarding the problem of causal discovery, see Section 5.1.1. In Section 5.1.2, we present the multivariate data generating process that we will use in the subsequent analyses. In Section 5.1.3, we see that, even for sample sizes that are too small for the confidence regions to achieve the correct coverage, our overall method (ICPH) is able to keep the type I error control.

Section 5.1.4 contains a power analysis. In Section 5.1.5, we test the robustness of ICPH against a range of different model violations, and include a comparison with two alternative causal discovery methods. The performance of ICPH for non-binary latent variables, for large numbers of predictor variables, or under violations of the h-invariance assumption, can be found in Appendix E.

5.1.1. Empirical coverage properties of adjusted confidence regions

The finite sample coverage properties of the confidence regions (12) depend on the true distribution over $(\mathbf{Y}, \mathbf{X})$ (i.e., on the parameters of the $\mathcal{SR}$ model as well on the marginal distribution of $\mathbf{X}$) and on the sample size. We here illustrate this sensitivity in the i.i.d. setting. Consider a joint distribution $\mathbb{P}$ over $(Y, X, H) \in \mathbb{R}^{1+p} \times \{1, \ldots, \ell\}$ which induces an $\mathcal{SR}$ model over (Y, X). For every $j \in \{1, \ldots, \ell\}$ let $p_j(y, x) = \mathbb{P}(H = j \mid Y = y, X = x)$ denote the posterior probability of state j based on the data (y, x). We then use the geometric mean of expected posterior probabilities

$$\text{GMEP} := \left(\prod_{j=1}^{\ell} \mathbb{E}[p_j(Y, X) \mid H = j] \right)^{1/\ell} \in [0, 1] \tag{16}$$

as a measure of difficulty of fitting the $\mathcal{SR}$ model induced by $\mathbb{P}$.[4] We expect smaller values of GMEP to correspond to more difficult estimation problems, which negatively affect the convergence rate of (13) and result in low finite sample coverage. If the between-states differences in the regression parameters of X are small, for example, we expect the unobserved states to be difficult to infer from the observed data (i.e., for every j, the expected posterior probabilities $\mathbb{E}[p_i(Y, X) \mid H = j]$ are close to uniform in i), resulting in small GMEP.

We now perform the following simulation study. For different model parameters and sample sizes, we generate i.i.d. data sets from the SCM

$$H := N_\lambda^H, \quad X := \mu_X + \sigma_X N^X, \quad Y := \mu_Y + \beta_1 X \cdot \mathbb{1}_{\{H=1\}} + \beta_2 X \cdot \mathbb{1}_{\{H=2\}} + \sigma_Y N^Y, \tag{17}$$

where all noise variables are jointly independent with marginal distributions $N_\lambda^H \sim \text{Ber}(\lambda)$, $N^X, N^Y \sim \mathcal{N}(0, 1)$. We construct adjusted confidence regions (12) for the vector of regression parameters $\theta^0 = (\mu_Y, \beta_1, \mu_Y, \beta_2, \sigma_Y^2)$ using the likelihood function (10) with parametrizations $(\mathbf{\Theta}^=, \mathcal{T}^=)$ and $(\mathbf{\Gamma}^{\text{IID}}, \mathcal{G}^{\text{IID}})$ (see Appendix B). We first sample 50 different sets of parameters independently as $\mu_X, \mu_Y, \beta_1, \beta_2 \sim \text{Uniform}(-1, 1)$, $\sigma_X \sim \text{Uniform}(0.1, 1)$, $\sigma_Y \sim \text{Uniform}(0.1, 0.5)$ and $\lambda \sim \text{Uniform}(0.3, 0.7)$. For each setting, we compute empirical coverage degrees based on 1000 independent data sets, each consisting of 100 independent replications from (17), and compare them to the GMEP of the underlying models, see Figure 4 (left). For the same simulations, we also compare the p-values

$$p := \max\{\alpha \in [0, 1] : \theta^0 \notin C_{\text{adjusted}}^\alpha(\hat{\theta})\} \tag{18}$$

---

4. If each of the distributions $\mathbb{P}_{(Y,X)|H=j}$, $j \in \{1, \ldots, \ell\}$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^{1+p}$, each factor in (16) is given as an integral over $\mathbb{R}^{1+p}$. In practice, we approximate these integrals by numerical integration.
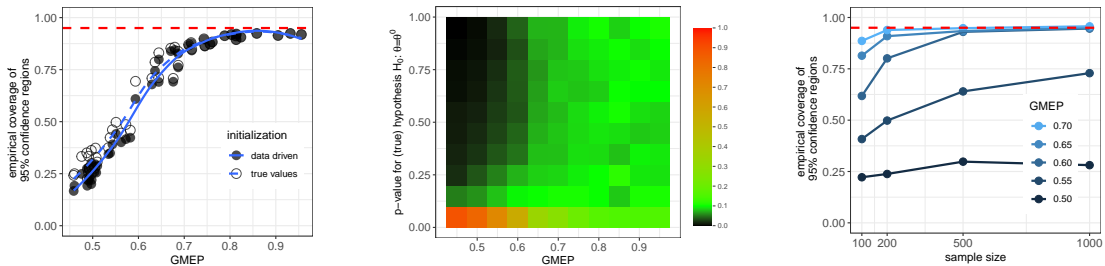
Figure 4: Empirical coverage properties of the adjusted confidence regions (12) using data simulated from the model (17). The left panel shows empirical coverage of 95%-confidence regions for different model parameters (see Equation 16 for a definition of GMEP), and a fixed sample size of 100. We see that the coverage properties strongly depend on GMEP, and that the poor performance for low GMEP is not an optimization problem (the likelihood scores obtained from starting the algorithm in the true values exceed those obtained from data driven initialization in less than 0.2% of simulations). In the middle panel, we use the same simulations, but only consider data-driven initialization. Each column corresponds to a histogram of p-values (18). For increasing GMEP, the p-value distribution approximates the desired uniform distribution. For 5 different parameter settings, we further increase the sample size (right). As suggested by Corollary 1, the empirical coverage gradually improves, although very low GMEP demand large amounts of data to obtain satisfactory coverage.

for the (true) hypotheses $H_0 : \theta = \theta^0$ to a uniform distribution (Figure 4 middle). For 5 models of different degrees of difficulty (GMEP $\approx 0.50, 0.55, 0.60, 0.65, 0.70$), we then compute empirical coverage degrees for increasing sample size (Figure 4 right).

For difficult estimation problems (i.e., low GMEP), the finite sample variance of the MLE is inflated, resulting in low empirical coverage and too small p-values (Figure 4 left and middle). Although there is no proof that NLM finds the global optimum, it is assuring that there is little difference when we start the algorithm at the (usually unknown) true values (Figure 4 left, hollow circles). Indeed, the thus obtained likelihood scores exceed those obtained from data driven initialization in less than 0.2% of simulations. As seen in Figure 4 (right), coverage properties improve with increasing sample size, although in models with low GMEP, we require large amounts of data in order to obtain satisfactory performance. We will see in Section 5.1.3 that even cases where we cannot expect the confidence regions to obtain valid coverage, our overall causal discovery method maintains type I error control.
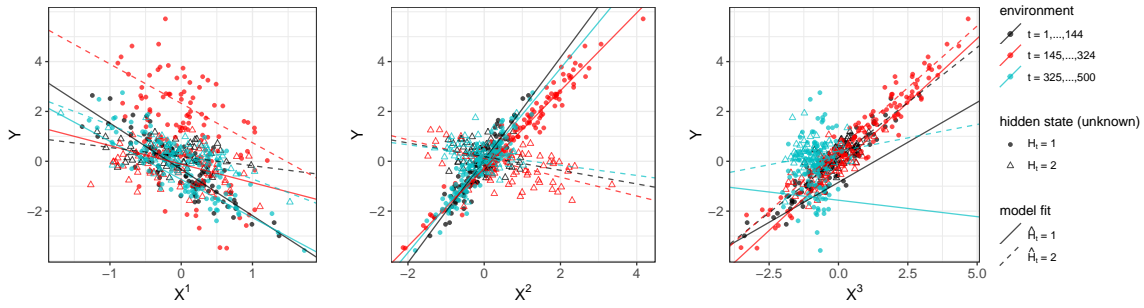
Figure 5: Data generated from the SCM described in Section 5.1.2 for each of the three environments (black, red, blue). Here, the only h-invariant set is $S^* = \{1, 2\}$ and we would therefore like our method to correctly identify the violations of the h-invariance of the sets $\{1\}, \{2\}$ and $\{3\}$. These violations are indicated by the different model fits (colored lines), which for none of the three variables are stable across all environments. For numerical results on such data sets, see Sections 5.1.3 and 5.1.4. The issue of label permutations can be seen from the occasional mismatch between the true latent states ($\bullet$: $H_t = 1$, $\triangle$: $H_t = 2$) and the estimated labels (—: $\hat{H}_t = 1$, - -: $\hat{H}_t = 2$).

5.1.2. DATA GENERATING PROCESS

We now specify the data generating process used in the following sections. We consider an SCM over the system of variables (Y, X^1, X^2, X^3, H) given by the structural assignments

$$X^1 := N^1, \quad H := N^H, \quad X^2 := \beta^{21} X^1 + N^2$$

$$Y := \sum_{j=1}^{\ell} (\mu_j^Y + \beta_{1j}^Y X^1 + \beta_{2j}^Y X^2 + \sigma_{Yj} N^Y) \mathbb{1}_{\{H=j\}}$$

$$X^3 := \beta^{3Y} Y + N^3,$$

where $N^H \sim \text{Multinomial}(1, \lambda)$, $N^Y \sim \mathcal{N}(0, 1)$ and $N^j \sim \mathcal{N}(\mu^j, \sigma_j^2)$. In Sections 5.1.3–5.1.5, the latent variable H is assumed to be binary, while Appendix E.1 treats the more general case where $\ell \geq 2$. The different environments are constructed as follows. We first draw random change points $1 < t_1 < t_2 < n$ and then generate data as described below.

- $e_1 = \{1, \ldots, t_1\}$: Here, we sample from the observational distribution.
- $e_2 = \{t_1 + 1, \ldots, t_2\}$: Here, we set $X^2 := \beta^{21} X^1 + \tilde{N}^2$, where $\tilde{N}^2$ is a Gaussian random variable with mean sampled uniformly between 1 and 1.5 and variance sampled uniformly between 1 and 1.5. Also, the mixing proportions λ are resampled.
- $e_3 = \{t_2 + 1, \ldots, n\}$: We again sample data from the above SCM, but this time we intervene on X^3. The structural assignment is replaced by $X^3 := \tilde{N}^3$, where $\tilde{N}_3$ is a Gaussian random variable with mean sampled uniformly between -1 and -0.5 and the same variance as the noise N^3 from the observational setting. The mixing proportions λ are again resampled.

21

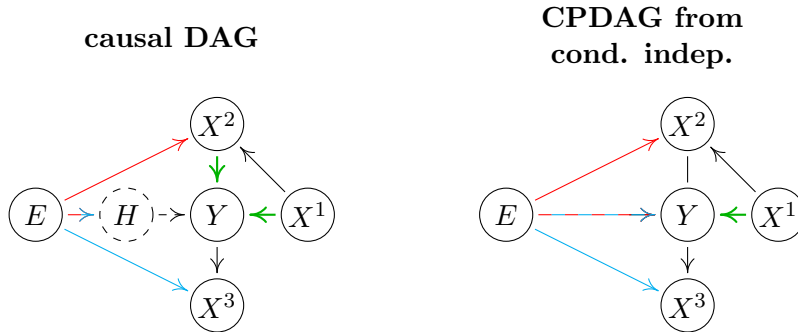**causal DAG**

CPDAG from cond. indep.

Figure 6: Left: the causal graph induced by the SCM in Section 5.1.2. The node E represents the different environments (E points into variables that have been intervened on, the color corresponds to the environments shown in Figure 5). Right: the CPDAG representing the Markov equivalence class of the graph where H is marginalized out. Since the edge $X^2 - Y$ is not oriented, the full set of causal parents $S^* = \{1, 2\}$ cannot be identified only from conditional independence statements. Our method exploits the simple form of the influence of H on Y. Note that in the case of an additional edge $E \to X^1$, none of the edges among the variables (Y, X^1, X^2, X^3) would be oriented in the CPDAG.

A sample data set can be seen in Figure 5, where points have been colored according to the above environments (black, red and blue for e_1, e_2 and e_3, respectively). The only h-invariant set is the set $S^* = \{1, 2\}$ of observable parents of Y. In the population case, our method therefore correctly infers $\tilde{S} = \{1, 2\}$, see Equation (7). The causal graph induced by the above data generating system can be seen in Figure 6 (left). Here, the environment is drawn as a random variable.[5] We also display the CPDAG representing the Markov equivalence class of the induced graph over the observed variables (right), showing that the full set of causal parents $S^* = \{1, 2\}$ cannot be identified only from conditional independence statements.

5.1.3. LEVEL ANALYSIS

Given that the theoretical coverage guarantees are only asymptotic, we cannot expect the tests (9) to satisfy type I error control for small sample sizes—especially if GMEP is low, see also Section 5.1.1. The following empirical experiments suggest, however, that even if the test level of the true hypothesis H_{0,S^*} is violated, ICPH may still keep the overall false discovery control. We use data sets $(Y_t, X_t^1, X_t^2, X_t^3)_{t \in \{1, \dots, n\}}$ generated as described in Section 5.1.2, and analyse the performance of ICPH for different sample sizes and different GMEP. Since the latter is difficult to control directly, we vary the between-states difference in regression coefficients for X^1 and X^2 in the structural assignment for Y, and report

---

5. To view the data set as i.i.d. realizations from such a model one formally adds a random permutation of the data set, which breaks the dependence of the realizations of the environment variable (this has no effect on the causal discovery algorithm, of course). Constantinou and Dawid (2017) discuss a non-stochastic treatment of such nodes.
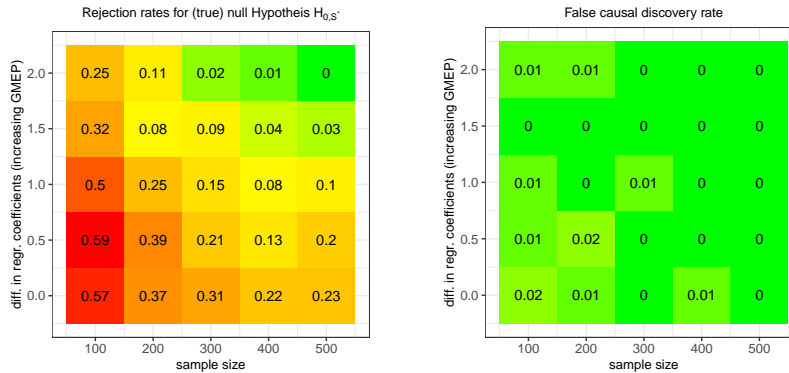
Figure 7: Estimates $\hat{\mathbb{P}}(\varphi_{S^*} \text{ rejects } H_{0,S^*})$ (left) and $\hat{\mathbb{P}}(\hat{S} \not\subseteq S^*)$ (right) of the type I error rates of the test φ_{S^*} and the overall method ICPH, respectively, based on the experiment described in Section 5.1.3 and 100 repetitions. The desired level is $\alpha = 0.05$. We have used NLM with parametrizations $\mathbf{\Theta}^=$ and $\mathbf{\Gamma}^{\text{IID}}$ (see Appendix B). The average GMEP values are $0.51, 0.56, 0.64, 0.66, 0.78$ (ordered in accordance to the vertical axis). For small sample sizes, and in particular for low GMEP, the type I error control of the test φ_{S^*} is violated. Even in these cases, however, the false causal discovery control of ICPH is satisfied.

the average GMEP for each setting. For every $n \in \{100, 200, 300, 400, 500\}$ and every $\Delta\beta \in \{0, 0.5, 1, 1.5, 2\}$, we simulate 100 independent data sets by drawing model parameters $\mu \sim^{iid} \text{Uniform}(-0.2, 0.2)$, $\sigma^2 \sim^{iid} \text{Uniform}(0.1, 0.3)$ (with the restriction that $\sigma_{Y1}^2 = \sigma_{Y2}^2$), $\beta \sim^{iid} \text{Uniform}([-1.5, -0.5] \cup [0.5, 1.5])$ and $\lambda \sim \text{Uniform}(0.3, 0.7)$. For $j \in \{1, 2\}$ we then assign $\beta_{j,2}^Y := \beta_{j,1}^Y + \text{sign}(\beta_{j,1}^Y)\Delta\beta$. The results are summarized in Figure 7. We see that even in settings for which the true hypothesis H_{0,S^*} is rejected for about every other simulation, ICPH stays conservative.

5.1.4. POWER ANALYSIS

Since the only h-invariant set is the set $S^* = \{1, 2\}$ of causal parents of Y, the population version of our method correctly infers $\tilde{S} = \{1, 2\}$, see Equation (7). For finite samples, identifiability of S^* is determined by the power of the tests for the hypotheses $H_{0,S}$. For a fixed value of $\Delta\beta = 1.5$ (average GMEP of 0.66) and increasing sample size, we generate i.i.d. data sets as described in Section 5.1.3 and analyze the performance of ICPH for two different variance constraints $\sigma_{Y1}^2 = \sigma_{Y2}^2$ and $\sigma_{Y1}^2, \sigma_{Y2}^2 \geq 10^{-4}$. The results in Figure 8 suggest that the former constraint results in higher performance, and it will therefore be our default setting for the rest of this section. As the sample size increases, ICPH tends to identify the set S^* (larges shares of green in bar plots).

For the same data that generated Figure 8, we compute rejection rates for non-causality (i.e., empirical proportions of not being contained in $\hat{S}$) for each of the predictors X^1, X^2 and X^3. Here, we also add a comparison to other methods. We are not aware of any other method that is suitable for inferring S^*, but we nevertheless add two approaches as baseline.
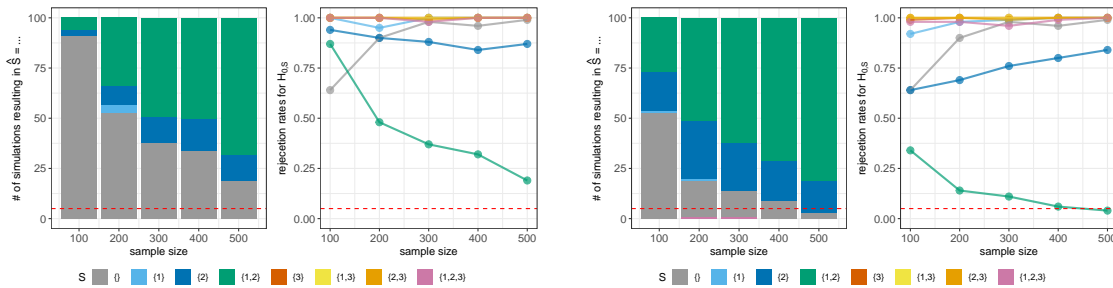
Figure 8: Output of ICPH (bar plots) and rejection rates for individual hypotheses (curve plots) for the experiment in Section 5.1.4 with parameter constraint $\sigma_{Y1}^2, \sigma_{Y2}^2 \geq 10^{-4}$ (left) and $\sigma_{Y1}^2 = \sigma_{Y2}^2$ (right). The larger the proportion of blue and green colors in the bar plots, the more power our method has. Simulations are ordered such that, within each bar, the bottom colors (yellow, light orange, dark orange, purple) correspond to false positives, i.e., cases where $\hat{S} \nsubseteq S^*$. Even though the level of the test for H_{0,S^*} is violated in the finite sample case, ICPH controls the empirical type I error rate at $\alpha = 0.05$ (indicated by a dashed horizontal line). Enforcing equality on error variances is beneficial, especially for small data sets. For both settings, the identification of S^* improves with increasing sample size.

- "k-means ICP": Pool data points from all environments and infer estimates $\hat{H}$ of the hidden states using 2-means clustering. Run the ordinary ICP algorithm (Peters et al., 2016) on each of the data sets $\{(Y_t, X_t) : \hat{H}_t = j\}$, $j \in \{1, 2\}$, testing all hypotheses at level $\alpha/2$, and obtain $\hat{S}_1$ and $\hat{S}_2$. Output the final estimate $\hat{S} = S_1 \cup S_2$.

- "JCI-PC": We use a modified version of the PC algorithm (Spirtes et al., 2000), which exploits our background knowledge of E being a source node: in between skeleton search and edge orientation, we orient all edges connecting E to another node. The resulting algorithm may be viewed as a variant of the of JCI algorithm (Magliacane et al., 2016). We apply it to the full system of observed variables (E, Y, X^1, X^2, X^3), and output the set of variables (among $\{X^1, X^2, X^3\}$) which have a directed edge to Y in the resulting PDAG.[6]

In the JCI-PC algorithm, we use conditional independence tests based on partial correlations. Since we apply it to a system of mixed variables (i.e., continuous as well as discrete), the assumptions underlying some of the involved tests will necessarily be violated. We are not aware of any family of tests which is more suitable. However, even in the population case, we cannot expect constraint-based methods such as JCI-PC to infer the set full S^*, see Figure 6. ICPH solves a specific problem and is the only method which exploits the simple influence of H on Y. The results in Figure 9 (black curves) confirm our previous findings:

---

6. Note that H can be marginalized out, so it is not necessary to use FCI. Furthermore, since we do not assume the intervention targets to be known, search algorithms for interventional data such as the GIES algorithm (Hauser and Bühlmann, 2012) are not applicable.
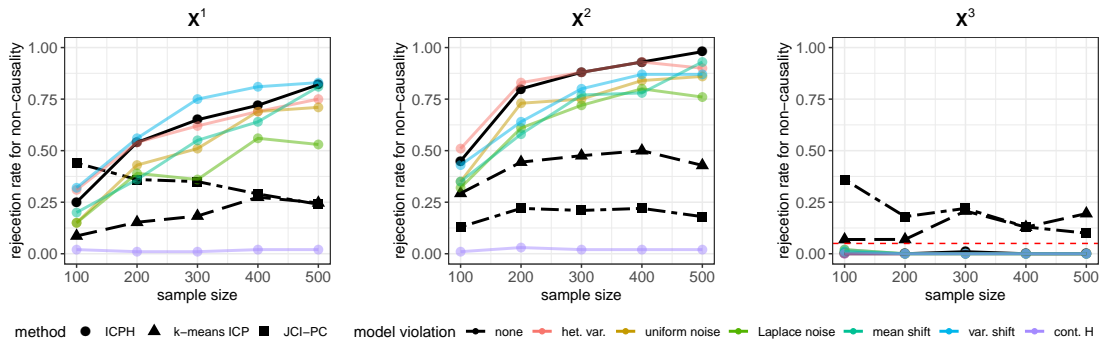
Figure 9: Rejection rates for non-causality. This figure contains two comparisons (one among all black curves, and another among all curves with round points). For data generated from the SCM in Section 5.1.2 (black), we compare the performance of ICPH (●) against the two alternative methods k-means ICP (▲) and JCI-PC (■) described in Section 5.1.4. For increasing sample size, ICPH outperforms both methods in terms level and power. As a robustness analysis, we further we apply ICPH to simulated data sets from the modified SCMs described in Section 5.1.5 (colored). Each of the modified SCMs yields a misspecification of the model for $P_{Y|X^{S^*}}$ that is assumed by our method. Most of these model misspecifications do not qualitatively affect the results: for increasing sample size, both causal parents X^1 and X^2 tend to be identified. For a continuous hidden variable, none of the variables is identified as causal (which is not incorrect, but uninformative). In all scenarios, ICPH maintains empirical type I error control.

causal discovery improves with increasing sample size, and our method stays conservative. ICPH outperforms both other methods in terms of level and power.

5.1.5. Robustness analysis

Our results are based on various assumptions, and we now investigate the robustness of ICPH against different kinds of model violations. We use data generated from the following modified versions of the SCM in Section 5.1.2. Unless mentioned otherwise, parameters are sampled as described in Section 5.1.3.

- Heterogeneous variances: The error variances σ_{Y1}^2 and σ_{Y2}^2 are sampled independently.
- Non-Gaussian noise: We generate errors N^Y from (i) a uniform distribution and (ii) a Laplace distribution.
- A direct effect $H \to X^1$: We allow for an influence of H on X^1 through binary shifts in (i) the mean value and (ii) the error variance. Parameters are sampled independently as $\mu_1^1, \mu_2^1 \sim \text{Uniform}(-1, 1)$ and $\sigma_{11}^2, \sigma_{12}^2 \sim \text{Uniform}(0.1, 1)$.
- A continuous hidden variable: We substitute the structural assignment for Y by $Y := (\mu^Y + \beta_1^Y X^1 + \beta_2^Y X^2)H + \sigma_Y N^Y$, where $H \sim \mathcal{N}(0, 1)$. The distribution of H does not change across environments.

25

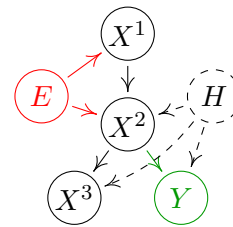| variable	description
Y	sun-induced fluorescence (SIF)
X^1	incident shortwave radiation (SW)
X^2	absorbed photosynthetically active radiation (APAR$_{chl}$)
X^3	gross primary productivity (GPP)
H	vegetation type

Figure 10: Variable descriptions (left) and causal graph constructed from background knowledge (right). In our analysis, we use the temporal ordering of data to construct the environment variable E. Due to seasonal cycles of aggradation and degradation of chlorophyll, APAR$_{chl}$ is not a constant fraction of SW (which itself is time-heterogeneous). The environment therefore "acts" on the variables X^1 and X^2. Furthermore, different vegetation types differ not only in their chlorophyll composition (and thus in APAR$_{chl}$), but also in their respective efficiencies of converting APAR$_{chl}$ into GPP and SIF—hence the arrows from H to X^2, X^3 and Y.

We now repeat the power analysis from Section 5.1.4 for data sets generated in the above way (Figure 9, colored curves). Most model violations do not qualitatively affect the results. Only the assumption on the state space of H is crucial for the power (not the level) of our method; for a continuous hidden variable, we mostly output the empty set.

5.2. Sun-Induced Fluorescence and Land Cover Classification

We now consider a real world data set for which we can compare our method's output against a plausible causal model constructed from background knowledge. The data set is related to the study of global carbon cycles, which are determined by the movement of carbon between land, atmosphere and ocean. Carbon is emitted, e.g., during fossil fuel combustion, land use change or cellular respiration, and assimilated back into the Earth's surface by processes of carbon fixation. A major component hereof is photosynthesis, where inorganic carbon is converted into organic compounds by terrestrial ecosystems. Direct measurements of carbon fluxes can be obtained from fluxtowers (`http://fluxnet.fluxdata.org`), but are only available at single locations. Constructing reliable global models for predicting photosynthesis using satellite data is an active line of research. While most of the commonly used models (e.g., Jung et al., 2009; Running and Zhao, 2015) use sunlight as the predominant driver, recent work (e.g., Guanter et al., 2012; Zhang et al., 2016) explores the predictive potential of sun-induced fluorescence (SIF), a (remotely sensible) electromagnetic radiation that is emitted by plants during the photosynthetic process.

Here, we take SIF as the target variable. As predictors, we include the incident shortwave radiation (SW), the photosynthetically active radiation absorbed by the plants' chlorophyll cells (APAR$_{chl}$), and the gross primary productivity (GPP), the latter of which is a measure of photosynthesis. Since GPP cannot be directly measured, we use spatially upscaled measurements from a network of fluxtowers (Jung et al., 2009). Background knowledge suggests that out of these three variables, only APAR$_{chl}$ is a direct causal parent of the target SIF.

Zhang et al. (2016) suggest evidence for a linear relationship between SIF and APAR_{chl}, and show that this relationship strongly depends on the type of vegetation. Estimates of the vegetation type can be obtained from the IGBP global land cover data base (Loveland et al., 2000). We use the IGBP classification to select data coming from two different vegetation types only. In the resulting data set, we thus expect the causal influence of SIF on APAR_{chl} to be confounded by a binary variable. When applying our method to these data, we remove information on vegetation type, so that this binary variable becomes latent. The data and the ground truth we consider is shown in Figure 10.

In Section 5.2.1, we use our causal discovery method to identify the causal predictor of SIF. In Section 5.2.2, we explore the possibility to reconstruct the vegetation type from the observed data (Y, X^1, X^2, X^3) when assuming that we have inferred the correct causal model. We believe that such estimates may be used to complement conventional vegetation type classifications.

5.2.1. Causal discovery

We denote the observed variables by (Y, X^1, X^2, X^3) as described in Figure 10 (left). The data are observed along a spatio-temporal grid with a temporal resolution of 1 month (Jan 2010 – Dec 2010), and a spatial resolution of $0.5° \times 0.5°$ covering the North American continent. The setup is directly taken from Zhang et al. (2016), and we refer to their work for a precise description of the data preprocessing for the variables (Y, X^2, X^3). The data for X^1 is publicly available at `https://search.earthdata.nasa.gov`. We select pixels classified as either *Cropland (CRO)* or *Evergreen Needleleaf Forest (ENF)*. These vegetation types are expected to differ in their respective relationships $X^2 \to Y$ (Zhang et al., 2016). As environments we use the periods Feb – Jul and Aug – Jan.[7]

The goal of the statistical analysis is to identify the set $S^* = \{2\}$ of causal parents of Y among the vector (X^1, X^2, X^3). Since the variables X^1 and X^2 are closely related, we regard distinguishing between their respective causal relevance for Y as a difficult problem. We analyze the data for different sample sizes. To do so, we gradually lower the spatial resolution in the following way. For every $c \in \{1, \ldots, 16\}$, we construct a new data set by increasing the pixel size of the original data set by a factor of c^2, and then averaging observations within each pixel. Grid cells that do not purely contain observations from either of the two vegetation types are discarded. We then apply our causal discovery method to each of the generated data sets, allowing for a binary hidden variable. The results are illustrated in Figure 11.[8] Indeed, for several sample sizes ($n \leq 390$), the true hypothesis H_{0,S^*} is accepted, and our method mostly correctly infers $\hat{S} = \{2\}$ (left plot). In all experiments, the variable X^2 is attributed the highest significance as a causal parent of Y (right plot). Also, we consistently do not reject the only non-ancestrial variable X^3, and the causal ordering implied by the right hand plot is in line with the assumed causal

---

7. We also conducted the experiments with alternative constructions of the environments. Since switching regression models are hard to fit if the distribution of the predictors strongly differs between states, some choices of environments make our method output the empty set—a result that is not incorrect, but uninformative.

8. We omit all intercept terms, impose an equality constraint on the error variances, and assume an i.i.d. structure on the hidden variables. For estimation, we use the NLM optimizer. In our implementation of the test (9), the lowest attainable p-value is 10^{-4}.
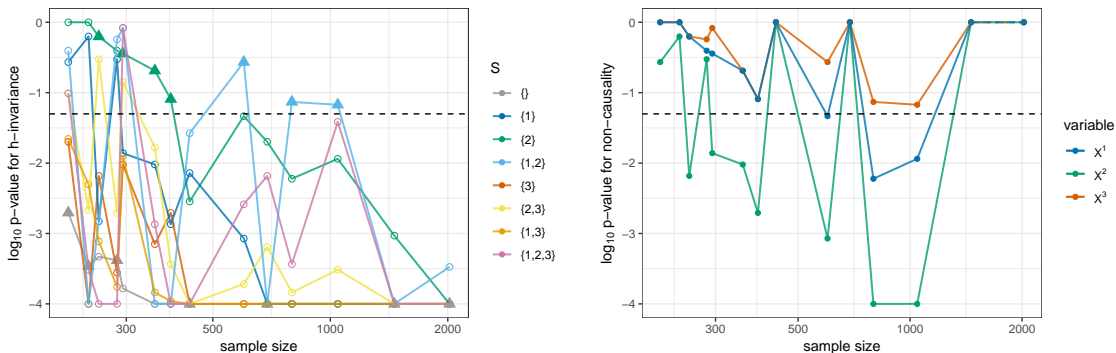
Figure 11: P-values for h-invariance of different sets $S \subseteq \{1, 2, 3\}$ (left) and p-values for non-causality (see Section 2.4.1) of the individual variables X^1, X^2 and X^3 (right). For every experiment, the estimated set $\hat{S}$ in the left plot is indicated by a triangle. For several sample sizes, our method correctly infers $\hat{S} = \{2\}$ (left), and the causal parent X^2 consistently obtains the lowest p-value for non-causality (right). Experiments for which all p-values for non-causality are equal to 1 correspond to instances in which all sets have been rejected. For large amounts of data, this is always the case (the two largest sample sizes are not shown here). At sample sizes 436, 797 and 1045, our method infers the set $\hat{S} = \{1, 2\}$. This finding may be due to imperfect measurements of the variable X^2, that do not contain all information from X^1 that is relevant for Y.

structure from Figure 10. As the sample size grows, the power of our tests of the hypotheses $H_{0,S}$ increases, and even small differences in regression coefficients are detected. For sample sizes above 1459 (the two largest sample sizes are not shown here), all hypotheses $H_{0,S}$ are rejected, and our method returns the uninformative output $\hat{S} = \emptyset$. At sample sizes 436, 797 and 1045, our method infers the set $\hat{S} = \{1, 2\}$, that is, the two predictors APAR_{chl} and SW. A possible explanation is that the true chlorophyll content is unknown, and that APAR_{chl} therefore itself is estimated (on the basis of the greenness index EVI (Huete et al., 2002)). Due to these imperfect measurements, X^1 may still contain information about Y that cannot be explained by X^2.

5.2.2. RECONSTRUCTION OF THE VEGETATION TYPE

We know that (Y, X^2) follows a switching regression model (see Figure 10), and that the hidden variable in this model corresponds to the true vegetation type. We can thus obtain estimates of the vegetation type by reconstructing the values of the hidden variable in the fitted model. We use the data set at its highest resolution, and exploit the background knowledge that H does not change throughout the considered time span. All observations obtained from one spatial grid cell are therefore assumed to stem from the same under-lying regime. Let $\mathcal{S} \subseteq \mathbb{R}^2$ and $\mathcal{T} = \{1, \ldots, 12\}$ be the spatial and the temporal grid, respectively, along which data are observed. We then classify each grid cell $s \in \mathcal{S}$ as $\hat{H}_s := \arg\max_{j \in \{1,2\}} \sum_{t \in \mathcal{T}} \hat{\mathbb{P}}(H_{st} = j \mid Y_{st}, X_{st})$, where $\hat{\mathbb{P}}$ refers to the fitted model. Our
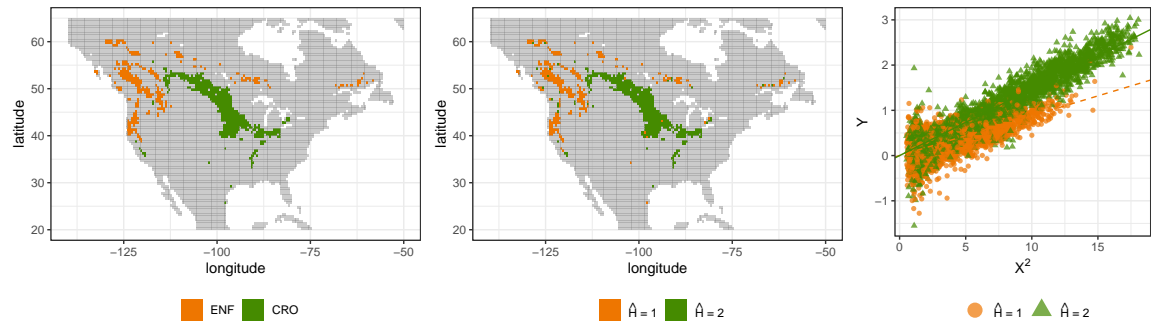
Figure 12: Vegetation type by IGBP (left) and estimates obtained from reconstructing the values of the hidden variable, as described in Section 5.2.2 (middle). We correctly classify more than 95% of the pixels. The right hand plot illustrates the vegetation-dependent linear relationship between Y and X^2. Switching regression model fits are indicated by straight lines, and points are colored according the reconstructed value of $\hat{H}$. Since the data are not well-clustered in the X^2-Y space, classifying observations based on data from (Y, X^2) is generally not a straight-forward task.

method correctly reconstructs the hidden variable in more than 95% of the grid cells (Figure 12, left and middle). As seen in Figure 12 (right), reconstructing H based on data from (Y, X^2) is not an easy classification problem.

So far, we have assumed that the IGBP classification corresponds to the true vegetation type. In reality, it is an estimate based on greenness indices that are constructed from remotely sensed radiation reflected from the Earth's surface. The outcome of our method may be viewed as an alternative ecosystem classification scheme, which additionally comes with a process-based interpretation: each cluster corresponds to a different slope parameter in the linear regression of SIF on $APAR_{chl}$. This parameter represents the efficiency at which absorbed energy is quenched as fluorescence, and is referred to as *fluorescence yield*.

6. Conclusions and Future Work

This paper discusses methodology for causal discovery that is applicable in the presence of discrete hidden variables. If the data set is time-ordered, the hidden variables may follow a Markov structure. The method is formulated in the framework of invariant causal prediction. It aims at inferring causal predictors of a target variable and comes with the following coverage guarantee: whenever the method's output is non-empty, it is correct with large probability. Our algorithm allows for several user choices and is tested on a wide range of simulations. We see that also in small sample regimes and under a variety of different model violations, the coverage is not negatively affected. Our implementation allows for using either the EM-algorithm or a numerical maximization technique. In our experiments, we find that the two options yield very similar results, but that the latter is computationally faster and more suitable for handling parameter constraints. The power of

both methods decreases with an increasing number of hidden states. This conforms to the theoretical result that, in general, identifiability of causal predictors cannot be achieved if the hidden variable may take arbitrarily many states, for example.

As part of the method, we propose a test for the equality of two switching regression models; to the best of our knowledge this is the first example of such a test and may be of interest in itself. We prove the asymptotic validity of this test by providing sufficient conditions for the existence, the consistency and the asymptotic normality of the maximum likelihood estimator in switching regression models.

On the real world data, the true causal parent is consistently attributed the highest significance as a causal predictor of the target variable. Switching regression models can also be used for classifying data points based on reconstructed values of the hidden variables.

For large sample sizes, most goodness of fits test are usually rejected in real data. Since the h-invariance assumption may not hold exactly either, it may be interesting to explore relaxations of this assumption. For example, Pfister et al. (2019a) propose a causal ranking, and Rothenhäusler et al. (2018) interpolate between prediction and invariance. Our robustness analysis in Section 5.1.5 suggests that the performance of our method is not negatively affected when allowing for a dependence between X and H, and we believe that our theoretical results could be extended to such scenarios (possibly adding mild assumptions). To widen the range of applicability of our method, it might also be worthwhile to consider non-linear models. In particular, it would be interesting to construct conditional independence tests that are able to take into account a mixture model structure.

Acknowledgments

Appendix A. Structural Causal Models

Below, we formally define structural causal models (Pearl, 2009; Bollen, 2014), and use a presentation similar to Peters et al. (2017, Chapter 6).

Definition 3 (Structural causal model) *A* structural causal model *(SCM) over variables $(Z_1, \ldots, Z_p)$ consists of a family of structural assignments*

$$Z_j := f_j(\mathbf{PA}_j, N_j), \qquad j = 1, \ldots, p,$$

where for each $j \in \{1, \ldots, p\}$, $\mathbf{PA}_j \subseteq \{Z_1, \ldots, Z_p\} \setminus \{Z_j\}$ denotes the parent set *of variable Z_j, and a product distribution over the noise variables $(N_1, \ldots, N_p)$. Every SCM induces a graph over the nodes in $\{Z_1, \ldots, Z_p\}$: for every j, one draws an arrow from each of the variables in $\mathbf{PA}_j$ to Z_j. We here require this graph to be acyclic. A variable Z_i is a* cause *of Z_j, if there exists a directed path from Z_i to Z_j. The variables in $\mathbf{PA}_j$ are said to be the* direct causes *of Z_j.*

Due to the acyclicity of the graph, an SCM induces a distribution over the variables $Z_1, \ldots, Z_p$. An *intervention* on Z_j corresponds to replacing the corresponding assignment. (We still require joint independence of all noise variables, as well as the acyclicity of the induced graph to be preserved under interventions.) This yields another SCM and another distribution, the intervention distribution.

Appendix B. Parametrizations of the Models IID and HMM

Define $\mathcal{G}^{\mathrm{IID}} := [0,1]^{\ell-1}$ and $\mathcal{G}^{\mathrm{HMM}} := \{\gamma \in [0,1]^{(\ell-1)\ell} \mid \text{ for all } j \in \{1, \ldots, \ell\} : \sum_{k=1}^{\ell-1} \gamma_{j\ell+k} \leq 1\}$ and parametrize the transition matrix via the maps $\boldsymbol{\Gamma}^{\mathrm{IID}} : \mathcal{G}^{\mathrm{IID}} \to [0,1]^{\ell \times \ell}$ and $\boldsymbol{\Gamma}^{\mathrm{HMM}} : \mathcal{G}^{\mathrm{HMM}} \to [0,1]^{\ell \times \ell}$, for all $i, j \in \{1, \ldots, \ell\}$ given by

$$\boldsymbol{\Gamma}^{\mathrm{IID}}_{ij}(\gamma) = \begin{cases} \gamma_j & j < \ell \\ 1 - \sum_{k=1}^{\ell-1} \gamma_k & j = \ell \end{cases} \qquad \text{and} \qquad \boldsymbol{\Gamma}^{\mathrm{HMM}}_{ij}(\gamma) = \begin{cases} \gamma_{i\ell+j} & j < \ell \\ 1 - \sum_{k=1}^{\ell-1} \gamma_{i\ell+k} & j = \ell. \end{cases}$$

For the regression matrix Θ, we consider the two types of parameter constraints discussed in Section 3.2. For $c > 0$, let $\mathcal{T}^c := (\mathbb{R}^p \times [c, \infty))^\ell$ and $\mathcal{T}^= := \mathbb{R}^{p\ell} \times (0, \infty)$ and parametrize the regression matrix via the maps $\boldsymbol{\Theta}^c : \mathcal{T}^c \to \mathbb{R}^{p \times \ell}$ and $\boldsymbol{\Theta}^= : \mathcal{T}^= \to \mathbb{R}^{p \times \ell}$, for all $i \in \{1, \ldots, p+1\}$ and $j \in \{1, \ldots, \ell\}$ given by

$$\boldsymbol{\Theta}^c_{ij}(\theta) = \theta_{(j-1)(p+1)+i} \qquad \text{and} \qquad \boldsymbol{\Theta}^=_{ij}(\theta) = \begin{cases} \theta_{(j-1)p+i} & i \leq p \\ \theta_{p\ell+1} & i = p+1. \end{cases}$$

Both of the parameter constraints induced by $(\boldsymbol{\Theta}^c, \mathcal{T}^c)$ and $(\boldsymbol{\Theta}^=, \mathcal{T}^=)$ ensure the existence of the maximum likelihood estimator, see Theorem 1. Since all of the above coordinate mappings are linear in θ and γ, Assumption (A4) in Section 3.5 is satisfied for any pair $(\boldsymbol{\Theta}, \boldsymbol{\Gamma})$ with $\boldsymbol{\Theta} \in \{\boldsymbol{\Theta}^c, \boldsymbol{\Theta}^=\}$ and $\boldsymbol{\Gamma} \in \{\boldsymbol{\Gamma}^{\mathrm{IID}}, \boldsymbol{\Gamma}^{\mathrm{HMM}}\}$.

Appendix C. Proofs

C.1. Proof of Proposition 2

Recall that by Definition 3, we require the underlying causal graph to be acyclic. For every $t \in \{1, \ldots, n\}$, we can therefore recursively substitute structural assignments to express $(X_t^{\mathrm{PA}^0(Y)}, H_t^*)$ as a function of all noise variables appearing in the structural assignments of the ancestors of Y_t. Using the joint independence of all noise variables (see Definition 3), it follows that $(X_t^{\mathrm{PA}^0(Y)}, H_t^*) \perp\!\!\!\perp N_t$. Using the i.i.d. assumption on $(N_t)_{t \in \{1, \ldots, n\}}$, we have that for all t and for all x, h, the distribution of $Y_t \mid (X_t^{\mathrm{PA}^0(Y)} = x, H_t^* = h) \overset{d}{=} f(x, h, N_t)$ does not depend on t, which shows that $S^* = \mathrm{PA}^0(Y)$ satisfies (4). By writing $Y_t = \sum_{h=1}^\ell f(X_t^{\mathrm{PA}^0(Y)}, h, N_t) \mathbb{1}_{\{H_t^* = h\}}$ and using the linearity of the functions $f(\cdot, h, \cdot)$, it follows that $S^* = \mathrm{PA}^0(Y)$ is h-invariant with respect to $(\mathbf{Y}, \mathbf{X})$. $\blacksquare$

C.2. Proof of Theorem 1

We first introduce some notation. Since neither of the parametrizations in question impose any constraints on the regression coefficients, we will throughout this proof write $\theta = (\beta, \delta)$,

where $\beta = (\beta_1, \ldots, \beta_\ell) \in \mathcal{B} := \mathbb{R}^{p \times \ell}$ and $\delta \in \mathcal{D}$ is the part of θ that parametrizes the error variances, i.e., $\mathcal{D}^= = (0, \infty)$ and $\mathcal{D}^c = [c, \infty)^\ell$. Also, we will use $\bar{\mathcal{D}}^= = [0, \infty]$, $\bar{\mathcal{D}}^c = [c, \infty]^\ell$, $\bar{\mathcal{B}} = (\mathbb{R} \cup \{-\infty, +\infty\})^{p \times \ell}$ to denote the "compactifications" of $\mathcal{D}^c$, $\mathcal{D}^=$ and $\mathcal{B}$, respectively. For every $h \in \{1, \ldots, \ell\}^m$ and every $j \in \{1, \ldots, \ell\}$ define $T_{h=j} := \{t \in \{1, \ldots, m\} : h_t = j\}$ and write the likelihood function as $G = \sum_{h \in \{1, \ldots, \ell\}^m} g_h$, where

$$g_h(\phi) = p(\mathbf{x}) \lambda(\gamma)_{h_1} \prod_{s=2}^m \mathbf{\Gamma}_{h_{s-1} h_s}(\gamma) \prod_{j=1}^\ell \prod_{t \in T_{h=j}} \mathcal{N}(y_t \mid x_t \beta_{h_t}, \sigma_{h_t}^2(\delta)),$$

where the product over an empty index set is defined to be 1.

Let $G^* := \sup_{\phi \in \mathcal{P}} G(\phi) \in (0, \infty]$. We want to show that there exists $\phi^* \in \mathcal{P}$ with $G(\phi^*) = G^*$ (which in particular shows that $G^* < \infty$). The idea of the proof is as follows. We first show that given an arbitrary point $\bar{\phi}$ in the compactification $\bar{\mathcal{P}}$ and an arbitrary sequence $(\phi^n)_{n \in \mathbb{N}}$ in $\mathcal{P}$ that converges to $\bar{\phi}$, we can construct a sequence $(\tilde{\phi}^n)_{n \in \mathbb{N}}$ with limit point $\tilde{\phi} \in \mathcal{P}$, such that $\lim_{n \to \infty} G(\tilde{\phi}^n) \geq \lim_{n \to \infty} G(\phi^n)$. We then let $(\phi^{*n})_{n \in \mathbb{N}}$ be a sequence with $\lim_{n \to \infty} G(\phi^{*n}) = G^*$. By compactness of $\bar{\mathcal{P}}$, we can wlog assume that $(\phi^{*n})_{n \in \mathbb{N}}$ is convergent in $\bar{\mathcal{P}}$ (otherwise we may choose a convergent subsequence). By the first part of the proof, there exists a sequence $(\tilde{\phi}^{*n})_{n \in \mathbb{N}}$ that is convergent to some $\phi^* \in \mathcal{P}$, and with $\lim_{n \to \infty} G(\tilde{\phi}^{*n}) = G^*$. By continuity of G, $G(\phi^*) = G^*$.

Let $\bar{\phi} = (\bar{\beta}, \bar{\delta}, \gamma) \in \bar{\mathcal{P}}$ and let $(\phi^n)_{n \in \mathbb{N}} = (\beta^n, \delta^n, \gamma^n)_{n \in \mathbb{N}}$ be such that $\lim_{n \to \infty} \phi^n = \bar{\phi}$. If $\bar{\phi} \in \mathcal{P}$, there is nothing to prove. Assume therefore $\bar{\phi} \in \bar{\mathcal{P}} \setminus \mathcal{P}$. Since $\mathcal{G}$ was assumed to be compact, $\bar{\mathcal{P}} = \bar{\mathcal{B}} \times \bar{\mathcal{D}} \times \mathcal{G}$. The problem can therefore be divided into the two cases $\bar{\delta} \in \bar{\mathcal{D}} \setminus \mathcal{D}$ and $\bar{\beta} \in \bar{\mathcal{B}} \setminus \mathcal{B}$, which are treated in Lemma 1 and Lemma 2, respectively. Together, they imply the existence of a sequence $(\tilde{\phi}^n)_{n \in \mathbb{N}}$ with $\lim_{n \to \infty} \tilde{\phi}^n \in \mathcal{P}$ and $\lim_{n \to \infty} G(\tilde{\phi}^n) \geq \lim_{n \to \infty} G(\phi^n)$, thereby completing the proof of Theorem 1.

We first consider the case where $\bar{\delta} \in \bar{\mathcal{D}} \setminus \mathcal{D}$.

Lemma 1 *Let $(\phi^n)_{n \in \mathbb{N}}$ be a sequence in $\mathcal{P}$ that converges to a point $\bar{\phi} = (\bar{\beta}, \bar{\delta}, \gamma) \in \bar{\mathcal{B}} \times (\bar{\mathcal{D}} \setminus \mathcal{D}) \times \mathcal{G}$ and assume that the limit $\lim_{n \to \infty} G(\phi^n)$ exists in $[0, \infty]$. Then, there exists a sequence $(\tilde{\phi}^n)_{n \in \mathbb{N}}$ with limit point $(\bar{\beta}, \delta, \gamma) \in \bar{\mathcal{B}} \times \mathcal{D} \times \mathcal{G}$, such that $\limsup_{n \to \infty} G(\tilde{\phi}^n) \geq \lim_{n \to \infty} G(\phi^n)$.*

Proof We treat the two parametrizations $(\mathbf{\Theta}^c, \mathcal{T}^c)$ and $(\mathbf{\Theta}^=, \mathcal{T}^=)$ separately.

If $\mathcal{D} = \mathcal{D}^c$, then $\bar{\mathcal{D}} \setminus \mathcal{D} = \{(\bar{\delta}_1, \ldots, \bar{\delta}_\ell) \in [c, \infty]^\ell : \bar{\delta}_j = \infty$ for at least one $j\}$. Let j be such that $\bar{\delta}_j = \infty$. Since for every $h \in \{1, \ldots, \ell\}^m$,

$$g_h(\phi^n) \begin{cases} \to 0 \text{ as } n \to \infty & \text{if } T_{h=j} \neq \emptyset \\ \text{does not depend on } \delta_j^n & \text{otherwise,} \end{cases} \tag{19}$$

we can simply substitute $(\delta_j^n)_{n \in \mathbb{N}}$ by the sequence $(\tilde{\delta}_j^n)_{n \in \mathbb{N}}$ that is constantly equal to c, to obtain $(\tilde{\phi}^n)_{n \in \mathbb{N}}$ with $\limsup_{n \to \infty} G(\tilde{\phi}^n) \geq \lim_{n \to \infty} G(\phi^n)$. By repeating this procedure for all j with $\bar{\delta}_j = \infty$, we obtain a sequence $(\tilde{\phi}^n)_{n \in \mathbb{N}}$ with $\limsup_{n \to \infty} G(\tilde{\phi}^n) \geq \lim_{n \to \infty} G(\phi^n)$ and such that $\delta = \lim_{n \to \infty} \delta^n \in \mathcal{D}$.

If $\mathcal{D} = \mathcal{D}^=$, then $\bar{\mathcal{D}} \setminus \mathcal{D} = \{0, \infty\}$. If $\bar{\delta} = \infty$, then $\lim_{n \to \infty} G(\phi^n) = 0$ and the result is trivial. Assume therefore that $\bar{\delta} = 0$. Let $h \in \{1, \ldots, \ell\}^m$ be fixed. By the assumption

on the sample $(\mathbf{y}, \mathbf{x})$, there exists no set of parameters that yield a perfect fit. We may therefore find a sequence $(s(n))_{n \in \mathbb{N}}$ of elements in $\{1, \dots, m\}$ such that $y_{s(n)} - x_{s(n)} \beta^n_{h_{s(n)}}$ is bounded away from zero for all n large enough. For every $n \in \mathbb{N}$ we have

$$g_h(\phi^n) \le p(\mathbf{x})(2\pi\sigma_1^2(\delta^n))^{-m/2} \exp\left(-\frac{1}{2\sigma_1^2(\delta^n)}(y_{s(n)} - x_{s(n)} \beta^n_{h_{s(n)}})^2\right).$$

Since the last factor on the right hand side goes to zero exponentially fast in $\sigma_1^2(\delta^n)$, it follows that $\lim_{n \to \infty} g_h(\phi^n) = 0$. Since h was arbitrary, we have that $\lim_{n \to \infty} G(\phi^n) = 0$, and the result follows. ∎

We now turn to the case where $\bar{\beta} \in \bar{\mathcal{B}} \setminus \mathcal{B}$.

Lemma 2 *Let $(\phi^n)_{n \in \mathbb{N}}$ be a sequence in $\mathcal{P}$ that converges to a point $\bar{\phi} = (\bar{\beta}, \delta, \gamma) \in (\bar{\mathcal{B}} \setminus \mathcal{B}) \times \mathcal{D} \times \mathcal{G}$. Then, there exists a sequence $(\tilde{\phi}^n)_{n \in \mathbb{N}}$ with limit point $(\beta, \delta, \gamma) \in \mathcal{B} \times \mathcal{D} \times \mathcal{G}$, such that $\lim_{n \to \infty} G(\tilde{\phi}^n) \ge \limsup_{n \to \infty} G(\phi^n)$.*

Proof The idea of the proof is as follows. We construct a bounded sequence $(\tilde{\beta}^n)_{n \in \mathbb{N}}$, such that the sequence $(\tilde{\phi}^n)_{n \in \mathbb{N}}$ obtained from $(\phi^n)_{n \in \mathbb{N}}$ by substituting $(\beta^n)_{n \in \mathbb{N}}$ by $(\tilde{\beta}^n)_{n \in \mathbb{N}}$ satisfies that $\lim_{n \to \infty} G(\tilde{\phi}^n) \ge \limsup_{n \to \infty} G(\phi^n)$. Since $(\delta^n)_{n \in \mathbb{N}}$ was assumed to be convergent in $\mathcal{D}$ (and hence bounded) and by compactness of $\mathcal{G}$, the whole sequence $(\tilde{\phi}^n)_{n \in \mathbb{N}}$ is bounded. We can therefore find a compact set $\mathcal{K} \subseteq \mathcal{P}$, such that $\{\tilde{\phi}^n : n \in \mathbb{N}\} \subseteq \mathcal{K}$. Consequently, we can wlog assume that $(\tilde{\phi}^n)_{n \in \mathbb{N}}$ is convergent in $\mathcal{K}$ (otherwise we may choose a convergent subsequence). The sequence $(\tilde{\phi}^n)_{n \in \mathbb{N}}$ then fulfills the requirements in Lemma 2, thereby completing the proof.

The crucial part that remains is the construction of the sequence $(\tilde{\beta}^n)_{n \in \mathbb{N}}$. This is done by induction. Let $(\phi^n)_{n \in \mathbb{N}} = (\beta_1^n, \dots, \beta_\ell^n, \delta^n, \gamma^n)$ be as stated in Lemma 2 and let K^∞ be the set of states k, for which $\|\beta_k^n\| \to \infty$ as $n \to \infty$. We then construct $(\tilde{\beta}^n)_{n \in \mathbb{N}}$ in the following way. Pick an arbitrary $k \in K^\infty$ and construct a bounded sequence $(\tilde{\beta}_k^n)_{n \in \mathbb{N}}$ (this construction is described below), such that the sequence $(\tilde{\phi}^n_{(k)})_{n \in \mathbb{N}}$ obtained from $(\phi^n)_{n \in \mathbb{N}}$ by substituting $(\beta_k^n)_{n \in \mathbb{N}}$ by $(\tilde{\beta}_k^n)_{n \in \mathbb{N}}$ satisfies that $\limsup_{n \to \infty} G(\tilde{\phi}^n_{(k)}) \ge \limsup_{n \to \infty} G(\phi^n)$. We then take $k' \in K^\infty \setminus \{k\}$ and similarly construct $(\tilde{\phi}^n_{(k,k')})_{n \in \mathbb{N}}$ from $(\tilde{\phi}^n_{(k)})_{n \in \mathbb{N}}$ such that $\limsup_{n \to \infty} G(\tilde{\phi}^n_{(k,k')}) \ge \limsup_{n \to \infty} G(\tilde{\phi}^n_{(k)})$. By inductively repeating this procedure for all elements of K^∞, we obtain a bounded sequence $(\tilde{\beta}^n)_{n \in \mathbb{N}}$, such that $(\tilde{\phi}^n)_{n \in \mathbb{N}} = (\tilde{\beta}^n, \delta^n, \gamma^n)_{n \in \mathbb{N}}$ satisfies that $\limsup_{n \to \infty} G(\tilde{\phi}^n) \ge \lim_{n \to \infty} G(\phi^n)$. Once again, we can wlog assume that $(G(\tilde{\phi}^n))_{n \in \mathbb{N}}$ converges, since otherwise we can choose a convergent subsequence $(G(\tilde{\phi}^{n_i}))_{i \in \mathbb{N}}$ with $\lim_{i \to \infty} G(\tilde{\phi}^{n_i}) = \limsup_{n \to \infty} G(\tilde{\phi}^n)$.

We now prove the induction step. Assume that we have iteratively constructed sequences for $k_1, \dots, k_j \in K^\infty$ (if $j = 0$, this corresponds to the base case). For simplicity write $(\check{\phi}^n)_{n \in \mathbb{N}} = (\check{\phi}^n_{(k_1, \dots, k_j)})_{n \in \mathbb{N}}$. Pick an arbitrary $k \in K^\infty \setminus \{k_1, \dots, k_j\}$. If for all $t \in \{1, \dots, m\}$, $|x_t \beta_k^n| \to \infty$ as $n \to \infty$, we could (similar to the proof of Lemma 1) take $(\tilde{\beta}_k^n)_{n \in \mathbb{N}}$ to be a constant sequence. Since in general, there might exist s such that $|x_s \beta_k^n| \not\to \infty$ as $n \to \infty$, we divide the problem as follows. Define $\mathcal{S}_1 := \{s \in \{1, \dots, m\} : |x_s \beta_k^n| \to \infty \text{ as } n \to \infty\}$, $\mathcal{S}_2 := \{1, \dots, m\} \setminus \mathcal{S}_1$, $\mathcal{H}_1 := \{h \in \{1, \dots, \ell\}^m : T_{h=k} \cap \mathcal{S}_1 \ne \emptyset\}$ and $\mathcal{H}_2 := \{1, \dots, \ell\}^m \setminus \mathcal{H}_1$, and write the likelihood function as $G = G_1 + G_2$, where $G_1 := \sum_{h \in \mathcal{H}_1} g_h$ and $G_2 :=$

$\sum_{h \in \mathcal{H}_2} g_h$. We now show that $\lim_{n \to \infty} G_1(\check{\phi}^n) = 0$. We formulate a slightly more general result, which we will also make use of later in the proof:

(*) Let $h \in \{1, \ldots, \ell\}^m$ and assume there exists a sequence $(s(n))_{n \in \mathbb{N}}$ of elements in $T_{h=k}$, such that $|x_{s(n)} \beta_k^n| \to \infty$ as $n \to \infty$. Then, $\lim_{n \to \infty} g_h(\phi^n) = 0$.

Proof [*] Since $(\delta^n)_{n \in \mathbb{N}}$ was assumed to be convergent in $\mathcal{D}$, all sequences $\{\sigma_j^2(\delta^n)\}_{n \in \mathbb{N}}$, $j \in \{1, \ldots, \ell\}$, are bounded from above and bounded away from 0. Since for all $n \in \mathbb{N}$,

$$g_h(\phi^n) \leq p(x)(2\pi)^{-n/2} \prod_{t=1}^m (\sigma_{h_t}^2(\delta^n))^{-1/2} \exp \underbrace{\left(-\frac{1}{2\sigma_k^2(\delta^n)} (y_{s(n)} - x_{s(n)} \beta_k^n)^2 \right)}_{\to -\infty},$$

we are done. ∎

For $h \in \mathcal{H}_1$, we can simply pick $s_0 \in T_{h=k} \cap \mathcal{S}_1$ and consider the sequence $(s(n))_{n \in \mathbb{N}}$ that is constantly equal to s_0. The result (*) therefore shows that $\lim_{n \to \infty} G_1(\check{\phi}^n) = 0$. It thus suffices to construct $(\tilde{\phi}_k^n)_{n \in \mathbb{N}}$ from $(\check{\phi}^n)_{n \in \mathbb{N}}$ such that $\limsup_{n \to \infty} G_2(\tilde{\phi}_k^n) \geq \limsup_{n \to \infty} G_2(\check{\phi}^n)$. Since for every $h \in \mathcal{H}_2$ we have $T_{h=k} \subseteq \mathcal{S}_2$, we take a closer look at $\mathcal{S}_2$. For every $s \in \mathcal{S}_2$, the sequence $(|x_s \beta_k^n|)_{n \in \mathbb{N}}$ is either bounded or can be decomposed into two sequences, one of which is bounded and one of which converges to infinity. For every $s \in \mathcal{S}_2$, let therefore I_s^b and I_s^∞ be disjoint subsets of $\mathbb{N}$ with $I_s^b \cup I_s^\infty = \mathbb{N}$, such that $(|x_s \beta_k^n|)_{n \in I_s^b}$ is bounded and such that either $I_s^\infty = \emptyset$ or $|I_s^\infty| = \infty$ with $(|x_s \beta_k^n|)_{n \in I_s^\infty}$ converging to infinity. Let $I^b := \cup_{s \in \mathcal{S}_2} I_s^b$ and define a sequence $(\tilde{\beta}_k^n)_{n \in \mathbb{N}}$ by

$$\tilde{\beta}_k^n := \begin{cases} \text{the projection of } \beta_k^n \text{ onto } \operatorname{span}_{\mathbb{R}}(\{x_s : s \text{ satisfies } n \in I_s^b\}) & \text{if } n \in I^b \\ 0 & \text{otherwise.} \end{cases}$$

We now show that the above defines a bounded sequence.

(∘) The sequence $(\tilde{\beta}_k^n)_{n \in \mathbb{N}}$ is bounded.

Proof [∘] For every $\mathcal{S} \subseteq \mathcal{S}_2$, define $I_{\mathcal{S}}^b := \{n \in \mathbb{N} : n \in I_s^b \Leftrightarrow s \in \mathcal{S}\}$ (where $I_{\emptyset}^b := \mathbb{N} \setminus I^b$). We can then decompose $(\tilde{\beta}_k^n)_{n \in \mathbb{N}}$ into the subsequences $(\tilde{\beta}_k^n)_{n \in I_{\mathcal{S}}^b}$, $\mathcal{S} \subseteq \mathcal{S}_2$, and prove that each of these sequences is bounded. Let $\mathcal{S} \subseteq \mathcal{S}_2$ and let $\{u_1, \ldots, u_d\}$ be an orthonormal basis for $\operatorname{span}_{\mathbb{R}}(\{x_s : s \in \mathcal{S}\})$. Since all sequences in $\{(|x_s \tilde{\beta}_k^n|)_{n \in I_{\mathcal{S}}^b} : s \in \mathcal{S}\}$ are bounded, then so are the sequences $(|u_1 \tilde{\beta}_k^n|)_{n \in I_{\mathcal{S}}^b}, \ldots, (|u_d \tilde{\beta}_k^n|)_{n \in I_{\mathcal{S}}^b}$ (this follows by expressing each of the u_is as a linear combination of elements in $\{x_s : s \in \mathcal{S}\}$). The result now follows from the identities $\|\tilde{\beta}_k^n\|^2 = \sum_{j=1}^d |u_j \tilde{\beta}_k^n|^2$, $n \in I_{\mathcal{S}}^b$. ∎

Let $(\tilde{\phi}_k^n)_{n \in \mathbb{N}}$ be the sequence obtained from $(\check{\phi}^n)_{n \in \mathbb{N}}$ by substituting $(\beta_k^n)_{n \in \mathbb{N}}$ by $(\tilde{\beta}_k^n)_{n \in \mathbb{N}}$. Finally, we show the following result.

(△) $\limsup_{n \to \infty} G(\tilde{\phi}_k^n) \geq \limsup_{n \to \infty} G(\check{\phi}^n)$.

Proof [△] Let $h \in \mathcal{H}_2$ and define $I_h^\infty := \bigcup_{s \in T_{h=k}} I_s^\infty$ (if $T_{h=k} = \emptyset$, we define $I_h^\infty := \emptyset$). The idea is to decompose $(\check{\phi}^n)_{n \in \mathbb{N}}$ into $(\check{\phi}^n)_{n \in I_h^\infty}$ and $(\check{\phi}^n)_{n \notin I_h^\infty}$ and to treat both sequences separately.

We start by considering $(\check{\phi}^n)_{n \notin I_h^\infty}$. First, observe that for every s, $\mathcal{N}(y_s \,|\, x_s \beta_k, \sigma_k^2(\delta))$ only depends on β_k via the inner product $x_s \beta_k$. By construction of I_h^∞ and $(\check{\beta}_k^n)_{n \in \mathbb{N}}$, we thus have that for all $n \notin I_h^\infty$ and for all $s \in T_{h=k}$, the function values $\mathcal{N}(y_s \,|\, x_s \check{\beta}_k^n, \sigma_k^2(\delta^n))$ and $\mathcal{N}(y_s \,|\, x_s \tilde{\beta}_k^n, \sigma_k^2(\delta^n))$ coincide. Consequently, we have that for all $n \notin I_h^\infty$, $g_h(\check{\phi}_k^n) = g_h(\tilde{\phi}^n)$. In particular, the sequences $(\check{g}_{h,b}^n)_{n \in \mathbb{N}}$ and $(\tilde{g}_{h,b}^n)$, for every $n \in \mathbb{N}$ defined by $\check{g}_{h,b}^n := g_h(\check{\phi}^n)\mathbb{1}_{\{n \notin I_h^\infty\}}$ and $\tilde{g}_{h,b}^n := g_h(\tilde{\phi}^n)\mathbb{1}_{\{n \notin I_h^\infty\}}$, coincide.

We now consider $(\check{\phi}^n)_{n \in I_h^\infty}$. By construction of the sets $I_s^\infty, s \in T_{h=k}$, either $I_h^\infty = \emptyset$ or $|I_h^\infty| = \infty$. If $|I_h^\infty| = \infty$, then for every $n \in \mathbb{N}$, there exists $\check{s}(n) \in T_{h=k}$ such that $n \in I_{\check{s}(n)}^\infty$. By applying (*) to the sequence $(\check{\phi}^n)_{n \in I^\infty}$ with $(s(n))_{n \in I_h^\infty} = (\check{s}(n))_{n \in I_h^\infty}$, it follows that $\lim_{n \to \infty, n \in I_h^\infty} g_h(\check{\phi}^n) = 0$. In particular, the sequences $(\check{g}_{h,\infty}^n)_{n \in \mathbb{N}}$ and $(\tilde{g}_{h,\infty}^n)_{n \in \mathbb{N}}$, for every $n \in \mathbb{N}$ defined by $\check{g}_{h,\infty}^n := g_h(\check{\phi}^n)\mathbb{1}_{\{n \in I_h^\infty\}}$ and $\tilde{g}_{h,\infty}^n := g_h(\tilde{\phi}^n)\mathbb{1}_{\{n \in I_h^\infty\}}$, converge to 0 as $n \to \infty$ (this holds also if $I^\infty = \emptyset$).

By combing the above results for all $h \in \mathcal{H}_2$, we finally have

$$
\begin{aligned}
\limsup_{n \to \infty} G_2(\check{\phi}^n) &= \limsup_{n \to \infty} \left(\sum_{h \in \mathcal{H}_2} \check{g}_{h,b}^n + \sum_{h \in \mathcal{H}_2} \check{g}_{h,\infty}^n \right) = \limsup_{n \to \infty} \left(\sum_{h \in \mathcal{H}_2} \check{g}_{h,b}^n \right) \\
&= \limsup_{n \to \infty} \left(\sum_{h \in \mathcal{H}_2} \tilde{g}_{h,b}^n \right) \leq \limsup_{n \to \infty} \left(\sum_{h \in \mathcal{H}_2} \tilde{g}_{h,b}^n + \sum_{h \in \mathcal{H}_2} \tilde{g}_{h,\infty}^n \right) \\
&= \limsup_{n \to \infty} G_2(\tilde{\phi}_k^n).
\end{aligned}
$$

Since $\limsup_{n \to \infty} G_1(\tilde{\phi}_k^n) \geq 0 = \limsup_{n \to \infty} G_1(\check{\phi}^n)$, the result follows. ∎

This completes the proof of Lemma 2. ∎

C.3. Proof of Theorem 2

We start by introducing some notation to be used in the proofs of Theorem 2 and Theorem 3. Let $\mathcal{K} := \mathbb{R}^p \times (0, \infty)$ be the full parameter space for a single pair $\kappa = (\beta^T, \sigma^2)^T$ of regression parameters. In analogy to previous notation, we will use $\kappa_j(\theta)$ to denote the jth pair of regression parameters of a parameter vector $\theta \in \mathcal{T}$. If the conditional distribution of $Y_t \,|\, (X_t = x, H_t = j)$ is a normal distribution with regression parameters κ, we will denote the conditional density of $(X_t, Y_t) \,|\, (H_t = j)$ by $f(x, y \,|\, \kappa)$. We use $\mathbb{P}_0$ for the distribution $\mathcal{SR}(\phi^0 \,|\, X_1)$ and $\mathbb{E}_0$ for the expectation with respect to $\mathbb{P}_0$. Finally, for every $k \in \mathbb{N}$, let $\mathcal{SR}^k(\cdot \,|\, X_1)$ denote the unconstrained class of mixture distributions of degree k (i.e., all parameters can vary independently within their range).

Theorem 2 now follows from Leroux (1992, Theorem 3). To prove the applicability of their result, we first state slightly adapted versions of their conditions (L1)–(L6) and prove

afterwards that they are satisfied. (L1) Γ^0 is irreducible, (L2) for each (x, y), $\kappa \mapsto f(x, y \mid \kappa)$ is continuous and vanishes at infinity (see the last paragraph of Section 2 in Leroux (1992)), (L3) for all $j, k \in \{1, \ldots, \ell\}$, the maps $\theta \mapsto \kappa_j(\theta)$ and $\gamma \mapsto \boldsymbol{\Gamma}_{jk}(\gamma)$ are continuous, (L4) for all $j \in \{0, \ldots, \ell\}$, $\mathbb{E}_0[|\log f(X_1, Y_1 \mid \kappa_j(\theta^0))|] < \infty$, (L5) for all $\kappa \in \mathcal{K}$, there exists a $\delta > 0$ such that $\mathbb{E}_0[\sup_{\kappa': \|\kappa - \kappa'\| < \delta} (\log f(X_1, Y_1 \mid \kappa'))^+] < \infty$, and (L6) for every $k \in \{1, \ldots, \ell\}$, the class $\mathcal{SR}^k(\cdot \mid X_1)$ satisfies the following identifiability property. Define

$$\Lambda^k := \{(\lambda_1, \ldots, \lambda_k) : \sum_{j=1}^k \lambda_j = 1\}, \text{ and}$$

$$\mathcal{Q}^k := \left\{ \{(\lambda_1, \kappa_1), \ldots, (\lambda_k, \kappa_k)\} : \begin{array}{l} (\lambda_1, \ldots, \lambda_k) \in \Lambda^k \text{ and } \kappa_1, \ldots, \kappa_k \in \mathcal{K} \\ \text{with all } \kappa_j\text{s being distinct} \end{array} \right\}$$

and consider the mapping $\varphi^k : \mathcal{Q}^k \to \mathcal{SR}(\cdot \mid X_1)$ that sends $q = \{(\lambda_1, \kappa_1), \ldots, (\lambda_k, \kappa_k)\}$ into the mixture distribution $P_q \in \mathcal{SR}(\cdot \mid X_1)$ with density

$$f_q(x, y) := \sum_{j=1}^k \lambda_j f(x, y \mid \kappa_j) = f(x) \sum_{j=1}^k \lambda_j f(y \mid x, \kappa_j).$$

Then, for every $k \in \{1, \ldots, \ell\}$, φ^k is a one-to-one map of $\mathcal{Q}^k$ onto $\mathcal{SR}^k(\cdot \mid X_1)$. It is therefore the set $\{(\lambda_1, \kappa_1), \ldots, (\lambda_k, \kappa_k)\}$, rather than the parameters $(\kappa_1, \ldots, \kappa_k)$ and $(\lambda_1, \ldots, \lambda_k)$ themselves, that is required to be identifiable.

We now show that (L1)–(L6) are satisfied. Condition (L1) is implied by (A3). Condition (L2) follows by the continuity of $\kappa \mapsto \mathcal{N}(y \mid x, \kappa)$ and (L3) is implied by (A4). For (L4), we see that for all $j \in \{0, \ldots, \ell\}$,

$$\log f(X_1, Y_1 \mid \kappa_j(\theta^0)) = \log(2\pi\sigma_j^2(\theta^0)) - \frac{1}{2\sigma_j^2(\theta^0)}(Y_1 - X_1\beta_j(\theta^0))^2 + \log f(X_1) \in \mathcal{L}^1(\mathbb{P}_0),$$

by (A7) and by moment-properties of the normal distribution. For (L5), let $\kappa = (\beta, \sigma^2) \in \mathcal{K}$ and choose $\delta := \sigma^2/2$. We then have

$$\mathbb{E}_0 \left[\sup_{\kappa': \|\kappa' - \kappa\| < \delta} (\log f(X_1, Y_1 | \kappa'))^+ \right] \leq \mathbb{E}_0 \left[\sup_{\kappa': \|\kappa' - \kappa\| < \delta} (\log f(Y_1 | X_1, \kappa'))^+ + |\log f(X_1)| \right]$$

$$\leq \mathbb{E}_0 \left[\sup_{\sigma': \|\sigma'^2 - \sigma^2\| < \delta} (-\frac{1}{2} \log(2\pi\sigma'^2))^+ + |\log f(X_1)| \right]$$

$$\leq \mathbb{E}_0 \left[\frac{1}{2} |\log(\pi\sigma^2)| + |\log f(X_1)| \right] < \infty.$$

It is left to prove (L6), the identifiability of the classes $\mathcal{SR}^k(\cdot \mid X_1)$. Teicher (1963, Proposition 1) shows an analogous result for mixtures of univariate normal distributions, that are parametrized by their mean and variance. His result will be the cornerstone for our argument. Consider a fixed $k \in \{1, \ldots, \ell\}$, let $q = \{(\lambda_1, \beta_1, \sigma_1^2), \ldots, (\lambda_k, \beta_k, \sigma_k^2)\}, q' = \{(\lambda_1', \beta_1', \sigma_1'^2), \ldots, (\lambda_k', \beta_k', \sigma_k'^2)\} \in \mathcal{Q}^k$ and assume that the induced mixtures P_q and $P_{q'}$ are identical. Collect q and q' into two matrices Q, Q' with columns $Q_{\cdot j} = (\lambda_j, \sigma_j^2, \beta_j^T)^T$ and

$Q'_{\cdot j} = (\lambda'_j, \sigma'^2_j, \beta'^T_j)^T$ for $j \in \{1, \ldots, k\}$. We wish to show that Q and Q' are equal up to a permutation of their columns. Because the densities f_q and $f_{q'}$ coincide Lebesgue-almost everywhere, it holds that, for all $x \in \mathrm{int}(\mathrm{supp}(X_1))$,

$$f_q(y \mid x) = \sum_{j=0}^{k} \lambda_j f(y \mid x, \kappa_j) = \sum_{j=0}^{k} \lambda'_j f(y \mid x, \kappa'_j) = f_{q'}(y \mid x) \qquad \text{for almost all } y.$$

It now follows from Teicher (1963, Proposition 1) that, for all $x \in \mathrm{int}(\mathrm{supp}(X_1))$,

$$\{(\lambda_1, \sigma_1^2, x\beta_1), \ldots, (\lambda_k, \sigma_k^2, x\beta_k)\} = \{(\lambda'_1, \sigma'^2_1, x\beta'_1), \ldots, (\lambda'_k, \sigma'^2_k, x\beta'_k)\}. \qquad (20)$$

In the remainder of the proof, we will consider several x simultaneously (rather than a fixed x). This will help us to draw conclusions about the betas. Equation (20) means that for every $z \in \mathcal{Z} := \mathbb{R}^2 \times \mathrm{int}(\mathrm{supp}(X_1))$, the vectors zQ and zQ' are equal up to a permutation of their entries (this permutation may depend on z). Let Σ denote the (finite) family of permutation matrices of size $k \times k$ and consider the partition

$$\mathcal{Z} = \bigcup_{M \in \Sigma} \mathcal{Z}_M, \qquad \text{where} \qquad \mathcal{Z}_M = \{z \in \mathcal{Z} : zQ = zQ'M^T\}.$$

Since $\mathcal{Z}$ is an open subset of $\mathbb{R}^{p+2}$, there exists an element $M_0 \in \Sigma$, such that $\mathcal{Z}_{M_0}$ contains an open subset of $\mathbb{R}^{p+2}$. We can therefore choose $p + 2$ linearly independent elements $z_1, \ldots, z_{p+2} \in \mathcal{Z}_{M_0}$ and construct the invertible matrix $\mathbf{Z} = [z_1^T, \ldots, z_{p+2}^T]^T$. Since $\mathbf{Z}Q = \mathbf{Z}Q'M_0^T$, it follows that $Q = Q'M_0^T$. $\blacksquare$

C.4. Proof of Theorem 3

Throughout the proof, we make use of the notation introduced in the first paragraph of Appendix C.3. Theorem 3 follows if both the below statements hold true.

(i) $m^{-1}\mathcal{J}(\hat{\phi}_m) \to \mathcal{I}_0$ as $m \to \infty$ in $\mathbb{P}_0$-probability.

(ii) $\sqrt{m}(\hat{\phi}_m - \phi^0)\mathcal{I}_0^{1/2} \xrightarrow{d} \mathcal{N}(0, I)$ as $m \to \infty$ under $\mathbb{P}_0$.

These results correspond to slightly adapted versions of Lemma 2 and Theorem 1, respectively, in Bickel et al. (1998) (here referred to as L2 and T1). L2 builds on assumptions (B1)–(B4) to be stated below. T1 additionally assumes that $\phi^0 \in \mathrm{int}(\mathcal{P})$ and that the Fisher information matrix $\mathcal{I}_0$ is positive definite, i.e., our (A2) and (A5). Assumptions (B1)–(B4) state local regularity conditions for a neighborhood of the true parameter ϕ^0. We therefore need to verify that there exists an open neighborhood $\mathcal{T}_0$ of θ^0, such that the following conditions are satisfied.

(B1) The transition matrix Γ^0 is irreducible and aperiodic.

(B2) For all $j, k \in \{1, \ldots, \ell\}$ and for all (x, y), the maps $\gamma \mapsto \mathbf{\Gamma}_{jk}(\gamma)$ and $\theta \mapsto f(x, y|\kappa_j(\theta))$ (for $\theta \in \mathcal{T}_0$) have two continuous derivatives.

(B3) Write $\theta = (\theta_1, \ldots, \theta_K)$. For all $n \in \{1, 2\}$, $i_1, \ldots, i_n \in \{1, \ldots, K\}$ and $j \in \{1, \ldots, \ell\}$, it holds that

(i)

$$\int \sup_{\theta \in \mathcal{T}_0} \left| \frac{\partial^n}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} f(x, y | \kappa_j(\theta)) \right| d(x, y) < \infty, \quad \text{and}$$

(ii)

$$\mathbb{E}_0 \left[\sup_{\theta \in \mathcal{T}_0} \left| \frac{\partial^n}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} \log f(X_1, Y_1 | \kappa_j(\theta)) \right|^{3-n} \right] < \infty.$$

(B4) For all (x, y), define

$$\rho(x, y) = \sup_{\theta \in \mathcal{T}_0} \max_{0 \le i, j \le \ell} \frac{f(x, y | \kappa_i(\theta))}{f(x, y | \kappa_j(\theta))}.$$

Then for all $j \in \{1, \ldots, \ell\}$, $\mathbb{P}_0(\rho(X_1, Y_1) = \infty \mid H_1 = j) < 1$.

We first construct the set $\mathcal{T}_0$. Let therefore $\varepsilon > 0$ and choose $\mathcal{T}_0$ so small that there exists $c > 0$, such that for all $\theta \in \mathcal{T}_0$ and for all $j \in \{1, \ldots, \ell\}$ and $k \in \{1, \ldots, d\}$, it holds that $\beta_{jk}(\theta) \in (\beta_{jk}(\theta^0) - \varepsilon, \beta_{jk}(\theta^0) + \varepsilon)$ and $\sigma_j^2(\theta) \ge c$. We can now verify the conditions (B1)–(B4).

Assumption (B1) is satisfied by (A3). For every (x, y), the maps $\kappa \mapsto f(x, y | \kappa)$ are two times continuously differentiable on $\mathbb{R}^p \times (0, \infty)$. Together with (A4), this implies (B2), independently of the choice of $\mathcal{T}_0$.

For the proof of (B3)(i)–(ii) we will make use of the following result. Let g be a polynomial of (x, y) of degree at most 4, i.e., a sum of functions on the form $b x_i^r x_k^s y^t$ for some $i, k \in \{1, \ldots, p\}$ and $r, s, t \in \{0, \ldots, 4\}$ with $r + s + t \le 4$. Then, for every $\kappa \in \mathcal{K}$, $\int g(|x|, |y|) f(x, y \mid \kappa) d(x, y) < \infty$, where $|x| = (|x_1|, \ldots, |x_p|)$. This result follows from the fact that for every x, $\int |y|^t f(y \mid x, \kappa) dy$ is a polynomial of $|x|$ of degree t, and the assumption that, for all $j \in \{1, \ldots, p\}$, $\mathbb{E}[|X_1^j|^4] < \infty$.

For (B3)(i), we treat all derivatives simultaneously. Let $n \in \{1, 2\}$, $i_1, \ldots, i_n \in \{1, \ldots, K\}$ and $j \in \{1, \ldots, \ell\}$ be fixed. Let $\{g_\theta\}_{\theta \in \mathcal{T}_0}$ be the functions, for all (x, y) and for all $\theta \in \mathcal{T}_0$ defined by

$$\frac{\partial^n}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} f(x, y \mid \kappa_j(\theta)) = g_\theta(x, y) \exp\left(-\frac{1}{2\sigma_j^2(\theta)} (y - x\beta_j(\theta))^2 \right) f(x),$$

(note that $f(x) = 0$ implies $f(x, y | \kappa_j(\theta)) = 0$). Then, for all (x, y), $\theta \mapsto g_\theta(x, y)$ is continuous, and for all $\theta \in \mathcal{T}_0$, $(x, y) \mapsto g_\theta(x, y)$ is a polynomial of degree at most 4. By the compactness of $\bar{\mathcal{T}}_0$, the closure of $\mathcal{T}_0$, and by the continuity of $\theta \mapsto g_\theta(x, y)$, there exists a polynomial g of degree 4, such that, for all (x, y), $\sup_{\theta \in \mathcal{T}_0} |g_\theta(x, y)| \le g(|x|, |y|)$.

Consider now a fixed $k \in \{1, \ldots, p\}$. By choice of $\mathcal{T}_0$, we have that for all x_k and for all $\theta \in \mathcal{T}_0$, it holds that $x_k(\beta_{jk}(\theta^0) - \text{sign}(x_k)\varepsilon) \le x_k \beta_{jk}(\theta) \le x_k(\beta_{jk}(\theta^0) + \text{sign}(x_k)\varepsilon)$. By writing $s(x) = (\text{sign}(x_1), \ldots, \text{sign}(x_p))$ it follows that for all (x, y) and all $\theta \in \mathcal{T}_0$, $y - x(\beta_j(\theta^0) - \text{diag}(s(x))\varepsilon) \le y - x\beta_j(\theta) \le y - x(\beta_j(\theta^0) + \text{diag}(s(x))\varepsilon)$. Consequently, we may for every (x, y) find $s(x, y) \in \{-1, 1\}^p$ (either $s(x)$ or $-s(x)$) such that for all $\theta \in \mathcal{T}_0$,

$$-(y - x\beta_j(\theta))^2 \le -(y - x\underbrace{(\beta_j(\theta^0) + \text{diag}(s(x, y))\varepsilon)}_{=:\beta_s})^2.$$

By choosing $C > 0$ small enough, it follows that for all (x, y) and for all $\theta \in \mathcal{T}_0$ it holds that

$$\exp\left(-\frac{1}{2\sigma_j^2(\theta)}(y - x\beta_j(\theta))^2\right) \leq \exp\left(-C(y - x\beta_j(\theta))^2\right) \leq \sum_{s \in \{-1,1\}^p} \exp\left(-C(y - x\beta_s)^2\right).$$

Since all integrals $\int g(|x|, |y|) \exp(-C(y - x\beta_s)^2) f(x) d(x, y)$, $s \in \{-1, 1\}^p$, are finite, this completes the proof of (B3)(i).

The proof of (B3)(ii) is similar to that of (B3)(i). Fix $n \in \{1, 2\}$, $i_1, \ldots, i_n \in \{1, \ldots, K\}$ and $j \in \{1, \ldots, \ell\}$. Let $\{h_\theta\}_{\theta \in \mathcal{T}_0}$ be the functions, for all (x, y) and for all $\theta \in \mathcal{T}_0$ defined by

$$\frac{\partial^n}{\partial\theta_{i_1} \cdots \partial\theta_{i_n}} \log f(x, y \mid \kappa_j(\theta)) = h_\theta(x, y).$$

Then, for all (x, y), $\theta \mapsto h_\theta(x, y)$ is continuous, and for all $\theta \in \mathcal{T}_0$, $(x, y) \mapsto h_\theta(x, y)$ is a polynomial of degree at most 2. We can therefore find a dominating polynomial h of degree 2, such that, for all (x, y), $\sup_{\theta \in \mathcal{T}_0} |h_\theta(x, y)| \leq h(|x|, |y|)$. Since $h(|X_1|, |Y_1|) \in \mathcal{L}^2(\mathbb{P}_0)$, this completes the proof of (B3)(ii).

(B4) is easily verified. Since the support $\mathcal{S}$ of the functions $f(\cdot \mid \kappa)$ does not depend on κ, it is enough to consider $(x, y) \in \operatorname{int}(\mathcal{S})$. For all $(x, y) \in \operatorname{int}(\mathcal{S})$ and for all $j \in \{1, \ldots, \ell\}$, $\theta \mapsto f(x, y \mid \kappa_j(\theta))$ is bounded from above and bounded away from zero (by choice of $\mathcal{T}_0$). The function ρ is therefore finite everywhere. ∎

C.5. Proof of Corollary 1

Let (A1)–(A7) hold true. By Theorem 2, we can decompose $(\hat{\phi}_m)_{m \in \mathbb{N}} = ((\hat{\theta}_m, \hat{\gamma}_m))_{m \in \mathbb{N}}$ into one or more subsequences, each of which is convergent to a permutation of ϕ^0. We can therefore find a sequence $(\pi_{\mathcal{P}}^m)_{m \in \mathbb{N}} = ((\pi_{\mathcal{T}}^m, \pi_{\mathcal{G}}^m))_{m \in \mathbb{N}}$ of permutations on $\mathcal{P}$, such that, $\mathbb{P}_0$-almost surely, the sequence of maximum likelihood estimators $(\pi_{\mathcal{P}}^m(\hat{\phi}_m))_{m \in \mathbb{N}}$ converges to ϕ^0 as $m \to \infty$. For $\alpha \in (0, 1)$ and for every $m \in \mathbb{N}$, we then have

$$\mathbb{P}_0^m(\theta^0 \in C_{\text{adjusted}}^\alpha(\hat{\theta}_m)) \geq \mathbb{P}_0^m(\theta^0 \in C^\alpha(\pi_{\mathcal{T}}^m(\hat{\theta}_m))) = \mathbb{P}_0^m(\phi^0 \in C^\alpha(\pi_{\mathcal{T}}^m(\hat{\theta}_m)) \times \mathcal{G}).$$

By Theorem 3, the right hand side converges to $1 - \alpha$ as $m \to \infty$. ∎

C.6. Proof of Theorem 4

By Corollary 1, the adjusted confidence regions within each environment all achieve the correct asymptotic coverage, ensuring the asymptotic validity of the test φ_{S^*} of H_{0,S^*}. Since, for every n, $\mathbb{P}_0^n(\hat{S}_n \subseteq S^*) \geq \mathbb{P}_0^n(\varphi_{S^*}^n \text{ accepts } H_{0,S^*}^n)$, the result follows. ∎

Appendix D. Further Details on Likelihood Optimization

Below, we describe the two optimization methods NLM and EM. Since the loglikelihood function (10) is non-convex, the performance of these routines depend on the initialization. In practice, we restart the algorithms in 5 different sets of starting values (using the `regmix.init` function from the R package `mixtools`).

D.1. Method I ("NLM"): Non-Linear Maximization

This method maximizes the loglikelihood function (10) numerically. We use the R optimizer `nlm`, which is a non-linear maximizer based on a Newton-type optimization routine (e.g., Schnabel et al., 1985). The method also outputs an estimate of the observed Fisher information, which is used for the construction of the confidence regions (11). An equality constraint on the error variances can be enforced directly by using the parametrization $(\mathbf{\Theta}^=, \mathcal{T}^=)$ described in Appendix B. A lower bound (we use 10^{-4} as a default value) can be imposed by suitable reparametrization of all error variances (e.g., Zucchini et al., 2016, Section 3.3.1).

D.2. Method II ("EM"): The EM-algorithm

Given starting values $\phi^{(0)} \in \mathcal{P}$, the EM-algorithm operates by alternating between the following two steps until a convergence criterion is met. (1) The E-step: Compute the posterior distribution $P_{(\mathbf{y},\mathbf{x})}^{(t)}$ of $\mathbf{H} \,|\, (\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}, \phi^{(t)})$ given the current parameters $\phi^{(t)}$. (2) The M-step: Maximize the expected complete data loglikelihood

$$Q(\phi \,|\, \phi^{(t)}) := \mathbb{E}_{P_{(\mathbf{y},\mathbf{x})}^{(t)}} \left[\ell_{\text{complete}}(\mathbf{y}, \mathbf{x}, \mathbf{H} \,|\, \phi) \right] \tag{21}$$

to obtain updates $\phi^{(t+1)} \in \arg\max_{\phi \in \mathcal{P}} Q(\phi \,|\, \phi^{(t)})$. Here, ℓ_{complete} is the loglikelihood function of the complete data $(\mathbf{y}, \mathbf{x}, \mathbf{h})$. The explicit forms of $P_{(\mathbf{y},\mathbf{x})}^{(t)}$ and Q depend on the choice of model. In model IID, $P_{(\mathbf{y},\mathbf{x})}^{(t)}$ is a product distribution which can be computed by simple applications of Bayes' theorem. In model HMM, the posterior distribution is obtained by the forward-backward algorithm. In both cases, (21) can be maximized analytically (e.g., Bishop, 2006, Chapters 9 and 13). The observed Fisher information $\mathcal{J}(\hat{\phi})$ can be computed analytically from the derivatives of (21), see Oakes (1999). In our R package, the EM-algorithm is only implemented for model IID and makes use of the package `mixreg`. An equality constraint on the error variances can be accommodated using the parametrization $(\mathbf{\Theta}^=, \mathcal{T}^=)$ from Appendix B. A lower bound on the error variances is enforced by restarting the algorithm whenever an update $\phi^{(t)}$ contains a variance component that deceeds the lower bound (`mixreg` uses the threshold 10^{-16}).

Figure 13 shows numerical results for ICPH when using the EM-algorithm as optimization routine. The results should be compared to Figure 8, where NLM has been applied to the same data. The two methods perform very similarly, although NLM is computationally faster (by approximately a factor of 6), and better suited for handling the lower bound constraint on the error variances.

Appendix E. Additional Numerical Experiments

In this section, we present additional experimental results. In all simulations, we use slight adaptations of the SCM in Section 5.1.2, and measure the performance of ICPH using rejection rates for non-causality (similar to Figure 9). All results are summarized in Figure 14.
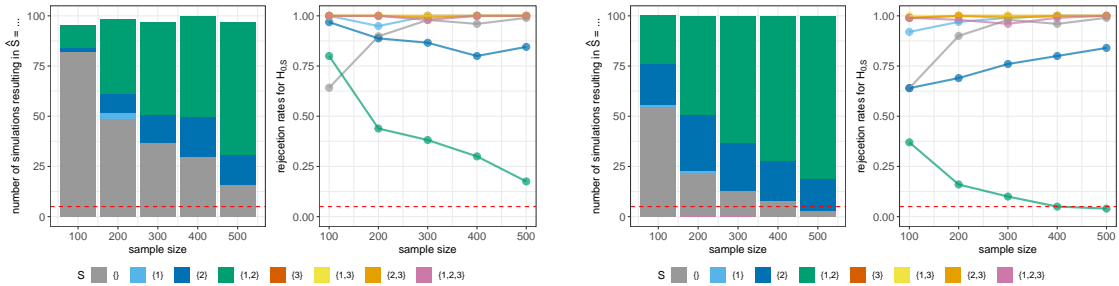
Figure 13: Output of ICPH (bar plots) and rejection rates for individual hypotheses (curve plots) for the experiment in Section 5.1.4 with parameter constraint $\sigma_{Y1}^2, \sigma_{Y2}^2 \geq 10^{-16}$ (left) and $\sigma_{Y1}^2 = \sigma_{Y2}^2$ (right), using the EM-algorithm as optimization routine. The results are very similar to those presented in Figure 8, where NLM is applied to the same data. The only notable differences are the missing values in the bar plots (left). These simulations correspond to instances in which the EM-algorithm, after trying several different starting values, failed to converge to a solution which satisfies the variance constraints.

E.1. Non-Binary Latent Variables and Unknown Number of States

ICPH requires the number of states as an input parameter—we test for h-invariance of degree ℓ in line 8 of Algorithm 1. If ℓ is unknown, we propose the following modification. Let $K \geq 3$ be some predefined integer (e.g., $K = 5$), and let for every $S \subseteq \{1, \ldots, d\}$ and every $k \in \{2, \ldots, K\}$, p_S^k be a p-value for the hypothesis $H_{0,S}^k$ of h-invariance of degree k of the set S, obtained from the test (9). We then substitute the p-value p_S in line 8 of Algorithm 1 by $p_S' := \max\{p_S^k : 2 \leq k \leq K\}$. By construction, the test defined by p_S' is a valid test of $H_{0,S}^\ell$ for any (unknown) $\ell \in \{2, \ldots, K\}$. Our code package automatically performs this procedure when the supplied argument `number.of.states` is a vector of length greater than one. We now investigate this procedure numerically. For a fixed sample size of $n = 500$ and for every $\ell \in \{2, 3, 4, 5\}$, we generate 100 i.i.d. data sets from the SCM in Section 5.1.2 with parameters sampled as in Section 5.1.3. The probabilities $\lambda_j = P(H = j)$, $j \in \{0, \ldots, \ell\}$ are sampled uniformly between 0.1 and $1/(\ell+1)$ and standardized correctly. In Figure 14 (left), we compare three different approaches: (i) we always test for h-invariance of degree 2 (circles), (ii) we always test for h-invariance of degree less than or equal to 5, using the approach described above (triangles), and (iii) we test for h-invariance using the true number of states ℓ (squares). For all methods, ICPH maintains the type I error control, but drops in power as the number of latent states increases. Even if the number of latent states is unknown (but small), ICPH often recovers the causal parents X^1 and X^2. In general, we propose to limit the application of ICPH to cases where the hidden variables is expected to take only a few different values.
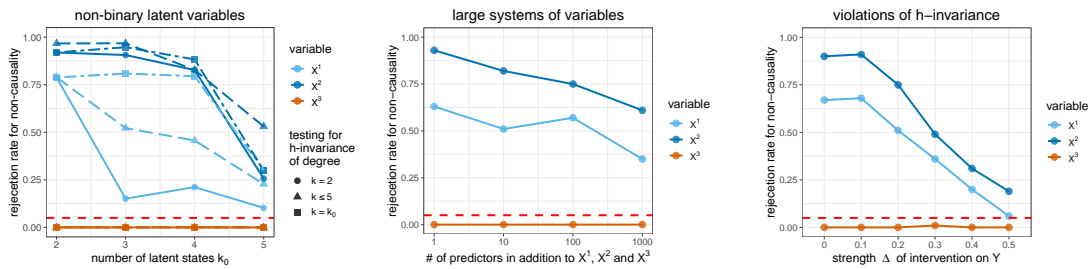
Figure 14: Rejection rates for non-causality of the variables X^1, X^2 and X^3 for the experiments described in Appendix E. We investigate the performance of ICPH for non-binary variables (left), for large numbers of predictors (middle), and under violations of the h-invariance assumption (right). By simultaneously testing for h-invariance of different degrees (see Appendix E.1 for details), we can recover X^1 and X^2 even if the true number of latent states is unknown (left figure, triangles). Our algorithm can be combined with an upfront variable screening (here using Lasso), which results in satisfactory performance even for large number of predictor variables (middle). Under violations of Assumption 1, the population version of ICPH is not able to infer $S^* = \{1, 2\}$. In the finite sample case we still identify X^1 and X^2 if H_{0,S^*} is only mildly violated (right).

E.2. Systems with Large Numbers of Variables

For a fixed sample size of $n = 300$, we simulate data (Y, X^1, X^2, X^3, H) as described in Section 5.1.2. For increasing $m \in \{1, 10, 100, 1000\}$, we generate additional predictor variables $(Z^1, \ldots, Z^m)$ from the structural assignments $Z^j := \alpha_j X^3 + N_j^Z$, $j = 1, \ldots, m$, where $N_1^Z, \ldots, N_m^Z$ are i.i.d. standard Gaussian noise variables, and all α_j are drawn independently from a Uniform$(-1, 1)$ distribution. We then perform variable screening by selecting the first 5 predictors included along the Lasso selection path (Tibshirani, 1994), and run ICPH on the reduced data set. The results in Figure 14 (middle) suggest that even for a large number of predictors, ICPH is generally able to infer S^* (provided that S^* contains only few variables).

E.3. Violations of the h-Invariance Assumption

The theoretical guarantees of our method rely on the existence of an h-invariant set (Assumption 1). We now empirically investigate the performance of ICPH under violations of this assumption. For a fixed sample size of $n = 300$, we generate data as described in Section 5.1.2, but include direct interventions on Y. For increasing values of $\Delta \in \{0, 0.1, \ldots, 0.5\}$, we change the coefficients $(\beta_{11}^Y, \beta_{21}^Y)$ in the structural assignment of Y to $(\beta_{11}^Y + \Delta, \beta_{21}^Y + \Delta)$ in environment e_2, and to $(\beta_{11}^Y - \Delta, \beta_{21}^Y - \Delta)$ in environment e_3. As expected, the power of our method drops with the strength of intervention (Figure 14 right).

References

P. J. Bickel, Y. Ritov, and T. Ryden. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4):1614–1635, 1998.

C. M. Bishop. *Machine Learning and Pattern Recognition.* Springer, New York, USA, 2006.

K. A. Bollen. *Structural equations with latent variables*, volume 210. John Wiley & Sons, 2014.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, 2011.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

W. K. Ching and M. K. Ng. *Markov Chains: Models, Algorithms and Applications.* Springer, New York, USA, 2nd edition, 2006.

T. Claassen, J. M. Mooij, and T. Heskes. Learning sparse causal models is not NP-hard. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 172–181, 2013.

P. Constantinou and A. P. Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, 45(6):2618–2653, 2017.

J. Cussens. Bayesian network learning with cutting planes. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 153–160, 2011.

R. De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8 (3):227–245, 02 1989.

S. Goldfeld and R. Quandt. The estimation of structural shifts by switching regressions. In *Annals of Economic and Social Measurement, Volume 2, number 4*, pages 475–485. NBER, 1973.

L. Guanter, C. Frankenberg, A. Dudhia, P. E. Lewis, J. Gómez-Dans, A. Kuze, H. Suto, and R. G. Grainger. Retrieval and global assessment of terrestrial chlorophyll fluorescence from gosat space measurements. *Remote Sensing of Environment*, 121:236–251, 2012.

R. J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795–800, 1985.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13 (Aug):2409–2464, 2012.

A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira. Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote sensing of environment*, 83(1-2):195–213, 2002.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, New York, USA, 2015.

D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 2012.

J. L. Jensen and N. V. Petersen. Asymptotic normality of the maximum likelihood estimator in state space models. *The Annals of Statistics*, 27(2):514–535, 1999.

M. Jung, M. Reichstein, and A. Bondeau. Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. *Biogeosciences*, 6(10):2001–2013, 2009.

N. M. Kiefer. Discrete parameter variation: Efficient estimation of a switching regression model. *Econometrica: Journal of the Econometric Society*, 46(2):427–434, 1978.

M. Koivisto. Advances in exact Bayesian structure discovery in Bayesian networks. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 241–248, 2006.

R. Langrock, T. Kneib, R. Glennie, and T. Michelot. Markov-switching generalized additive models. *Statistics and Computing*, 27(1):259–270, 2017.

E. L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer Science & Business Media, New York, USA, 2nd edition, 2006.

B. G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40(1):127–143, 1992.

T. R. Loveland, B. C. Reed, J. F. Brown, D. O. Ohlen, Z. Zhu, L. W. M. J. Yang, and J. W. Merchant. Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *International Journal of Remote Sensing*, 21(6-7): 1303–1330, 2000.

M. H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

S. Magliacane, T. Claasen, and J. M. Mooij. Joint causal inference on observational and experimental datasets. *arXiv preprint arXiv:1611.10351*, 2016.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

D. Oakes. Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):479–482, 1999.

J. M. Ogarrio, P. Spirtes, and J. Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the 8th International Conference on Probabilistic Graphical Models PGM*, pages 368–379, 2016.

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 by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press, Cambridge, MA, USA, 2017.

N. Pfister, S. Bauer, and J. Peters. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019a.

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

T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.

D. Rothenhäusler, P. Bühlmann, N. Meinshausen, and J. Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

S. W. Running and M. Zhao. Daily GPP and annual NPP (MOD17A2/A3) products NASA Earth observing system MODIS land algorithm. *MOD17 Users Guide*, 2015.

M. Sander. Market timing over the business cycle. *Journal of Empirical Finance*, 46: 130–145, 2018.

R. B. Schnabel, J. E. Koonatz, and B. E. Weiss. A modular system of algorithms for unconstrained minimization. *ACM Transactions on Mathematical Software (TOMS)*, 11 (4):419–440, 1985.

E. Sgouritsa, D. Janzing, J. Peters, and B. Schölkopf. Identifying finite mixtures of nonparametric product distributions and causal inference of confounders. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 556–565, 2013.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

T. Silander and P. Myllymak. A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 445–452, 2006.

R. Silva and R. Evans. Causal inference through a witness protection program. *Journal of Machine Learning Research*, 17(56):1–53, 2016. URL `http://jmlr.org/papers/v17/15-130.html`.

R. Silva and Z. Ghahramani. The hidden life of latent variables: Bayesian learning with mixed graph models. *Journal of Machine Learning Research*, 10:1187–1238, 2009.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *In Proceedings of 11th Conference on Uncertainty in Artificial Intelligence UAI*, pages 499–506, 1995.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, Massachusetts, USA, 2nd edition, 2000.

B. Steudel, D. Janzing, and B. Schölkopf. Causal Markov condition for submodular information measures. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 464–476, 2010.

H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4): 1265–1269, 1963.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

M. Tsagris, G. Borboudakis, V. Lagani, and I. Tsamardinos. Constraint-based causal discovery with mixed data. *International Journal of Data Science and Analytics*, 6(1):19–30, 2018.

T. R. Turner. Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3):371–384, 2000.

Y. Zhang, X. Xiao, C. Jin, J. Dong, S. Zhou, P. Wagle, J. Joiner, L. Guanter, Y. Zhang, G. Zhang, et al. Consistency between sun-induced chlorophyll fluorescence and gross primary production of vegetation in North America. *Remote Sensing of Environment*, 183:154–169, 2016.

W. Zucchini, I. L. MacDonald, and R. Langrock. *Hidden Markov Models for Time Series: An Introduction Using R.* CRC Press, Taylor & Francis, Boca Raton, Florida, USA, 2nd edition, 2016.