Rune Christiansen

Causal Inference in the Presence of Hidden Variables: Structure Learning, Effect Estimation and Distribution Generalization

June 2020

PhD thesis

Department of Mathematical Sciences
University of Copenhagen

Rune Christiansen
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark

Principal supervisor:	Prof. Jonas Peters University of Copenhagen
Co-supervisor:	Prof. Niels Richard Hansen University of Copenhagen
Assessment committee:	Prof. Susanne Ditlevsen (chair) University of Copenhagen
	Prof. Nicolai Meinshausen ETH Zurich
	Prof. Pierre Pinson Technical University of Denmark

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen. It was supported by a research grant (18968) from VILLUM fonden.

Acknowledgments

I am grateful to my supervisor, Jonas Peters, for introducing me to the fascinating field of causality. I thank him for his contagious excitement, for his many encouraging words, and for always giving every question, may many of them have been naive or insignificant, his undivided attention. I thank him for not only being an attentive supervisor and a mentor for good scientific practice, but also a caring neighbor and an enthusiastic discgolf buddy.

I am grateful to my co-supervisor, Niels Richard Hansen, for helpful scientific discussions throughout my graduate and PhD studies, and even more so for his honest professional advice. It is fair to say to that, without his encouragement, I probably would not have applied for this position.

I further thank all members of the assessment committee, Susanne Ditlevsen, Nicolai Meinshausen and Pierre Pinson, for careful reading of this thesis, and for helpful typographical corrections.

During my PhD studies, I was fortunate to spend several months at the Max Planck Institute for Biogeochemistry in Jena, Germany, and later at the National Institute for Applied Statistics Research Australia, Wollongong, Australia. For granting me these opportunities, for warmly welcoming and integrating me at their departments, and for high-quality academic supervision, I am grateful to Markus Reichstein, Miguel D. Mahecha, Noel Cressie and Andrew Zammit-Mangion.

To all my colleagues at the department, in particular to my fellow junior staff members Aleksander, Angélica, Frederik, Gin, Laura, Lasse, Mads, Niels, Nikolaj, Phillip and Yijing, I thank you for an open and collaborative working environment. In particular, I wish to thank Søren for proofreading parts of my introduction, Martin, Nicola and Niklas for running some of the last miles with me, and Lydia for handing me those much-needed gel packs.

I wish to thank Matthias Baumann for a fun and exciting research collaboration, for countless (more or less formal) remote meetings, and for baring with my pedantic graphical perfectionism.

As recent circumstances demanded, my thesis was completed at my personal desk at home, rather than at work. What could have become an exercise in radical social isolation turned into a flourishing

social biosphere, with the initiation of several gardening-, cooking-, baking- and brewing projects. For all the swims, the workouts, the yoga, the good foods and drinks, the (attempted) movie nights, the dart and discgolf sessions, the respect for my busy schedule and the willingness to adjust theirs accordingly, for the emotional support, the many laughs, the silliness and the good vibes, I am deeply grateful to my roommates Tiffany and Mikkel. Without you, this final stretch would not have been half as fun.

I feel extremely fortunate to have a wealth of close friendships in my life. I do not intend to offend any of them by saying that very few will understand as much as the introduction of this thesis. I thank them exactly for this. For pulling me away from my desk and for filling my life with all those non-academic things that make it worth living. I am grateful to my old buddies Aj, Alex, Allan, Christian, Fred, Jebbe, Jesper, Joscha, Krister, Malte, Marvin, Maxi, Olmo, Rasmus and Tue, many of which have stood side by side with me for half my life; and to new friends, Norge and Simon, for proving that strong connections need not be built on long histories. I thank Bahne, Flemming and Jorge for their close friendship and for our mutual journey on the spiritual path. I further thank Johanna and Liv for many hours of personal talk, and for allowing me to tap into their, at times slightly ridiculous, but always warm and positive, energy.

At last, I wish to thank my family. I am grateful to my parents for their lovingly liberal upbringing, for always letting me make my own mistakes, and for their trust in my ability to learn from these. I thank Jan, Lea, Levi and Viggo for many playground excursions, sunny backyard afternoons, delicious homemade dinners and solid theme-parties; and all of Røde Bjarne for an ever-open door to their urban oasis. I am grateful to Inge, Lisa, Luca and Annika for always making me feel at home. Finally, I thank my granny for her sincere interest in my work, and for her impressively sharp and relevant questions about it.

Abstract

This thesis aims at advancing the field of statistical causality. Causal modeling is relevant whenever one seeks an understanding not only of how a system evolves by itself, but also how it may respond if some of its components are altered or replaced ('intervened on'). Arguably, such situations are frequently encountered. Inferring causal knowledge from data is a notoriously hard problem, since, even in the limit of infinitely many data, there are typically several compatible causal explanations. Often, this issue is further compounded by incomplete access to all relevant parts of the system (i.e., by the existence of 'hidden variables').

This work addresses several open problems related to causal learning in the presence of hidden variables. It consists of three main theoretical contributions. Chapter 2 considers the task of learning causal relations (the 'causal structure') from heterogeneous data in cases where these are not known a priori. We exploit a fundamental invariance property which is often assumed of causal regression models. In Chapter 3, we present a causal approach to the problem of distributional robustness, where one aims to learn prediction models that perform well not only on the training data, but also on test data that may come from a different distribution. We use the concept of interventions to model the differences in training and test distribution. Chapter 4 emerged from discussions with environmental scientists, and is motivated by the question of a causal relationship between armed conflict and tropical forest loss. It resulted in the development of a novel causal framework for spatio-temporal stochastic processes, and a procedure for drawing causal inference from observational spatio-temporal data.

Resumé

Denne afhandling bidrager til udviklingen af forskningsfeltet statistisk kausalitet. Kausal modellering er relevant i tilfælde hvor man udover at forstå hvordan et system udvikler sig af sig selv, samtidig er interesseret i, hvordan det reagerer, hvis nogle af dets komponenter bliver forandret eller udskiftet ('interveneret på'). Der kan argumenteres for, at sådanne situationer indtræder ofte. At inferere kausal viden er en notorisk svær opgave da der, selv i grænsen af uendeligt meget data, typisk findes adskillige kompatible kausale forklaringer. Dette problem er yderligere forstærket af ufuldstændig adgang til alle systemets relevante dele (dvs. af tilstedeværelsen af 'uobserverbare variable').

Denne opgave adresserer flere åbne spørgsmål relateret til kausal læring i tilstedeværelsen af uobserverbare variable. Den består af tre hovedbidrag. Kapitel 2 betragter problemet af at lære kausale relationer (den 'kausale struktur') fra heterogent data i situationer, hvor denne ikke er givet på forhånd. Vi udnytter en fundamental invarians egenskab, som ofte bliver antaget omkring kausale regressionsmodeller. I Kapitel 3 præsenteres en kausal tilgang til spørgsmålet om fordelingsmæssig robusthed, hvilket beskriver en prediktionsmodels egenskab til ikke blot at fungere godt på træningsdata, men lige såvel på testdata som måtte komme fra en anden fordeling. Vi anvender konceptet interventioner til at modellere forskellene mellem trænings- og test fordelingen. Kapitel 4 opstod ud fra samtaler med miljøforskere, og er motiveret af spørgsmålet omkring en kausal relation mellem væbnet konflikt og tropisk skovrydning i Colombia. Projektet resulterede i udviklingen af en ny kausal ramme for rumlig-tidslige stokastiske processer, og en procedure til at drage kausal inferens på baggrund af observationel rumlig-tidslig data.

Contents

Contribution

This thesis contains three main methodological contributions, each of which consists (up to minor aesthetic modifications) of a previously published or submitted article.

Chapter 2: R. Christiansen and J. Peters. Switching regression models and causal inference in the presence of discrete latent variables. *Journal of Machine Learning Research*, 21(41), 2020

Chapter 3: R. Christiansen, N. Pfister, M. E. Jakobsen, N. Gnecco, and J. Peters. The difficult task of distribution generalization in nonlinear models. *arXiv preprint arXiv:2006.07433*, 2020b

Chapter 4: R. Christiansen, M. Baumann, T. Kuemmerle, M. D. Mahecha, and J. Peters. Towards causal inference for spatio-temporal data: Conflict and forest loss in Colombia. *arXiv preprint arXiv:2005.08639*, 2020a

Two additional contributions to the applied sciences, which are not included in this thesis, are cited below.

- M. D. Mahecha, F. Gans, G. Brandt, R. Christiansen, S. E. Cornell, N. Fomferra, G. Kraemer, J. Peters, P. Bodesheim, G. Camps-Valls, et al. Earth system data cubes unravel global multivariate dynamics. *Earth System Dynamics Discussions*, 11:201–234, 2020

- D. Martini, J. Pacheco-Labrador, O. Perez-Priego, C. van der Tol, T. S. El-Madany, T. Julitta, M. Rossini, M. Reichstein, R. Christiansen, U. Rascher, et al. Nitrogen and phosphorus effect on sun-induced fluorescence and gross primary productivity in mediterranean grassland. *Remote Sensing*, 11(21): 2562, 2019

1 | Introduction

On an individual level, anticipating the effect of a change in one's behavior (an 'intervention'), is key for decision making. The ability to do so relies on a causal understanding (a 'causal model') of reality. In causal data science, one aims at embedding such an understanding into the workings of automated processes. In the absence of causal intuitions to consult, this integration requires proper mathematical formalism. Statistical causal models provide such formalism. They are *statistical* since they allow for probabilistic (rather than deterministic) relations between variables, and they are *causal* in that they model not only the distribution of data that are passively observed (the 'observational distribution'), but also how the system responds to changes in the data generating mechanism (i.e., they model 'intervention distributions', too).

This chapter contains a brief introduction to causal data science. We present important concepts which will be used throughout this work, survey a few existing methods for causal learning, and discuss how the contributions of this thesis fit into the current scientific landscape. The introduction is structured as follows. In Section 1.1, we introduce structural causal models, which provide a formal framework for discussing questions of causality, and which will be our starting point for causal reasoning throughout most of this thesis. Section 1.2 addresses the fact that, in practice, causal models usually need to be learned from data. We present methods for inferring causal relations in cases where these are not known a priori, and further explain how background knowledge of a system's causal structure can be exploited for estimating effects of hypothetical interventions. Section 1.3 introduces the problem of distributional robustness, where one aims to learn prediction models that are robust against distributional changes, and argues how causal concepts may play a role for this task. In Section 1.4, we discuss the inadequacy of classical statistical causal models for modeling complex data structures such as spatio-temporal data.

1.1. Causal reasoning

Causal reasoning describes the procedure of drawing conclusions based on a causal model. Throughout the last decades, many statistical causal models have been proposed. Among the most widely used are structural causal models [Bollen, 1989, Pearl, 2009], causal graphical models [Spirtes et al., 2000], and the framework of potential outcomes [Rubin, 1974]. Below, we formally define structural causal models and introduce the notion of interventions. We further discuss the role of autonomy, which plays a central part in our contribution in Chapter 4. To describe causal relations among random variables, we will rely on graphical representations. We therefore start with a few definitions.

1.1.1. Graph terminology

A *graph* is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{X_1, \dots, X_p\}$ is a set of *vertices* or *nodes* and $\mathcal{E}$ is a set of *edges*. For now, we assume that $\mathcal{E}$ contains only *directed* edges, which are of the form $X_i \to X_j$, for $X_i, X_j \in \mathcal{V}$ with $X_i \neq X_j$. Such a graph is called a *directed graph*. If $X_i \to X_j \in \mathcal{E}$ then X_i is a *parent* of X_j, and X_j is a *child* of X_i. A *path* in $\mathcal{G}$ is a sequence of (at least two) distinct nodes $X_{i_1}, \dots, X_{i_m} \in \mathcal{V}$ such that for all $k = 1, \dots, m-1$, there is an edge between X_{i_k} and $X_{i_{k+1}}$. If all these edges are of the form $X_{i_k} \to X_{i_{k+1}}$, the path is a *directed path*. In that case, X_{i_1} is an *ancestor* of X_{i_k} and X_{i_k} a *descendant* of X_{i_1}. A *(directed) cycle* is a (directed) path $X_{i_1}, \dots, X_{i_k}$ with $X_{i_1} = X_{i_k}$. A directed graph that contains no directed cycles is called a *directed acyclic graph* (DAG).

1.1.2. Structural causal models

Below, we formally define structural causal models, also called structural equation models.

Definition 1.1 (Structural causal model)**.** *A* structural causal model *(SCM) over variables $X_1, \dots, X_p$ is a pair $M = (\mathcal{S}, Q)$ consisting of*

- *a family $\mathcal{S}$ of structural assignments*

$$X_j := f_j(\mathrm{PA}_j, \varepsilon_j), \qquad j = 1, \dots, p,$$

$$
\begin{aligned}
X_1 &:= f_1(X_4, \varepsilon_1) \\
X_2 &:= f_2(X_1, \varepsilon_2) \\
X_3 &:= f_3(X_2, X_4, \varepsilon_3) \\
X_4 &:= \varepsilon_4,
\end{aligned}
$$

where $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4) \sim Q$.

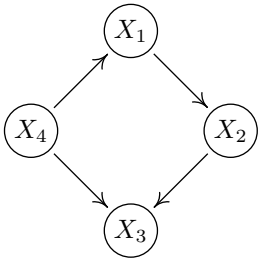


FIGURE 1.1. Our running example for all of this chapter: an SCM over variables (X_1, X_2, X_3, X_4). For every node in the graph (right), the set of parents coincides with the set of variables that appear in the respective structural assignment (left).

where for each $j \in \{1, \dots, p\}$, f_j is some real-valued function, and $\mathrm{PA}_j \subseteq \{X_1, \dots, X_p\} \setminus \{X_j\}$ denotes the parent set *of variable X_j, and*

- *a product distribution Q over the noise variables $(\varepsilon_1, \dots, \varepsilon_p)$.*

The structural assignments $\mathcal{S}$ induce a directed graph $\mathcal{G}$ with nodes $X_1, \dots, X_p$: for every j, one draws an arrow from each of the variables in PA_j to X_j. We require this graph to be acyclic.

Further details about SCMs are provided by Bongers et al. [2016], for example. We refer to the graph $\mathcal{G}$ induced by an SCM as the *causal graph*. Whenever $\mathcal{G}$ contains a directed path from X_j to X_k, then X_j is said to be *a cause* of X_k, or simply to *cause* X_k. For each $k \in \{1, \dots, p\}$, the variables in PA_k are the *direct causes* or the *causal parents* of X_k. A variable X_ℓ which causes both X_j and X_k is a *confounder* of the pair (X_j, X_k), and the (potential) causal effect of X_j on X_k is said to be *confounded* by X_ℓ. In Figure 1.1, we present the SCM which will serve as a running example throughout this introduction. Here, X_2 causes X_3, but this causal effect is confounded by the variable X_4.

Every SCM entails a joint distribution over its variables. This can be seen by iteratively substituting structural assignments into each

other. Due to the assumed acyclicity of $\mathcal{G}$, this procedure terminates in a unique expression of $(X_1, \ldots, X_p)$ in terms of the noise variables $(\varepsilon_1, \ldots, \varepsilon_p)$. We call the induced distribution the *observational distribution*, since it is regarded a statistical model for data obtained under passive observation. Apart from the observational distribution, an SCM further models changes in the data generating mechanism via the concept of interventions.

1.1.3. Interventions

An intervention in an SCM is a formal abstraction of a change in mechanism occurring in a real-world process (e.g., a change in political policy). Mathematically, it is a mapping between model classes. Let $\mathcal{M}$ be a fixed class of SCMs over $(X_1, \ldots, X_p)$. An *intervention* is a mapping from $\mathcal{M}$ into a possibly larger set of SCMs, which takes as input a model $M = (\mathcal{S}, Q) \in \mathcal{M}$, and outputs another model $M(i) = (\mathcal{S}^i, Q^i)$ over $(X_1, \ldots, X_p)$, the *intervened model*. We require all interventions to preserve the joint independence of the error variables as well as the acyclicity of the causal graph. The latter restriction ensures that each intervened model $M(i)$ induces a unique distribution over $(X_1, \ldots, X_p)$, the *intervention distribution*. A variable whose structural assignment or noise distribution is altered by an intervention is said to having been *intervened on*. As such, an intervention may modify the structural assignment of a variable in a way which depends on the input model itself. An example of such a modification is a so-called *shift-intervention*, which shifts the structural assignment of a variable, say X_j, by a constant $c \in \mathbb{R}$. That is, if $X_j := f_j(\mathrm{PA}_j, \varepsilon_j)$ is the structural assignment in the original model, then $X_j := f_j(\mathrm{PA}_j, \varepsilon_j) + c$ is the structural assignment in the intervened model (and ε_j has the same distribution in both models). An intervention may also completely break the structural dependence of X_j on its causal parents by assigning it to an independent noise variable (X_j is being 'randomized'). In cases where this noise variable is degenerate, say the intervention assigns $X_j := x_j$ for some $x_j \in \mathbb{R}$, we simply write $M(x_j)$ to denote the intervened model (assuming no other parts of the model are altered by the intervention).

In Chapter 3, we return to a more detailed discussion on differ-

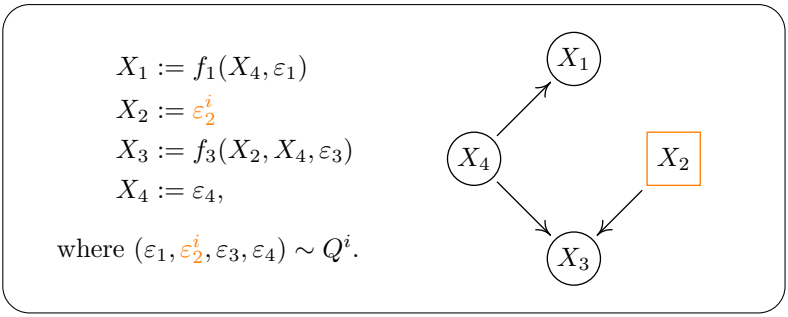


FIGURE 1.2. An SCM obtained from the model in Figure 1.1 by intervening on X_2. This model induces a (potentially) different distribution and a different causal graph.

ent types of interventions. Here, we focus on the key idea. Figure 1.2 shows the SCM from Figure 1.1 after randomizing X_2. In the intervened model, X_2 has no causal parents, and is statistically independent of X_1.

1.1.4. Autonomy

A fundamental assumption that is often made about real-world physical processes is the assumption of independent mechanisms. It states that *the causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other* [Peters et al., 2017, Principle 2.1]. Consider, for instance, the generative process of a simple cause-effect pair (X, Y). Let A be the generating mechanism for the cause X, and let B be the mechanism which produces Y from X. Then, in general, the output of A and the output of B are statistically dependent. The assumption of independent mechanisms states that the mechanism A itself, that is, the means by which A operates, carries no information about B. In other words, assuming that it is possible to replace A by a different mechanism A', we expect that, after such an operation, mechanism B would still be in place.

The above assumption is embedded in our definition of interven-

tions, which allows for the isolated modification of a single structural assignment, while leaving the remaining parts of the model unaffected. Since each assignment defines the conditional distribution of a variable given its causal parents, this property induces a factorization of the entailed distribution into several autonomous units. We illustrate this fact using the model M from Figure 1.1. Consider two interventions i_1 and i_2 which randomize X_1 and X_2, respectively. For simplicity, assume that the distributions entailed by the models M, $M(i_1)$ and $M(i_2)$ have densities with respect to a product measure, and let p, p^{i_1} and p^{i_2} denote such densities. It follows by straight-forward computations, that these densities satisfy the factorizations

$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= p(x_1 \mid x_4)p(x_2 \mid x_1)p(x_3 \mid x_2, x_4)p(x_4) \\ p^{i_1}(x_1, x_2, x_3, x_4) &= p^{i_1}(x_1) \quad p(x_2 \mid x_1)p(x_3 \mid x_2, x_4)p(x_4) \\ p^{i_2}(x_1, x_2, x_3, x_4) &= p(x_1 \mid x_4)p^{i_2}(x_2) \quad p(x_3 \mid x_2, x_4)p(x_4), \end{aligned}$$

where each equality holds for all (x_1, x_2, x_3, x_4) outside a nullset of the respective distribution. Although the joint distribution may be altered by each of the interventions, it differs from the original distribution only by the factor related to the intervened variable. The conditionals of each non-intervened variable given its direct causes remains unaffected. While this property is simply a consequence of the way we have defined interventions in SCMs, it can serve as a starting point for causal formalism in situations where SCMs are not applicable. We revisit this discussion in Section 1.4.

1.2. Causal learning

Causal models can be used to formally define and quantify causal relationships among a system of variables. In most cases, the true causal model is not fully known, and needs to be learned from data. Causal learning can be tackled effectively by performing interventions in the system of interest. This is the rationale behind controlled randomized trials [Peirce, 1883], which remove the influence of confounders via randomized treatment allocation, and are often considered the gold standard for inferring intervention effects. In

many situations, however, it is infeasible to embed the system of interest within a controlled study design, and causal knowledge thus needs to be inferred from observational data (or naturally occurring interventional data). Such cases will be the focus of this section. We distinguish between the learning of causal *relations* (e.g., parts of the causal graph) and the learning of causal *effects* (e.g., the expectation of Y under interventions on X). These topics are treated in Sections 1.2.2 and 1.2.3, respectively. First, we discuss some of the fundamental difficulties associated with causal learning.

1.2.1. A three-fold inference problem

In classical statistical learning, one aims to infer properties of some unknown (observational) distribution based on i.i.d. replications from it. Given access to only finitely many data, this can be a challenging task in and of itself. In causal learning, we are faced with two additional layers of difficulty. First, there is the problem of identifiability. We are trying to infer *causal* knowledge, which often involves targets of inference that are defined in terms of intervention distributions or in terms of the causal graph. In cases where the system of interest is only passively observed, this poses a fundamental challenge: typically, several different causal mechanisms can lead to the same observed behavior. Without suitable constraints on the underlying causal model class, the inferential target can thus remain unknown, even in the limit of infinitely many data. Second, there is the problem of learnability. Even in cases where the inferential target is identified from the observational distribution, it may not be obvious how to compute it in a finite number of computational steps. In many causal learning tasks, the definition of an identifiable causal target is therefore followed by the construction of an algorithmic procedure (a 'population algorithm') which maps the observational distribution to the target of inference. After formulating this target as a computable property of the observational distribution, we are, thirdly, left with the well-known statistical problems of estimating distributional features from finite samples of data (i.e., the construction of a corresponding 'sample algorithm'). In what follows, we will mostly discuss causal learning problems on population level.

1.2.2. Learning causal relations

It is well-known that, without suitable restrictions on the model class, the causal graph is not identified from the observational distribution. Nevertheless, the problem of causal structure learning from observational data has been addressed in several lines of work. Among them are constraint-based methods, score-based methods, methods based on restricted SCMs, and methods based on the independence of causal mechanisms. A detailed overview of recent methodological advancements is provided by Spirtes and Zhang [2016], Guo et al. [2018]. Below, we treat constraint-based methods in more detail, and discuss the difficulties that can arise from the existence of hidden variables. We then introduce the novel causal discovery method ICPH, which will be the content of Chapter 2.

1.2.2.1. Constraint-based causal discovery

Constraint-based methods aim to learn a set of graphs that are compatible with the conditional independencies embedded in the data. They make use of two central assumptions, which relate properties of the unknown graph to properties of the observed distribution. In the following two definitions, $\mathbb{P}$ denotes a generic distribution over a random vector $\mathbb{X}$, and $\mathcal{G}$ a generic graph over the variables in $\mathbb{X}$. By slight abuse of notation, we sometimes treat $\mathbb{X}$ as a set.

Definition 1.2 (Markov property)**.** $\mathbb{P}$ *is said to satisfy the* Markov property*, or to be* Markovian*, with respect to* $\mathcal{G}$*, if for all disjoints sets* $A, B, C \subseteq \mathbb{X}$*, it holds that*

$$A \perp\!\!\!\perp_{\mathcal{G}} B \,|\, C \quad \Rightarrow \quad A \perp\!\!\!\perp_{\mathbb{P}} B \,|\, C.$$

The above condition is sometimes referred to as the *global* Markov property, as opposed to its local or pairwise version. Here, $\perp\!\!\!\perp_{\mathcal{G}}$ denotes a graphical independence relation among sets of vertices in $\mathcal{G}$, and $\perp\!\!\!\perp_{\mathbb{P}}$ denotes conditional independence in $\mathbb{P}$. The Markov property states that certain graphical relations in $\mathcal{G}$ imply properties about the observational distribution $\mathbb{P}$. In order to exploit the Markov property for structure learning, we also require the reverse implication, which is known as the *faithfulness condition*.

Definition 1.3 (Faithfulness)**.** $\mathbb{P}$ *is said to satisfy* faithfulness*, or to be* faithful*, with respect to* $\mathcal{G}$*, if for all disjoints sets* $A, B, C \subseteq \mathbb{X}$*, it holds that*

$$A \perp\!\!\!\perp_{\mathbb{P}} B \,|\, C \quad \Rightarrow \quad A \perp\!\!\!\perp_{\mathcal{G}} B \,|\, C.$$

Together, Definitions 1.2 and 1.3 define a one-to-one correspondence between the graphical independence relations in $\mathcal{G}$ and the conditional independence patterns in $\mathbb{P}$. If $\mathbb{P}$ is both Markovian and faithful with respect to $\mathcal{G}$, then $\mathbb{P}$ identifies the Markov equivalence class of $\mathcal{G}$, which consists of all graphs $\tilde{\mathcal{G}}$ which agree with $\mathcal{G}$ on the set of graphical independence relations they imply. The objective of constraint-based methods is to infer this equivalence class from observational data.

Obviously, the Markov equivalence class of a graph depends on the graphical independence relation $\perp\!\!\!\perp_{\mathcal{G}}$. In the case where all variables $\mathbb{X} = (X_1, \ldots, X_p)$ are observed, the inferential target is the true causal DAG, and $\perp\!\!\!\perp_{\mathcal{G}}$ typically denotes d-separation in $\mathcal{G}$ [Pearl, 2009]. With this graphical independence relation, it can be shown that the Markov property is satisfied if $(\mathbb{P}, \mathcal{G})$ are induced by an SCM [Pearl, 2009, Theorem 1.4.1]. Further, any two Markov equivalent graphs have the same skeleton [Verma and Pearl, 1988]. The Markov equivalence class $[\mathcal{G}]_d$ of $\mathcal{G}$ can therefore be conveniently represented in terms of a *completed partially directed acyclic graph* (CPDAG), which may contain both directed and undirected ($-$) edges. A directed edge $X_i \to X_j$ in this graph is an edge common to all DAGs in $[\mathcal{G}]_d$. Likewise, the absence of an edge between X_i and X_j means that these variables are not connected by an edge in any DAG from $[\mathcal{G}]_d$. Both these cases therefore correspond to a direct causal relation (or the absence of one) which, under the assumption of faithfulness, is identified from $\mathbb{P}$. An undirected edge $X_i - X_j$ indicates that both orientations $X_i \to X_j$ and $X_i \leftarrow X_j$ can be found among some DAGs in $[\mathcal{G}]_d$, and hence corresponds to a direct causal relation which cannot be inferred via the above procedure.

For the graph $\mathcal{G}_M$ in our running example, the only two d-separation statements that hold true are

$$\{X_2\} \perp\!\!\!\perp_{\mathcal{G}_M} \{X_4\} \,|\, \{X_1\} \quad \text{and} \quad \{X_1\} \perp\!\!\!\perp_{\mathcal{G}_M} \{X_3\} \,|\, \{X_2, X_4\}. \tag{1.2.1}$$

This means that $\mathbb{P}_M$ is Markovian and faithful with respect to $\mathcal{G}_M$ if

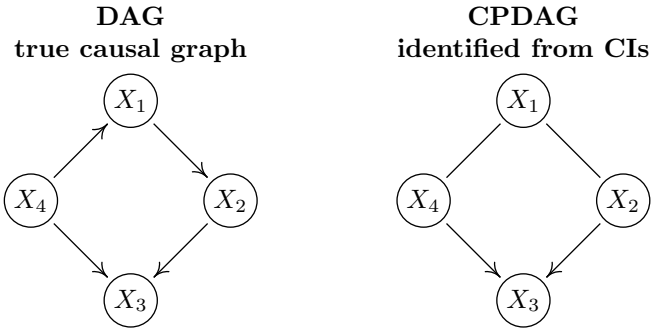


FIGURE 1.3. True causal graph $\mathcal{G}_M$ (left), and the corresponding Markov equivalence class $[\mathcal{G}_M]_d$ represented in terms of a CPDAG (right). If $\mathbb{P}_M$ is Markovian and faithful with respect to $\mathcal{G}_M$, the Markov equivalence class of $\mathcal{G}_M$ can be computed by checking for conditional independencies ('CIs') in $\mathbb{P}_M$. In this example, $\mathbb{P}_M$ identifies the two directed edges $X_4 \to X_3$ and $X_2 \to X_3$. For the edges $X_4 - X_1$ and $X_1 - X_2$, both orientations are compatible with the conditional independencies in $\mathbb{P}_M$.

and only if the following (and only those) conditional independence statements are satisfied:

$$X_2 \perp\!\!\!\perp_{\mathbb{P}_M} X_4 \,|\, X_1 \quad \text{and} \quad X_1 \perp\!\!\!\perp_{\mathbb{P}_M} X_3 \,|\, (X_2, X_4). \tag{1.2.2}$$

Figure 1.3 shows the true causal graph $\mathcal{G}_M$ alongside with the associated CPDAG. Here, all graphs in $[\mathcal{G}_M]_d$ agree on the oriented edges $X_4 \to X_3$ and $X_2 \to X_3$, while the causal relations $X_4 - X_1$ and $X_1 - X_2$ remain unidentified.

1.2.2.2. Hidden variables

If some of the variables in $\mathbb{X}$ remain unobserved, there may not exist a DAG over the observed variables $\mathbb{X}^* \subseteq \mathbb{X}$ which correctly represents their conditional independence relations. In such cases, Richardson et al. [2002] propose the use of an extended space of graphs called maximal ancestral graphs (MAGs), which, in addition to directed and undirected edges, also may include bidirected ($\leftrightarrow$)

edges. Here, we do not formally define MAGs, and rather focus on the overall idea behind their use for causal discovery. Unlike in CPDAGs, an undirected edge in a MAG does not reflect uncertainty about the true edge orientation, but rather represents a statistical dependence arising from an implicit conditioning variable. Here, we assume that no conditioning variables exist, and therefore disregard such edges from the below discussion about the causal interpretation of MAGs.

Every DAG $\mathcal{G}$ over the full set of variables $\mathbb{X}$ can be transformed into a unique MAG $\mathcal{G}^*$ over the observed variables $\mathbb{X}^*$ by 'marginalizing out' the latent variables [Richardson et al., 2002, p. 981]. This MAG then serves as the causal inferential target. It encodes ancestral causal information about the underlying DAG in the following sense. Whenever $X_i \to X_j$ in $\mathcal{G}^*$, then X_i is an ancestor of X_j in $\mathcal{G}$, and whenever $X_i \leftarrow X_j$ or $X_i \leftrightarrow X_j$ in $\mathcal{G}^*$, then X_i is *not* an ancestor of X_j in $\mathcal{G}$. In words, every arrow tail indicates a cause and every arrow head indicates a non-cause. Further, $\mathcal{G}^*$ encodes, via a graphical independence relation called m-separation, the same graphical independence statements that hold true among the observed variables in $\mathcal{G}$. That is, for any disjoint subsets $A, B, C \subseteq \mathbb{X}^*$, it holds that

$$A \perp\!\!\!\perp^m_{\mathcal{G}^*} B \,|\, C \quad \Leftrightarrow \quad A \perp\!\!\!\perp^d_{\mathcal{G}} B \,|\, C, \tag{1.2.3}$$

where $\perp\!\!\!\perp^m_{\mathcal{G}^*}$ and $\perp\!\!\!\perp^d_{\mathcal{G}}$ refer to m-separation in $\mathcal{G}^*$ and d-separation in $\mathcal{G}$, respectively [Richardson et al., 2002, Theorem 4.18]. Consequently, any distribution $\mathbb{P}$ over $\mathbb{X}$ that is Markovian and faithful w.r.t. $\mathcal{G}$ (in terms of d-separation) induces a marginal $\mathbb{P}^*$ over $\mathbb{X}^*$ which is Markovian and faithful w.r.t. $\mathcal{G}^*$ (in terms of m-separation). In such a case, $\mathbb{P}^*$ identifies the Markov equivalence class $[\mathcal{G}^*]_m$ of $\mathcal{G}^*$. This equivalence class can be represented in terms of a *partial ancestral graph* (PAG), which may contain directed, undirected, bidirected, nondirected ($\circ\!\!-\!\!\circ$), partially undirected ($\circ\!\!-$) and partially directed ($\circ\!\!\to$) edges. Every non-circle mark in a PAG is a mark common to all MAGs in the associated equivalence class $[\mathcal{G}^*]_m$, and therefore corresponds to an ancestral causal relation which is identified from $\mathbb{P}^*$. If every circle mark in a PAG corresponds to a mark that is *not* common to all MAGs in $[\mathcal{G}^*]_m$, then this PAG is called *maximally informative*. This PAG represents the maximal amount of

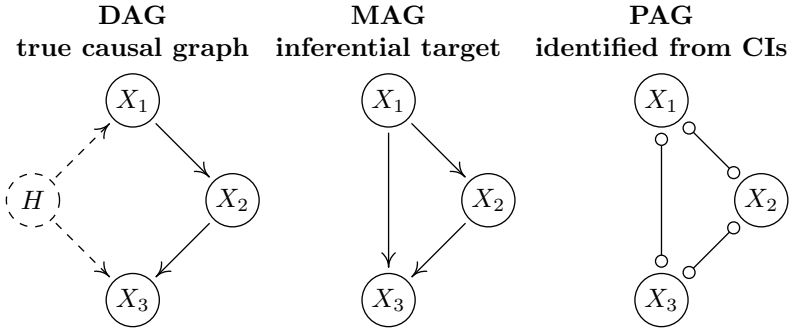


FIGURE 1.4. Left: true causal graph $\mathcal{G}_M$. Since no d-separation holds true among the observed variables in $\mathcal{G}_M$, no fully connected MAG over $\{X_1, X_2, X_3\}$ can be rejected as a candidate for the true inferential target $\mathcal{G}_M^*$ (middle). This results in an uninformative Markov equivalence class $[\mathcal{G}_M^*]_m$ with an associated maximally informative PAG that contains only circle edge marks (right). In this example, it is thus not possible to infer any ancestral causal relations among $\{X_1, X_2, X_3\}$ solely based on conditional independencies.

ancestral causal information identifiable from $\mathbb{P}^*$ via m-separation.

Let us return to our running example. Suppose that we do not observe the variable X_4. This situation is depicted in Figure 1.4 (left), where we have replaced X_4 by H (for 'hidden'). The corresponding MAG $\mathcal{G}_M^*$ (Figure 1.4 middle) represents ancestral causal relations among the observed variables in $\mathcal{G}_M$. For example, the edge $X_1 \to X_3$ in $\mathcal{G}_M^*$ says that X_1 is an ancestor of X_3 in $\mathcal{G}_M$ and that X_3 is not an ancestor of X_1 in $\mathcal{G}_M$. In this example, no d-separation statement holds true among the observed variables $\{X_1, X_2, X_3\}$ in $\mathcal{G}_M$, implying the non-existence of any m-separations in $\mathcal{G}_M^*$. In the equivalence class $[\mathcal{G}_M^*]_m$, no edge mark is common to all MAGs, resulting in a maximally informative PAG that contains no ancestral causal information (Figure 1.4 right).

1.2.2.3. Invariant causal prediction

We now briefly introduce the causal discovery method which will be the subject of study in Chapter 2. It is based on the principle of invariant causal prediction first proposed by Peters et al. [2016], and extends the existing methodology to a setting with hidden variables. Rather than aiming at learning all of the causal structure, our method tries to infer the set of (observable) causal parents of a certain target variable Y in cases where some of the direct causes remain unobserved. It makes a simplicity assumption on the hidden variables and exploits the existence of an exogenous 'environmental variable' E. Formally, our method does not rely on data being generated by an SCM. In particular, the environmental 'variable' is not assumed to be drawn from a probability distribution, and may simply be an index describing different experimental settings or conditions ('environments') under which the data were generated. To match up with the rest of this chapter, we here formulate our method in terms of an SCM.

Consider an SCM over variables (X, Y, H, E), where $X \in \mathbb{R}^d$ is a vector of observed covariates, Y is a real-valued target variable of interest, H is a discrete unobserved direct cause of Y, say $H \in \{1, \dots, \ell\}$, for some (small) integer $\ell \geq 2$, and E is a real-valued variable that is known to be a source node in the causal graph over (X, Y, H, E). Our method then aims to infer the set $S^* \subseteq \{1, \dots, d\}$ of causal parents of Y among the observed covariates X (here, we identify each predictor X_j by its index $j \in \{1, \dots, d\}$). The variable E is allowed to directly influence all parts of the system except for the target variable Y itself. Under this assumption, the set S^* satisfies the following key property. For all x, e, it holds that

$$\begin{aligned}\mathbb{P}_{Y|(X_{S^*}=x,E=e)} &= \sum_{j=1}^{\ell} \mathbb{P}_{Y|(X_{S^*}=x,H=j,E=e)} \mathbb{P}(H = j \,|\, X_{S^*} = x, E = e) \\ &= \sum_{j=1}^{\ell} \mathbb{P}_{Y|(X_{S^*}=x,H=j)} \mathbb{P}(H = j \,|\, X_{S^*} = x, E = e),\end{aligned} \tag{1.2.4}$$

where the second equality follows from the conditional independence $Y \perp\!\!\!\perp E \,|\, (X_{S^*}, H)$. That is, the distribution of $Y \,|\, (X_{S^*} = x, E = e)$

can be expressed as an ℓ-fold mixture of the distributions $Y \,|\, (X_{S^*} = x, H = j)$, $j \in \{1, \dots, \ell\}$, each of which does not depend on the value e of the environmental variable E. This property can be exploited for causal discovery. The population version of our proposed algorithm ICPH ('invariant causal prediction in the presence of hidden variables') proceeds as follows: (1) run through all possible subsets of predictors $S \subseteq \{1, \dots, d\}$ and check whether (1.2.4) holds true for $S^* = S$, and (2) output the set

$$\tilde{S} := \bigcap_{S \text{ satisfies (1.2.4)}} S \tag{1.2.5}$$

of predictors necessary for (1.2.4) to hold.[1] We construct an estimator $\hat{S}$ of $\tilde{S}$ by taking the intersection over all sets $S \subseteq \{1, \dots, d\}$ which pass a statistical test for the hypothesis of (1.2.4) being satisfied. Our main theoretical result is the asymptotic false discovery control of our method. In words, it says that, as the number of data points tends to infinity, the estimator $\hat{S}$ is, with controllable large probability, a subset of the true set S^* of observable causal parents.

Apart from obtaining provable false discovery control, our method is applicable in situations where purely constraint-based methods cannot be expected to work. To see this, consider once again our running example. Let $Y = X_3$ be the target variable of interest, and assume that $H = X_4$ is discrete-valued and unobserved. Assume further that the variable E directly affects X_1, X_2 and H. The causal graph for this system of variables can be seen in Figure 1.5 (left), where the inferential target $S^* = \{2\}$ is indicated in green. Assuming faithfulness (this assumption is not necessary for the above false discovery control to hold true), the only sets satisfying (1.2.4) are $\{2\}$ and $\{1, 2\}$. In the population case, our estimator therefore correctly infers $\tilde{S} = \{2\}$. It is not possible to make the same inference solely based on conditional independencies, see Figure 1.5 (right).

1.2.3. Learning causal effects

Often, we are interested not only in learning the existence of causal relations (such as 'X causes Y'), but also in quantifying the strength

[1] We define the intersection over an empty index set to be the empty set.

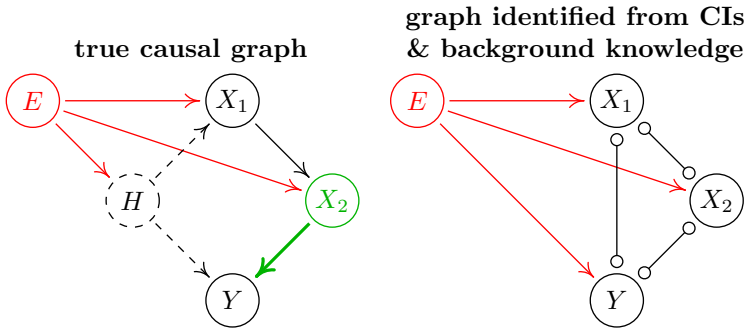


FIGURE 1.5. Left: true causal graph for the SCM in Figure 1.1, with an additional exogenous variable E. Our method tries to infer set $S^* = \{2\}$ of observable causal parents of Y (green). In this example, under the assumption of faithfulness, the population version (1.2.5) of our estimator correctly identifies S^*. Right: graph representing the maximal causal information that can be obtained via m-separation. This graph is constructed in three steps: (1) compute the Markov equivalence class of the MAG associated with the true causal DAG, (2) disregard all MAGs not compatible with E being a source node (i.e., all MAGs containing edges of the form $E - \bullet$ or $E \leftarrow \bullet$ for some generic node $\bullet$), and (3) for every pair of observed variables, use the same edge marks as for PAGs to represent the variation of the corresponding edge mark among all plausible MAGs. In this example, no causal relations among the variables (X_1, X_2, Y) can be identified via this procedure.

of these relations. For example, we may be interested in assessing the expected value of Y under a range of different interventions on X. In other words, we want to learn properties of intervention distributions. This task requires causal background knowledge. In general, however, full identification of the causal structure is not necessary for computing intervention effects. E.g., it suffices to have access to a suitable 'adjustment set', a definition of which we give below. We further discuss the influence of hidden variables, and introduce the instrumental variables approach, which plays a central role in Chapter 3.

1.2.3.1. Covariate adjustment

Some intervention distributions can be directly expressed as computable functionals of the observational distribution and (parts of) the causal graph. Formulas describing such functional relationships are called 'adjustment formulas' and exist in various forms. Below, we give an example of an adjustment formula due to Pearl [2009]. For simplicity, we assume the existence of densities, which we denote by p.

Proposition 1.1 (Backdoor adjustment)**.** *Consider an SCM M over a set of variables $\mathbb{X}$, and let $X, Y \in \mathbb{X}$. Assume that Y is not a causal parent of any of the variables from X. Let $Z \subseteq \mathbb{X} \setminus \{X, Y\}$ be a set of variables satisfying that (i) Z contains no descendant of X, and (ii) Z blocks all paths from X to Y that are of the form $X \leftarrow \cdots Y$ (they enter X 'through the backdoor'). Then, for all x, y, it holds that,*

$$p_{M(x)}(y) = \int_z p_M(y \,|\, x, z)\, p_M(z) dz, \tag{1.2.6}$$

i.e., the distribution of Y in the intervened model $M(x)$ can be computed from the observational distribution over (X, Y, Z).

A set Z which satisfies Equation (1.2.6) is called a 'valid adjustment set' for the pair (X, Y). A proof of Proposition 1.1 can be found in [Peters et al., 2017, Proposition 6.41], for example. Here, we refrain from formally defining the notion of path blocking, and rather illustrate the above result using our running example. Assume that we are interested in computing the intervention distribution $\mathbb{P}^{X_3}_{M(x_2)}$, for some $x_2 \in \mathbb{R}$. As seen in Figure 1.6, this distribution can be obtained by adjusting for X_1, for X_4, or for both of these variables.

1.2.3.2. Hidden variables

If relevant parts of the system remain unobserved, there may not exist a valid adjustment set among the set of observed variables. This poses a major statistical challenge, since in general, it is impossible to distinguish between the statistical dependencies originating from

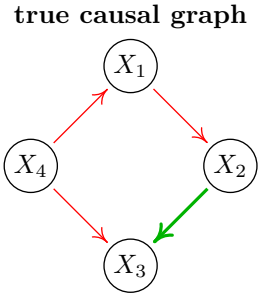


FIGURE 1.6. True causal graph (left), and valid adjustment sets for the pair (X_2, X_3) (right). The only backdoor path from X_2 to X_3 is the path $X_2 \leftarrow X_1 \leftarrow X_4 \rightarrow X_3$ (red), which is blocked by each of the displayed sets of nodes. Any such set therefore serves as a valid adjustment set for computing the intervention distribution $\mathbb{P}^{X_3}_{M(x_2)}$, see Proposition 1.1.

causal relations, and those induced by hidden confounders. To illustrate this point, assume that the model class $\mathcal{M}$ in our running example corresponds to the set of SCMs with linear structural assignments and zero-mean Gaussian noise variables. For simplicity, assume that in the true model M from Figure 1.1, all linear coefficients are equal to one, and all noise variables have unit variance. Say we are interested in the causal coefficient $\beta := \frac{d}{dx_2}\mathbb{E}_{M(x_2)}[X_3] = 1$, but we do not have access to the variables X_1 and X_4, i.e., no valid adjustment set exists for the pair (X_2, X_3). This situation is illustrated in Figure 1.7 (left). The figure additionally displays two alternative models from $\mathcal{M}$ which agree with M on the causal structure as well as on the entailed distribution over (X_2, X_3), but which induce different causal effects $X_2 \rightarrow X_3$. Even if the true causal structure of M is known, β can thus not be identified from the observational distribution over (X_2, X_3). This problem can sometimes be remedied by the use of instrumental variables.

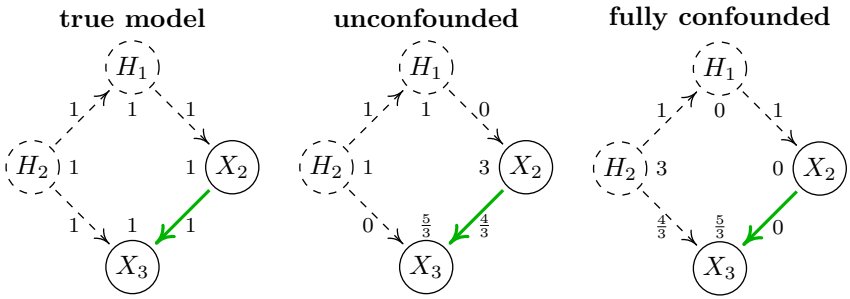


FIGURE 1.7. True causal model M (left) and two alternative models (middle and right) which agree with M on the causal structure as well as on the observational distribution over (X_2, X_3). Numbers next to arrows and nodes indicate linear coefficients and error variances, respectively, in the corresponding structural assignments. The noise variables in the two alternative models have been chosen carefully to induce the correct observational distribution. Even given access to the full observational distribution over (X_2, X_3) along with the true causal graph, it is impossible to infer the strength of the causal relation $X_2 \to X_3$.

1.2.3.3. Instrumental variables

The instrumental variables method [e.g., Theil, 1953, Fuller, 1977] is a popular approach for adjusting for latent confounders. It exploits the existence of observed exogenous variables, which are facilitated as 'instruments' for inferring causal effects. Although the method does not rely on data being generated by an SCM, we here make this assumption to employ the causal formalism introduced in Section 1.1. Consider an SCM over a real-valued response $Y \in \mathbb{R}$, a vector of observed covariates $X \in \mathbb{R}^d$, additional observed variables $A \in \mathbb{R}^r$, and hidden variables $H \in \mathbb{R}^q$. We focus on the case of a linear causal influence $X \to Y$. Assume that there exists a noise variable $\varepsilon_Y \perp\!\!\!\perp (X, H)$, a measurable function $h : \mathbb{R}^{q+1} \to \mathbb{R}$, and a vector of coefficients $\beta \in \mathbb{R}^d$, such that the structural assignment for Y is given as

$$Y := X^\top \beta + h(H, \varepsilon_Y),$$

where $\xi_Y := h(H, \varepsilon_Y)$ has mean zero.[2] We do not state the remaining structural assignments, but we explicitly allow for the hidden variables H to enter the assignments for X. As highlighted in the previous section, in such a case, the causal coefficient β is generally not identified from the observational distribution over (X, Y). This is where the variables A come into play. Under the assumption that (i) $\mathbb{E}[A\xi_Y] = 0$ and (ii) $\mathbb{E}[AA^\top]$, $\mathbb{E}[AX^\top]$ are all of full column rank, β is uniquely determined by the observational distribution over (X, Y, A). Indeed, under (i) and (ii), we have the well-defined expression

$$\begin{aligned}
&\left(\mathbb{E}[XA^\top]\mathbb{E}[AA^\top]\mathbb{E}[AX^\top]\right)^{-1}\mathbb{E}[XA^\top]\mathbb{E}[AA^\top]\mathbb{E}[AY] \\
&= \left(\mathbb{E}[XA^\top]\mathbb{E}[AA^\top]\mathbb{E}[AX^\top]\right)^{-1}\mathbb{E}[XA^\top]\mathbb{E}[AA^\top]\mathbb{E}[AX^\top]\beta \\
&\quad + \left(\mathbb{E}[XA^\top]\mathbb{E}[AA^\top]\mathbb{E}[AX^\top]\right)^{-1}\mathbb{E}[XA^\top]\mathbb{E}[AA^\top]\mathbb{E}[A\xi_Y] \\
&= \beta,
\end{aligned}$$

where the left-most side is a function only of the observational distribution over (X, Y, A). If (i) and (ii) are satisfied, the variables in A are said to be *instruments* for (X, Y). In the terminology from Section 1.2.1, the above equation defines a population algorithm for inferring β. A consistent estimator for β (a sample algorithm) can be obtained by substituting all expectations by their empirical counterparts, resulting in the so-called two stage least squares estimator (TSLS) [Theil, 1953].

The instrumental variables method has been studied for several decades, and is particularly prominent in the econometrics literature, where several alternatives to the TSLS have been proposed [e.g., Anderson and Rubin, 1949, Theil, 1958, Fuller, 1977]. Most of these methods assume a linear causal relationship $X \to Y$, as is done in the formulation above. The question arises what can be done in the nonlinear case, i.e., where the structural assignment for Y is given by $Y := f(X) + h(H, \varepsilon_Y)$ for some nonlinear function f. If f belongs to some known parametric class of C^2 functions, identifiability can be ensured by assuming that $(i)'$ $\mathbb{E}[\xi_Y \mid A] = 0$ together with a generalized version of the rank conditions in (ii), which we

[2] This can be assumed w.l.o.g. by including an intercept into X.

make explicit in Appendix B.2. These conditions are inspired by the work of Amemiya [1974], Jorgenson and Laffont [1974], Kelejian [1971], who study consistency of estimators in such function classes.

Unless precise background knowledge is available, it is usually hard to justify the membership of the true causal function f to some priorly specified (low-dimensional) nonlinear function class. We therefore require procedures with data-driven model complexities. While a plethora of such methods exists for standard regression problems, only few are applicable in the instrumental variables setting. Moreover, most existing methods, such as the one proposed by Racine and Hayfield [2018], focus on in-sample estimation of the causal function, and cannot be expected to provide causal predictions for values of X that lie outside the range of the training data. Extrapolating estimates outside of this domain requires additional assumptions on the causal function class. As part of our contribution in Chapter 3, we propose a novel nonlinear instrumental variables estimator called the NILE ('Nonlinear Intervention-robust Linear Extrapolator'). It achieves flexible in-sample estimation via a penalized B-spline approach, and exploits a linear extrapolation assumption on f to obtain out-of-sample estimates.

1.3. Distributional robustness

The NILE estimator inherits its name from the well-known fact that a prediction model which uses the true causal relationship f to predict Y from X remains valid under arbitrary interventions on the covariates (i.e., it is 'robust' w.r.t. such interventions). This result may be seen as a special case of distributional robustness, which describes the property of a regression model of obtaining predictive guarantees across a range of test distributions that differ from the training distribution. This property is of particular interest in situations where we require predictions for Y under unprecedented circumstances, e.g., when spatially or temporally extrapolating climate models, when applying learned marketing strategies to a newly evolving industry, or in reinforcement learning tasks, when transferring knowledge between two different episodes of an arcade game. In all such cases, the test distribution for (X, Y) may differ from its

training distribution.

The problem of finding distributionally robust regression models has been well-studied, and is known in different variants under a range of different names, e.g., domain generalization, out-of-distribution prediction or covariate shift; see Section 3.1.2 for a detailed list of references. The general objective is to minimize the worst-case prediction risk $\sup_{\tilde{\mathbb{P}} \in \mathcal{N}(\mathbb{P})} \mathbb{E}_{\tilde{\mathbb{P}}}[(Y - f_\diamond(X))^2]$ across all test distributions $\tilde{\mathbb{P}}$ in some suitable neighborhood $\mathcal{N}(\mathbb{P})$ of the training distribution $\mathbb{P}$ over (X, Y). This neighborhood is often taken to be a ball around the training distribution with respect to some suitable metric, e.g., the Wasserstein metric.

In Chapter 3, we consider test distributions that arise from interventions in a causal model; an approach that is motivated by the idea that distributional changes may have causal explanations. A set of test distributions induced in such a way may differ substantially from any neighborhood that can be realized as a ball with respect to a commonly used metric, and arguably, in some scenarios, describes a more realistic class of distributions to be encountered in practice. We consider a class of SCMs $\mathcal{M}$ which, in addition to X and Y, allow for the existence of exogenous variables $A \in \mathbb{R}^r$ and hidden variables $H \in \mathbb{R}^q$. Interventions may occur on either X or A. If $\mathcal{I}$ is a set of such interventions, $M \in \mathcal{M}$ is the true data generating model, and $\mathcal{F}$ is the class of permitted prediction functions, then we consider the minimax problem

$$\arg\min_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2]. \tag{1.3.1}$$

A solution to this problem is called a *minimax solution*. We prove that, under a large class of interventional settings, the causal prediction function is a minimax solution, and that any alternative prediction function is not robust to misspecifications of the intervention set $\mathcal{I}$.

Identifying minimax solutions from observational data is closely related to the overall inference challenges discussed in Section 1.2.1: if $\psi(M)$ is the set of solutions to (1.3.1), then, based on $\mathbb{P}_M$, it is only possible to identify the class

$$\left\{ \psi(\tilde{M}) \,:\, \tilde{M} \in \mathcal{M} \text{ and } \mathbb{P}_{\tilde{M}} = \mathbb{P}_M \right\}$$

of plausible solution sets. This motivates the definition of distribution generalization, which requires the existence of a function f^* that is a member of $\psi(\tilde{M})$ for all $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$.[3] We provide sufficient conditions on $\mathbb{P}_M$, $\mathcal{I}$ and $\mathcal{M}$ which allow for distribution generalization, and present corresponding impossibility results proving the necessity of some of these conditions.

1.4. Modeling real data

So far, we have sojourned in the idealized realm of mathematical models. We used the framework of SCMs to formally quantify causal relationships among random variables, and discussed various techniques for estimating such relationships from i.i.d. observational or partly intervened data. In reality, true i.i.d. replications are hard to find. Any two distinct phenomena occur under distinct conditions, and different observations must, in principle, be regarded as realizations from different distributions. Often, these differences are negligible enough to justify an i.i.d. assumption. If the data consist of sufficiently heterogeneous or dependent measurements, however, it may be more sensible to regard them as a single outcome of some joint distribution over all observations in the data set.

Figure 1.8 (right) shows a real-world spatial data set $\mathbf{Z}_{\mathcal{S}} = (Z_s)_{s\in\mathcal{S}}$ containing multivariate measurements from a vector of variables $Z = (Z^1, Z^2, Z^3, Z^4) \in \mathbb{R}^4$ observed at several spatial locations $s \in \mathcal{S} \subseteq \mathbb{R}^2$. (We here use Z rather than X to align with the notation used in Chapter 4.) When modeling these data with a classical SCM over (Z^1, Z^2, Z^3, Z^4), e.g., the one from Figure 1.1, we make the implicit assumption that there are no causal or statistical relations between any observations obtained from different locations. That is, we assume a causal structure as shown in Figure 1.8 (left), where each of the five graphs corresponds to a different spatial location. As visually indicated by the strong spatial autocorrelation patterns in the data (Figure 1.8 right), this does not seem like a sensible assumption. To allow for spatial dependencies, we

[3] We only require f^* to *approximately* solve the minimax problems associated with the models $\tilde{M}$ (see Chapter 3). Here, we neglect this detail in favor of notational simplicity.

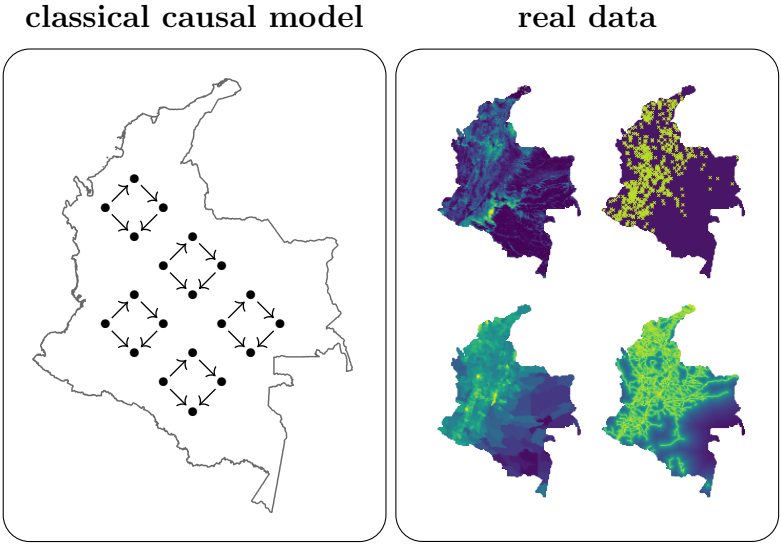


FIGURE 1.8. Inadequacy of classical causal models for modeling the causal dynamics of real-world spatial data sets. When using the SCM from Figure 1.1 to model the multivariate spatial data set shown in the right panel, we make the implicit assumption that the causal graph over all measurements decomposes into several disconnected copies of the same graph; one copy for each observed location. As becomes evident from the strong correlation structures in the data, this spatial independence assumption can hardly be justified.

need a joint causal model across all observations in the data set. SCMs are not easily applicable, since they would require the individual specification of $4 \cdot |\mathcal{S}|$ structural assignments, each of which possibly depends on $4 \cdot |\mathcal{S}| - 1$ variables. There are two main difficulties associated with this task. First, it may not be obvious how to resolve the problem of causal cycles. In a time series setting, one can ensure the existence of a well-defined observational distribution by only allowing for causal relations to be directed forward in time. Without the notion of a 'forward'-direction, this idea cannot be directly transferred to spatial data. Here, one may reasonably

assume that every variable can be causally influenced by all other variables within a suitable spatial neighborhood. This assumption directly leads to causal cycles. An SCM specified in this way may not entail an observational distribution. Second, drawing inference about an SCM based on a single observation requires homogeneity assumptions on its structural assignments. For example, one may assume that the functional dependence of each variable on its spatial neighbors remains the same across space. While effectively reducing the degrees of freedom, such an approach demands a large number of complicated boundary conditions, in particular for irregularly shaped domains like the one in Figure 1.8.

For simplicity, we have highlighted the above challenges using a purely spatial data set. In fact, the maps in Figure 1.8 are aggregations of a spatio-temporal data set related to armed conflict and tropical forest loss in Colombia, which we analyze in Chapter 4. To quantify the causal influence of conflict on forest loss, we require a class of causal models for spatio-temporal data. As part of our contribution, we develop such a class of models, formulated in terms of stochastic processes. Within this framework, a spatio-temporal data set may be viewed as a sample from an underlying causal process, observed at discrete points in space and time. Our approach is based on the principle of autonomy discussed in Section 1.1.4. We now illustrate it using a spatio-temporal version of our running example. To that end, let $\mathbf{Z}$ be spatio-temporal stochastic process with coordinate processes denoted by $\mathbf{Z}^{(j)}$, $j \in \{1, 2, 3, 4\}$. We can define causal relations among these coordinate processes by specifying that the joint density of $\mathbf{Z}$ admits the factorization[4]

$$p(\mathbf{Z}) = p(\mathbf{Z}^{(1)} \mid \mathbf{Z}^{(4)})\, p(\mathbf{Z}^{(2)} \mid \mathbf{Z}^{(1)})\, p(\mathbf{Z}^{(3)} \mid \mathbf{Z}^{(2)}, \mathbf{Z}^{(4)})\, p(\mathbf{Z}^{(4)}),$$

and that, in addition, the above factors correspond to autonomous units which allow for localized interventions. That is, we assume that an intervention i in the data generating mechanism for $\mathbf{Z}$, which intervenes on all of a certain coordinate process, say on $\mathbf{Z}^{(2)}$, changes the joint distribution of $\mathbf{Z}$ only by replacing the conditional for $\mathbf{Z}^{(2)}$ by some new distribution $p^i(\mathbf{Z}^{(2)} \mid \mathbf{Z}^{(1)})$. For a data generating pro-

[4] For simplicity, we here assume the existence of densities, but this is not formally required by our model class (see Chapter 4).

cess which follows such an interventional behavior, we may reasonably think of the above conditionals as models for causal mechanisms.

Unlike SCMs, our models only accommodate interventions on certain bundles of variables at once. As such, it allows for causal relations between variables within each bundle without modeling these explicitly. It is a useful class of models if the main objective lies in quantifying causal relations among the different coordinate processes, and if the causal dynamics within each of these processes are only of secondary interest. In Chapter 4, we consider the problem of quantifying the causal influence of a vector of covariates $X \in \mathbb{R}^d$ on a real-valued response Y in the presence of some latent confounders $H \in \mathbb{R}^\ell$. Within the above model class, this means that we are considering a stochastic process $(\mathbf{Y}, \mathbf{X}, \mathbf{H})$ whose density admits the causal factorization

$$p(\mathbf{Y}, \mathbf{X}, \mathbf{H}) = p(\mathbf{Y} \mid \mathbf{X}, \mathbf{H})\, p(\mathbf{X} \mid \mathbf{H})\, p(\mathbf{H}).$$

By imposing additional assumptions on the conditional $p(\mathbf{Y} \mid \mathbf{X}, \mathbf{H})$, our framework allows us to formally define a notion of the 'causal effect' of $\mathbf{X}$ on $\mathbf{Y}$. We show how to estimate the causal effect from observational data, and develop a procedure for testing the hypothesis of this effect being zero. Our methods exploit the assumption of $\mathbf{H}$ being time-invariant.

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

Joint work with
Jonas Peters

Abstract

Given a response Y and a vector $X = (X^1, \dots, X^d)$ 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.

2.1. Introduction

2.1.1. 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-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 2.5.2. 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 2.1.3.

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 2.1.2). 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.

2.1.2. 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 2.1. 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 2.3. Additionally, Sections 2.2.5 and 2.2.6 contain our proposed test for the equality of switching regression models that is available in our code package as the function `test.equality.sr`.

2.1.3. The principle of invariant causal prediction

This section follows the presentation provided by Pfister et al. [2019b]. Suppose that we observe several instances $(Y_1, X_1), \dots, (Y_n, X_n)$ 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 $e \subseteq \{1, \dots, n\}$, and use $\mathcal{E}$ to denote the collection of these, i.e., $\dot{\bigcup}_{e\in\mathcal{E}} e = \{1, \dots, n\}$. 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 $S^* \subseteq \{1, \dots, d\}$ 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 $S^* \subseteq \{1, \dots, d\}$, such that for all x and all $1 \leq s, t \leq n$, we have

$$Y_s \,|\, (X_s^{S^*} = x) \overset{d}{=} Y_t \,|\, (X_t^{S^*} = x), \tag{2.1.1}$$

where $X_t^{S^*}$ denotes the covariates in S^* at instance t. For simplicity, the reader may think about (2.1.1) 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.1), and different interventional settings, a sufficient condition for (2.1.1) 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 2.2.3, 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 (2.1.1) holds. We therefore make use of the environments. Given a set $S \subseteq \{1, \dots, d\}$, 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 $P^e_{Y|X^S}$, and check whether for all $e, f \in \mathcal{E}$, we have that

$$P^e_{Y|X^S} = P^f_{Y|X^S}. \tag{2.1.2}$$

In the population case, Equation (2.1.2) can be used to recover (parts of) S^* from the conditional distributions $P^e_{Y|X^S}$: for each subset $S \subseteq \{1, \dots, d\}$ of predictors we check the validity of (2.1.2) and output the set

$$\tilde{S} := \bigcap_{S \text{ satisfies (2.1.2)}} S \tag{2.1.3}$$

of variables that are necessary to obtain predictive stability. Under assumption (2.1.1), $\tilde{S}$ only contains variables from S^*. For purely observational data, i.e., $(Y_t, X_t) \overset{d}{=} (Y_s, X_s)$ for all s, t, Equation (2.1.2) is trivially satisfied for any set $S \subseteq \{1, \dots, d\}$ and thus $\tilde{S} = \emptyset$. It is 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

[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.

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 $S^* \subseteq \{1, \dots, d\}$. Since the invariance assumption (2.1.1) 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 (2.1.1). If the latent variables are constrained to take only few values, the h-invariance property can, similarly to (2.1.3), be used for the inference of S^*.

2.1.4. Organization of the paper

The remainder of the paper is organized as follows. Section 2.2 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 2.3, 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 2.4, 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 2.5 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 A.3. All our code is available as an `R` pack-

age 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.

2.2. Invariant causal prediction in the presence of latent variables

Consider a collection $(\mathbf{Y}, \mathbf{X}, \mathbf{H}) = (Y_t, X_t, H_t)_{t\in\{1,\dots,n\}}$ 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 2.3.1 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 2.1.3, 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 $S^* \subseteq \{1,\dots,d\}$ 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 2.2.1 shows why the original version of invariant causal prediction is inapplicable. In Sections 2.2.2 and 2.2.4 we introduce the formal concept of h-invariance and present an adapted version of the inference principle discussed in Section 2.1.3. In Sections 2.2.5 and 2.2.6 we then present tests for h-invariance of sets $S \subseteq \{1,\dots,d\}$, 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 2.2.3.

2.2.1. Latent variables and violation of invariance

The inference principle described in Section 2.1.3 relies on the invariance assumption (2.1.1). 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 2.1 (Violation of invariance due to latent variables)**.** *We consider a linear model for the data* $(Y_t, X_t^1, X_t^2, H_t^*)_{t\in\{1,\dots,n\}} \in \mathbb{R}^{n\times 4}$*. Assume there exist i.i.d. zero-mean noise variables* $\varepsilon_1, \dots, \varepsilon_n$ *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 homogeneous in time, 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* (2.1.1)*.*

The above example shows that in the presence of hidden variables, assumption (2.1.1) 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]

$$Y_t \,|\, (X_t^{S^*} = x, H_t^* = h) \overset{d}{=} Y_s \,|\, (X_s^{S^*} = x, H_s^* = h). \tag{2.2.1}$$

Ideally, we would like to directly exploit this property for the inference of S^*. Given a candidate set $S \subseteq \{1, \dots, d\}$, we need to check if there exist $H_1^*, \dots, H_n^*$ such that (2.2.1) holds true for $S^* = S$. Similarly to (2.1.3), 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. (2.2.1) holds for all $(x, h) \in \mathrm{supp}((X_t^{S^*}, H_t^*)) \cap \mathrm{supp}((X_s^{S^*}, H_s^*))$.

Proposition 2.1 (Necessity of constraining the influence of H^*)**.** *Let $S \subseteq \{1, \dots, d\}$ be an arbitrary subset of the predictors X_t. Then, there exist variables $H_1, \dots, H_n$ such that (2.2.1) is satisfied for $S^* = S$ and $(H_t^*)_{t \in \{1,\dots,n\}} = (H_t)_{t \in \{1,\dots,n\}}$.*

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 2.1 shows that without constraining the influence of H^*, (2.2.1) cannot be used to identify S^*. Identifiability improves, however, for univariate, discrete latent variables $H^* \in \{1, \dots, \ell\}$ with relatively few states $\ell \geq 2$. Equation (2.2.1) then translates into the following assumption on the observed conditional distributions $P_{Y_t \,|\, X_t^{S^*}}$: for all t, x it holds that

$$P_{Y_t | (X_t^{S^*} = x)} = \sum_{j=1}^{\ell} \lambda_{xt}^j P_x^j, \tag{2.2.2}$$

for some $\lambda_{xt}^1, \dots, \lambda_{xt}^\ell \in (0, 1)$ with $\sum_{j=1}^{\ell} \lambda_{xt}^j = 1$ and distributions $P_x^1, \dots, P_x^\ell$ that do not depend on t. This fact can be seen by expressing the conditional density of $P_{Y_t | (X_t^{S^*} = x)}$ as $f_t(y \,|\, x) = \int f_t(y \,|\, x, h) f_t(h \,|\, x) dh$. By (2.2.1), $f_t(y \,|\, x, h)$ does not depend on t. Property (2.2.2) then follows by taking $\lambda_{xt}^j = P(H_t^* = j \,|\, X_t^{S^*} = x)$ and letting P_x^j denote the distribution of $Y_1 \,|\, (X_1^{S^*} = x, H_1^* = j)$.

The conditional distributions of $Y_t \,|\, (X_t^{S^*} = x)$ 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 (2.2.2) into the framework of mixtures of linear regressions with Gaussian noise. The invariance assumption on $P_x^1, \dots P_x^\ell$ then corresponds to time-homogeneity of the regression parameters of all mixture components.

2.2.2. Hidden invariance property

As motivated by Proposition 2.1, 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 2.3.1. Since the inference principle described in Section 2.1.3 requires us to evaluate (2.2.2) 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 2.2).

Definition 2.1 (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 $j \in \{1, \dots, \ell\}$. 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 $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2), j \in \{1, \dots, \ell\}$, with $(\varepsilon_1, \dots, \varepsilon_\ell) \perp\!\!\!\perp X$, 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 A.5.1.) Second, it will be convenient to parametrize the matrix Θ by a map $\theta \mapsto \mathbf{\Theta}(\theta)$, $\theta \in \mathcal{T}$, 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 $\mathcal{SR}_{\mathbf{\Theta}}(\theta, \lambda \,|\, X)$ ("Switching Regression") to denote the distribution P over (Y, X) satisfying Definition 2.1 with parameters $(\mathbf{\Theta}(\theta), \lambda)$, although we will often omit the implicit dependence on $\mathbf{\Theta}$ and simply write $\mathcal{SR}(\theta, \lambda \,|\, X)$. For now, the reader may think of $(\mathbf{\Theta}, \mathcal{T})$ as the unconstrained parametrization, where $\mathcal{T} = (\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}_{>0})^\ell$ and where $\mathbf{\Theta}$ consists of the coordinate projections $\mathbf{\Theta}_{ij}(\theta) = \theta_{(j-1)(p+2)+i}$. 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 (2.2.2) into the model class $\mathcal{SR}$.

Definition 2.2 (h-invariance)**.** *A set $S \subseteq \{1, \dots, d\}$ is called* h-invariant *w.r.t.* $(\mathbf{Y}, \mathbf{X}) = (Y_t, X_t)_{t \in \{1,\dots,n\}}$ *if there exist* θ *and* $\lambda_1, \dots, \lambda_n$ *such that, for all* t, $P_{(Y_t, X_t^S)} = \mathcal{SR}(\theta, \lambda_t \,|\, X_t^S)$.

Definition 2.2 describes an invariance in the regression parameters θ and makes no restriction on the mixing proportions $\lambda_1, \dots, \lambda_n$. 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 2.1. *There exists a set $S^* \subseteq \{1, \dots, d\}$ 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 2.2.3 we show that if the data $(\mathbf{Y}, \mathbf{X}, \mathbf{H})$ are generated by different interventions in an SCM (see Appendix A.1), in which the variable $H^* \in \{1, \dots, \ell\}$ acts on Y, Assumption 2.1 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 2.1 and 2.2.

2.2.3. Relation to causality

Assumption 2.1 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 2.2 (Causal interpretation of S^*)**.** *Consider an SCM over the system of variables $(Y_t, X_t, H_t^*)_{t \in \{1,\dots,n\}}$, where for every t, $(Y_t, X_t, H_t^*) \in \mathbb{R}^1 \times \mathbb{R}^d \times \{1, \dots, \ell\}$. Assume that the structural assignment of Y is fixed across time, and for every $t \in \{1, \dots, n\}$ given by*

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

where $(N_t)_{t \in \{1,\dots,n\}}$ are i.i.d. noise variables. Here, $\mathrm{PA}^0(Y) \subseteq \{1, \dots, d\}$ denotes the set of parents of Y_t among $(X_t^1, \dots, X_t^d)$. The structural assignments for the remaining variables $X^1, \dots, X^d, H^*$ are allowed

to change between different time points. Then, property (2.2.1) *is satisfied for* $S^* = \mathrm{PA}^0(Y)$*. If furthermore the assignment* $f(\cdot, h, \cdot)$ *is linear for all* $h \in \{1, \dots, \ell\}$ *and the noise variables* N_t *are normally distributed, then, Assumption 2.1 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 2.2 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 2.2. *Consider an SCM over* $(Y_t, X_t, H_t^*)_{t \in \{1,\dots,n\}}$*, 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 2.1. 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 2.2, the set* $S^* = \{2\}$ *of observable parents of* Y *is* h*-invariant w.r.t.* $(\mathbf{Y}, \mathbf{X})$*, see Figure 2.1.*

2.2.4. Inference of the h-invariant set

In general, Definition 2.2 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

$$H_{0,S} : S \text{ is } h\text{-invariant with respect to } (\mathbf{Y}, \mathbf{X}), \text{ and} \tag{2.2.3}$$

$$\tilde{S} := \bigcap_{S:\, H_{0,S} \text{ true}} S, \tag{2.2.4}$$

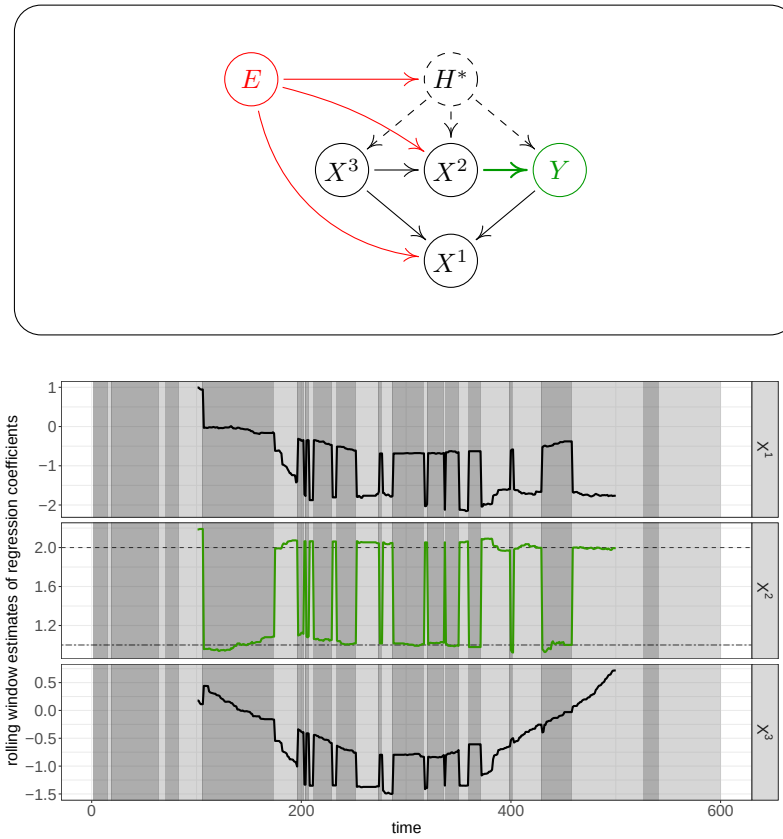


FIGURE 2.1. An illustration of the h-invariance property based on simulated data from the SCM in Example 2.2. The causal graph (top) and rolling window estimates of regression coefficients in the linear interaction model for the conditional distribution of Y given (X^1, H^*), (X^2, H^*) and (X^3, H^*), respectively (bottom). 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 2.2.5.

where S runs over subsets $S \subseteq \{1, \dots, d\}$. In (2.2.4), 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 $(\varphi_S)_{S \subseteq \{1,\dots,d\}}$ of the hypotheses $(H_{0,S})_{S \subseteq \{1,\dots,d\}}$, we therefore define an empirical version of (2.2.4) by

$$\hat{S} := \bigcap_{S:\, \varphi_S \text{ accepts } H_{0,S}} S. \tag{2.2.5}$$

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

Proposition 2.3 (Coverage property)**.** *Under Assumption 1 and given a family of tests $(\varphi_S)_{S \subseteq \{1,\dots,d\}}$ of $(H_{0,S})_{S \subseteq \{1,\dots,d\}}$ 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* (2.2.5) *is controlled at level α.*

The set S^* in Proposition 2.3 may not be uniquely determined by the h-invariance property. But since our output is the *intersection* (2.2.5) 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 $(\varphi_S)_{S \subseteq \{1,\dots,d\}}$, which we will discuss in Section 2.2.5.

2.2.4.1. Tests for non-causality of individual predictors

Proposition 2.3 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 $j \in \{1, \dots, d\}$ 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 $H_{0,S}$, $S \subseteq \{1, \dots, d\}$, 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 2.3, and is a direct consequence of φ_{S^*} achieving correct level α.

2.2.5. 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 (2.2.5). Let $S \subseteq \{1, \dots, d\}$ 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 $\mathbf{Y}_e = (Y_t)_{t \in e}$ 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.

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

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 2.3.

(ii) Based on (i), test if $\theta_e = \theta_f$ for all $e, f \in \mathcal{E}$, see Section 2.2.6.

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 2.2 illustrates step (i) for the two candidate 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 $r_t^j = Y_t - X_t^S \hat{\beta}_j$, $j \in \{1, \dots, \ell\}$.

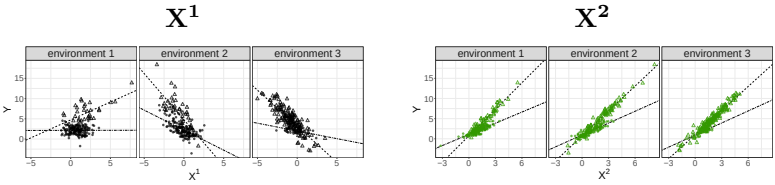


FIGURE 2.2. Testing procedure for $H_{0,S}$, here illustrated for the sets $\{1\}$ (black; not h-invariant) and $\{2\}$ (green; h-invariant) using the same data that generated Figure 2.1. First, we split data up into several environments, here $e_1 = \{1, \ldots, 200\}$, $e_2 = \{201, \ldots, 400\}$ 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.)

2.2.6. 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

$$T_S(\mathbf{Y}, \mathbf{X}^S) := \max \left\{ \alpha \in [0,1] : \bigcap_{e \in \mathcal{E}} C_e^{\alpha/|\mathcal{E}|}(\mathbf{Y}_e, \mathbf{X}_e^S) \neq \emptyset \right\}, \quad (2.2.6)$$

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.

2.3. 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 2.3.1–2.3.2 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 2.3.3–2.3.4 we construct confidence regions based on the maximum likelihood estimator, and in Section 2.3.5 we show that these confidence regions attain the correct asymptotic coverage. As a corollary, we obtain that the test defined in (2.2.6) satisfies asymptotic type I error control.

Let $S \subseteq \{1, \dots, d\}$ and consider a fixed environment e, say $e = \{1, \dots, m\}$. 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 $(\mathbf{Y}_e, \mathbf{X}_e^S)$.

2.3.1. Time dependence and time independence

Assume there exist parameters θ and $\lambda_1, \dots, \lambda_m$ such that, for all $t \in \{1, \dots, m\}$, $(Y_t, X_t) \sim \mathcal{SR}(\theta, \lambda_t \,|\, X_t)$. Let $\mathbf{H} = (H_t)_{t \in \{1,\dots,m\}} \in \{1, \dots, \ell\}^m$ be such that for every $t \in \{1, \dots, m\}$, the distributional statement in Definition 2.1 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 $t = 1, \dots, m$ are jointly independent and the marginal distribution of $\mathbf{H}$ is time-homogeneous. Furthermore, for every $t \in \{1, \dots, m\}$, 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 2.3. The Markov chain $\mathbf{H}$ is initiated in its stationary distribution. Fur-

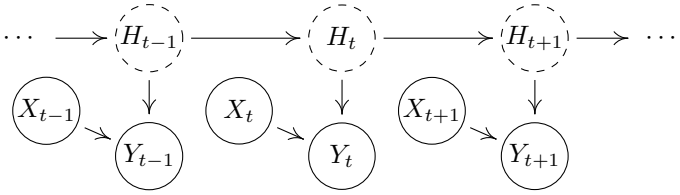


FIGURE 2.3. A hidden Markov model for $(\mathbf{Y}, \mathbf{X})$. All observations (across different $t \in \{1, \dots, m\}$) are conditionally independent given $\mathbf{H}$, and (Y_t, X_t) only depends on $\mathbf{H}$ through the present state H_t. Moreover, the variables in $\mathbf{H}$ resemble a first order Markov chain, that is, $(H_1, \dots, H_{t-1}) \perp\!\!\!\perp H_{t+1} \,|\, H_t$ for all $t \in \{2, \dots, m-1\}$.

thermore, for every $t \in \{1, \dots, m\}$, 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 2.5.1.5 suggests, however, that violations of this assumption do not negatively affect the performance of our causal discovery method.

For $i, j \in \{1, \dots, \ell\}$, 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 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 A.2.

2.3.1.1. 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 $\mathcal{SR}_{(\mathbf{\Theta},\mathbf{\Gamma})}(\theta, \gamma \,|\, \mathbf{X})$ to denote the joint distribution P over $(\mathbf{Y}, \mathbf{X})$ with marginals $(Y_t, X_t) \sim \mathcal{SR}_{\mathbf{\Theta}}(\theta, \lambda(\gamma) \,|\, X_t)$ 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 $\mathcal{SR}(\theta, \gamma \,|\, \mathbf{X})$. 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 $\mathcal{P} := \mathcal{T} \times \mathcal{G}$.

2.3.2. 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 $\mathcal{SR}(\theta, \gamma \,|\, \mathbf{X})$ is given by

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

where $p(\mathbf{x})$ is the (unspecified) density of $\mathbf{X}$, and where, for $j \in \{1, \ldots, \ell\}$, $\mathcal{N}(y_t \,|\, x_t \beta_j, \sigma_j^2)$ 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 $\{\mathcal{SR}(\theta, \gamma \,|\, \mathbf{X}) : (\theta, \gamma) \in \mathcal{T} \times \mathcal{G}\}$ is then given by, for every $(\theta, \gamma) \in \mathcal{T} \times \mathcal{G}$,

$$\ell_{(\mathbf{\Theta},\mathbf{\Gamma})}(\mathbf{y}, \mathbf{x} \,|\, \theta, \gamma) = \log \sum_{h_1} \cdots \sum_{h_m} p_{(\mathbf{\Theta},\mathbf{\Gamma})}(\mathbf{y}, \mathbf{x}, \mathbf{h} \,|\, \theta, \gamma). \tag{2.3.1}$$

It is well known that, in general, the loglikelihood function (2.3.1) is non-concave and may have several local maxima. For unconstrained parametrizations $(\mathbf{\Theta}, \mathcal{T})$ and $(\mathbf{\Gamma}, \mathcal{G})$, 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 $\mathbf{\Gamma}(\gamma)$ 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, $p_{(\mathbf{\Theta},\mathbf{\Gamma})}(\mathbf{y}, \mathbf{x}, \mathbf{h} \,|\, \theta, \gamma)$ 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 A.2. In the following theorem, we show that either of these parametrizations ensures the existence of the maximum likelihood estimator.

Theorem 2.1 (Existence of the MLE)**.** *Let* $(\mathbf{y}, \mathbf{x})$ *be a sample of* $(\mathbf{Y}, \mathbf{X}) = (Y_t, X_t)_{t\in\{1,\dots,m\}}$ *and assume that the set* $\{(y_t, x_t) \,|\, t \in \{1, \dots, m\}\}$ *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 A.2), 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 A.4 for details.

2.3.3. 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

$$C^\alpha(\hat{\theta}) := \left\{ \hat{\theta} + \mathcal{J}^{-1/2}(\hat{\theta}) v \,:\, \|v\|_2^2 \le q_{\chi^2(\dim(\theta))}(\alpha) \right\}, \qquad (2.3.2)$$

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

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.

2.3.4. 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 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 $\pi_\mathcal{T} := \mathbf{\Theta}^{-1} \circ (\Theta \mapsto \Theta M_\pi^T) \circ \mathbf{\Theta}$, $\pi_\mathcal{G} := \mathbf{\Gamma}^{-1} \circ (\Gamma \mapsto M_\pi \Gamma M_\pi^T) \circ \mathbf{\Gamma}$ 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 (2.3.2) 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

$$C^\alpha_{\text{adjusted}}(\hat{\theta}) := \bigcup_{\pi \in \Pi} C^\alpha(\pi_\mathcal{T}(\hat{\theta})). \tag{2.3.3}$$

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

2.3.5. Asymptotic coverage of adjusted confidence regions

Assume that the distribution of X_t is stationary across $e = \{1, \dots, m\}$ 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 $(\mathbf{Y}_m, \mathbf{X}_m) = (Y_t, X_t)_{t \in \{1,\dots,m\}}$, 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 $i \in \{1, \dots, p+1\}$ and $j, k \in \{1, \dots, \ell\}$, 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, with

$$\eta := \lim_{m \to \infty} \frac{\partial}{\partial \phi} f_m(Y_m, X_m \,|\, \mathbf{Y}_{m-1}, \mathbf{X}_{m-1}, \phi)\bigg|_{\phi = \phi^0},$$

the Fisher information matrix $\mathcal{I}_0 := \mathbb{E}_0[\eta\eta^T]$ 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 A.2, see Theorem 2.1. 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 (2.3.3) 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 2.3.4. 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 (2.3.3). The main work is contained in Theorems 2.2 and 2.3. 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 2.2 (Consistency of the MLE)**.** *Assume that (A1), (A3), (A4) and (A7) hold true. Then,* $\mathbb{P}_0$*-a.s.,* $[\hat{\phi}_m] \to [\phi^0]$ *as* $m \to \infty$*.*

Theorem 2.2 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. We now prove a central limit theorem for these subsequences.

Theorem 2.3 (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* $\mathbb{P}_0$*.*

Theorems 2.2 and 2.3 imply the following coverage guarantee.

Corollary 2.1 (Asymptotic coverage of adjusted confidence regions)**.** *Under Assumptions (A1)–(A7), the adjusted confidence regions* (2.3.3) *achieve the correct asymptotic coverage. That is, for any* $\alpha \in (0, 1)$*,*

$$\liminf_{m\to\infty} \mathbb{P}_0^m(\theta^0 \in C^{\alpha}_{\text{adjusted}}(\hat{\theta}_m)) \geq 1 - \alpha. \tag{2.3.4}$$

As another corollary, the asymptotic type I error control of the tests defined by (2.2.6) follows by applying Corollary 2.1 to each environment separately.

2.4. ICPH: Algorithm and false discovery control

We now summarize the above sections into our overall method. In Section 2.4.1 we provide a pseudo code for this procedure, and Section 2.4.2 presents our main theoretical result—an asymptotic version of Proposition 2.3, which states that our method is consistent.

2.4.1. Algorithm

Given data $(\mathbf{Y}, \mathbf{X})$ and a collection $\mathcal{E}$ of environments, we run through all $S \subseteq \{1, \dots, d\}$, test the hypothesis $H_{0,S}$ with the test defined by (2.2.6) using the adjusted confidence regions (2.3.3), 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")

Input: response $\mathbf{Y} \in \mathbb{R}^n$, covariates $\mathbf{X} \in \mathbb{R}^{n \times d}$, environment indicator $\mathbf{E} \in \{1, \ldots, |\mathcal{E}|\}^n$ (i.e., $\mathbf{E}_t = k \Leftrightarrow t \in e_k$);
Options: `model` $\in$ {"IID", "HMM"}, `method` $\in$ {"EM", "NLM"}, `variance.constraint` $\in$ {"lower bound", "equality"}, `number.of.states` $\in \mathbb{N}_{\geq 2}$, `intercept` $\in$ {TRUE, FALSE}, `test.parameters` $\subseteq$ {"intercept", "beta", "sigma"}, `alpha` $\in (0, 1)$;
for $S \subseteq \{1, \ldots, d\}$ **do**
    for $e \in \mathcal{E}$ **do**
        Fit an $\mathcal{SR}$ model to $(\mathbf{Y}_e, \mathbf{X}_e^S)$, see Section 2.3.2;
        Construct the perm.-adjusted conf. region (2.3.3);
    end
    Compute a p-value p_S for $H_{0,S}$ using the test defined by (2.2.6);
end
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 2.1.4). 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 2.4.2 for more details.

2.4.2. Asymptotic false discovery control of ICPH

The cornerstone for the false discovery control of ICPH is given in Corollary 2.1. 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 2.3. We will now summarize this line of reasoning into out main theoretical result.

Assume we have data $((\mathbf{Y}_n, \mathbf{X}_n))_{n\in\mathbb{N}} = \big((Y_{n,t}, X_{n,t})_{t\in\{1,\dots,n\}}\big)_{n\in\mathbb{N}}$ 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 $\mathcal{E}_n = \{e_{n,1}, \dots, e_{n,K}\}$, such that, for all n, $e_{n,1}, \dots, e_{n,K}$ are disjoint with $\cup_k e_{n,k} = \{1, \dots, n\}$ 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 $\{(\mathbf{\Theta}^S, \mathcal{T}^S)\}_{S\subseteq\{1,\dots,d\}}$, i.e., for every $S \subseteq \{1, \dots, d\}$, $\mathbf{\Theta}^S$ 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 $S \subseteq \{1, \dots, d\}$, let $H^n_{0,S}$ denote the hypothesis (2.2.3) for the data $(\mathbf{Y}_n, \mathbf{X}^S_n)$ and let φ^n_S be the corresponding test defined by (2.2.6) with the confidence regions (2.3.3). Finally, define for every n the estimator

$$\hat{S}_n := \bigcap_{S:\, \varphi^n_S \text{ accepts } H^n_{0,S}} S. \tag{2.4.1}$$

We then have the following result.

Theorem 2.4 (Asymptotic false discovery control)**.** *Assume that Assumption 2.1 is satisfied. That is, there exists a set $S^* \subseteq \{1, \dots, d\}$ 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 $(\mathbf{Y}_{n,k}, \mathbf{X}^{S^*}_{n,k})$ with parametrizations $(\mathbf{\Theta}^{S^*}, \mathcal{T}^{S^*})$ and $(\mathbf{\Gamma}, \mathcal{G})$. Then, the estimator $\hat{S}_n$ enjoys the following coverage property*

$$\liminf_{n\to\infty} \mathbb{P}^n_0(\hat{S}_n \subseteq S^*) \geq 1 - \alpha, \tag{2.4.2}$$

where $\mathbb{P}^n_0$ 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

$(\hat{S}^n_{\text{screening}})_{n\in\mathbb{N}}$ of screening estimators, we can for every $n \in \mathbb{N}$ construct an estimator $\bar{S}_n$ of S^* analogously to (2.4.1), but where the intersection is taken only over those S additionally satisfying that $S \subseteq \hat{S}^n_{\text{screening}}$. Given that $\liminf_{n\to\infty} \mathbb{P}^n_0(S^* \subseteq \hat{S}^n_{\text{screening}}) \geq 1-\alpha$, it then follows from

$$\begin{aligned}\mathbb{P}^n_0(\bar{S}_n \not\subseteq S^*) &= \mathbb{P}^n_0([\bar{S}_n \not\subseteq S^*] \cap [S^* \subseteq \hat{S}^n_{\text{screening}}]) \\ &\quad + \mathbb{P}^n_0([\bar{S}_n \not\subseteq S^*] \cap [S^* \not\subseteq \hat{S}^n_{\text{screening}}]) \\ &\leq \mathbb{P}^n_0(\varphi^n_{S^*} \text{ rejects } H^n_{0,S^*}) + \mathbb{P}^n_0(S^* \not\subseteq \hat{S}^n_{\text{screening}}),\end{aligned}$$

that the estimator $(\bar{S}_n)_{n\in\mathbb{N}}$ satisfies the asymptotic false discovery control (2.4.2) 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].

2.5. Experiments

In this section, we apply our method to simulated data (Section 2.5.1) and to a real world data set on photosynthetic activity and sun-induced fluorescence (Section 2.5.2). 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 A.4.2.

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 2.5.1.1. In Section 2.5.1.2, we present the multivariate data generating process that we will use in the subsequent analyses. In Section 2.5.1.3, 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 2.5.1.4 contains a power analysis. In Section 2.5.1.5, 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 A.5.

2.5.1.1. Empirical coverage properties of adjusted confidence regions

The finite sample coverage properties of the confidence regions (2.3.3) 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 $(Y, X, H) \in \mathbb{R}^{1+p} \times \{1, \dots, \ell\}$ which induces an $\mathcal{SR}$ model over (Y, X). For every $j \in \{1, \dots, \ell\}$ let $p_j(y, x) = \mathbb{P}(H = j \,|\, Y = y, X = x)$ denote the posterior probability of state j based on the data (y, x). We then use the geometric mean of expected posterior probabilities

$$\text{GMEP} := \left(\prod_{j=1}^{\ell} \mathbb{E}[p_j(Y, X) \,|\, H = j] \right)^{1/\ell} \in [0, 1] \tag{2.5.1}$$

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 (2.3.4) 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 $\mathbb{E}[p_i(Y, X) \,|\, H = j]$ 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

$$\begin{aligned} H &:= N_\lambda^H, \quad X := \mu_X + \sigma_X N^X, \\ Y &:= \mu_Y + \beta_1 X \cdot \mathbb{1}_{\{H=1\}} + \beta_2 X \cdot \mathbb{1}_{\{H=2\}} + \sigma_Y N^Y, \end{aligned} \tag{2.5.2}$$

[4] If each of the distributions $\mathbb{P}_{(Y,X)|H=j}$, $j \in \{1, \dots, \ell\}$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^{1+p}$, each factor in (2.5.1) is given as an integral over $\mathbb{R}^{1+p}$. In practice, we approximate these integrals by numerical integration.

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 (2.3.3) for the vector of regression parameters $\theta^0 = (\mu_Y, \beta_1, \mu_Y, \beta_2, \sigma_Y^2)$ using the likelihood function (2.3.1) with parametrizations $(\mathbf{\Theta}^=, \mathcal{T}^=)$ and $(\mathbf{\Gamma}^{\text{IID}}, \mathcal{G}^{\text{IID}})$ (see Appendix A.2). We sample 50 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 (2.5.2), and compare them to the GMEP of the underlying models, see Figure 2.4 (left). For the same simulations, we also compare the p-values

$$p := \max\{\alpha \in [0,1] : \theta^0 \notin C^\alpha_{\text{adjusted}}(\hat{\theta})\} \tag{2.5.3}$$

for the (true) hypotheses $H_0 : \theta = \theta^0$ to a uniform distribution (Figure 2.4 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 2.4 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 2.4 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 2.4 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 2.4 (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 2.5.1.3 that even cases where we cannot expect the confidence regions to obtain valid coverage, our overall causal discovery method maintains type I error control.

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

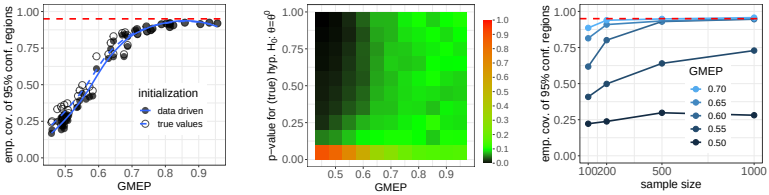


FIGURE 2.4. Empirical coverage properties of the adjusted confidence regions (2.3.3) using data simulated from the model (2.5.2). The left panel shows empirical coverage of 95%-confidence regions for different model parameters (see Equation 2.5.1 for a definition of GMEP), and a fixed sample size of 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 (2.5.3). 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 2.1, the empirical coverage gradually improves, although very low GMEP demand large amounts of data to obtain satisfactory coverage.

(Y, X^1, X^2, X^3, H) given by the structural assignments

$$\begin{aligned} X^1 &:= N^1, \quad H := N^H, \quad X^2 := \beta^{21} X^1 + N^2 \\ Y &:= \sum_{j=1}^{\ell} (\mu_j^Y + \beta_{1j}^Y X^1 + \beta_{2j}^Y X^2 + \sigma_{Yj} N^Y) \mathbb{1}_{\{H=j\}} \\ X^3 &:= \beta^{3Y} Y + N^3, \end{aligned}$$

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 2.5.1.3–2.5.1.5, the latent variable H is assumed to be binary, while Appendix A.5.1 treats the more general case where $\ell \geq 2$. The different environments are constructed as follows. We

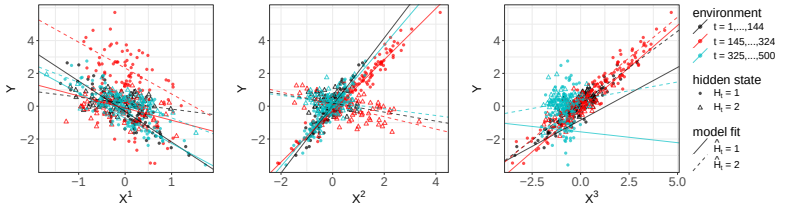


FIGURE 2.5. Data generated from the SCM described in Section 2.5.1.2 for each of the three environments (black, red, blue). Here, the only h-invariant set is $S^* = \{1, 2\}$ and we would therefore like our method to correctly identify the violations of the h-invariance of the sets $\{1\}, \{2\}$ and $\{3\}$. 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 2.5.1.3 and 2.5.1.4. 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 (—: $\hat{H}_t = 1$, - -: $\hat{H}_t = 2$).

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.

A sample data set can be seen in Figure 2.5, 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 (2.2.4). The causal graph induced by the above data generating system can be seen in Figure 2.6 (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.

2.5.1.3. Level analysis

Given that the theoretical coverage guarantees are only asymptotic, we cannot expect the tests (2.2.6) to satisfy type I error control for small sample sizes—especially if GMEP is low, see also Section 2.5.1.1. 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 2.5.1.2, 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 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 $\sigma^2_{Y1} = \sigma^2_{Y2}$), $\beta \sim^{iid} \text{Uniform}([-1.5, -0.5] \cup [0.5, 1.5])$ and $\lambda \sim \text{Uniform}(0.3, 0.7)$. For $j \in \{1, 2\}$ we then assign $\beta^Y_{j,2} := \beta^Y_{j,1} + \text{sign}(\beta^Y_{j,1})\Delta\beta$. The results are summarized in Figure 2.7. 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] 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.

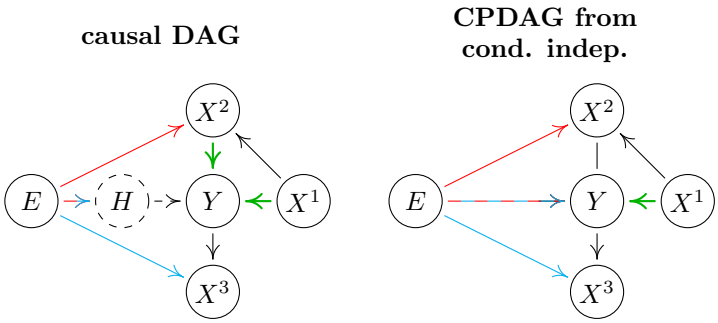


FIGURE 2.6. Left: the causal graph induced by the SCM in Section 2.5.1.2. The node E represents the different environments (E points into variables that have been intervened on, the color corresponds to the environments shown in Figure 2.5). Right: the CPDAG representing the Markov equivalence class 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.

2.5.1.4. 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 (2.2.4). 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 2.5.1.3 and analyze the performance of ICPH for two different variance constraints $\sigma^2_{Y1} = \sigma^2_{Y2}$ and $\sigma^2_{Y1}, \sigma^2_{Y2} \geq 10^{-4}$. The results in Figure 2.8 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).

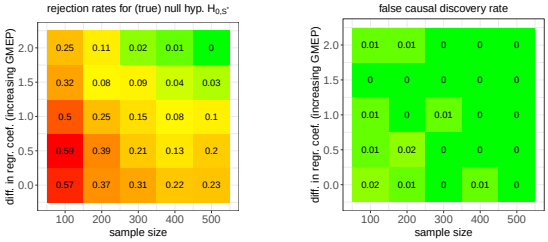


FIGURE 2.7. Estimates $\hat{\mathbb{P}}(\varphi_{S^*}$ rejects $H_{0,S^*})$ (left) and $\hat{\mathbb{P}}(\hat{S} \not\subseteq S^*)$ (right) of the type I error rates of the test φ_{S^*} and the overall method ICPH, respectively, based on the experiment described in Section 2.5.1.3 and 100 repetitions. The desired level is $\alpha = 0.05$. We have used NLM with parametrizations $\mathbf{\Theta}^=$ and $\mathbf{\Gamma}^{\text{IID}}$ (see Appendix A.2). 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.

For the same data that generated Figure 2.8, 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.

- "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

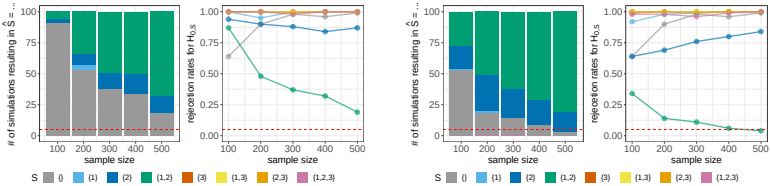


FIGURE 2.8. Output of ICPH (bar plots) and rejection rates for individual hypotheses (curve plots) for the experiment in Section 2.5.1.4 with parameter constraint $\sigma_{Y1}^2, \sigma_{Y2}^2 \geq 10^{-4}$ (left) and $\sigma_{Y1}^2 = \sigma_{Y2}^2$ (right). The larger the proportion of blue and green 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 $\hat{S} \not\subseteq S^*$. 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.

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 2.6. ICPH

[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.

solves a specific problem and is the only method which exploits the simple influence of H on Y. The results in Figure 2.9 (black curves) confirm our previous findings: causal discovery improves with increasing sample size, and our method stays conservative. ICPH outperforms both other methods in terms of level and power.

2.5.1.5. 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 2.5.1.2. Unless mentioned otherwise, parameters are sampled as described in Section 2.5.1.3.

- Heterogeneous variances: The error variances σ^2_{Y1} and σ^2_{Y2} 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 $\mu^1_1, \mu^1_2 \sim \text{Uniform}(-1, 1)$ and $\sigma^2_{11}, \sigma^2_{12} \sim \text{Uniform}(0.1, 1)$.
- A continuous hidden variable: We substitute the structural assignment for Y by $Y := (\mu^Y + \beta^Y_1 X^1 + \beta^Y_2 X^2)H + \sigma_Y N^Y$, where $H \sim \mathcal{N}(0, 1)$. The distribution of H does not change across environments.

We now repeat the power analysis from Section 2.5.1.4 for data sets generated in the above way (Figure 2.9, 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.

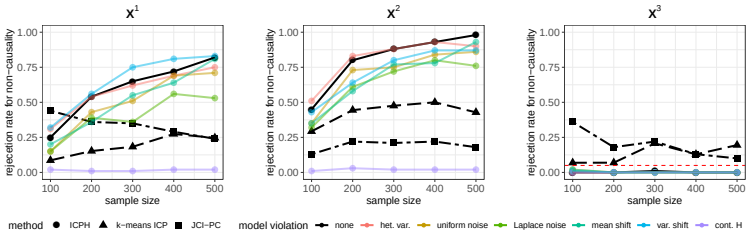


FIGURE 2.9. Rejection rates for non-causality. This figure contains two comparisons (one among all black curves, and another among all curves with round points). For data generated from the SCM in Section 2.5.1.2 (black), we compare the performance of ICPH (●) against the two alternative methods k-means ICP (▲) and JCI-PC (■) described in Section 2.5.1.4. For increasing sample size, ICPH outperforms both methods in terms level and power. As a robustness analysis, we further apply ICPH to simulated data sets from the modified SCMs described in Section 2.5.1.5 (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.

2.5.2. 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 ecosys-

variable	description
Y	sun-induced fluorescence (SIF)
X^1	incident shortwave rad. (SW)
X^2	abs. phot. active rad. (APAR_{chl})
X^3	gross primary prod. (GPP)
H	vegetation type

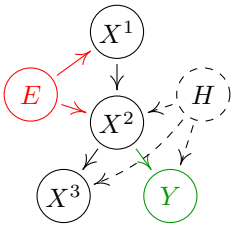


FIGURE 2.10. Variable descriptions (left) and causal graph constructed from background knowledge (right). In our analysis, we use the temporal ordering of data to construct the environment variable E. Due to seasonal cycles of aggradation and degradation of chlorophyll, APAR_{chl} is not a constant fraction 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.

tems. 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 2.10.

In Section 2.5.2.1, we use our causal discovery method to identify the causal predictor of SIF. In Section 2.5.2.2, 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.

2.5.2.1. Causal discovery

We denote the observed variables by (Y, X^1, X^2, X^3) as described in Figure 2.10 (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 $0.5^\circ \times 0.5^\circ$ 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 $S^* = \{2\}$

[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.

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 2.11.[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 structure from Figure 2.10. 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.

2.5.2.2. Reconstruction of the vegetation type

We know that (Y, X^2) follows a switching regression model (see Figure 2.10), and that the hidden variable in this model corresponds to

[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 (2.2.6), the lowest attainable p-value is 10^{-4}.

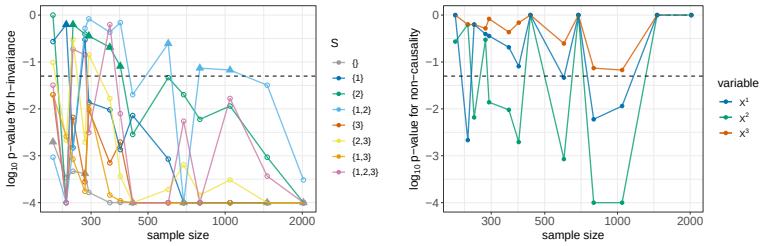


FIGURE 2.11. P-values for h-invariance of different sets $S \subseteq \{1, 2, 3\}$ (left) and p-values for non-causality (see Section 2.2.4.1) of the individual variables X^1, X^2 and X^3 (right). For every experiment, the estimated set $\hat{S}$ in the left plot is indicated by a triangle. For several 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.

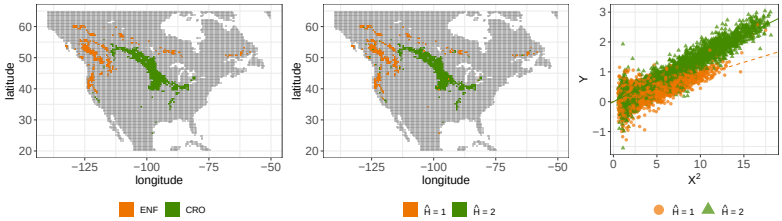


FIGURE 2.12. Vegetation type by IGBP (left) and estimates obtained from reconstructing the values of the hidden variable, as described in Section 2.5.2.2 (middle). We correctly classify more than 95% of the pixels. The right hand plot illustrates the vegetation-dependent linear relationship between Y 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.

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 underlying 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 method correctly reconstructs the hidden variable in more than 95% of the grid cells (Figure 2.12, left and middle). As seen in Figure 2.12 (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*.

2.6. 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 2.5.1.5 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

We thank Roland Langrock for insightful comments and providing parts of the code; Jens Ledet Jensen, Miguel Mahecha and Markus Reichstein for helpful discussions; and Yao Zhang and Xiangming Xiao for providing parts of the data used in Section 2.5.2. We thank two anonymous referees and the AE for many helpful and constructive comments. This research was supported by a research grant (18968) from VILLUM FONDEN.

3 | The Difficult Task of Distribution Generalization in Nonlinear Models

Joint work with

Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco and Jonas Peters

Abstract

We consider the problem of predicting a response from a set of covariates when the test distribution differs from the training distribution. Here, we consider robustness against distributions that emerge as intervention distributions. Causal models that regress the response variable on all of its causal parents have been suggested for the above task since they remain valid under arbitrary interventions on any subset of covariates. However, in linear models, for a set of interventions with bounded strength, alternative approaches have been shown to be minimax prediction optimal. In this work, we analyze minimax solutions in nonlinear models for both direct and indirect interventions on the covariates. We prove that the causal function is minimax optimal for a large class of interventions. We introduce the notion of distribution generalization, which is motivated by the fact that, in practice, minimax solutions need to be identified from observational data. We prove sufficient conditions for distribution generalization and present corresponding impossibility results. To illustrate the above findings, we propose a practical method, called NILE, that achieves distribution generalization in a nonlinear instrumental variable setting with linear extrapolation. We prove consistency, present empirical results and provide code.

3.1. Introduction

Large-scale learning systems, particularly those focusing on prediction tasks, have been successfully applied in various domains of application. Since inference is usually done during training time, any difference between training and test distribution poses a challenge for prediction methods Quionero-Candela et al. [2009], Pan and Yang [2010], Csurka [2017], Arjovsky et al. [2019]. Dealing with differences in training and test distribution is of great importance in fields such as many environmental sciences, where methods need to extrapolate both in space and time. Tackling this task requires restrictions on how the distributions may differ, since, clearly, generalization becomes impossible if the test distribution may be arbitrary. Given a response Y and some covariates X, existing procedures often aim to find a function f which minimizes the worst-case risk $\sup_{P \in \mathcal{N}} \mathbb{E}_P[(Y - f(X))^2]$ across distributions contained in a small neighborhood $\mathcal{N}$ of the training distribution. The neighborhood $\mathcal{N}$ should be representative of the difference between the training and test distributions, and often mathematical tractability is taken into account, too [Abadeh et al., 2015, Sinha et al., 2017]. A typical approach is to define a ρ-ball of distributions $\mathcal{N}_\rho(P_0) := \{P : D(P, P_0) \leq \rho\}$ around the training distribution P_0, with respect to some divergence measure D, such as the Kullback-Leibler divergence or the χ^2 divergence [Hu and Hong, 2013, Ben-Tal et al., 2013, Bertsimas et al., 2018, Lam, 2019, Duchi et al., 2016]. While these divergence functions only consider distributions with the same support as P_0, the Wasserstein distance allows to define a neighborhood of distributions around P_0 with possibly different supports [Abadeh et al., 2015, Sinha et al., 2017, Esfahani and Kuhn, 2018, Blanchet et al., 2019]. In our analysis, we do not start from a divergence measure, but we construct a neighborhood of distributional changes by using the concept of interventions [Pearl, 2009, Peters et al., 2017].

We will see that, depending on the considered setup, one can find models that perform well under interventions which yield distributions that are considered far away from the observational distribution in any commonly used metric. Using causal concepts for the

above problem has been motivated by the following observation. A causal prediction model, that uses only the direct causes of the response Y as covariates, is known to be invariant under interventions on variables other than Y: the conditional distribution of Y given its causes does not change (this principle is known as invariance, autonomy or modularity) [Aldrich, 1989, Haavelmo, 1944, Pearl, 2009]. Such a model yields the minimal worst-case prediction error when considering all interventions on variables other than Y [e.g., Rojas-Carulla et al., 2018, Theorem 1, Appendix]. It has therefore been suggested to use causal models in problems of domain generalization or distributional shifts [Schölkopf et al., 2012, Rojas-Carulla et al., 2018, Heinze-Deml and Meinshausen, 2017, Magliacane et al., 2018, Meinshausen, 2018, Arjovsky et al., 2019, Pfister et al., 2019c]. One may argue, however, that causal methods are too conservative in that the interventions which induce the test distributions may not be arbitrarily strong. As a result, methods which focus on a trade-off between predictability and causality have been proposed for linear models [Rothenhäusler et al., 2018, Pfister et al., 2019a], see also Section 3.5.1. In this work, we consider the problem of characterizing and finding minimax optimal models in a more general, nonlinear framework.

3.1.1. Contribution

We assume that the true data generating process can be described by a model M that belongs to a class of models $\mathcal{M}$ and induces an observational distribution $\mathbb{P}_M$. We then consider the risk of a prediction function $f_\diamond$ from a function class $\mathcal{F}$ under a modified model $M(i)$ that is obtained from M by an intervention i, which belongs to a set of interventions $\mathcal{I}$. Here, interventions can either act directly on X or indirectly, via an exogenous variable A, if the latter exists (precise definitions are provided in Section 3.2 below). Our work has four main contributions. (1) We analyze the relation between the causal function (defined formally in Section 3.2) and the minimizer of $\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2]$. Our findings go beyond existing results in that the causal function is shown to be minimax optimal already for relatively small intervention classes. We further prove that, in general, the difference between a minimax solution

intervention	$\text{supp}_{\mathcal{I}}(X)$	assumptions	result
on X (well-behaved)	$\subseteq \text{supp}(X)$	Ass. 3.1	Prop. 3.6
on X (well-behaved)	$\not\subseteq \text{supp}(X)$	Ass. 3.1 & 3.2	Prop. 3.8
on A	$\subseteq \text{supp}(X)$	Ass. 3.1 & 3.3	Prop. 3.12
on A	$\not\subseteq \text{supp}(X)$	Ass. 3.1, 3.2 & 3.3	Prop. 3.12

TABLE 3.1. Summary of conditions under which generalization is possible. Corresponding impossibility results are shown in Propositions 3.7, 3.11 and 3.13.

and the causal function can be bounded and that any minimax solution different from the causal function is not robust with respect to misspecification of the intervention class. (2) In practice, we usually have to learn the minimax solution from an observational distribution, in the absence of causal background knowledge. We therefore introduce the concept of distribution generalization, which requires the existence of a prediction model f^* which (approximately) solves the minimax problem $\arg\min_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f_\diamond(X))^2]$ for all $\tilde{M}$ with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$. To the best of our knowledge, the considered setup is novel. (3) We then investigate explicit conditions on $\mathcal{M}$, $\mathcal{I}$ and $\mathbb{P}_M$ that allow us to use the observational distribution of (X, Y, A) to identify a function $f^* : \mathbb{R}^d \to \mathbb{R}$ that generalizes to $\mathcal{I}$, i.e., it (approximately) solves the above minimax problem. We prove several results. E.g., if the interventions are such that the support of X does not increase with respect to the training distribution, then identifiability of the causal function — a well-studied problem in causality — is in general sufficient for generalization. We furthermore give sufficient conditions for generalization to interventions on either A or X that extend the support of X. Table 3.1 summarizes some of these results. We also prove that, without these assumptions, generalization is impossible; (4) In Section 3.5, we discuss how minimax functions can be learned from finitely many data and explain how existing methodology fits into our framework. We propose a novel estimator, the NILE, that is applicable in a nonlinear instrumental variables (IV) setting and achieves distribution generalization with linear extensions. We prove consistency and pro-

vide empirical results. Our code is available as an `R`-package at `https://runesen.github.io/NILE`. Scripts generating all our figures and results can be found at the same url.

3.1.2. Further related work

That the causal function is minimax optimal under the set of all interventions on the covariates has been shown by Rojas-Carulla et al. [2018], for example, where the additional assumption of no hidden variables is made. In Section 3.2, we extend this result in various ways. The question of distributional robustness, sometimes also referred to as out-of-distribution generalization, aims to develop procedures that are robust to changes between training and testing distribution. Empirically, this problem is often studied using adversarial attacks, where small digital [Goodfellow et al., 2014] or physical [Evtimov et al., 2017] perturbations of pictures can deteriorate the performance of a model; arguably, these procedures are not yet fully understood theoretically. Unlike the procedures mentioned in Section 3.1.1 that aim to minimize the worst-case risk across distributions contained in a neighborhood of the training distribution, e.g., in the Wasserstein metric, [Sinha et al., 2017], we assume these neighborhoods to be generated by interventions. To the best of our knowledge, the characterization of distribution generalization that we consider in Section 3.4 is novel.

In settings of covariate shift, one usually assumes that the training and test distribution of the covariates are different, while the conditional distribution of the response given the covariates remains invariant [Daume III and Marcu, 2006, Bickel et al., 2009, David et al., 2010, Muandet et al., 2013]. Sometimes, it is additionally assumed that the support of the training distribution covers the one of the test distribution [Shimodaira, 2000]. In this work, the conditional distribution of the response given the covariates is allowed to change between interventions, due to the existence of a hidden confounder, and we consider settings where the test observations lie outside the training support. Data augmentation methods increase the diversity of the training dataset by changing the geometry and the color of the images (e.g., by rotation, cropping or changing saturation) [Zhang et al., 2017, Shorten and Khoshgoftaar, 2019]. This

allows the user to create models that generalize better to unseen environments [e.g., Volpi et al., 2018]. We view these approaches as a way to enlarge the support of the covariates, which comes with theoretical advantages, see Section 3.4.

Minimizing the worst-case prediction error can also be formulated in terms of minimizing the regret in a multi-armed bandit problem [Lai and Robbins, 1985, Auer et al., 2002, Bartlett et al., 2008]. In that setting, the agent can choose the distribution which generates the data. In our setting, though, we do not assume to have control on the interventions and hence on the distribution of the sampled data.

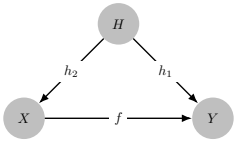
3.1.3. Structure of this work

We introduce our framework for generating a collection of intervention distributions in Section 3.2. In Section 3.3, we formalize the problem considered in this work, namely to find a model that predicts well under a set of intervention distributions. We prove that for a wide range of intervention classes, this is achieved by the causal function. In reality, we are not given the full causal model, but only the observational distribution. This problem is considered in Section 3.4, where we provide sufficient conditions under which distribution generalization is possible and prove corresponding impossibility results. The condition whether the intervened X values are inside the support of the training distribution will play an important role. Section 3.5 considers the problem of learning models from a finite amount of data. In particular, we propose a method, called NILE, that learns a generalizing model in a nonlinear IV setting. We prove consistency and compare our method to state-of-the art approaches empirically. In Appendix B.1, we comment on the different model classes that are contained in our framework. Appendix B.2 summarizes existing results on identifiability in IV models and Appendix B.3 provides details on the test statistic that we use in NILE. Appendix B.4 contains an additional experiment and all proofs are provided in Appendix B.5.

3.2. Modeling intervention induced distributions

We now specify the statistical model used throughout this paper. For a real-valued response variable $Y \in \mathbb{R}$ and predictors $X \in \mathbb{R}^d$, we consider the problem of estimating a regression function that works well not only on the training data, but also under distributions that we will model by interventions. We require a model that is able to model an observational distribution of (X, Y) (training) and the distribution of (X, Y) under a class of interventions on (parts of) X (testing). We will do so by means of a structural causal model (SCM) [Bollen, 1989, Pearl, 2009]. More precisely, denoting by $H \in \mathbb{R}^q$ some additional (unobserved) variables, we consider the SCM

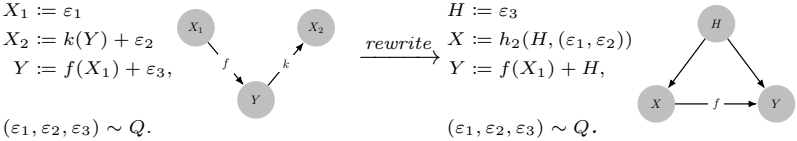
$$
\begin{aligned}
H &\coloneqq \varepsilon_H && q \text{ assignments} \\
X &\coloneqq h_2(H, \varepsilon_X) && d \text{ assignments} \\
Y &\coloneqq f(X) + h_1(H, \varepsilon_Y) && 1 \text{ assignment}
\end{aligned}
$$



Here, f, h_1 and h_2 are measurable functions, the innovation terms ε_X, ε_Y and ε_H are independent vectors with possibly dependent coordinates. Two comments are in order. The joint distribution of (X, Y) is constrained only by requiring that X and $h_1(\varepsilon_Y, H)$ enter the equation of Y additively. This constraint affects the allowed conditional distributions of Y given X but does not make any restriction on the marginal distributions of either X or Y. Furthermore, we do not assume that the above SCM represents the true causal relationships between the random variables. We do not assume any causal background knowledge of the system. Instead, the SCM is used only to construct the test distributions (by considering interventions on X) for which we are analyzing the predictive performance of different methods – similar to how one could have considered a ball around the training distribution. If causal background knowledge exists, however, e.g., in the form of an SCM over variables X and Y, it can be made to fit into the above framework. As such, our framework includes a large variety of models, including SCMs in which some of the X are not ancestors but descendants of Y (this requires

adapting the set of interventions appropriately), see Appendix B.1 for details. The following remark shows such an example, and may be interesting to readers with a special interest in causality. It can be skipped at first reading.

Remark 3.1 (Rewriting causal background knowledge)**.** *If a priori causal background knowlededge is available, e.g., in form of an SCM, our framework is still applicable after an appropriate transformation. The following example shows a reformulation of an SCM over variables X_1, X_2 and Y.*



Here, $h_2(H, (\varepsilon_1, \varepsilon_2)) := (\varepsilon_1, k(f(\varepsilon_1) + H) + \varepsilon_2)$. Both SCMs induce the same observational distribution over (X_1, X_2, Y) and any intervention on the covariates in the SCM on the left-hand side can be rewritten as an intervention on the covariates in the SCM on the right-hand side. Details and a more general treatment are provided in Appendix B.1.

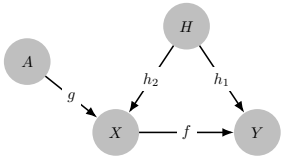
Sometimes, the vector X contains variables that are independent of H and that enter additively into the assignments of the other covariates. If such covariates exist, it can be useful to explicitly distinguish them from the other covariates. We will denote them by A and call them exogenous variables. Such variables are interesting for two reasons. (i) Under additional assumptions, they can be used as instrumental variables [e.g., Bowden and Turkington, 1985, Greene, 2003], a well-established tool for ensuring that f can be uniquely recovered from the observational distribution of (X, Y). And (ii), we will see below that in general, interventions on such variables lead to intervention distributions with desirable properties. In the remainder of this article, we will therefore consider a slightly larger class of SCMs that also includes exogenous variables A. It contains the SCM presented at the beginning of Section 3.2 as a special case.[1]

[1] This follows by choosing A as an indep. noise variable and a constant g.

3.2.1. Model

Formally, we consider a response $Y \in \mathbb{R}^1$, covariates $X \in \mathbb{R}^d$, exogenous variables $A \in \mathbb{R}^r$, and unobserved variables $H \in \mathbb{R}^q$. Let further $\mathcal{F} \subseteq \{f : \mathbb{R}^d \to \mathbb{R}\}$, $\mathcal{G} \subseteq \{g : \mathbb{R}^r \to \mathbb{R}^d\}$, $\mathcal{H}_1 \subseteq \{h_1 : \mathbb{R}^{q+1} \to \mathbb{R}\}$ and $\mathcal{H}_2 \subseteq \{h_2 : \mathbb{R}^{q+d} \to \mathbb{R}^d\}$ be fixed sets of measurable functions. Moreover, let $\mathcal{Q}$ be a collection of probability distributions on $\mathbb{R}^{d+1+r+q}$, such that for all $Q \in \mathcal{Q}$ it holds that if $(\varepsilon_X, \varepsilon_Y, \varepsilon_A, \varepsilon_H) \sim Q$, then $\varepsilon_X, \varepsilon_Y, \varepsilon_A$ and ε_H are jointly independent, and for all $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ it holds that $\xi_Y := h_1(\varepsilon_H, \varepsilon_Y)$ and $\xi_X := h_2(\varepsilon_H, \varepsilon_X)$ have mean zero.[2] Let $\mathcal{M} := \mathcal{F} \times \mathcal{G} \times \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}$ denote the model class. Every model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$ then specifies an SCM by[3]

$$
\begin{aligned}
A &:= \varepsilon_A && r \text{ assignments} \\
H &:= \varepsilon_H && q \text{ assignments} \\
X &:= g(A) + h_2(H, \varepsilon_X) && d \text{ assignments} \\
Y &:= f(X) + h_1(H, \varepsilon_Y) && 1 \text{ assignment}
\end{aligned}
$$



with $(\varepsilon_X, \varepsilon_Y, \varepsilon_A, \varepsilon_H) \sim Q$. For each model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$, we refer to f as the *causal function* (for the pair (X, Y)) and assume that the entailed distribution has finite second moments. Furthermore, we denote by $\mathbb{P}_M$ the joint distribution over the observed variables (X, Y, A) induced by the SCM specified by M. If no exogenous variables A exist, one can think of the function g as being a constant function.

3.2.2. Interventions

Each SCM $M \in \mathcal{M}$ can now be modified by the concept of interventions [e.g., Pearl, 2009, Peters et al., 2017]. An intervention corresponds to replacing one or more of the structural assignments of the SCM. For example, we intervene on all covariates X by replacing the d assignments with, e.g., a random variable, which is indepen-

[2] This can be assumed without loss of generality if $\mathcal{F}$ and $\mathcal{G}$ are closed under addition and scalar multiplication, and contain the constant function.

[3] For appropriate choices of h_2, the model includes settings in which (some of) the A directly influence Y.

dent of the other noise variables and has a multivariate Gaussian distribution. Importantly, an intervention on some of the variables does not change the assignment of any other variable. In particular, an intervention on X does not change the conditional distribution of Y, given X and H (this is an instance of the invariance property mentioned in Section 3.1). More generally, we denote by $M(i)$ the intervened SCM over the variables (X^i, A^i, Y^i, H^i), obtained by performing the intervention i in model M. We do not require that the intervened model $M(i)$ belong to the model class $\mathcal{M}$, but we require that $M(i)$ induces a joint distribution over (X^i, Y^i, A^i, H^i), which has finite second moments. We use $\mathcal{I}$ to denote a collection of interventions.

In this work, we only consider interventions on the covariates X and A. More specifically, for a given model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$ and an intervention $i \in \mathcal{I}$, the intervened SCM $M(i)$ takes one of two forms. First, for an intervention on X it is given by

$$\begin{aligned} A^i &:= \varepsilon_A^i, \quad H^i := \varepsilon_H^i, \\ X^i &:= \psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i), \\ Y^i &:= f(X^i) + h_1(H^i, \varepsilon_Y^i), \end{aligned}$$

and, second, for an intervention on A it is given by

$$\begin{aligned} A^i &:= \psi^i(I^i, \varepsilon_A^i), \quad H^i := \varepsilon_H^i, \\ X^i &:= g(A^i) + h_2(H^i, \varepsilon_X^i), \\ Y^i &:= f(X^i) + h_1(H^i, \varepsilon_Y^i). \end{aligned}$$

In both cases, $(\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i) \sim Q$, the (possibly degenerate) random vector I^i is independent of $(\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i)$, and ψ^i is a measurable function, whose arguments are all part of the structural assignment of the intervened variable in model M. We will see below that this class of interventions is rather flexible. It does, however, not allow for arbitrary manipulations of M. For example, the noise variable ε_Y is not allowed to enter the structural assignment of the intervened variable. Interventions on A will generally be easier to analyze than interventions on X. We therefore distinguish between the following different types of interventions on X. Let i be an intervention on X with intervention map ψ^i. The intervention is then

called

confounding-preserving if there exists a map φ^i, such that
$\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = \varphi^i(A^i, g(A^i), h_2(H^i, \varepsilon_X^i), I^i)$

and it is called

confounding-removing if for all models $M \in \mathcal{M}$,
$\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) \perp\!\!\!\perp H^i$ under $M(i)$.

Furthermore, we call a set of interventions $\mathcal{I}$ *well-behaved* either if it consists only of confounding-preserving interventions or if it contains at least one confounding-removing intervention. Confounding-preserving interventions contain, for example, *shift interventions* on X, which linearly shift the original assignment by I^i, that is, $\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = g(A^i) + h_2(H^i, \varepsilon_X^i) + I^i$. The naming 'confounding-preserving' stems from the fact that the unobserved (confounding) variables H only enter the intervened structural assignment of X via the term $h_2(H^i, \varepsilon_X^i)$, which is the same as in the original model. Some interventions are confounding-removing and confounding-preserving, but not every confounding-removing intervention is confounding-preserving. For example, the intervention $\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = \varepsilon_X^i$ is confounding-removing but, in general, not confounding-preserving. Similarly, not all confounding-preserving are confounding-removing.

If the context does not allow for any ambiguity, we omit the superscript i and write expressions such as $\mathbb{E}_{M(i)}[(Y - f(X))^2]$. The support of random variables under interventions will play an important role for the analysis of distribution generalization. Throughout this paper, $\text{supp}^M(Z)$ denotes the support of the random variable $Z \in \{A, X, H, Y\}$ under the distribution $\mathbb{P}_M$, which is induced by the SCM $M \in \mathcal{M}$. Moreover, $\text{supp}_{\mathcal{I}}^M(Z)$ denotes the union of $\text{supp}^{M(i)}(Z)$ over all interventions $i \in \mathcal{I}$. We call a collection of interventions on Z *support-reducing* (w.r.t. M) if $\text{supp}_{\mathcal{I}}^M(Z) \subseteq \text{supp}^M(Z)$ and *support-extending* (w.r.t. M) if $\text{supp}_{\mathcal{I}}^M(Z) \not\subseteq \text{supp}^M(Z)$. Whenever it is clear from the context which model is considered, we may drop the indication of M altogether and simply write $\text{supp}(Z)$.

3.3. Interventional robustness and the causal function

Let $\mathcal{M}$ be a fixed model class, let $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$ be the true data generating model, and let $\mathcal{I}$ be a class of interventions. In this work, we aim to find a function $f^* : \mathbb{R}^d \to \mathbb{R}$, such that the predictive model $\hat{Y} = f^*(X)$ has low worst-case risk over all distributions induced by the interventions in $\mathcal{I}$. We therefore consider the optimization problem

$$\arg\min_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right], \tag{3.3.1}$$

where $\mathbb{E}_{M(i)}$ is the expectation in the intervened model $M(i)$. In general, this optimization problem is neither guaranteed to have a solution, nor is the solution, if it exists, ensured to be unique. Whenever a solution f^* exists, we refer to it as a *minimax solution* (for model M w.r.t. $(\mathcal{F}, \mathcal{I})$).

If, for example, $\mathcal{I}$ consists only of the trivial intervention, that is, $\mathbb{P}_M = \mathbb{P}_{M(i)}$, we are looking for the best predictor on the observational distribution. In that case, the minimax solution is obtained by any conditional mean function, $f^* : x \mapsto \mathbb{E}[Y|X = x]$ (provided that $f^* \in \mathcal{F}$). For larger classes of interventions, however, the conditional mean may become sub-optimal in terms of prediction. To see this, it is instructive to decompose the risk under an intervention. Since the structural assignment for Y remains unchanged for all interventions that we consider in this work, it holds for all $f_\diamond \in \mathcal{F}$ and all interventions i on either A or X that

$$\begin{aligned}&\mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \\ &= \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] + 2\mathbb{E}_{M(i)}[\xi_Y(f(X) - f_\diamond(X))].\end{aligned}$$

Here, the middle term does not depend on i since $\xi_Y = h_1(H, \varepsilon_Y)$ remains fixed. If i is a confounding-removing intervention, then $\xi_Y \perp\!\!\!\perp X$ under $\mathbb{P}_{M(i)}$, and, because of $\mathbb{E}_M[\xi_Y] = 0$, the last term in the above equation vanishes. Therefore, if $\mathcal{I}$ consists only of confounding-removing interventions, the causal function is a solution to the minimax-problem (3.3.1). The following proposition

shows that an even stronger statement holds: The causal function is already a minimax solution if $\mathcal{I}$ contains at least one confounding-removing intervention on X.

Proposition 3.1 (confounding-removing interventions on X)**.** *If $\mathcal{I}$ is a set of interventions on X or A and at least one of these is a confounding-removing intervention, then the causal function f is a minimax solution.*

One step in the proof of this proposition is to show that the minimal worst-case loss is attained at a confounding-removing intervention. That is,

$$\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] = \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}_{\text{cr}}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big],$$

where $\mathcal{I}_{\text{cr}} \subseteq \mathcal{I}$ denotes the non-empty subset of confounding-removing interventions. This observation will also be used in the proofs of some of the results that follow below.

We now prove that when restricting ourselves to linear functions only, the causal function is also a minimax solution with respect to the set of all shift interventions on X – interventions that appear in linear IV models and recently gained further attention in the causal community [Rothenhäusler et al., 2018, Sani et al., 2020]. The proposition below also makes precise in which sense shift interventions are related to linear model classes. Intuitively, when the causal relation between X and Y is linear, shift interventions are sufficient to create unbounded variability in all directions of the covariance matrix of X (more precisely, the unbounded eigenvalue condition below is satisfied if $\mathcal{I}$ is the set of all shift interventions on X). As the following proposition shows, under this condition, the causal function is a minimax solution.

Proposition 3.2 (unbounded interventions on X with linear $\mathcal{F}$)**.** *Let $\mathcal{F}$ be the class of all linear functions, and let $\mathcal{I}$ be a set of interventions on X or A s.t. $\sup_{i \in \mathcal{I}} \lambda_{\min}\big(\mathbb{E}_{M(i)}\big[XX^\top\big]\big) = \infty$, where $\lambda_{\min}$ denotes the smallest eigenvalue (assuming that the considered moments exist). Then, the causal function f is the unique minimax solution.*

Even if the causal function f does not solve the minimax problem (3.3.1), the difference between the minimax solution and the causal function cannot be arbitrarily large. The following proposition shows that the worst-case L_2-distance between f and any function $f_\diamond$ that performs better than f (in terms of worst-case risk) can be bounded by a term which is related to the strength of the confounding.

Proposition 3.3 (difference between causal function and minimax solution)**.** *Let $\mathcal{I}$ be a set of interventions on X or A. Then, for any function $f_\diamond \in \mathcal{F}$ which satisfies that*

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \leq \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2],$$

it holds that

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] \leq 4 \operatorname{Var}_M(\xi_Y).$$

Even though the difference can be bounded, it may be non-zero, and one may benefit from choosing a function that differs from the causal function f. This choice, however, comes at a cost: it relies on the fact that we know the class of interventions $\mathcal{I}$. In general, being a minimax solution is not entirely robust with respect to misspecification of $\mathcal{I}$. In particular, if the set $\mathcal{I}_2$ of interventions describing the test distributions is misspecified by a set $\mathcal{I}_1 \neq \mathcal{I}_2$, then the considered minimax solution with respect to $\mathcal{I}_1$ may perform worse than the causal function on the test distributions.

Proposition 3.4 (properties of the minimax solution under mis-specified interventions)**.** *Let $\mathcal{I}_1$ and $\mathcal{I}_2$ be any two sets of interventions on X, and let $f_1^* \in \mathcal{F}$ be a minimax solution w.r.t. $\mathcal{I}_1$. Then, if $\mathcal{I}_2 \subseteq \mathcal{I}_1$ it holds that*

$$\sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\big[(Y - f_1^*(X))^2\big] \leq \sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\big[(Y - f(X))^2\big].$$

If $\mathcal{I}_2 \not\subseteq \mathcal{I}_1$, however, it can happen (even if $\mathcal{F}$ is linear) that

$$\sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\big[(Y - f_1^*(X))^2\big] > \sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\big[(Y - f(X))^2\big].$$

The second part of the proposition should be understood as a non-robustness property of non-causal minimax solutions. Improvements on the causal function are possible in situations, where one has reasons to believe that the test distributions do not stem from a set of interventions that is much larger than the specified set.

So far, the optimizer of the minimax problem (3.3.1) depends on the true model M. In practice, however, we do not have access to the true model M, but only to its observational distribution $\mathbb{P}_M$. This motivates the definition of distribution generalization.

3.4. Distribution generalization

Throughout this section, let $\mathcal{M}$ denote a fixed model class, let $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$ be the true (but unknown) data generating model, with observational distribution $\mathbb{P}_M$, and let $\mathcal{I}$ be a set of interventions on X or A. Depending on the model class $\mathcal{M}$, there may be several models $\tilde{M} \in \mathcal{M}$ that induce the observational distribution $\mathbb{P}_M$ but do not agree with M on all intervention distributions induced by $\mathcal{I}$. Each such model induces a potentially different minimax problem, with a potentially different set of solutions. Given knowledge only of $\mathbb{P}_M$, it is therefore generally not possible to identify a solution to (3.3.1). In this section, we study conditions on $\mathcal{M}$, $\mathbb{P}_M$ and $\mathcal{I}$, under which this becomes possible. More precisely, we aim to characterize under which conditions $(\mathbb{P}_M, \mathcal{M})$ generalizes to $\mathcal{I}$.

Definition 3.1 (distribution generalization)**.** *$(\mathbb{P}_M, \mathcal{M})$ is said to generalize to $\mathcal{I}$ if for every $\varepsilon > 0$ there exists a function $f^* \in \mathcal{F}$ such that, for all models $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, it holds that*

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} \left[(Y - f^*(X))^2 \right] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} \left[(Y - f_\diamond(X))^2 \right] \right| \leq \varepsilon. \tag{3.4.1}$$

Distribution generalization does not require the existence of a minimax solution in $\mathcal{F}$ (which would require further assumptions on the function class $\mathcal{F}$) and instead focuses on whether an approximate

solution can be identified based only on the observational distribution $\mathbb{P}_M$. If, however, there exists a function $f^* \in \mathcal{F}$ which, for every $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, is a minimax solution for $\tilde{M}$ w.r.t. $(\mathcal{F}, \mathcal{I})$, then, in particular, $(\mathbb{P}_M, \mathcal{M})$ generalizes to $\mathcal{I}$. As the next proposition shows, generalization is closely linked to the ability of identifying the joint intervention distributions of (X, Y) from the observational distribution.

Proposition 3.5 (Sufficient conditions for distribution generalization)**.** *Assume that for all $\tilde{M} \in \mathcal{M}$ it holds that*[4]

$$\mathbb{P}_{\tilde{M}} = \mathbb{P}_M \quad \Rightarrow \quad \mathbb{P}^{(X,Y)}_{\tilde{M}(i)} = \mathbb{P}^{(X,Y)}_{M(i)} \quad \forall i \in \mathcal{I},$$

where $\mathbb{P}^{(X,Y)}_{M(i)}$ is the joint distribution of (X, Y) under $M(i)$. Then, $(\mathbb{P}_M, \mathcal{M})$ generalizes to $\mathcal{I}$.

Proposition 3.5 provides verifiable conditions for distribution generalization, and is a useful result for proving possibility statements. It is, however, not a necessary condition. In Propositions 3.6 and 3.8, we give further conditions under which distribution generalization is possible for all well-behaved sets of interventions. In particular, if the set of interventions $\mathcal{I}$ contains at least one confounding-removing intervention it can be shown that the causal function always generalizes, even in cases where the interventional marginal of X is not identified. We will see that distribution generalization is closely linked to the relation between the support of $\mathbb{P}_M$ and the support of the intervention distributions. Below, we therefore distinguish between support-reducing interventions (Section 3.4.1) and support-extending interventions (Section 3.4.2) on X. In Section 3.4.3, we consider interventions on A. We will see that parts of the analysis carry over from the interventions on X.

3.4.1. Support-reducing interventions on X

In order to simplify the following analysis, we will constrain ourselves to cases in which the causal function is identified on the support of X. This condition is made precise in the following assumption.

[4] It is in fact sufficient if the marginal distribution of X, $\mathbb{E}_{\tilde{M}(i)}[Y \mid X]$ and $\mathbb{E}_{\tilde{M}(i)}[Y^2 \mid X]$ remain fixed for all $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$.

Assumption 3.1 (Identifiability of f on the support of X)**.** *For all* $\tilde{M} = (\tilde{f}, \dots) \in \mathcal{M}$ *with* $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$*, it holds that* $\tilde{f}(x) = f(x)$ *for all* $x \in \text{supp}(X)$*.*

Assumption 3.1 concerns identifiability of the causal function from the observational distribution on the support of X. This question has received a lot of attention in literature. In linear instrumental variable settings, for example, one assumes that the functions f and g are linear and the product moment between A and X has rank at least the dimension of X [e.g., Wooldridge, 2010]. In linear non-Gaussian models, one can identify the function f even if there are no instruments [Hoyer et al., 2008]. For nonlinear models, restricted structural causal models can be exploited, too. In that case, Assumption 3.1 holds under regularity conditions if $h_1(H, \varepsilon_Y)$ is independent of X [Zhang and Hyvärinen, 2009, Peters et al., 2014, 2017] and first attempts have been made to extend such results to non-trivial confounding cases [Janzing et al., 2009]. The nonlinear IV setting [e.g., Amemiya, 1974, Newey, 2013, Newey and Powell, 2003] is discussed in more detail in Appendix B.2, where we give a brief overview of identification results for linear, parametric and non-parametric function classes. There is also a technical aspect regarding identifiability: Assumption 3.1 states that f is identifiable, even on $\mathbb{P}_M$-null sets, which is usually achieved by placing further constraints on the function class, such as smoothness. Even though this issue seems technical, it becomes important when considering hard interventions that set X to a fixed value, for example.

Assumption 3.1 is not necessary for generalization. Rothenhäusler et al. [2018] show, for example, that if $\mathcal{F}$ and $\mathcal{G}$ consist of linear functions it is possible to generalize to a set of bounded interventions on A – even if Assumption 3.1 does not hold. If, however, Assumption 3.1 holds, then distribution generalization is possible even in nonlinear settings, under a large class of interventions if these are support-reducing.

Proposition 3.6 (Generalization to support-reducing interventions on X)**.** *Let* $\mathcal{I}$ *be a well-behaved set of interventions on* X*, and assume that* $\text{supp}_{\mathcal{I}}(X) \subseteq \text{supp}(X)$*. Then, under Assumption 3.1,* $(\mathbb{P}_M, \mathcal{M})$ *generalizes to the interventions* $\mathcal{I}$*.*

Proposition 3.6 states that Assumption 3.1 is a sufficient condition for generalization when $\mathcal{I}$ is a well-behaved set of support-reducing interventions. However, for an arbitrary set of interventions, generalization can become impossible, even if Assumption 3.1 is satisfied and all interventions are support-reducing.

3.4.1.1. Impossibility of generalization under changes in confounding

Consider, for example, a one-dimensional linear instrumental variable setting. Let therefore $\mathcal{Q}$ be a class of product distributions on $\mathbb{R}^4$, such that for all $Q \in \mathcal{Q}$, the coordinates of Q are non-degenerate, zero-mean with finite second moment. Let $\mathcal{M}$ be the class of all models of the form

$$A \coloneqq \varepsilon_A, \quad H \coloneqq \sigma\varepsilon_H, \quad X \coloneqq \gamma A + \varepsilon_X + \tfrac{1}{\sigma}H, \quad Y \coloneqq \beta X + \varepsilon_Y + \tfrac{1}{\sigma}H, \tag{3.4.2}$$

with $\gamma, \beta \in \mathbb{R}$, $\sigma > 0$ and $(\varepsilon_A, \varepsilon_X, \varepsilon_Y, \varepsilon_H) \sim Q \in \mathcal{Q}$. Assume that $\mathbb{P}_M$ is induced by some (unknown) model $M = M(\gamma, \beta, \sigma, Q)$ from the above model class (here, we slightly adapt the notation from Section 3.2). The following proposition shows that if the set of interventions $\mathcal{I}$ is not well-behaved, distribution generalization is not always ensured.

Proposition 3.7 (Impossibility of generalization to non-well-behaved interventions)**.** *Assume that $\mathcal{M}$ is given as defined above, and let $\mathcal{I} \subseteq \mathbb{R}_{>0}$ be a compact set of interventions on X defined by $\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = iH$, for $i \in \mathcal{I}$ (this set of interventions is not well-behaved). Then, $(\mathbb{P}_M, \mathcal{M})$ does not generalize to the interventions in $\mathcal{I}$ (even if Assumption 3.1 is satisfied). In addition, any prediction model other than the causal model may perform arbitrarily bad under the interventions $\mathcal{I}$. That is, for any $b \neq \beta$ and any $c > 0$, there exists a model $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, such that*

$$\Big| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - bX)^2\big] - \inf_{b_\diamond \in \mathbb{R}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b_\diamond X)^2\big] \Big| \geq c.$$

3.4.2. Support-extending interventions on X

If the interventions in $\mathcal{I}$ extend the support of X, i.e., $\text{supp}_{\mathcal{I}}(X) \not\subseteq \text{supp}(X)$, Assumption 3.1 is not sufficient for ensuring distribution generalization. This is because there may exist models $\tilde{M} \in \mathcal{M}$ which agree with M on the observational distribution, but whose corresponding causal function $\tilde{f}$ differs from f outside of the support of X. In that case, a support-extending intervention on X may result in different dependencies between X and Y in the two models, and therefore induce a different set of minimax solutions. The following assumption on the model class $\mathcal{F}$ ensures that any $f \in \mathcal{F}$ is uniquely determined by its values on $\text{supp}(X)$.

Assumption 3.2 (Extrapolation of $\mathcal{F}$)**.** *For all $\tilde{f}, \bar{f} \in \mathcal{F}$ with $\tilde{f}(x) = \bar{f}(x)$ for all $x \in \text{supp}(X)$, it holds that $\tilde{f} \equiv \bar{f}$.*

We will see that this assumption is sufficient (Proposition 3.8) for generalization with respect to well-behaved interventions on X. Furthermore, it is also necessary (Proposition 3.11) if $\mathcal{F}$ is sufficiently flexible. The following proposition can be seen as an extension of Proposition 3.6.

Proposition 3.8 (Generalization under support-extending interventions on X)**.** *Let $\mathcal{I}$ be a well-behaved set of interventions on X. Then, under Assumptions 3.1 and 3.2, $(\mathbb{P}_M, \mathcal{M})$ generalizes to $\mathcal{I}$.*

Because the interventions may change the marginal distribution of X, the preceding proposition includes examples, in which distribution generalization is possible even if some of the considered joint (test) distributions are arbitrarily far from the training distribution, in terms of any reasonable divergence measure over distributions, such as Wasserstein distance or f-divergence.

The proposition relies on Assumption 3.2. Even though this assumption is restrictive, it is satisifed by several reasonable function classes, which therefore allow for generalization under any set of well-behaved interventions. Below, we give two examples of such a function class.

3.4.2.1. Sufficient conditions for generalization

Assumption 3.2 states that every function in $\mathcal{F}$ is globally identified by its values on $\text{supp}(X)$. This is, for example, satisfied if $\mathcal{F}$ is a linear space of functions with domain $\mathcal{D} \subseteq \mathbb{R}^d$ which are linearly independent on $\text{supp}(X)$. More precisely,

$$\mathcal{F} \text{ is linearly closed} : f_1, f_2 \in \mathcal{F}, c \in \mathbb{R} \Rightarrow f_1 + f_2, cf_1 \in \mathcal{F}, \text{ and} \tag{3.4.3}$$

$$\mathcal{F} \text{ is lin. ind. on } \text{supp}(X) : f_1(x) = 0 \ \forall x \in \text{supp}(X) \Rightarrow f_1(x) \equiv 0. \tag{3.4.4}$$

Examples of such classes include (i) globally linear parametric function classes, i.e., $\mathcal{F}$ is of the form

$$\mathcal{F}^1 := \{f_\diamond : \mathcal{D} \to \mathbb{R} \mid \text{ there exists } \gamma \in \mathbb{R}^k \text{ s.t. } \forall x \in \mathcal{D} : f_\diamond(x) = \gamma^\top \nu(x)\},$$

where $\nu = (\nu_1, \ldots, \nu_k)$ consists of real-valued, linearly independent functions satisfying that $\mathbb{E}_M[\nu(X)\nu(X)^\top]$ is strictly positive definite, and (ii) the class of differentiable functions that extend linearly outside of $\text{supp}(X)$, that is, $\mathcal{F}$ is of the form

$$\mathcal{F}^2 := \{f_\diamond : \mathcal{D} \to \mathbb{R} \mid f_\diamond \in C^1, \forall x \notin \text{supp}(X) : f_\diamond(x) = f_\diamond(x_b) + \nabla f_\diamond(x_b)(x - x_b)\},$$

where $x_b := \arg\min_{z \in \text{supp}(X)} \|x - z\|$ and $\text{supp}(X)$ is assumed to be closed with non-empty interior. Clearly, both of the above function classes are linearly closed. To see that $\mathcal{F}^1$ satisfies (3.4.4), let $\gamma \in \mathbb{R}^k$ be s.t. $\gamma^\top \nu(x) = 0$ for all $x \in \text{supp}(X)$. Then, it follows that $0 = \mathbb{E}_M[(\gamma^\top \nu(X))^2] = \gamma^\top \mathbb{E}_M[\nu(X)\nu(X)^\top]\gamma$ and hence that $\gamma = 0$. To see that $\mathcal{F}^2$ satisfies (3.4.4), let $f_\diamond \in \mathcal{F}^2$ and assume that $f_\diamond(x) = 0$ for all $x \in \text{supp}(X)$. Then, $f_\diamond(x) = 0$ for all $x \in \mathcal{D}$ and thus $\mathcal{F}^2$ uniquely defines the function on the entire domain $\mathcal{D}$.

By Proposition 3.8, generalization with respect to these model classes is possible for any well-behaved set of interventions. In practice, it may often be more realistic to impose bounds on the higher order derivatives of the functions in $\mathcal{F}$. We now prove that this still allows for approximate distribution generalization, see Propositions 3.9 and 3.10.

3.4.2.2. Sufficient conditions for approximate distribution generalization

For differentiable functions, exact generalization cannot always be achieved. Bounding the first derivative, however, allows us to achieve approximate generalization. We therefore consider the following function class

$$\mathcal{F}^3 := \{f_\diamond : \mathcal{D} \to \mathbb{R} \,|\, f_\diamond \in C^1 \text{ with } \|\nabla f_\diamond\|_\infty \leq K\}, \tag{3.4.5}$$

for some fixed $K < \infty$, where $\nabla f_\diamond$ denotes the gradient and $\mathcal{D} \subseteq \mathbb{R}^d$. We then have the following result.

Proposition 3.9 (Approx. generalization with bdd. derivatives (confounding-removing))**.** *Let $\mathcal{F}$ be as defined in* (3.4.5)*. Let $\mathcal{I}$ be a set of interventions on X containing at least one confounding-removing intervention, and assume that Assumption 3.1 holds true. (In this case, the causal function f is a minimax solution.) Then, for all f^* with $f^* = f$ on* $\operatorname{supp}(X)$ *and all $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, it holds that*

$$\begin{aligned}\Big|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f^*(X))^2\big] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big]\Big| \\ \leq 4\delta^2 K^2 + 4\delta K \sqrt{\operatorname{Var}_M(\xi_Y)},\end{aligned}$$

where $\delta := \sup_{x \in \operatorname{supp}_{\mathcal{I}}^M(X)} \inf_{z \in \operatorname{supp}^M(X)} \|x - z\|$. If $\mathcal{I}$ consists only of confounding-removing interventions, the same statement holds when replacing the bound by $4\delta^2 K^2$.

Proposition 3.9 states that the deviation of the worst-case generalization error from the best possible value is bounded by a term that grows with the square of δ. Intuitively, this means that under the function class defined in (3.4.5), approximate generalization is reasonable only for interventions that are close to the support of X. We now prove a similar result for cases in which the minimax solution is not necessarily the causal function. The following proposition bounds the worst-case generalization error for arbitrary confounding-preserving interventions. Here, the bound additionally accounts for the approximation to the minimax solution.

Proposition 3.10 (Approx. generalization with bdd. derivatives (confounding-preserving))**.** *Let $\mathcal{F}$ be as defined in* (3.4.5)*. Let $\mathcal{I}$ be a set of confounding-preserving interventions on X, and assume that Assumption 3.1 is satisfied. Let $\varepsilon > 0$ and let $f^* \in \mathcal{F}$ be such that*

$$\left|\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f^*(X))^2\right] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right]\right| \leq \varepsilon.$$

Then, for all $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, it holds that

$$\begin{aligned}&\left|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f^*(X))^2\right] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right]\right| \\ &\quad\leq \varepsilon + 12\delta^2K^2 + 32\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} + 4\sqrt{2}\delta K\sqrt{\varepsilon}\end{aligned}$$

where $\delta := \sup_{x\in\mathrm{supp}_{\mathcal{I}}^M(X)} \inf_{z\in\mathrm{supp}^M(X)} \|x - z\|$.

We can take f^* to be the minimax solution if it exists. In that case, the terms involving ε disappear from the bound, which then becomes more similar to the one in Proposition 3.9.

3.4.2.3. Impossibility of generalization without restrictions on $\mathcal{F}$

If we do not constrain the function class $\mathcal{F}$, generalization is impossible. Even if we consider the set of all continuous functions $\mathcal{F}$, we cannot generalize to interventions outside the support of X. This statement holds even if Assumption 3.1 is satisfied.

Proposition 3.11 (Impossibility of extrapolation)**.** *Assume that $\mathcal{F} = \{f_\diamond : \mathbb{R}^d \to \mathbb{R} \mid f_\diamond \text{ is continuous}\}$. Let $\mathcal{I}$ be a well-behaved set of support-extending interventions on X, such that $\mathrm{supp}_{\mathcal{I}}(X)\backslash\mathrm{supp}(X)$ has non-empty interior. Then, $(\mathbb{P}_M, \mathcal{M})$ does not generalize to the interventions in $\mathcal{I}$, even if Assumption 3.1 is satisfied. In particular, for any function $\bar{f} \in \mathcal{F}$ and any $c > 0$, there exists a model $\tilde{M} \in \mathcal{M}$, with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, such that*

$$\left|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - \bar{f}(X))^2\right] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right]\right| \geq c.$$

3.4.3. Interventions on A

We can now derive corresponding results for interventions on A, for which, as we will see, parts of the analysis simplify. We will be able to employ several of the above results by realizing that any intervention on A can be written as an intervention on X, in which the structural assignment of X is altered in a way that depends on the functional relationship g between X and A. The effect of such an intervention on the prediction model is propagated by g. More formally, under such an intervention, a model $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q})$ with $\tilde{g} \neq g$ may induce a distribution over (X, Y) that differs from the one induced by M. Without further restrictions on the function class $\mathcal{G}$, this may happen even in cases where $\tilde{M}$ and M agree on the observational distribution. This motivates an assumption on the identifiability of g.

Assumption 3.3 (Identifiability of g)**.** *For all* $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ *with* $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$*, it holds that* $\tilde{g}(a) = g(a)$ *for all* $a \in \text{supp}(A) \cup \text{supp}_{\mathcal{I}}(A)$.

Since $g(A)$ is a conditional mean for X given A, the values of g are identified from $\mathbb{P}_M$ for $\mathbb{P}_M$-almost all a. If $\text{supp}_{\mathcal{I}}(A) \subseteq \text{supp}(A)$, Assumption 3.3 therefore holds if, for example, $\mathcal{G}$ contains continuous functions only. The point-wise identifiability of g is necessary, for example, if some of the test distributions are induced by hard interventions on A, which set A to some fixed value $a \in \mathbb{R}^r$. In the case where the interventions $\mathcal{I}$ extend the support of A, we additionally require the function class $\mathcal{G}$ to extrapolate from $\text{supp}(A)$ to $\text{supp}(A) \cup \text{supp}_{\mathcal{I}}(A)$; this is similar to the conditions on $\mathcal{F}$ which we made in Section 3.4.2 and requires further restrictions on $\mathcal{G}$. Under Assumption 3.3, we obtain a result corresponding to Propositions 3.6 and 3.8.

Proposition 3.12 (Generalization under interventions on A)**.** *Let* $\mathcal{I}$ *be a set of interventions on* A *and assume Assumption 3.3 is satisfied. Then,* $(\mathbb{P}_M, \mathcal{M})$ *generalizes to* $\mathcal{I}$ *if either* $\text{supp}_{\mathcal{I}}(X) \subseteq \text{supp}(X)$ *and Assumption 3.1 is satisfied or if both Assumptions 3.1 and 3.2 are satisfied.*

3.4.3.1. Impossibility of generalization without constraints on $\mathcal{G}$

Without restrictions on the model class $\mathcal{G}$, generalization to interventions on A is impossible. This holds true even under strong assumptions on the true causal function (such as f is known to be linear). Below, we give a formal impossibility result for hard interventions on A, which set A to some fixed value, where $\mathcal{G}$ is the set of all continuous functions.

Proposition 3.13 (Impossibility to generalize under interventions on A)**.** *Assume that* $\mathcal{F} = \{f_\diamond : \mathbb{R}^d \to \mathbb{R} \,|\, f_\diamond \text{ is linear}\}$ *and* $\mathcal{G} = \{g_\diamond : \mathbb{R}^r \to \mathbb{R}^d \,|\, g_\diamond \text{ is continuous}\}$*. Let* $\mathcal{A} \subseteq \mathbb{R}^r$ *be bounded, and let* $\mathcal{I}$ *denote the set of all hard interventions which set* A *to some fixed value from* $\mathcal{A}$*. Assume that* $\mathcal{A} \setminus \operatorname{supp}(A)$ *has nonempty interior. Assume further that* $\mathbb{E}_M[\xi_X \xi_Y] \neq 0$ *(this excludes the case of no hidden confounding). Then,* $\mathbb{P}_M$ *does not generalize to the interventions in* $\mathcal{I}$*. In addition, any function other than* f *may perform arbitrarily bad under the interventions in* $\mathcal{I}$*. That is, for any* $\bar{f} \neq f$ *and* $c > 0$*, there exists a model* $\tilde{M} \in \mathcal{M}$ *with* $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$ *such that*

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} \left[(Y - \bar{f}(X))^2 \right] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} \left[(Y - f_\diamond(X))^2 \right] \right| \geq c.$$

3.5. Learning generalizing models from data

So far, our focus has been on the possibility to generalize, that is, we have investigated under which conditions it is possible to identify generalizing models from the observational distribution. In practice, generalizing models need to be estimated from finitely many data. This task is challenging for several reasons. First, analytical solutions to the minimax problem (3.3.1) are only known in few cases. Even if generalization is possible, the inferential target thus often remains a complicated object, given as a well-defined but unknown function of the observational distribution. Second, we have seen that the ability to generalize depends strongly on whether the interventions extend the support of X, see Propositions 3.8 and 3.11. In a setting with a finite amount of data, the empirical support of the data lies within some bounded region, and suitable constraints

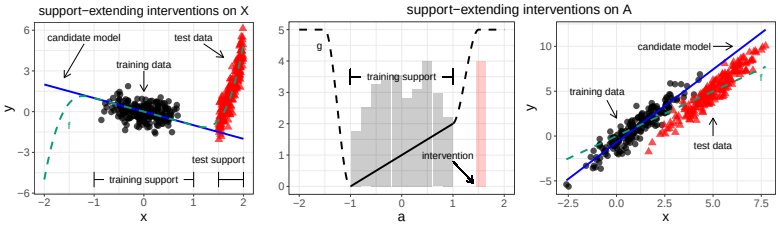


FIGURE 3.1. The left plot illustrates the straight-forward idea behind the impossibility result in Proposition 3.11. The plots in the middle and on the right-hand side illustrate the impossibility result in Proposition 3.13. All plots visualize the case of univariate variables. Under well-behaved interventions on X (left; here using confounding-removing interventions) which extend the support of X, generalization is impossible without further restrictions on the function class $\mathcal{F}$. This holds true even if Assumption 3.1 is satisfied. Indeed, although the candidate model (blue line) coincides with the causal model (green dashed curve) on the support of X, it may perform arbitrarily bad on test data generated under support-extending interventions. Under interventions on A (right and middle), generalization is impossible even under strong assumptions on the function class $\mathcal{F}$ (here, $\mathcal{F}$ is the class of all linear functions). Any support-extending intervention on A shifts the marginal distribution of X by an amount which depends on the (unknown) function g, resulting in a distribution of (X, Y) which cannot be identified from the observational distribution. Without further restrictions on the function class $\mathcal{G}$, any candidate model apart from the causal model may result in arbitrarily large worst-case prediction risk.

on the function class $\mathcal{F}$ are necessary when aiming to achieve empirical generalization outside this region, even if X comes from a distribution with full support. As we show in our simulations in Section 3.5.2.4, constraining the function class can also improve the prediction performance at the boundary of the support.

In Section 3.5.1, we survey existing methods for learning generalizing models. Often, these methods assume either a globally linear model class $\mathcal{F}$ or are completely non-parametric and therefore do not generalize outside the empirical support of the data. Motivated by this observation, we introduce in Section 3.5.2 a novel estimator, which exploits an instrumental variable setup and a particular extrapolation assumption to learn a globally generalizing model.

3.5.1. Existing methods

As discussed in Section 3.1.2, a wide range of methods have been proposed to guard against various types of distributional changes. Here, we review methods that fit into the causal framework in the sense that the distributions that in the minimax formulation the supremum is taken over are induced by interventions.

For well-behaved interventions on X which contain at least one confounding-removing intervention, estimating minimax solutions reduces to the well-studied problem of estimating causal relationships. One class of algorithms for this task is given by linear instrumental variable (IV) approaches. They assume that $\mathcal{F}$ is linear and require identifiability of the causal function (Assumption 3.1) via a rank condition on the observational distribution, see Appendix B.2. Their target of inference is to estimate the causal function, which by Proposition 3.1 will coincide with the minimax solution if the set $\mathcal{I}$ consists of well-behaved interventions with at least one of them being confounding-removing. A basic estimator for linear IV models is the two-stage least squares (TSLS) estimator, which minimizes the norm of the prediction residuals projected onto the subspace spanned by the observed instruments (TSLS objective). TSLS estimators are consistent but do not come with strong finite sample guarantees; e.g., they do not have finite moments in a just-identified setup [e.g., Mariano, 2001]. K-class estimators Theil [1958] have been proposed to overcome some of these issues. They minimize a linear combina-

model class	interventions	$\text{supp}_{\mathcal{I}}(X)$	assumptions	algorithm
$\mathcal{F}$ linear	on X or A (at least one confounding-removing)	–	Ass. 3.1	linear IV (e.g., TSLS, K-class or PULSE Theil [1958], Jakobsen and Peters [2020])
$\mathcal{F}, \mathcal{G}$ linear	on A	bounded strength	–	anchor regression (Rothenhäusler et al. [2018], see also Theil [1958])
$\mathcal{F}$ smooth	on X or A (at least one confounding-removing)	support-reducing	Ass. 3.1	nonlinear IV (e.g., NPREGIV Racine and Hayfield [2018])
$\mathcal{F}$ smooth & linearly extrapolates	on X or A (at least one confounding-removing)	–	Ass. 3.1 & 3.2	**NILE** Section 3.5.2

TABLE 3.2. List of algorithms to learn the generalizing function from data, the considered model class, types of interventions, support under interventions, and additional model assumptions. Sufficient conditions for Assumption 3.1 are given, for example, in the IV literature by generalized rank conditions, see Appendix B.2.

tion of the residual sum of squares (OLS objective) and the TSLS objective. K-class estimators can be seen as utilizing a bias-variance trade-off. For fixed and non-trivial relative weights, they have, in a Gaussian setting, finite moments up to a certain order that depends on the sample-size and the number of predictors used. If the weights are such that the OLS objective is ignored asymptotically, they consistently estimate the causal parameter [e.g., Mariano, 2001]. More recently, PULSE has been proposed [Jakobsen and Peters, 2020], a data-driven procedure for choosing the relative weights such that the prediction residuals 'just' pass a test for simultaneous uncorrelatedness with the instruments.

In cases where the minimax solution does not equal causal func-

tion, only few algorithms exist. Anchor regression [Rothenhäusler et al., 2018] is a procedure that can be used when $\mathcal{F}$ and $\mathcal{G}$ are linear and h_1 is additive in the noise component. It finds the minimax solution if the set $\mathcal{I}$ consists of all interventions on A up to a fixed intervention strength, and is applicable even if Assumption 3.1 is not necessarily satisfied.

In a linear setting, where the regression coefficients differ between different environments, it is also possible to minimize the worst-case risk among the observed environments [Meinshausen and Bühlmann, 2015]. In its current formulation, this approach does not quite fit into the above framework, as it does not allow for changing distributions of the covariates. A summary of the mentioned methods and their assumptions is given in Table 3.2.

If $\mathcal{F}$ is a nonlinear or non-parametric class of functions, the task of finding minimax solutions becomes more difficult. In cases where the causal function is among such solutions, this problem has been studied in the econometrics community. For example, Newey [2013], Newey and Powell [2003] treat the identifiability and estimation of causal functions in non-parametric function classes. Several non-parametric IV procedures exists, e.g., NPREGIV Racine and Hayfield [2018] contains modified implementations of Darolles et al. [2011] and Horowitz [2011]. Identifiability and estimation of the causal function using nonlinear IV methods in parametric function classes is discussed in Appendix B.2. Unlike in the linear case, most of the methods do not aim to extrapolate and only recover the causal function inside the support of X, that is, they cannot be used to predict interventions outside of this domain. In the following section, we propose a procedure that is able to extrapolate when $\mathcal{F}$ consists of functions which extend linearly outside of the support of X. In our simulations, we show that such an assumption can improve the prediction performance on the boundary of the support.

3.5.2. NILE

We have seen in Proposition 3.11 that in order to generalize to interventions which extend the support of X, we require additional assumptions on the function class $\mathcal{F}$. In this section, we start from such assumptions and verify both theoretically and practically that

they allow us to perform distribution generalization in the considered setup. Along the way, several choices can be made and usually several options are possible. We will see that our choices yield a method with competitive performance, but we do not claim optimality of our procedure. Several of our choices were partially made to keep the theoretical exposition simple and the method computationally efficient. We first consider the univariate case (i.e., X and A are real-valued) and comment later on the possibility to extend the methodology to higher dimensions. Unless specific background knowledge is given, it might be reasonable to assume that the causal function extends linearly outside a fixed interval $[a, b]$. By additionally imposing differentiability on $\mathcal{F}$, any function from $\mathcal{F}$ is uniquely defined by its values within $[a, b]$, see also Section 3.4.2.1. Given an estimate f on $[a, b]$, the linear extrapolation property then yields a global estimate on the whole of $\mathbb{R}$. In principle, any class of differentiable functions can be used. Here, we assume that, on the interval $[a, b]$, the causal function f is contained in the linear span of a B-spline basis. More formally, let $B = (B_1, ..., B_k)$ be a fixed B-spline basis on $[a, b]$, and define $\eta := (a, b, B)$. Our procedure assumes that the true causal function f belongs to the function class $\mathcal{F}_\eta := \{f_\eta(\cdot; \theta) : \theta \in \mathbb{R}^k\}$, where for every $x \in \mathbb{R}$ and $\theta \in \mathbb{R}^k$, $f_\eta(x; \theta)$ is given as

$$f_\eta(x;\theta) := \begin{cases} B(a)^\top\theta + B'(a)^\top\theta(x-a) & \text{if } x < a \\ B(x)^\top\theta & \text{if } x \in [a,b] \\ B(b)^\top\theta + B'(b)^\top\theta(x-b) & \text{if } x > b, \end{cases} \tag{3.5.1}$$

where $B' := (B'_1, \ldots, B'_k)$ denotes the component-wise derivative of B. In our algorithm, $\eta = (a, b, B)$ is a hyper-parameter, which can be set manually, or be chosen from data.

3.5.2.1. Estimation procedure

We now introduce our estimation procedure for fixed choices of all hyper-parameters. Section 3.5.2.2 describes how these can be chosen from data in practice. Let $(\mathbf{X}, \mathbf{Y}, \mathbf{A}) \in \mathbb{R}^{n\times 3}$ be n i.i.d. realizations sampled from a distribution over (X, Y, A), let $\eta = (a, b, B)$ be fixed and assume that $\text{supp}(X) \subseteq [a, b]$. Our algorithm aims

to learn the causal function $f_\eta(\cdot;\theta^0) \in \mathcal{F}_\eta$, which is determined by the linear causal parameter θ^0 of a k-dimensional vector of covariates $(B_1(X),\dots,B_k(X))$. From standard linear IV theory, it is known that at least k instrumental variables are required to identify the k causal parameters, see Appendix B.2. We therefore artificially generate such instruments by nonlinearly transforming A, by using another B-spline basis $C = (C_1,\dots,C_k)$. The parameter θ^0 can then be identified from the observational distribution under appropriate rank conditions, see Section 3.5.2.3. In that case, the hypothesis $H_0(\theta) : \theta = \theta^0$ is equivalent to the hypothesis $\tilde{H}_0(\theta) : \mathbb{E}[C(A)(Y - B(X)^\top\theta)] = 0$. Let $\mathbf{B} \in \mathbb{R}^{n\times k}$ and $\mathbf{C} \in \mathbb{R}^{n\times k}$ be the associated design matrices, for each $i \in \{1,\dots,n\}$, $j \in \{1,\dots,k\}$ given as $\mathbf{B}_{ij} = B_j(X_i)$ and $\mathbf{C}_{ij} = C_j(A_i)$. A straightforward choice would be to construct the standard TSLS estimator, i.e., $\hat{\theta}$ as the minimizer of $\theta \mapsto \|\mathbf{P}(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$, where $\mathbf{P}$ is the projection matrix onto the columns of $\mathbf{C}$, see also Hall [2005]. Even though this procedure may result in an asymptotically consistent estimator, there are several reasons why it may be suboptimal in a finite sample setting. First, the above estimator can have large finite sample bias, in particular if k is large. Indeed, in the extreme case where $k = n$, and assuming that all columns in $\mathbf{C}$ are linearly independent, $\mathbf{P}$ is equal to the identity matrix, and $\hat{\theta}$ coincides with the OLS estimator. Second, since θ corresponds to the linear parameter of a spline basis, it seems reasonable to impose constraints on θ which enforce smoothness of the resulting spline function. Both of these points can be addressed by introducing additional penalties into the estimation procedure. Let therefore $\mathbf{K} \in \mathbb{R}^{k\times k}$ and $\mathbf{M} \in \mathbb{R}^{k\times k}$ be the matrices that are, for each $i,j \in \{1,\dots,k\}$, defined as $\mathbf{K}_{ij} = \int B_i''(x)B_j''(x)dx$ and $\mathbf{M}_{ij} = \int C_i''(a)C_j''(a)da$, and let $\gamma,\delta > 0$ be the respective penalties associated with $\mathbf{K}$ and $\mathbf{M}$. For $\lambda \geq 0$ and with $\mu := (\gamma,\delta,C)$, we then define the estimator

$$\hat{\theta}^n_{\lambda,\eta,\mu} := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^k} \|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 + \gamma\theta^\top\mathbf{K}\theta, \quad (3.5.2)$$

where $\mathbf{P}_\delta := \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top$ is the 'hat'-matrix for a penalized regression onto the columns of $\mathbf{C}$. By choice of $\mathbf{K}$, the term $\theta^\top\mathbf{K}\theta$ is equal to the integrated squared curvature of the spline function

parametrized by θ. The above may thus be seen as a nonlinear extension of K-class estimators [Theil, 1958], with an additional penalty term which enforces linear extrapolation. In principle, the above approach extends to situations where X and A are higher-dimensional, in which case B and C consist of multivariate functions. For example, Fahrmeir et al. [2013] propose the use of tensor product splines, and introduce multivariate smoothness penalties based on pairwise first- or second order parameter differences of basis functions which are close-by with respect to some suitably chosen metric. Similarly to (3.5.2), such penalties result in a convex optimization problem. However, due to the large number of involved variables, the optimization procedure becomes computationally burdensome already in small dimensions.

Within the function class $\mathcal{F}_\eta$, the above defines the global estimate $f_\eta(\cdot; \hat{\theta}^n_{\lambda,\eta,\mu})$, for every $x \in \mathbb{R}$ given by

$$f_\eta(x; \hat{\theta}^n_{\lambda,\eta,\mu}) := \begin{cases} B(a)^\top \hat{\theta}^n_{\lambda,\eta,\mu} + B'(a)^\top \theta^n_{\lambda,\eta,\mu}(x-a) & \text{if } x < a \\ B(x)^\top \hat{\theta}^n_{\lambda,\eta,\mu} & \text{if } x \in [a,b] \\ B(b)^\top \theta^n_{\lambda,\eta,\mu} + B'(b)^\top \theta^n_{\lambda,\eta,\mu}(x-b) & \text{if } x > b. \end{cases} \tag{3.5.3}$$

We deliberately distinguish between three different groups of hyperparameters η, μ and λ. The parameter $\eta = (a, b, B)$ defines the function class to which the causal function f is assumed to belong. To prove consistency of our estimator, we require this function class to be correctly specified. In turn, the parameters λ and $\mu = (\gamma, \delta, C)$ are algorithmic parameters that do not describe the statistical model. Their values only affects the finite sample behavior of our algorithm, whereas consistency is ensured as long as C satisfies certain rank conditions, see Assumption (B2) in Section 3.5.2.3. In practice, γ and δ are chosen via a cross-validation procedure, see Section 3.5.2.2. The parameter λ determines the relative contribution of the OLS and TSLS losses to the objective function. To choose λ from data, we use an idea similar to the PULSE [Jakobsen and Peters, 2020].

3.5.2.2. Algorithm

Let for now η, μ be fixed. In the limit $\lambda \to \infty$, our estimation procedure becomes equivalent to minimizing the TSLS loss $\theta \mapsto \|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$, which may be interpreted as searching for the parameter θ which complies 'best' with the hypothesis $\tilde{H}_0(\theta) : \mathbb{E}[C(A)(Y - B(X)^\top\theta)] = 0$. For finitely many data, following the idea introduced in [Jakobsen and Peters, 2020], we propose to choose the value for λ such that $\tilde{H}_0(\hat{\theta}^n_{\lambda,\eta,\mu})$ is just accepted (e.g., at a significance level $\alpha = 0.05$). That is, among all $\lambda \geq 0$ which result in an estimator that is not rejected as a candidate for the causal parameter, we chose the one which yields maximal contribution of the OLS loss to the objective function. More formally, let for every $\theta \in \mathbb{R}^k$, $T(\theta) = (T_n(\theta))_{n\in\mathbb{N}}$ be a statistical test at (asymptotic) level α for $\tilde{H}_0(\theta)$ with rejection threshold $q(\alpha)$. That is, $T_n(\theta)$ does not reject $\tilde{H}_0(\theta)$ if and only if $T_n(\theta) \leq q(\alpha)$. The penalty $\lambda^\star_n$ is then chosen in the following data-driven way

$$\lambda^\star_n := \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda,\eta,\mu}) \leq q(\alpha)\}.$$

In general, $\lambda^\star_n$ is not guaranteed to be finite for an arbitrary test statistic T_n. Even for a reasonable test statistic it might happen that $T_n(\hat{\theta}^n_{\lambda,\eta,\mu}) > q(\alpha)$ for all $\lambda \geq 0$; see Jakobsen and Peters [2020] for further details. We can remedy the problem by reverting to another well-defined and consistent estimator, such as the TSLS (which minimizes the TSLS loss above) if $\lambda^\star_n$ is not finite. Furthermore, if $\lambda \mapsto T_n(\hat{\theta}^n_{\lambda,\eta,\mu})$ is monotonic, $\lambda^\star_n$ can be computed efficiently by a binary search procedure. In our algorithm, the test statistic T and rejection threshold q can be supplied by the user. Conditions on T that are sufficient to yield a consistent estimator $f_\eta(\cdot, \hat{\theta}_{\lambda^\star_n,\mu,\eta})$, given that $\mathcal{F}_\eta$ is correctly specified, are presented in Section 3.5.2.3. Two choices of test statistics which are implemented in our code package can be found in Appendix B.3.

For every $\gamma \geq 0$, let $\mathbf{Q}_\gamma = \mathbf{B}(\mathbf{B}^\top\mathbf{B} + \gamma\mathbf{K})^{-1}\mathbf{B}^\top$ be the 'hat'-matrix for the penalized regression onto $\mathbf{B}$. Our algorithm then proceeds as follows.

Algorithm 2: NILE ("Nonlinear Intervention-robust Linear Extrapolator")

input: data $(\mathbf{X}, \mathbf{Y}, \mathbf{A}) \in \mathbb{R}^{n \times 3}$;
options: k, T, q, α;
begin
  $a \leftarrow \min_i X_i$, $b \leftarrow \max_i X_i$;
  construct cubic B-spline bases $B = (B_1, \ldots, B_k)$ and $C = (C_1, \ldots, C_k)$ at equidistant knots, with boundary knots at respective extreme values of $\mathbf{X}$ and $\mathbf{A}$;
  define $\hat{\eta} \leftarrow (a, b, B)$;
  choose $\delta^n_{\mathrm{CV}} > 0$ by 10-fold CV to minimize the out-of-sample mean squared error of $\hat{\mathbf{Y}} = \mathbf{P}_\delta \mathbf{Y}$;
  choose $\gamma^n_{\mathrm{CV}} > 0$ by 10-fold CV to minimize the out-of-sample mean squared error of $\hat{\mathbf{Y}} = \mathbf{Q}_\gamma \mathbf{Y}$;
  define $\mu^n_{\mathrm{CV}} \leftarrow (\delta^n_{\mathrm{CV}}, \gamma^n_{\mathrm{CV}}, C)$;
  approximate $\lambda^\star_n = \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda, \mu^n_{\mathrm{CV}}, \hat{\eta}}) \leq q(\alpha)\}$ by binary search;
  update $\gamma^n_{\mathrm{CV}} \leftarrow (1 + \lambda^\star_n) \cdot \gamma^n_{\mathrm{CV}}$;
  compute $\hat{\theta}^n_{\lambda^\star_n, \mu^n_{\mathrm{CV}}, \hat{\eta}}$ using (3.5.2);
end
output: $\hat{f}^n_{\mathrm{NILE}} := f_{\hat{\eta}}(\cdot\,; \hat{\theta}^n_{\lambda^\star_n, \mu^n_{\mathrm{CV}}, \hat{\eta}})$ defined by (3.5.3);

The penalty parameter γ^n_{CV} is chosen to minimize the out-of-sample mean squared error of the prediction model $\hat{\mathbf{Y}} = \mathbf{Q}_\gamma \mathbf{Y}$, which corresponds to the solution of (3.5.2) for $\lambda = 0$. After choosing $\lambda^\star_n$, the objective function in (3.5.2) increases by the term $\lambda^\star_n \|\mathbf{P}_{\delta^n_{\mathrm{CV}}}(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$. In order for the penalty term $\gamma\theta^\top \mathbf{K}\theta$ to impose the same degree of smoothness in the altered optimization problem, the penalty parameter γ needs to be adjusted accordingly. The heuristic update in our algorithm is motivated by the simple observation that for all $\delta, \lambda \geq 0$, $\|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda \|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 \leq (1 + \lambda)\|\mathbf{Y} - \mathbf{B}\theta\|_2^2$.

3.5.2.3. Asymptotic generalization (consistency)

We now prove consistency of our estimator in the case where the hyper-parameters (η, μ) are fixed (rather than data-driven), and the

function class $\mathcal{F}_\eta$ is correctly specified. Fix any $a < b$ and a basis $B = (B_1, \ldots, B_k)$. Let $\eta_0 = (a, b, B)$ and let the model class be given by $\mathcal{M} = \mathcal{F}_{\eta_0} \times \mathcal{G} \times \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}$, where $\mathcal{F}_{\eta_0}$ is as described in Section 3.5.2. Assume that the data-generating model $M = (f_{\eta_0}(\cdot\,; \theta^0), g, h_1, h_2, Q) \in \mathcal{M}$ induces an observational distribution $\mathbb{P}_M$ such that $\text{supp}^M(X) \subseteq (a, b)$. Let further $\mathcal{I}$ be a set of interventions on X or A, and let $\alpha \in (0, 1)$ be a fixed significance level.

We prove asymptotic generalization (consistency) for an idealized version of the NILE estimator which utilizes η_0, rather than the data-driven values. Choose any $\delta, \gamma \geq 0$ and basis $C = (C_1, ..., C_k)$ and let $\mu = (\delta, \gamma, C)$. We will make use of the following assumptions.

(B1) $\forall \tilde{M} \in \mathcal{M}$ with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$ it holds that $\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[X^2] < \infty$ and $\sup_{i \in \mathcal{I}} \lambda_{\max}(\mathbb{E}_{\tilde{M}(i)}[B(X)B(X)^\top]) < \infty$.

(B2) $\mathbb{E}_M[B(X)B(X)^\top]$, $\mathbb{E}_M[C(A)C(A)^\top]$ and $\mathbb{E}_M[C(A)B(X)^\top]$ are of full rank.

(C1) $T(\theta)$ has uniform asymptotic power on any compact set of alternatives.

(C2) $\lambda_n^\star := \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda, \eta_0, \mu}) \leq q(\alpha)\}$ is almost surely finite.

(C3) $\lambda \mapsto T_n(\hat{\theta}^n_{\lambda, \eta_0, \mu})$ is weakly decreasing and $\theta \mapsto T_n(\theta)$ is continuous.

Assumptions (B1)–(B2) ensure consistency of the estimator as long as $\lambda_n^\star$ tends to infinity. Intuitively, in this case, we can apply arguments similar to those that prove consistency of the TSLS estimator. Assumptions (C1)–(C3) ensure that consistency is achieved when choosing $\lambda_n^\star$ in the data-driven fashion described in Section 3.5.2.2. In Assumption (B1), $\lambda_{\max}$ denotes the largest eigenvalue. In words, the assumption states that, under each model $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$, there exists a finite upper bound on the variance of any linear combination of the basis functions $B(X)$, uniformly over all distributions induced by $\mathcal{I}$. The first two rank conditions of (B2) enable certain limiting arguments to be valid and they guarantee

that estimators are asymptotically well-defined. The last rank condition of (B2) is the so-called rank condition for identification. It guarantees that θ^0 is identified from the observational distribution in the sense that the hypothesis $H_0(\theta) : \theta = \theta^0$ becomes equivalent with $\tilde{H}_0(\theta) : \mathbb{E}_M[C(A)(Y - B(X)^\top\theta)] = 0$. (C1) means that for any compact set $K \subseteq \mathbb{R}^k$ with $\theta^0 \notin K$ it holds that $\lim_{n\to\infty} P(\inf_{\theta\in K} T_n(\theta) \leq q(\alpha)) = 0$. If the considered test has, in addition, a level guarantee, such as pointwise asympotic level, the interpretation of the finite sample estimator discussed in Section 3.5.2.2 remains valid (such level guarantee may potentially yield improved finite sample performance, too). (C2) is made to simplify the consistency proof. As previously discussed in Section 3.5.2.2, if (C2) is not satisfied, we can output another well-defined and consistent estimator on the event $(\lambda_n^\star = \infty)$, ensuring that consistency still holds.

Under these conditions, we have the following asymptotic generalization guarantee.

Proposition 3.14 (Asymptotic generalization)**.** *Let $\mathcal{I}$ be a set of interventions on X or A of which at least one is confounding-removing. If assumptions (B1)–(B2) and (C1)–(C3) hold true, then, for any $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, and any $\varepsilon > 0$, it holds that*

$$\mathbb{P}_M\Big(\big|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\big] - \inf_{f_\diamond\in\mathcal{F}_{\eta_0}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big]\big| \leq \varepsilon\Big) \to 1,$$

as $n \to \infty$. In the above event, only $\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}$ is stochastic.

3.5.2.4. Experiments

We now investigate the empirical performance of our proposed estimator, the NILE, with $k = 50$ spline basis functions. To choose $\lambda_n^\star$, we use the test statistic T_n^2, which tests the slightly stronger hypothesis $\bar{H}_0$, see Appendix B.3. In all experiments use the significance level $\alpha = 0.05$. We include two other approaches as baseline: (i) the method NPREGIV (using its default options) introduced in

Section 3.5.1, and (ii) a linearly extrapolating estimator of the ordinary regression of Y on X (which corresponds to the NILE with $\lambda^\star \equiv 0$). In all experiments, we generate data sets of size $n = 200$ as independent replications from

$$\begin{aligned} A &:= \varepsilon_A, \quad H := \varepsilon_H, \\ X &:= \alpha_A A + \alpha_H H + \alpha_\varepsilon \varepsilon_X, \\ Y &:= f(X) + 0.3H + 0.2\varepsilon_Y, \end{aligned} \tag{3.5.4}$$

where $(\varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y)$ are jointly independent with $\mathcal{U}(-1, 1)$ marginals. To make results comparable across different parameter settings, we impose the constraint $\alpha_A^2 + \alpha_H^2 + \alpha_\varepsilon^2 = 1$, which ensures that in all models, X has variance $1/3$. The function f is drawn from the linear span of a basis of four natural cubic splines with knots placed equidistantly within the 90% inner quantile range of X. By well-known properties of natural splines, any such function extends linearly outside the boundary knots. Figure 3.2 (left) shows an example data set from (3.5.4), where the causal function is indicated in green. We additionally display estimates obtained by each of the considered methods, based on 20 i.i.d. datasets. Due to the confounding variable H, the OLS estimator is clearly biased. NPREGIV exploits A as an instrumental variable and obtains good results within the support of the observed data. Due to its non-parametric nature, however, it cannot extrapolate outside this domain. The NILE estimator exploits the linear extrapolation assumption on f to produce global estimates.

We further investigate the empirical worst-case mean squared error across several different models of the form (3.5.4). That is, for a fixed set of parameters $(\alpha_A, \alpha_H, \alpha_\varepsilon)$, we construct several models $M_1, \ldots, M_N$ of the form (3.5.4) by randomly sampling causal functions $f_1, \ldots, f_N$ (see Appendix B.4 for further details on the sampling procedure). For every $x \in [0, 2]$, let $\mathcal{I}_x$ denote the set of hard interventions which set X to some fixed value in $[-x, x]$. We then characterize the performance of each method using the average (across different models) worst-case mean squared error (across the

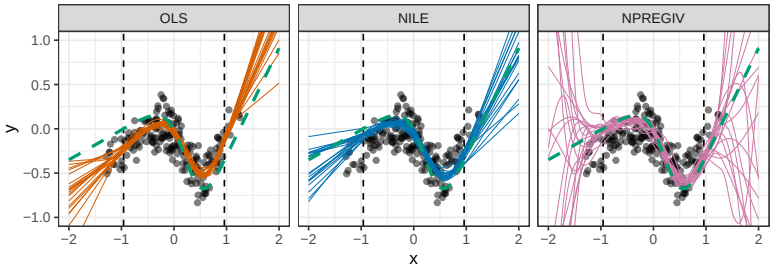


FIGURE 3.2. A sample dataset from the model (3.5.4) with $\alpha_A = \sqrt{1/3}$, $\alpha_H = \sqrt{2/3}$, $\alpha_\varepsilon = 0$. The true causal function is indicated by a green dashed line. For each method, we show 20 estimates of this function, each based on an independent sample from (3.5.4). For values within the support of the training data (vertical dashed lines mark the inner 90% quantile range), NPREGIV correctly estimates the causal function well. As expected, when moving outside the support of X, the estimates become unreliable, and we gain an increasing advantage by exploiting the linear extrapolation assumed by the NILE.

interventions in $\mathcal{I}_x$), i.e., for each estimator $\hat{f}$, we consider

$$\frac{1}{N}\sum_{j=1}^{N} \sup_{i \in \mathcal{I}_x} \mathbb{E}_{M_j(i)}\big[(Y - \hat{f}(X))^2\big] = \mathbb{E}[\xi_Y^2] + \frac{1}{N}\sum_{j=1}^{N} \sup_{\tilde{x} \in [-x,x]} (f_j(\tilde{x}) - \hat{f}(\tilde{x}))^2, \tag{3.5.5}$$

where $\xi_Y := 0.3H + 0.2\varepsilon_Y$ is the noise term for Y (which is fixed across all experiments). In practice, we evaluate the functions $\hat{f}$, $f_1, \ldots, f_N$ on a fine grid on $[-x, x]$ to approximate the above supremum. Figure 3.3 plots the average worst-case mean squared error versus intervention strength for different parameter settings. The optimal worst-case mean squared error $\mathbb{E}[\xi_Y^2]$ is indicated by a green dashed line. The results show that the linear extrapolation property of the NILE estimator is beneficial in particular for strong interventions. In the case of no confounding ($\alpha_H = 0$), the minimax solution coincides with the regression of Y on X, hence even the OLS estimator yields good predictive performance. In this case,

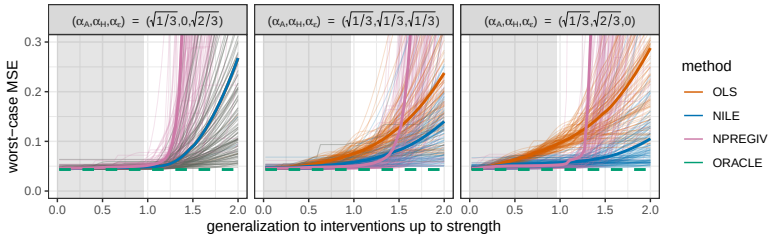


FIGURE 3.3. Predictive performance under confounding-removing interventions on X for different confounding- and intervention strengths (see alpha values in the grey panel on top). The right panel corresponds to the same parameter setting as in Figure 3.2. The plots in each panel are based on data sets of size $n = 200$, generated from $N = 100$ different models of the form (3.5.4). For each model, we draw a different function f, resulting in a different minimax solution (see Appendix B.4 for details on the sampling procedure). The performances under individual models are shown by thin lines; the average performance (3.5.5) across all models is indicated by thick lines. In all considered models, the optimal prediction error is equal to $\mathbb{E}[\xi_Y^2]$ (green dashed line). The grey area indicates the inner 90 % quantile range of X in the training distribution; the white area can be seen as an area of generalization.

the hypothesis $\bar{H}_0(\hat{\theta}^n_{\lambda,\delta^n_{\text{CV}},\gamma^n_{\text{CV}}})$ is accepted already for small values of λ (in this experiment, the empirical average of $\lambda^\star_n$ equals 0.015), and the NILE estimator becomes indistinguishable from the OLS. As the confounding strength increases, the OLS becomes increasingly biased, and the NILE objective function differs more notably from the OLS (average $\lambda^\star_n$ of 2.412 and 5.136, respectively). The method NPREGIV slightly outperforms the NILE inside the support of the observed data, but drops in performance for stronger interventions. We believe that the increase in extrapolation performance of the NILE for stronger confounding (increasing α_H) might stem from the fact that, as the $\lambda^\star_n$ increases, also the smoothness penalty γ increases, see Algorithm 2. While this results in slightly worse in-sample prediction, it seems beneficial for extrapolation (at

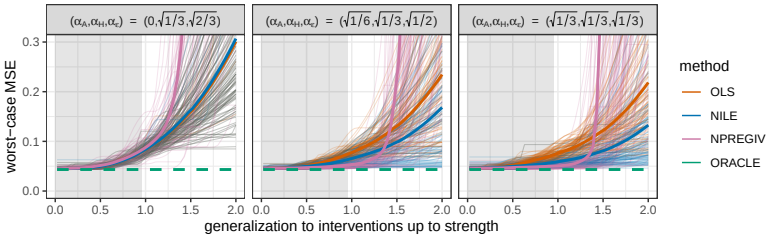


FIGURE 3.4. Predictive performance for varying instrument strength. If the instruments have no influence on X ($\alpha_A = 0$), the second term in the objective function (3.5.2) is effectively constant in θ, and the NILE therefore coincides with the OLS estimator (which uses $\lambda = 0$). This guards the NILE against the large variance which most IV estimators suffer from in a weak instrument setting. For increasing influence of A, it clearly outperforms both alternative methods for large intervention strengths.

least for the particular function class that we consider). We do not claim that our algorithm has theoretical guarantees which explain this increase in performance.

In the case, where all exogenous noise comes from the unobserved variable ε_X (i.e., $\alpha_A = 0$), the NILE coincides with the OLS estimator. In such settings, standard IV methods are known to perform poorly, although also the NPREGIV method seems robust to such scenarios. As the instrument strength increases, the NILE clearly outperforms OLS and NPREGIV for interventions on X which include values outside the training data.

3.6. Discussion and future work

In many real world problems, the test distribution may differ from the training distribution. This requires statistical methods that come with a provable guarantee in such a setting. It is possible to characterize robustness by considering predictive performance for distributions that are close to the training distribution in terms of standard divergences or metrics, such as KL divergences or Wasser-

stein distance. As an alternative view point, we have introduced a novel framework that formalizes the task of distribution generalization when considering distributions that are induced by a set of interventions. Based on the concept of modularity, interventions modify parts of the joint distribution and leave other parts invariant. Thereby, they impose constraints on the changes of the distributions that are qualitatively different from considering balls in a certain metric. As such, we see them as a useful language to describe realistic changes between training and test distributions. Our framework is general in that it allows us to model a wide range of causal models and interventions, which do not need to be known beforehand. We have proved several generalization guarantees, some of which show robustness for distributions that are not close to the training distribution by considering almost any of the standard metrics. We have further proved impossibility results that indicate the limits of what is possible to learn from the training distribution. In particular, in nonlinear models, strong assumptions are required for distribution generalization to a different support of the covariates. As such, methods such as anchor regression cannot be expected to work in nonlinear models, unless strong restrictions are placed on the function class $\mathcal{G}$.

Our work can be extended into several directions. It may, for example, be worthwhile to investigate the sharpness of the bounds we provide in Section 3.4.2.2 and other extrapolation assumptions on $\mathcal{F}$. While our results can be applied to situations where causal background knowledge is available, via a transformation of SCMs, our analysis is deliberately agnostic about such information. It would be interesting to see whether stronger theoretical results can be obtained by including causal background information. Finally, it could be worthwhile to investigate whether NILE, which outperforms existing approaches with respect to extrapolation, can be combined with non-parametric methods. This could yield an even better performance on estimating the causal function within the support of the covariates.

We view our work as a step towards understanding the problem of distribution generalization. We hope that considering the concepts of interventions may help to shed further light into the question

under which assumptions it is possible to generalize knowledge that was acquired during training to a different test distribution.

Acknowledgments

We thank Thomas Kneib for helpful discussions. RC and JP were supported by a research grant (18968) from VILLUM FONDEN.

4 Towards Causal Inference for Spatio-Temporal Data: Conflict and Forest Loss in Colombia

Joint work with

Matthias Baumann, Tobias Kuemmerle, Miguel D. Mahecha and Jonas Peters

Abstract

In many data scientific problems, we are interested not only in modeling the behavior of a system that is passively observed, but also in inferring how the system reacts to changes in the data generating mechanism. Given knowledge of the underlying causal structure, such behavior can be estimated from purely observational data. To do so, one typically assumes that the causal structure of the data generating mechanism can be fully specified. Furthermore, many methods assume that data are generated as independent replications from that mechanism. Both of these assumptions are usually hard to justify in practice: datasets often have complex dependence structures, as is the case for spatio-temporal data, and the full causal structure between all involved variables is hardly known. Here, we present causal models that are adapted to the characteristics of spatio-temporal data, and which allow us to define and quantify causal effects despite incomplete causal background knowledge. We further introduce a simple approach for estimating causal effects, and a non-parametric hypothesis test for these effects being zero. The proposed methods do not rely on any distributional assumptions on the data, and allow for arbitrarily many latent confounders, given that these confounders do not

vary across time (or, alternatively, they do not vary across space). Our theoretical findings are supported by simulations and code is available online. This work has been motivated by the following real-world question: how has the Colombian conflict influenced tropical forest loss? There is evidence for both enhancing and reducing impacts, but most literature analyzing this problem is not using formal causal methodology. When applying our method to data from 2000 to 2018, we find a reducing but insignificant causal effect of conflict on forest loss. Regionally, both enhancing and reducing effects can be identified.

4.1. Introduction

4.1.1. Spatio-temporal data analysis

In principle, all data are spatio-temporal data: Any observation of any phenomenon occurs at a particular point in space and time. If information on the spatio-temporal origin of data are available, this information can be exploited for statistical modeling in various ways; this is the study of spatio-temporal statistics [see, e.g., Sherman, 2011, Montero et al., 2015, Cressie and Wikle, 2015, Wikle et al., 2019]. Spatio-temporal statistical models find their application in many environmental and sustainability sciences, and have been used, for example, for the analysis of biological growth patterns [Chaplain et al., 1999], to identify hotspots of species co-occurrence [Ward et al., 2015], to model meteorological fields [Bertolacci et al., 2019], or to assess the development of land-use change [Liu et al., 2017] and sea level rise [Zammit-Mangion et al., 2015]. They are frequently used in epidemiology for prevalence mapping of infectious diseases [Giorgi et al., 2018], and have also been applied in the social sciences, for example, for the modeling of housing prices [Holly et al., 2010], or for election forecasting [Pavía et al., 2008]. In almost all of these domains, the abundance of spatio-temporal data has increased rapidly over the last decades. Several advances aim to improve the accessibility of such datasets, e.g., via 'data cube' approaches [Nativi et al., 2017, Giuliani et al., 2019, Appel and Pebesma, 2019, Mahecha et al., 2020].

Most spatio-temporal statistical models are models for the observational distribution, that is, they model processes that are passively observed. By allowing for spatio-temporal trends and dependence structures, such models can be accurate descriptions of complex processes, and have proven to be effective tools for spatio-temporal prediction (i.e., filtering and smoothing), inference and forecasting [Wikle et al., 2019]. However, to answer interventional questions such as "How does a certain policy change affect land-use patterns?", we require a model for the intervention distribution, that is, for data generated under a change in the data generating mechanism — we require a causal model for the data generating process.

4.1.2. Causality

For i.i.d. and time series data, that is, for data, for which the spatial information can be neglected, causal models have been well-studied. Among the most widely used approaches are structural causal models, causal graphical models, and the framework of potential outcomes [see, e.g., Bollen, 1989, Pearl, 2009, Peters et al., 2017, Rubin, 1974]. Knowledge of the causal structure of a system does not only provide us with cause-effect relationships; sometimes, it also allows us to quantify causal relations by estimating intervention effects from observational data. If, for example, we know that W is causing X and Y, and that X is causing Y, procedures such as variable adjustment can be used to estimate the causal influence of X on Y from i.i.d. replications from the model [Pearl, 2009, Rubin, 1974]. While using a slightly different language, the same underlying causal deliberations are the basis of many works in econometrics, e.g., work related to generalized methods of moments and identifiability of parameters [e.g., Hansen, 1982, Newey and McFadden, 1994]. In this field, data are often assumed to have a time series structure [e.g., Hall, 2005].

Existing causal models for i.i.d. or time series data do not apply easily to a spatio-temporal setting, since we cannot regard a spatio-temporal dataset as a collection of independent replications from some random vector or timeseries generated from the same underlying causal system. Nevertheless, several methods have been proposed for spatio-temporal causal modeling [e.g., Lozano et al., 2009, Luo et al., 2013, Zhu et al., 2017]. These are mostly algorithmical approaches that extend the concept of Granger causality [Granger, 1980, Wiener, 1956] to spatio-temporal data. They reduce the question of causality to predictability and a positive time lag. In particular, these methods assume that there are no relevant unobserved variables ('confounders') and they do not resolve the question of time-instantaneous causality between different points in space. Further, to the best of our knowledge, existing work does not provide a formal model for causality for spatio-temporal data. As a consequence, the precise definition of the target of inference, the causal effect, remains vague.

In this work, we introduce a class of causal models for multivari-

ate spatio-temporal stochastic processes. A spatio-temporal dataset may then be viewed as a single realization from such a model, observed at discrete points in space and time. The full causal structure among all variables of a spatio-temporal process can hardly be fully specified. In practice, however, a full causal specification may also not be necessary: we are often interested in quantifying only certain causal relationships, while being indifferent to other parts of the causal structure. The introduced causal models are well adapted to such settings. They allow us to model a causal influence of a vector of covariates X on a target variable Y while leaving other parts of the causal structure unspecified. In particular, the models accommodate largely unspecified autocorrelation patterns in the response variable, which are a common phenomena in spatio-temporal data.

The introduced framework allows us to formally talk about causality in a spatio-temporal context and can be used to construct well-defined targets of inference. As an example, we define the intervention effect ('causal effect') of X on Y. We show that this effect can be estimated from observational spatio-temporal data and introduce a corresponding estimator. We further construct a non-parametric hypothesis test for the effect being zero. Our methods do not rely on any distributional assumptions on the data generating process. They further allow for the influence of arbitrarily many latent confounders if these confounders do not vary across time. In principle, our method also allows to analyze problems where temporal and spatial dimensions are interchanged, meaning that confounders may vary in time but remain static across space.

Our work has been motivated by the following application.

4.1.3. Conflict and forest loss in Colombia

Tropical forests show the highest values of biodiversity for many organismic groups, e.g., in terms of vascular plants [Kreft and Jetz, 2007], or certain animal groups like amphibians [Hof et al., 2011]. Additionally, contiguous low-land tropical forests store large amounts of carbon [Avitabile et al., 2016], play an important role in climate-regulation, and provide livelihoods to millions of people [Lambin and Meyfroidt, 2011]. Yet, tropical forest loss remains a major global environmental problem, as many of these areas continue to be under

pressure due to agricultural expansion [Carlson et al., 2013, Angelsen and Kaimowitz, 1999], legal and illicit mining [Sonter et al., 2017], timber harvest [Pearson et al., 2014] or urban expansion [DeFries et al., 2010].

A problem that is still only partly understood is the interaction between forest loss and armed conflicts [Baumann and Kuemmerle, 2016], which are frequent events in tropical areas [Gleditsch et al., 2002, Pettersson and Wallensteen, 2015]. In particular, it has been reported that armed conflict may have both positive and negative impacts on forest loss. On the one hand, conflict can lead to increasing pressure on forests, as (illegal) timber exports may allow for financing warfare activities [Harrison, 2015]. Also, reduced law enforcement in conflict regions may lead to plundering natural resources or undertaking illegal mining activities, altogether leading to increasing forest loss [Irland, 2008, Butsic et al., 2015]. On the other hand, the outbreak of armed conflicts can also reduce the pressure on forest resources. This may happen, for example, when economic and political insecurity interrupt large-scale mining activities, or economic sanctions stopping international timber trade [Le Billon, 2000]. Investors may further be hesitant to invest in agricultural activities [Collier et al., 2000], thereby reducing the pressure on forest areas compared to peace times [Gorsevski et al., 2012]. In a global overview, Gaynor et al. [2016] call for a regional nuanced analysis of such interactions.

Along these lines, we here focus on the specific case of Colombia, where an armed conflict has been present for over 50 years, causing more than 200,000 fatalities, until a peace agreement was reached in 2016. Throughout this period, a variety of interacting social and political factors have regionally led to internal migration and changes in livelihoods and land-use that are all related to the overall conflict [Armenteras et al., 2013]. There is evidence that forest loss can be, at least regionally, attributed to the armed conflict [Castro-Nunez et al., 2017, Landholm et al., 2019]. At the same time, there are also arguments suggesting that the pressure on forests was partially reduced when armed conflict prevented logging [Dávalos et al., 2016], either directly (by demanding human resources) or indirectly (e.g., due to land-abandonment in the wake of local conflicts [Sánchez-

Cuervo et al., 2012, Negret et al., 2017]). Most papers report evidence that both positive and negative impacts of conflict on forest loss may happen in parallel, depending on the local conditions [e.g., Sánchez-Cuervo and Aide, 2013, Castro-Nunez et al., 2017]. Latest evidence suggests, however, that forest loss has increased substantially after the initiation of the peace process in protected areas [Armenteras et al., 2019, Clerici et al., 2020]. We believe that a purely data-driven approach can be a useful addition to this debate.

In our analysis, we use a spatio-temporal dataset containing measurements of the following variables.

- X_s^t : binary conflict indicator for location s at year t.
- Y_s^t : absolute forest loss in location s from year $t-1$ to year t, measured in square kilometers.
- W_s^t : distance from location s to the closest road, measured in kilometers.

Data are annually aggregated, covering the years from 2000 to 2018, and spatially explicit at a $10\text{km} \times 10\text{km}$-resolution. We provide a detailed description of the data processing in Section 4.4. A summary of the dataset can be seen in Figure 4.1. Visually, there is a strong positive dependence between the occurrence of a conflict and the loss of forest canopy. This observation is supported by simple summary statistics: the average forest loss across measurements classified as conflict events exceeds that from non-conflict events by almost 50%; a difference that is declared highly significant by a standard t-test (Figure 4.1, left). The strong signficance of the statistical dependence between forest loss and conflict has been reported before [e.g., Landholm et al., 2019]. When seeking a causal explanation for the observed data, however, we regard such an analysis as flawed in two ways. First, both conflicts and forest loss predominantly occur in areas with good transport infrastructure (Figure 4.1, right), indicating that the potential causal effect of X on Y is confounded by W. In fact, we expect the existence of several other confounders (e.g., population density, market infrastructure, mining operations, cocaine plantations, etc.), many of which may be unobserved. Failing to account for confounding variables leads to biased estimates of the causal effect. Second, strong spatial dependencies in X and Y

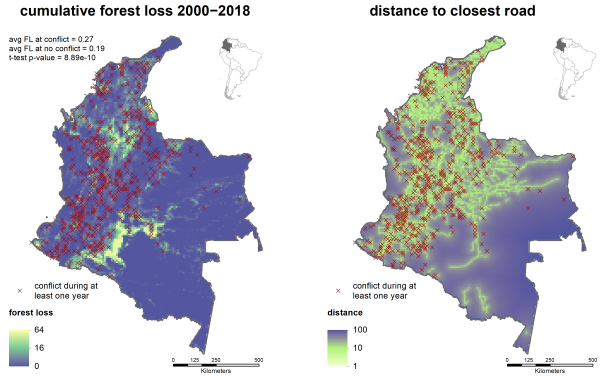


FIGURE 4.1. Temporally aggregated summary of the dataset described in Section 4.1.3. For visual purposes, the above color scales are square root- and $\log_{10}$-transformed, respectively. Conflicts are predictive of exceedances in forest loss (left), but this dependence is partly induced by a common dependence on transport infrastructure, which we measure by the mean distance to a road (right). Failing to account for this variable and other confounders biases our estimate of the causal influence of conflict on forest loss. Also, since both conflicts and forest loss exhibit complicated spatial dependence patterns, the independence assumptions underlying a standard two-sample t-test are likely to be violated. To correctly assess statements of statistical significance, we need a test which acknowledges the spatial dependence in the data.

reduce the effective sample size, and a standard t-test thus exaggerates the significance of the observed difference in sample averages. To test hypotheses about X and Y, we need statistical tests which are adapted to the spatio-temporal nature of data.

4.1.4. Contributions and structure of the paper

Apart from the case study, this paper contains three main theoretical contributions: the definition of a causal model for spatio-temporal data, a method for estimating causal effects, and a hypothesis test for the overall existence of such effects. Our class of

causal models is introduced in Section 4.2. It translates the situation of a vector of covariates X that causally affect a real-valued response variable Y into a spatio-temporal setting. It allows for arbitrary influences of latent confounders, as long as these confounders do not vary across time. It further accommodates largely unspecified spatio-temporal dependence structures in the data. Within our model class, we conceptually define the causal effect of X and Y, propose an estimator of this quantity, and prove consistency. This finding is supported by a simulation study. In Section 4.3, we introduce a non-parametric hypothesis test for the overall existence of a causal effect, and prove that this test obtains valid level in finite samples. Section 4.4 applies our methodology to the above example. All data used for our analysis are publicly available. A description of how it can be obtained, along with an implementation of our method and reproducing scripts for all our figures and results, can be found at `github.com/runesen/spatio_temporal_causality`. All our proofs are contained in Appendix C.2.

4.2. Quantifying causal effects for spatio-temporal data

A spatio-temporal dataset may be viewed as a single realization of a spatio-temporal stochastic process, observed at discrete points in space and time. In this section, we provide a formal framework to quantify causal relationships among the components of a multivariate spatio-temporal process. This framework is presented in Section 4.2.1. In Section 4.2.2, we define the class of latent spatial confounder models (LSCMs) which will be the subject of study in this work, Section 4.2.3 shows how to estimate causal effects within this model class, and Section 4.2.4 discusses several extensions of our methodology.

4.2.1. Causal models for spatio-temporal processes

Throughout this section, let $(\Omega, \mathcal{A}, P)$ be some background probability space. A p-dimensional *spatio-temporal process* $\mathbf{Z}$ is a ran-

dom variable taking values in the sample space $\mathcal{Z}_p$ of all $(\mathcal{B}(\mathbb{R}^2 \times \mathbb{N}), \mathcal{B}(\mathbb{R}^p))$-measurable functions, where $\mathcal{B}(\cdot)$ denotes the Borel σ-algebra. We equip $\mathcal{Z}_p$ with the σ-algebra $\mathcal{F}_p$, defined as the smallest σ-algebra such that for all $B \in \mathcal{B}(\mathbb{R}^2 \times \mathbb{N})$, the mapping $\mathcal{Z}_p \ni z \mapsto \int_B z(x)dx$ is $(\mathcal{F}_p, \mathcal{B}(\mathbb{R}^p))$-measurable. The induced probability measure $\mathbb{P}$ on the measurable space $(\mathcal{Z}_p, \mathcal{F}_p)$, for every $F \in \mathcal{F}_p$ defined by $\mathbb{P}(F) := P(\mathbf{Z}^{-1}(F))$, is said to be the *distribution of* $\mathbf{Z}$. Throughout this paper, we use the notation Z_s^t to denote the random vector obtained from marginalizing $\mathbf{Z}$ at spatial location s and temporal instance t. We use $\mathbf{Z}_s$ for the time series $(Z_s^t)_{t\in\mathbb{N}}$, $\mathbf{Z}^t$ for the spatial process $(Z_s^t)_{s\in\mathbb{R}^2}$, and $\mathbf{Z}^{(S)}$ for the spatio-temporal process corresponding to the coordinates in $S \subseteq \{1, \ldots, p\}$. We call a spatio-temporal process *weakly stationary* if the marginal distribution of Z_s^t is the same for all $(s,t) \in \mathbb{R}^2 \times \mathbb{N}$, and *time-invariant* if $\mathbb{P}(\mathbf{Z}^1 = \mathbf{Z}^2 = \cdots) = 1$.

Multivariate spatio-temporal processes are used for the joint modeling of different phenomena, each of which corresponds to a coordinate process. Let us consider a decomposition of these coordinate processes into disjoint 'bundles'. We are interested in specifying causal relations among these bundles while leaving the causal structure among variables within each bundle unspecified. Similarly to a graphical model [Lauritzen, 1996], our approach relies on a factorization of the joint distribution of $\mathbf{Z}$ into a number of components, each of which models the conditional distribution for one bundle given several others. This approach induces a graphical relation among the different bundles. We will equip these relations with a causal interpretation by additionally specifying the distribution of $\mathbf{Z}$ under certain interventions on the data generating process. More formally, we have the following definition.

Definition 4.1 (Causal graphical models for spatio-temporal processes)**.** *A causal graphical model for a p-dimensional spatio-temporal process* $\mathbf{Z}$ *is a triplet* $(\mathcal{S}, \mathcal{G}, \mathcal{P})$ *consisting of*

- *a family* $\mathcal{S} = (S_j)_{j=1}^k$ *of non-empty, disjoint sets* $S_1, \ldots, S_k \subseteq \{1, \ldots, p\}$ *with* $\bigcup_{j=1}^k S_j = \{1, \ldots, p\}$,
- *a directed acyclic graph* $\mathcal{G}$ *with vertices* $S_1, \ldots, S_k$*, and*

- *a family* $\mathcal{P} = (\mathcal{P}^j)_{j=1}^k$ *of collections* $\mathcal{P}^j = \{\mathbb{P}^j_z\}_{z \in \mathcal{Z}_{|\mathrm{PA}_j|}}$ *of distributions on* $(\mathcal{Z}_{|S_j|}, \mathcal{F}_{|S_j|})$*, where for every* j*,* $\mathrm{PA}_j := \bigcup_{i:S_i \to S_j \in \mathcal{G}} S_i$*. Whenever* $\mathrm{PA}_j = \emptyset$*,* $\mathcal{P}^j$ *consists only of a single distribution which we denote by* $\mathbb{P}^j$*.*

Since $\mathcal{G}$ *is acyclic, we can without loss of generality assume that* $S_1, \ldots, S_k$ *are indexed such that* $S_i \not\to S_j$ *in* $\mathcal{G}$ *whenever* $i > j$*. The above components induce a unique joint distribution* $\mathbb{P}$ *over* $\mathbf{Z}$*. For every* $F = \bigtimes_{j=1}^k F_j$*, it is defined by*

$$\mathbb{P}(F) = \int_{F_1} \cdots \int_{F_k} \mathbb{P}^k_{z^{(\mathrm{PA}_k)}}(dz^{(S_k)}) \cdots \mathbb{P}^1(dz^{(S_1)}). \tag{4.2.1}$$

We call $\mathbb{P}$ *the* observational distribution*. For each* $j \in \{1, \ldots, k\}$*, the conditional distribution of* $\mathbf{Z}^{(S_j)}$ *given* $\mathbf{Z}^{(\mathrm{PA}_j)}$ *as induced by* $\mathbb{P}$ *equals* $\mathcal{P}^j$*. We define an* intervention on $\mathbf{Z}^{(S_j)}$ *as replacing* $\mathcal{P}^j$ *by another model* $\tilde{\mathcal{P}}^j$*. This operation results in a new graphical model* $(\mathcal{S}, \mathcal{G}, \tilde{\mathcal{P}})$ *for* $\mathbf{Z}$ *which induces, via* (4.2.1)*, a new distribution* $\tilde{\mathbb{P}}$*, the* interventional distribution*.*

Assume that we perform an intervention on $\mathbf{Z}^{(S_i)}$. By definition, the resulting interventional distribution differs from the observational distribution only in the way in which $\mathbf{Z}^{(S_i)}$ depends on $\mathbf{Z}^{(\mathrm{PA}_i)}$, while all other conditional distributions $\mathbf{Z}^{(S_j)} \,|\, \mathbf{Z}^{(\mathrm{PA}_j)}$, $j \neq i$, remain the same. This property is analogous to the modularity property of structural causal models [Haavelmo, 1944, Aldrich, 1989, Pearl, 2009, Peters et al., 2017] and justifies a causal interpretation of the conditionals in $\mathcal{P}$. We refer to the graph $\mathcal{G}$ as the *causal structure of* $\mathbf{Z}$, and sometimes write $\mathbf{Z} = [\mathbf{Z}^{(S_k)} \,|\, \mathbf{Z}^{(\mathrm{PA}_k)}] \cdots [\mathbf{Z}^{(S_1)}]$ to emphasize this structure.

4.2.2. Latent spatial confounder model

Motivated by the example on conflict and forest loss introduced in Section 4.1.3, we are particularly interested in scenarios where a target variable Y is causally influenced by a vector of covariates X, and where (X, Y) are additionally affected by some latent variables H. In general, inferring causal effects under arbitrary influences of

latent confounders is impossible, and we therefore need to impose additional restrictions on the variables in H. We here make the fundamental assumption that they do not vary across time (alternatively, one can assume that the hidden variables are invariant over space, see Section 4.2.4.3).

Definition 4.2 (Latent spatial confounder model)**.** *Consider a spatio-temporal process* $(\mathbf{X}, \mathbf{Y}, \mathbf{H}) = (X_s^t, Y_s^t, H_s^t)_{(s,t)\in\mathbb{R}^2\times\mathbb{N}}$ *over a real-valued response* Y*, a vector of covariates* $X \in \mathbb{R}^d$ *and a vector of latent variables* $H \in \mathbb{R}^\ell$*. We call a causal graphical model over* $(\mathbf{X}, \mathbf{Y}, \mathbf{H})$ *with causal structure* $[\mathbf{Y} \,|\, \mathbf{X}, \mathbf{H}][\mathbf{X} \,|\, \mathbf{H}][\mathbf{H}]$ *a* latent spatial confounder model (LSCM) *if both of the following conditions hold true for the observational distribution.*

- *The latent process* $\mathbf{H}$ *is weakly stationary and time-invariant.*
- *There exists a function* $f : \mathbb{R}^{d+\ell+1} \to \mathbb{R}$ *and an i.i.d. sequence* $\boldsymbol{\varepsilon}^1, \boldsymbol{\varepsilon}^2, \ldots$ *of weakly-stationary spatial error processes, independent of* $(\mathbf{X}, \mathbf{H})$*, such that*

$$Y_s^t = f(X_s^t, H_s^t, \varepsilon_s^t) \qquad \text{for all } (s,t) \in \mathbb{R}^2 \times \mathbb{N}. \tag{4.2.2}$$

Throughout this section, we assume that $(\mathbf{X}, \mathbf{Y}, \mathbf{H})$ come from an LSCM. The above definition says that for every s, t, Y_s^t depends on $(\mathbf{X}, \mathbf{H})$ only via (X_s^t, H_s^t), and that this dependence remains the same for all points in space-time. Together with the weak stationarity of $\mathbf{H}$ and $\boldsymbol{\varepsilon}$, this assumption ensures that the average causal effect of X_s^t on Y_s^t (which we introduce below) remains the same for all s, t. Our model class imposes no restrictions on the marginal distribution of $\mathbf{X}$. The spatial dependence structure of the error process $\boldsymbol{\varepsilon}$ must have the same marginal distributions everywhere, but is otherwise unspecified (in particular, $\boldsymbol{\varepsilon}$ is not required to be stationary). The temporal independence assumption on $\boldsymbol{\varepsilon}$ is necessary for our construction of resampling tests, see Section 4.3. We now formally define our inferential target.

Definition 4.3 (Average causal effect)**.** *The* average causal effect *of* $\mathbf{X}$ *on* $\mathbf{Y}$ *is defined as the function* $f_{\mathrm{AVE}(X\to Y)} : \mathbb{R}^d \to \mathbb{R}$*, for every* $x \in \mathbb{R}^d$ *given by*

$$f_{\mathrm{AVE}(X\to Y)}(x) := \mathbb{E}[f(x, H_0^1, \varepsilon_0^1)]. \tag{4.2.3}$$

Here, the causal effect is an average effect in that it takes the expectation over both the noise variable (as opposed to making counterfactual statements [Rubin, 1974]) and the hidden variables (see also Remark 4.1). The following proposition justifies $f_{\mathrm{AVE}(X\to Y)}$ as a quantification of the causal influence of $\mathbf{X}$ on $\mathbf{Y}$.

Proposition 4.1 (Causal interpretation)**.** *Let $(s,t) \in \mathbb{R}^2 \times \mathbb{N}$ and $x \in \mathbb{R}^d$ be fixed, and consider any intervention on $\mathbf{X}$ such that $X_s^t = x$ holds almost surely in the induced interventional distribution $\mathbb{P}_x$. We then have that*

$$\mathbb{E}_{\mathbb{P}_x}[Y_s^t] = f_{\mathrm{AVE}(X\to Y)}(x),$$

i.e., $f_{\mathrm{AVE}(X\to Y)}(x)$ is the expected value of Y_s^t under any intervention that enforces $X_s^t = x$.

In many practical applications, we do not have explicit knowledge of, or data from, the interventional distributions $\mathbb{P}_x$. If we have access to the causal graph, however, we can sometimes compute intervention effects from the observational distribution. In the i.i.d. setting, depending on which variables are observed, this can be done by covariate adjustment or G-computation [Pearl, 2009, Rubin, 1974, Shpitser et al., 2010], for example. The following proposition shows a similar result in the case of a latent spatial confounder model. It follows directly from Fubini's theorem.

Proposition 4.2 (Covariate adjustment)**.** *Let $f_{Y|(X,H)}$ denote the regression function $(x,h) \mapsto \mathbb{E}[Y_s^t \mid X_s^t = x, H_s^t = h]$ (by definition of an LSCM, this function is the same for all s,t). For all $x \in \mathbb{R}^d$, it holds that*

$$f_{\mathrm{AVE}(X\to Y)}(x) = \mathbb{E}[f_{Y|(X,H)}(x, H_0^1)]. \tag{4.2.4}$$

Proposition 4.2 shows that $f_{\mathrm{AVE}(X\to Y)}$ is identified from the full observational distribution over $(\mathbf{X}, \mathbf{Y}, \mathbf{H})$ (given that the LSCM structure is known). Since $\mathbf{H}$ is unobserved, the main challenge is to estimate (4.2.4) merely based on data from $(\mathbf{X}, \mathbf{Y})$, see Section 4.2.3. (We discuss in Section 4.2.4.1 how to further include observable covariates that are allowed to vary over time.)

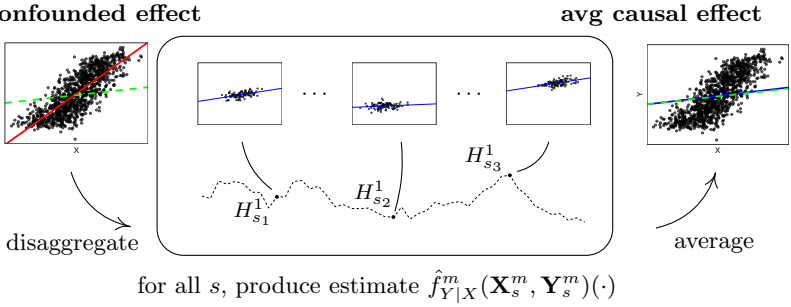


FIGURE 4.2. Conceptual idea for estimating the average causal effect (green dashed line) defined in (4.2.3). In both the left and right panel, we do not display the information of time and space. The figure in the middle shows the data at different locations $s \in \{s_1, \ldots, s_n\}$, i.e., every small plot corresponds to a single time series. The dashed curve illustrates the unobserved realization of the spatial confounder H^t_s (for visual purposes, we here consider one-dimensional space). Due to this confounder, regressing Y^t_s on X^t_s (red line in left plot) leads to a biased estimator of the average causal effect. By exploiting the time-invariance of H^t_s, our estimator (blue line in right plot) removes this bias. This procedure is illustrated in the middle figure: at every location s, we observe several time instances (X^t_s, Y^t_s), $t = 1, \ldots, m$, with the same conditionals $Y^t_s \,|\, (X^t_s, H^t_s)$ and the same (unobserved) value of H^t_s. For each realization h_s of H^1_s, we can therefore estimate the regression $f_{Y|(X,H)}(\cdot, h_s)$ only using the data $(\mathbf{X}^m_s, \mathbf{Y}^m_s)$ (blue lines in middle figure). A final estimate of the average causal effect (blue line in right plot) is obtained by approximating the expectation in (4.2.5) by a sample average over all spatial locations. We make this approach precise in Section 4.2.3.

Remark 4.1 (An alternative definition of the average causal effect)**.** *In our definition of the average causal effect* (4.2.3)*, we take the expectation with respect to the hidden variables H. By the assumption of time-invariance, however, there is only a single replication of the spatial process $\mathbf{H}^1$. One may argue that it is more relevant to define the inferential target in terms of that one realization, rather than in terms of a distribution over possible alternative outcomes which will never manifest themselves. This leads to the alternative definition of the average causal effect*

$$\begin{aligned} x \mapsto & \lim_{S\to\infty} \frac{1}{(2S)^2} \int_{[-S,S]^2} \mathbb{E}[f(x, h_s^1, \varepsilon_0^1)]\, ds \\ = & \lim_{S\to\infty} \frac{1}{(2S)^2} \int_{[-S,S]^2} f_{Y|(X,H)}(x, h_s^1)\, ds, \end{aligned}$$

assuming that the above limits exist. Under the assumption of ergodicity of $\mathbf{H}^1$, the above expression coincides with Definition 4.3, but it is learnable from data, via the estimator introduced in Section 4.2.3, even if this is not the case.[1] *Here, we choose the formulation in Definition 4.3 because we found that it results in a more comprehensible theory.*

4.2.3. Estimation of the average causal effect

4.2.3.1. Definition and consistency

In practice, we only observe the process $(\mathbf{X}, \mathbf{Y})$ at a finite number of points in space and time. We assume that at every temporal instance, we observe the process at the same spatial locations $s_1, \ldots, s_n \in \mathbb{R}^2$ (these locations need not lie on a regular grid). To simplify notation, we further assume that the observed time points are $t = 1, 2, \ldots, m$, i.e., we have access to a dataset $(\mathbf{X}_n^m, \mathbf{Y}_n^m) = (X_s^t, Y_s^t)_{(s,t)\in\{s_1,\ldots,s_n\}\times\{1,\ldots,m\}}$. The proposed method is based on the following key idea: for every $s \in \{s_1, \ldots, s_n\}$, we observe several time instances (X_s^t, Y_s^t), $t \in \{1, \ldots, m\}$, all with the same conditionals $Y_s^t \,|\, (X_s^t, H_s^t)$. Since $\mathbf{H}$ is time-invariant, we can, for every s, estimate $f_{Y|(X,H)}(\cdot, h_s)$ for the (unobserved) realization

[1] We are grateful to Steffen Lauritzen for emphasizing this viewpoint.

h_s of H_s^1 using the data (X_s^t, Y_s^t), $t \in \{1, \dots, m\}$. The expectation in (4.2.4) is then approximated by averaging estimates obtained from different spatial locations. This idea is visualized in Figure 4.2. More formally, our method requires as input a model class for the regressions $f_{Y|(X,H)}(\cdot, h)$, $h \in \mathbb{R}^\ell$, alongside with a suitable estimator $\hat{f}_{Y|X} = (\hat{f}_{Y|X}^m)_{m\in\mathbb{N}}$, and returns

$$\hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(\mathbf{X}_n^m, \mathbf{Y}_n^m)(x) := \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{Y|X}^m(\mathbf{X}_{s_i}^m, \mathbf{Y}_{s_i}^m)(x), \qquad (4.2.5)$$

an estimator of the average causal effect (4.2.3) within the given model class. In Section 4.3, we further provide a statistical test for the overall existence of a causal effect. Our approach may be seen as summarizing the output of a spatially varying regression model [e.g., Gelfand et al., 2003] that is allowed to change arbitrarily from one location to the other (within the model class dictated by $\hat{f}_{Y|X}$). By permitting such flexibility, our method does not rely on observing data from a continuous or spatially connected domain, and accommodates complex influences of the latent variables. An implementation can be found in our code package, see Section 4.1.4.

To prove consistency of our estimator, we let the number of observable points in space-time increase. Let therefore $(s_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^2$ be a sequence of spatial coordinates, and consider the array of data $(\mathbf{X}_n^m, \mathbf{Y}_n^m)_{n,m\geq 1}$, where for every $n, m \in \mathbb{N}$, $(\mathbf{X}_n^m, \mathbf{Y}_n^m) = (X_s^t, Y_s^t)_{(s,t)\in\{s_1,\dots,s_n\}\times\{1,\dots,m\}}$. We want to prove that the corresponding sequence of estimators (4.2.5) consistently estimates (4.2.3). To obtain such a result, we need two central assumptions.

Assumption 4.1 (Law of Large Numbers for the latent process)**.** *For all measurable functions $\varphi : \mathbb{R}^\ell \to \mathbb{R}$ with $\mathbb{E}[|\varphi(H_0^1)|] < \infty$ it holds that $\frac{1}{n}\sum_{i=1}^n \varphi(H_{s_i}^1) \to \mathbb{E}[\varphi(H_0^1)]$ in probability as $n \to \infty$.*

The above assumption ensures that, for an increasing number of spatial locations at which data are observed, the spatial average in (4.2.5) approximates the expectation in (4.2.3). As the number of observed time points tends to infinity, we require the estimators $\hat{f}_{Y|X}^m$ to converge to the integrand in (4.2.3), at least in some area $\mathcal{X} \subseteq \mathbb{R}^d$.

Assumption 4.2 (Consistent estimators of the conditional expectations)**.** *There exists $\mathcal{X} \subseteq \mathbb{R}^d$ s.t. for all $x \in \mathcal{X}$ and $s \in \mathbb{R}^2$, it holds that $\hat{f}^m_{Y|X}(\mathbf{X}^m_s, \mathbf{Y}^m_s)(x) - f_{Y|(X,H)}(x, H^1_s) \to 0$ in probability as $m \to \infty$.*

A slightly stronger, but maybe more intuitive formulation is to require the above consistency to hold conditionally on $\mathbf{H}$, i.e., assuming that for all $x \in \mathcal{X}$, $s \in \mathbb{R}^2$ and almost all $\mathbf{h}$, $\hat{f}^m_{Y|X}(\mathbf{X}^m_s, \mathbf{Y}^m_s)(x) \to f_{Y|(X,H)}(x, h^1_s)$ as $m \to \infty$, in probability under $\mathbb{P}(\cdot \mid \mathbf{H} = \mathbf{h})$. It follows from the dominated convergence theorem that this assumption implies Assumption 4.2.

Under Assumptions 4.1 and 4.2, we obtain the following consistency result.

Theorem 4.2 (Consistent estimator of the average causal effect)**.** *Let $(\mathbf{X}, \mathbf{Y}, \mathbf{H})$ come from an LSCM as defined in Definition 4.2. Let $(s_n)_{n\in\mathbb{N}}$ be a sequence of spatial coordinates such that the marginalized process $(H^1_{s_n})_{n\in\mathbb{N}}$ satisfies Assumption 4.1, and assume that for all $x \in \mathcal{X}$, $\mathbb{E}[|f_{Y|(X,H)}(x, H^1_0)|] < \infty$. Let furthermore $\hat{f}_{Y|X} = (\hat{f}^m_{Y|X})_{m\in\mathbb{N}}$ be an estimator satisfying Assumption 4.2. We then have the following consistency result. For all $x \in \mathcal{X}$, $\delta > 0$ and $\alpha > 0$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$ we can find $M_n \in \mathbb{N}$ such that for all $m \geq M_n$ we have that*

$$\mathbb{P}\left(\left|\hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(\mathbf{X}^m_n, \mathbf{Y}^m_n)(x) - f_{\mathrm{AVE}(X\to Y)}(x)\right| > \delta\right) \leq \alpha. \quad (4.2.6)$$

Apart from the LSCM structure, the above result does not rely on any particular distributional properties of the data. Assumptions 4.1 and 4.2 do not impose strong restrictions on the data generating process and hold true for several model classes, including linear and nonlinear models. Below, we provide sufficient conditions under which these assumptions are true.

4.2.3.2. Sufficient conditions for Assumptions 4.1 and 4.2

For Assumption 4.1, we consider a stationary Gaussian setup. By considering a regular spatial sampling scheme, we can make use of standard ergodic theorems for stationary and ergodic time series.

Proposition 4.3 (Sufficient conditions for Assumption 4.1)**.** *Assume that* $\mathbf{H}^1$ *is a stationary multivariate Gaussian process with covariance function* $C : \mathbb{R}^2 \to \mathbb{R}^{\ell\times\ell}$*, i.e.,* $C(h) = \mathrm{Cov}(H^1_s, H^1_{s+h})$ *for all* $s, h \in \mathbb{R}^2$*. Assume that* $C(h) \to 0$ *entrywise as* $\|h\|_2 \to \infty$*. Consider a regular grid* $\{s^1_1, s^1_2, \dots\} \times \{s^2_1, s^2_2, \dots, s^2_m\} \subseteq \mathbb{R}^2$*, where* $s^1_1 < s^1_2 < \cdots$ *are equally spaced, and let* $(s_n)_{n\in\mathbb{N}}$ *be the spatial sampling scheme for every* $i \in \mathbb{N}$ *and* $j \in \{1, \dots, m\}$ *given as* $s_{(i-1)m+j} = (s^1_i, s^2_j)$*. Then, the process* $(H^1_{s_n})_{n\in\mathbb{N}}$ *satisfies Assumption 4.1.*

For Assumption 4.2, we consider the slightly stronger version formulated conditionally on $\mathbf{H}$. We let $\mathcal{H} \subseteq \mathcal{Z}_\ell$ denote the set of all functions $\mathbf{h} : \mathbb{R}^2 \times \mathbb{N} \to \mathbb{R}^\ell$ that are constant in the time-argument. Since $\mathbf{H}$ is time-invariant, we have that $\mathbb{P}(\mathbf{H} \in \mathcal{H}) = 1$, and it therefore suffices to prove the statement for all $\mathbf{h} \in \mathcal{H}$. Below, we use, for every $\mathbf{h} \in \mathcal{H}$, $\mathbb{P}_\mathbf{h}$ to denote the conditional distribution $\mathbb{P}(\cdot \mid \mathbf{H} = \mathbf{h})$ and $\mathbb{E}_\mathbf{h}$ for the expectation with respect to $\mathbb{P}_\mathbf{h}$.

We now make some structural assumptions on the function f in (4.2.2), which allow us to parametrically estimate the regressions $x \mapsto f_{Y|(X,H)}(x, h)$. Let $\{\varphi_1, \dots, \varphi_p\}$ be a known basis of continuous functions on $\mathbb{R}^d$, and with $\varphi_1 \equiv 1$ an intercept term. With $\varphi := (\varphi_1, \dots, \varphi_p)$, we make the following assumptions on the underlying LSCM.

(L1) There exist functions $f_1 : \mathbb{R}^\ell \to \mathbb{R}^p$ and $f_2 : \mathbb{R}^{\ell+1} \to \mathbb{R}$ such that Equation (4.2.2) splits into

$$Y^t_s = \varphi(X^t_s)^\top f_1(H^t_s) + f_2(H^t_s, \varepsilon^t_s) \qquad \text{for all } (s,t) \in \mathbb{R}^2 \times \mathbb{N},$$

and such that for all $h \in \mathbb{R}^\ell$, $f_2(h, \varepsilon^1_0)$ has finite second moment.

For every s, t, define $\xi^t_s = f_2(H^t_s, \varepsilon^t_s)$. We can w.l.o.g. assume that for all s, t and h, $\mathbb{E}[\xi^t_s \mid H^t_s = h] = 0$. (Since $\varphi_1 \equiv 1$, this can always be accommodated by adding $h \mapsto \mathbb{E}[\xi^t_s \mid H^t_s = h]$ to the first coordinate of f_1.) For every fixed $\mathbf{h} \in \mathcal{H}$ and $s \in \mathbb{R}^2$, assumption (L1) says that, under $\mathbb{P}_\mathbf{h}$, $(\mathbf{X}_s, \mathbf{Y}_s)$ follows a simple regression model, where $\mathbb{E}[Y^t_s \mid X^t_s]$ depends linearly on $\varphi(X^t_s)$. For arbitrary but fixed h^1_s, we can therefore estimate $x \mapsto \mathbb{E}[Y^1_s \mid X^1_s = x, H^1_s = h^1_s]$ using standard

OLS estimation. For every $s \in \mathbb{R}^2$ and $m \in \mathbb{N}$, let $\mathbf{\Phi}_s^m \in \mathbb{R}^{m \times p}$ be the design matrix given by $(\mathbf{\Phi}_s^m)_{ij} = \varphi_j(X_s^i)$. We define an estimator $\hat{f}_{Y|X} = (\hat{f}_{Y|X}^m)_{m \in \mathbb{N}}$, for every $x \in \mathbb{R}^d$ and $m \in \mathbb{N}$ by

$$\hat{f}_{Y|X}^m(\mathbf{X}_s^m, \mathbf{Y}_s^m)(x) = \varphi(x)^\top \hat{\gamma}_s^m, \tag{4.2.7}$$

where $\hat{\gamma}_s^m := ((\mathbf{\Phi}_s^m)^\top \mathbf{\Phi}_s^m)^{-1} (\mathbf{\Phi}_s^m)^\top \mathbf{Y}_s^m$. To formally prove consistency of (4.2.7), we need some regularity conditions on the predictors $\mathbf{X}$.

(L2) For all $\mathbf{h} \in \mathcal{H}$, $s \in \mathbb{R}^2$ and $\delta > 0$, it holds that

$$\lim_{m \to \infty} \mathbb{P}_{\mathbf{h}}(\| \tfrac{1}{m} (\mathbf{\Phi}_s^m)^\top \boldsymbol{\xi}_s^m \|_2 > \delta) = 0.$$

(L3) For all $\mathbf{h} \in \mathcal{H}$, $s \in \mathbb{R}^2$, there exists $c > 0$ such that

$$\lim_{m \to \infty} \mathbb{P}_{\mathbf{h}}(\lambda_{\min} \left(\tfrac{1}{m} (\mathbf{\Phi}_s^m)^\top \mathbf{\Phi}_s^m \right) \leq c) = 0,$$

where $\lambda_{\min}$ denotes the minimal eigenvalue.

We first state the result and discuss assumptions (L2) and (L3) afterwards.

Proposition 4.4 (Sufficient conditions for Assumption 4.2)**.** *Assume that* $(\mathbf{X}, \mathbf{Y}, \mathbf{H})$ *come from an LSCM satisfying (L1)–(L3). Then, Assumption 4.2 is satisfied with* $\mathcal{X} = \mathbb{R}^d$ *and with* $\hat{f}_{Y|X}^m$ *as defined in* (4.2.7)*.*

Since for every $(s,t) \in \mathbb{R}^2 \times \mathbb{N}$ and $\mathbf{h} \in \mathcal{H}$, X_s^t and ξ_s^t are independent under $\mathbb{P}_{\mathbf{h}}$ with $\mathbb{E}_{\mathbf{h}}[\xi_s^t] = 0$, (L2) states a natural LLN-type condition, which is satisfied under suitable constraints on the temporal dependence structure in $\mathbf{X}$, and on its variance. Assumption (L3) says that, with probability tending to one, the matrix $\frac{1}{m}(\mathbf{\Phi}_s^m)^\top \mathbf{\Phi}_s^m$ is bounded away from singularity as $m \to \infty$. This is in particular satisfied if $\frac{1}{m}(\mathbf{\Phi}_s^m)^\top \mathbf{\Phi}_s^m$ converges in probability entrywise to some matrix which is strictly positive definite. In Appendix C.1, we give two examples in which this is the case.

4.2.3.3. An example LSCM

To illustrate the consistency result in Theorem 4.2, we now consider a simple example with one covariate ($d = 1$) and two hidden variables ($\ell = 2$).

Example 4.1 (Latent Gaussian process and a linear average causal effect)**.** *Let $\boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\xi}^t, \boldsymbol{\varepsilon}^t$, $t \in \mathbb{N}$, be independent versions of a univariate stationary spatial Gaussian process with mean 0 and covariance function $u \mapsto e^{-\frac{1}{2}\|u\|_2}$. For notational simplicity, let $\bar{\mathbf{H}}$ and $\tilde{\mathbf{H}}$ denote the respective first and second coordinate process of $\mathbf{H}$. We define a marginal distribution over $\mathbf{H}$ and conditional distributions $\mathbf{X} \,|\, \mathbf{H}$ and $\mathbf{Y} \,|\, (\mathbf{X}, \mathbf{H})$ by specifying that for all $(s,t) \in \mathbb{R}^2 \times \mathbb{N}$,*

$$H_s^t = (\bar{H}_s^t, \tilde{H}_s^t) = (\zeta_s, 1 + \tfrac{1}{2}\zeta_s + \tfrac{\sqrt{3}}{2}\psi_s),$$
$$X_s^t = \exp(-\|s\|_2^2/1000) + (0.2 + 0.1 \cdot \sin(2\pi t/100)) \cdot \bar{H}_s^t \cdot \tilde{H}_s^t + 0.5 \cdot \xi_s^t,$$
$$Y_s^t = (1.5 + \bar{H}_s^t \cdot \tilde{H}_s^t) \cdot X_s^t + (\bar{H}_s^t)^2 + |\tilde{H}_s^t| \cdot \varepsilon_s^t.$$

Interventions on $\mathbf{X}$, $\mathbf{Y}$ or $\mathbf{H}$ are defined as in Definition 4.1. In this LSCM, the average causal effect $f_{\mathrm{AVE}(X \to Y)}$ is the linear function $x \mapsto \beta_0 + \beta_1 x$, with $\beta_0 := \mathbb{E}[(\bar{H}_0^1)^2] = 1$ and $\beta_1 := 1.5 + \mathbb{E}[\bar{H}_0^1 \cdot \tilde{H}_0^1] = 2$. We define a spatial sampling scheme $(s_i)_{i \in \mathbb{N}}$ for every $j \in \mathbb{N}$ and $k \in \{1, \ldots, 25\}$ by $s_{25 \cdot (j-1)+k} = (j,k)$. Given a sample $(\mathbf{X}_n^m, \mathbf{Y}_n^m) = (X_s^t, Y_s^t)_{(s,t) \in \{s_1, \ldots, s_n\} \times \{1, \ldots, m\}}$ from $(\mathbf{X}, \mathbf{Y})$, we construct an estimator of $f_{\mathrm{AVE}(X \to Y)}$ by

$$\hat{f}^{nm}_{\mathrm{AVE}(X \to Y)}(\mathbf{X}_n^m, \mathbf{Y}_n^m)(x) = \frac{1}{n} \sum_{i=1}^{n} (1 \;\; x)\, \hat{\beta}^m_{\mathrm{OLS}}(\mathbf{X}_{s_i}^m, \mathbf{Y}_{s_i}^m), \qquad (4.2.8)$$

where $\hat{\beta}^m_{\mathrm{OLS}}(\mathbf{X}_{s_i}^m, \mathbf{Y}_{s_i}^m) \in \mathbb{R}^2$ is the OLS estimator for the linear regression at spatial location s_i, that is of $\mathbf{Y}_{s_i}^m = (Y_{s_i}^1, \ldots, Y_{s_i}^m)$ on $\mathbf{X}_{s_i}^m = (X_{s_i}^1, \ldots, X_{s_i}^m)$ (we assume that the regression includes an intercept term). It follows by Propositions 4.3 and 4.4 (see in particular Example C.2 and Remark C.2 in Appendix C.1) that Assumptions 4.1 and 4.2 are satisfied.[2] *Hence,* (4.2.8) *is a consistent estimator of $f_{\mathrm{AVE}(X \to Y)}$.*

[2] Strictly speaking, Example C.2 and Remark C.2 show that (L1)–(L3) are satisfied for bounded basis functions. We are confident that the same holds true in the current example.

Figure 4.3 shows results from a numerical experiment based on Example 4.1. The left panel shows the simulated dataset, the plot in the middle represents our method, and the right panel illustrates that the estimator is consistent. More details are provided in the figure caption. The example shows that we can estimate causal effects even under complex influences of the latent process $\mathbf{H}$. To construct the estimator $\hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}$, we have used that the influence of $(\mathbf{X}, \mathbf{H})$ on $\mathbf{Y}$ is linear in $\mathbf{X}$. It is worth noting, however, that we do not assume knowledge of the particular functional dependence of $\mathbf{Y}$ on $\mathbf{H}$; we obtain consistency under any influence of the form $Y^t_s = f_1(H^t_s) \cdot X^t_s + f_2(H^t_s, \varepsilon^t_s)$, see Proposition 4.4.

4.2.4. Extensions

4.2.4.1. Observed confounders

For simplicity, we have until now assumed that the only confounders of (X, Y) are the variables in H. Our method naturally extends to settings with observed (time- and space-varying) confounders. Let $W \in \mathbb{R}^p$ be a vector of observed covariates, and consider a causal graphical model over $(\mathbf{X}, \mathbf{W}, \mathbf{Y}, \mathbf{H})$ with causal structure $[\mathbf{Y} \,|\, \mathbf{X}, \mathbf{W}, \mathbf{H}][\mathbf{X} \,|\, \mathbf{W}, \mathbf{H}][\mathbf{W}, \mathbf{H}]$. Similarly to Definition 4.2, assume that $\mathbf{W}$ and $\mathbf{H}$ are weakly stationary, $\mathbf{H}$ is time-invariant, and there exists a function $f : \mathbb{R}^{d+p+\ell+1} \to \mathbb{R}$ and an i.i.d. sequence $\boldsymbol{\varepsilon}^1, \boldsymbol{\varepsilon}^2, \ldots$ of weakly stationary error processes, independent of $(\mathbf{X}, \mathbf{W}, \mathbf{H})$, such that

$$Y^t_s = f(X^t_s, W^t_s, H^t_s, \varepsilon^t_s) \quad \text{for all } (s,t) \in \mathbb{R}^2 \times \mathbb{N}. \tag{4.2.9}$$

We define the average causal effect of $\mathbf{X}$ on $\mathbf{Y}$, for every $x \in \mathbb{R}^d$, by

$$f_{\mathrm{AVE}(X\to Y)}(x) = \mathbb{E}[f(x, W^1_0, H^1_0, \varepsilon^1_0)].$$

It is straight-forward to show that this function enjoys the same causal interpretation as given in Proposition 4.1. Similarly to Proposition 4.2, we have that for all $x \in \mathbb{R}^d$, it holds that

$$f_{\mathrm{AVE}(X\to Y)}(x) = \mathbb{E}[f_{Y|(X,W,H)}(x, W^1_0, H^1_0)],$$

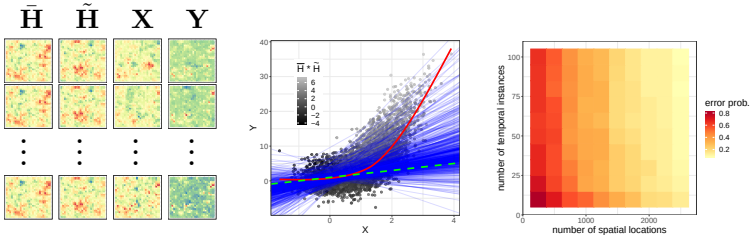


FIGURE 4.3. Results for applying our methodology to the LSCM in Example 4.1. The left panel shows a sample dataset of the process $(\mathbf{X}, \mathbf{Y}, \mathbf{H})$ observed at the spatial grid $\{1, \ldots, 25\} \times \{1, \ldots, 25\} \subseteq \mathbb{R}^2$ and at several temporal instances. For the sake of illustration, each square has its own colorscale, and colors are therefore not comparable across plots. The middle panel illustrates the output from our method applied to the same dataset. The average causal effect, our inferential target, is indicated by a dashed green line. Due to confounding by $\mathbf{H}$, a standard nonlinear regression (red curve) severely overestimates the causal influence of $\mathbf{X}$ on $\mathbf{Y}$. By estimating the dependence between $\mathbf{X}$ and $\mathbf{Y}$ in each spatial location separately (thin blue lines), and aggregating the results into a final estimate (thick blue line), all spatial confounding is removed. In the right panel, we investigate the consistency result from Theorem 4.2 empirically. For increasing numbers of spatial locations n (shown on the x-axis) and temporal instances m (shown on the y-axis), we generate several datasets $(\mathbf{X}^m_{n,i}, \mathbf{Y}^m_{n,i})$, $i = 1, \ldots, 100$, compute estimates $\hat{\beta}^{nm}_i$ of the causal coefficients β, and use these to compute empirical error probabilities $\hat{\mathbb{P}}(\|\hat{\beta}^{nm} - \beta\|_2 > \delta)$. In the above plot, we have chosen $\delta = 0.2$. As n and m increase, the error probability tends towards zero.

where $f_{Y|(X,W,H)}$ is the regression function of Y_s^t onto (X_s^t, W_s^t, H_s^t). As an estimator for the case where $\mathbf{H}$ remains unobserved, we then use

$$\hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(\mathbf{X}_n^m, \mathbf{Y}_n^m, \mathbf{W}_n^m)(x) := \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}_W[\hat{f}^m_{Y|(X,W)}(\mathbf{X}_{s_i}^m, \mathbf{Y}_{s_i}^m, \mathbf{W}_{s_i}^m)(x, W_0^1)],$$

where $\hat{\mathbb{E}}_W$ is the expectation w.r.t. some estimate of the marginal distribution of W.

4.2.4.2. Temporally lagged causal effects

We can incorporate temporally lagged causal effects by allowing the function f in (4.2.2) to depend on past values of the predictors. That is, we model a joint causal influence of the past $k \geq 1$ temporal instances of the predictors by assuming the existence of a function $f : \mathbb{R}^{d\cdot k+\ell+1} \to \mathbb{R}$ and an i.i.d. sequence $\boldsymbol{\varepsilon}^1, \boldsymbol{\varepsilon}^2, \ldots$ such that

$$Y_s^t = f(X_s^{t-k+1}, \ldots, X_s^t, H_s^t, \varepsilon_s^t) \quad \text{for all } s \in \mathbb{R}^2 \text{ and } t \geq k.$$

In this case, the average causal effect (4.2.3) is a function $\mathbb{R}^{d\cdot k} \to \mathbb{R}$ which can be estimated, similarly to (4.2.5), using a regression estimator $\hat{f}^m_{Y|X}$ of Y_s^t onto $(X_s^{t-k+1}, \ldots, X_s^t)$.

4.2.4.3. Exchanging the role of space and time

We have assumed that the hidden confounders do not vary across time. This assumption allowed us to estimate the regression $x \mapsto \mathbb{E}[Y_s^t \mid X_s^t = x, H_s^t = h]$ at all unobserved values h. In fact, our method can be formulated in more general terms. If $(\mathbf{X}, \mathbf{Y}, \mathbf{H})$ is a multivariate process defined on some general, possibly random, index set $\mathcal{I} = I_1 \times \cdots \times I_p$ (see the definition of a data cube [Mahecha et al., 2020]), it is enough to require $\mathbf{H}$ to be invariant across one (or several) of the dimensions in $\mathcal{I}$. Similarly to (4.2.5), the idea is then to estimate the dependence of $\mathbf{Y}$ on $(\mathbf{X}, \mathbf{H})$ along these invariant dimensions, followed by an aggregation across the remaining dimensions. In case of a spatio-temporal process, for example, our method also applies if the hidden variables are constant across space, rather than time.

4.3. Testing for the existence of causal effects

The previous section has been concerned with the quantification and estimation of the causal effect of X on Y. In this section, we introduce hypothesis tests for this effect being non-zero. We consider the null hypothesis

$$H_0 : \begin{cases} (\mathbf{X}, \mathbf{Y}) \text{ come from an LSCM with a function } f \\ \text{that is constant with respect to } X_s^t, \end{cases}$$

which formalizes the assumption of "no causal influence of X on Y" within the framework of LSCMs. We construct a non-parametric hypothesis test for H_0 using data resampling. Our approach acknowledges the existence of spatial dependence in the data without modeling it explicitly. It thus does not rely on distributional assumptions apart from the LSCM structure.

For the construction of a resampling test, we closely follow the setup presented in Pfister et al. [2018]. We require a data permutation scheme which, under the null hypothesis, leaves the distribution of the data unaffected. In particular, it must preserve the dependence between $\mathbf{X}$ and $\mathbf{Y}$ that is induced by $\mathbf{H}$. The idea is to permute observations of $\mathbf{Y}$ corresponding to the same (unobserved) values of $\mathbf{H}$. Since $\mathbf{H}$ is assumed to be constant within every spatial location, this is achieved by permuting $\mathbf{Y}$ along the time axis. Let $(\mathbf{X}_n^m, \mathbf{Y}_n^m)$ be the observed data. For every $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{(d+1)\times n\times m}$ and every permutation σ of the elements in $\{1, \dots, m\}$, let $\sigma(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{(d+1)\times n\times m}$ denote the permuted array with entries $(\sigma(x, y))_s^t = (x_s^t, y_s^{\sigma(t)})$. We then have the following exchangeability property.

Proposition 4.5 (Exchangeability)**.** *For any permutation σ of the elements in $\{1, \dots, m\}$, we have that, under H_0,*

$$\sigma(\mathbf{X}_n^m, \mathbf{Y}_n^m) \text{ is equal in distribution to } (\mathbf{X}_n^m, \mathbf{Y}_n^m).$$

Proposition 4.5 is the cornerstone for the construction of a valid resampling test. Under the null hypothesis, we can compute pseudo-replications of the observed sample $(\mathbf{X}_n^m, \mathbf{Y}_n^m)$ using the permutation

scheme described above. Given any test statistic $\hat{T} : \mathbb{R}^{(d+1)\times n\times m} \to \mathbb{R}$, we obtain a p-value for H_0 by comparing the value of $\hat{T}$ calculated on the original dataset with the empirical null distribution of $\hat{T}$ obtained from the resampled datasets. The choice of $\hat{T}$ determines the power of the test. More formally, let $M := m!$ and let $\sigma_1, \dots, \sigma_M$ be all permutations of the elements in $\{1, \dots, m\}$. By Proposition 4.5, each of these permutations yields a new dataset with the same distribution as $(\mathbf{X}_n^m, \mathbf{Y}_n^m)$. Let $B \in \mathbb{N}$ and let $k_1, \dots, k_B$ be independent, uniform draws from $\{1, \dots, M\}$. For every $(\mathbf{x}, \mathbf{y})$, we define

$$p_{\hat{T}}(\mathbf{x}, \mathbf{y}) := \frac{1 + |\{b \in \{1, \dots, B\} : \hat{T}(\sigma_{k_b}(\mathbf{x}, \mathbf{y})) \geq \hat{T}(\mathbf{x}, \mathbf{y})\}|}{1 + B},$$

and construct for every $\alpha \in (0, 1)$ a test $\varphi_{\hat{T}}^{\alpha} : \mathbb{R}^{(d+1)\times n\times m} \to \{0, 1\}$ of H_0 defined by $\varphi_{\hat{T}}^{\alpha} = 1 :\Leftrightarrow p_{\hat{T}} \leq \alpha$.[3] The following level guarantee for $\varphi_{\hat{T}}^{\alpha}$ follows directly from [Pfister et al., 2018, Proposition B.4].

Corollary 4.1 (Level guarantee of resampling test)**.** *Assume that for all $k, \ell \in \{1, \dots, B\}$, $k \neq \ell$, it holds that, under H_0,*

$$\mathbb{P}(\hat{T}(\sigma_k(\mathbf{X}_n^m, \mathbf{Y}_n^m)) = \hat{T}(\sigma_\ell(\mathbf{X}_n^m, \mathbf{Y}_n^m))) = 0.$$

Then, for every $\alpha \in (0, 1)$, the test $\varphi_{\hat{T}}^{\alpha}$ has correct level α.

Corollary 4.1 ensures valid test level for a large class of test statistics. The particular choice of test statistic should depend on the alternative hypothesis that we seek to have power against. Within the LSCM model class, it makes sense to quantify deviations from the null hypothesis using functionals of the average causal effect, i.e., $T = \psi(f_{\mathrm{AVE}(X\to Y)})$ for some suitable function ψ. As a test statistic, we then use the plug-in estimator

$$\hat{T}(\mathbf{X}_n^m, \mathbf{Y}_n^m) = \psi(\hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(\mathbf{X}_n^m, \mathbf{Y}_n^m)).$$

An implementation of the above testing procedure is contained in our code package, see Section 4.1.4.

[3] Two-sided tests can be obtained using $p_{\hat{T},\text{2-sided}} := \min(1, 2 \cdot \min(p_{\hat{T}}, p_{-\hat{T}}))$, for example.

4.4. Conflict and forest loss in Colombia

We now return to the problem of conflict ($\mathbf{X}$) and forest loss ($\mathbf{Y}$) introduced in Section 4.1.3. We first describe our data sources in Section 4.4.1, and then apply our proposed methodology in Section 4.4.2. In Section 4.4.3, we introduce two alternative approaches for comparison, Section 4.4.4 contains our results, and Section 4.4.5 interprets these results in light of the Colombian peace process.

4.4.1. Data description and preprocessing

Our analysis is based on two main datasets: (1) a remote sensing-based forest loss dataset for the period 2000–2018, which identifies annual forest loss at a spatial resolution of 30m $\times$ 30m using Landsat satellites [Hansen et al., 2013]. Here, forest loss is defined as complete canopy removal. (2) Spatially explicit information on conflict events from 2000 to 2018, based on the Georeferenced Event Dataset (GED) from the Uppsala Conflict Data Program (UCDP) [Croicu and Sundberg, 2015]. In this dataset, a conflict event is defined as "an incident where armed force was used by an organized actor against another organized actor, or against civilians, resulting in at least one direct death at a specic location and a specifc date" [Sundberg and Melander, 2013]. Such events were identified through global newswire reporting, global monitoring of local news, and other secondary sources such as reports by non-governmental organizations (for information on the data collection as well as control for quality and consistency of the data, please refer to Sundberg and Melander [2013] and Croicu and Sundberg [2015]). We homogenized these datasets through aggregation to a spatial resolution of 10km $\times$ 10km by averaging the annual forest loss within each grid, and by counting all conflict events occurring in the same year and within the same grid. As a proxy for local transport infrastructure, we additionally calculated, for each spatial grid, the average Euclidean distance to the closest road segment, using spatial data from `https://diva-gis.org` containing all primary and secondary roads in Colombia. We regard this variable as relatively constant throughout the considered time-span.

4.4.2. Quantifying the causal influence of conflict on forest loss

We assume that $(\mathbf{X}, \mathbf{Y})$ come from an LSCM as defined in Defintion 4.2. Since X_s^t is binary, we can characterize the causal influence of $\mathbf{X}$ on $\mathbf{Y}$ by $T := f_{\mathrm{AVE}(X\to Y)}(1) - f_{\mathrm{AVE}(X\to Y)}(0)$, i.e., the difference in expected forest loss $\mathbb{E}[Y_s^t]$ under the respective interventions enforcing conflict ($X_s^t := 1$) and peace ($X_s^t := 0$). Positive values of T correspond to an augmenting effect of conflict on forest loss and negative values correspond to a reducing effect. Our goal is to estimate T, and to test the hypothesis $H_0 : T = 0$ (no causal effect of $\mathbf{X}$ on $\mathbf{Y}$). To construct an estimator of the average causal effect of the form (4.2.5), we require estimators of the conditional expectations $x \mapsto f_{Y|(X,H)}(x, h)$. Since X_s^t is binary, we use simple sample averages of the response variable. To make the resulting estimator of $f_{\mathrm{AVE}(X\to Y)}$ well-defined, we omit all locations which do not contain at least one observation from each of the regimes $X_s^t = 0$ and $X_s^t = 1$. More precisely, let $(\mathbf{X}_n^m, \mathbf{Y}_n^m)$ be the observed dataset. We then use the estimator, for every $x \in \{0, 1\}$ defined as

$$\hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(\mathbf{X}_n^m, \mathbf{Y}_n^m)(x) = \frac{1}{|\mathcal{I}_n^m|} \sum_{i\in\mathcal{I}_n^m} \frac{1}{|\{t : X_{s_i}^t = x\}|} \sum_{t:X_{s_i}^t=x} Y_{s_i}^t, \tag{4.4.1}$$

where $\mathcal{I}_n^m := \{i \in \{1, \dots, n\} : \exists t_0, t_1 \in \{1, \dots, m\} \text{ s.t. } X_{s_i}^{t_0} = 0 \text{ and } X_{s_i}^{t_1} = 1\}$. To test H_0, we use the resampling test introduced in Section 4.3 with test statistic $\hat{T} = \hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(1) - \hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(0)$.

The estimator (4.4.1) is constructed from a reduced dataset. The used data exclusion criterion is not independent of the assumed hidden confounders (i.e., the distribution of the hidden variables is expected to differ between the reduced data and the original data), and therefore results in a biased estimator. Under additional assumptions on the underlying LSCM, however, (4.4.1) may still be used to estimate T. We now give a population version argument. Assume that there is no interaction between X_s^t and H_s^t in the causal mechanism for Y_s^t, i.e., the function f in (4.2.2) splits into $f_1(X_s^t, \varepsilon_s^t) + f_2(H_s^t, \varepsilon_s^t)$. Then, the conditional expectation of $Y_s^t \,|\, (X_s^t, H_s^t)$ likewise splits additively into a function of X_s^t and a

function of H_s^t. Using Proposition 4.2, it follows that any two different models for the marginal distribution of the latent process $\mathbf{H}$ induce average causal effects $f_{\mathrm{AVE}(X \to Y)}$ that are equal up to an additive constant. In particular, every model for $\mathbf{H}$ induces the same value for T. By regarding the reduced dataset as a realization from a modified LSCM, in which the distribution of $\mathbf{H}$ has been altered, this argument justifies the use of (4.4.1) as an estimator for T.[4]

4.4.3. Comparison with alternative assumptions on the causal structure

To emphasize the relevance of the assumed causal structure, we compare our method with two alternative approaches based on different assumptions about the ground truth: Model 1 assumes no confounders of $(\mathbf{X}, \mathbf{Y})$ and Model 2 assumes that the only confounder is the observed process $\mathbf{W}$ (mean distance to a road). Even though none of the models may be a precise description of the data generating mechanism, we regard both Models 1 and 2 as less realistic than the LSCM. In both models we can, similarly to Definition 4.3, define the average causal effect of $\mathbf{X}$ on $\mathbf{Y}$. Under Model 1, $f_{\mathrm{AVE}(X \to Y)}$ coincides with the conditional expectation of Y_s^t given X_s^t, which can be estimated simply using sample averages (as is done in Figure 4.1 left). Under Model 2, $f_{\mathrm{AVE}(X \to Y)}$ can, analogously to Proposition 4.2, be computed by adjusting for the confounder $\mathbf{W}$. For each $x \in \{0, 1\}$, we obtain an estimate $\hat{f}^{nm}_{\mathrm{AVE}(X \to Y)}(x)$ by calculating sample averages of $\mathbf{Y}$ across different subsets $\{(s,t) \in \mathbb{R}^2 \times \mathbb{N} : X_s^t = x, W_s^t \in \mathcal{W}_j\}, j \in \mathcal{J}$ (we here construct these by considering 100 equidistant quantiles of $\mathbf{W}$), and subsequently averaging over the resulting values. In both models, we further test the hypothesis of no causal effect of $\mathbf{X}$ on $\mathbf{Y}$ using approaches similar to the ones presented in Section 4.3. Under the LSCM assumption, we have constructed a permutation scheme that permutes the values of $\mathbf{Y}$ along the time axis, to preserve the dependence between $\mathbf{Y}$ and the assumed time-invariant confounders $\mathbf{H}$, see Proposition 4.5. Similarly, we construct a permutation scheme for

[4] In practice, we use the values of $\mathbf{X}$ to exclude data points, and the above argument must thus be regarded a heuristic.

Model 2 by permuting observations of $\mathbf{Y}$ corresponding to similar values of the confounder $\mathbf{W}$ (i.e., values within the same quantile range). Under the null hypothesis corresponding to Model 1, $\mathbf{X}$ and $\mathbf{Y}$ are (unconditionally) independent, and we therefore permute the values of $\mathbf{Y}$ completely at random. Strictly speaking, the permutation schemes for Models 1 and 2 require additional exchangeability assumptions on $\mathbf{Y}$ in order to yield valid resampling tests. In Appendix C.3, we repeat the analysis for Model 1 using a spatial block-permutation to account for the spatial dependence in $\mathbf{Y}$, and obtain similar results.

4.4.4. Results

The results of applying our method and the two alternative approaches to the entire study region are depicted in Figure 4.4. Under Model 1, there is an enhancing, highly significant causal effect of conflict on forest loss ($\hat{T} = 0.073$, $P = 0.002$). When adjusting for transport infrastructure (quantified by $\mathbf{W}$, Model 2), the size of the estimated causal effect shrinks, and becomes insignificant ($\hat{T} = 0.049$, $P = 0.168$). (Note that we have considered other confounders, too, yet obtained similar results. For example, when adjusting for population density, which we consider as moderately temporally varying, we obtain $\hat{T} = 0.038$ and $P = 0.214$.) When applying the methodology proposed in this paper, that is, adjusting for all time-invariant confounders, the estimated effect swaps sign ($\hat{T} = -0.018$, $P = 0.578$), but is insignificant. One reason for this non-finding could be the time delay between the proposed cause (conflict) and effect (forest loss). To account for this potential issue, we also test for a causal effect of $\mathbf{X}$ on $\mathbf{Y}$ that is temporally lagged by one year, i.e., we use an estimator similar to (4.4.1), where we compare the average forest loss succeeding conflict events with the average forest loss succeeding non-conflict events. Again, the estimated influence of $\mathbf{X}$ on $\mathbf{Y}$ is negative and insignificant ($\hat{T} = -0.0293$, $P = 0.354$). Additionally, we perform alternative versions of the last two tests where we account for potential autocorrelation in the response variable, by adopting a temporal block-permutation scheme. In both cases, the test is insignificant, see Appendix C.3.

The above analysis provides an estimate for the average causal ef-

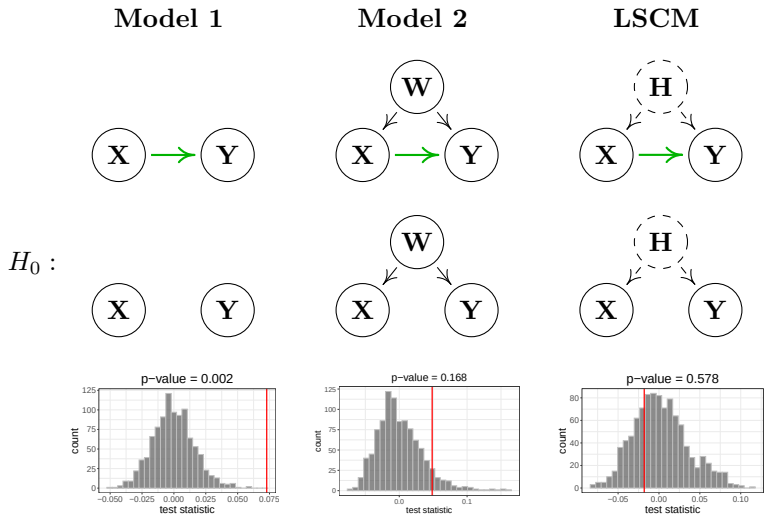


FIGURE 4.4. Testing for a causal influence of conflict (**X**) on forest loss (**Y**) using our method (right) and two alternative approaches (left and middle) which are based on different and arguably less realistic assumptions on the causal structure. The process **W** corresponds to the mean distance to a road, and **H** represents unobserved time-invariant confounders. Each of the above models gives rise to a different expression for the test statistic $\hat{T} = \hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(1) - \hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(0)$ (indicated by red vertical bars), see Sections 4.4.2 and 4.4.3. The gray histograms illustrate the empirical null distributions of $\hat{T}$ under the respective null hypotheses obtained from 999 resampled datasets. The results show that our conclusions about the causal influence of conflict on forest loss strongly depend on the assumed causal structure: under Model 1, there is a positive, highly significant effect ($\hat{T} = 0.073$). When adjusting for the confounder **W**, the effect size decreases and becomes insignificant ($\hat{T} = 0.049$). When applying our proposed methodology, the estimated effect is negative ($\hat{T} = -0.018$).

fect, see Equation (4.2.3), which, in particular, averages over space. Given that Colombia is a country with high ambiental and socio-economic heterogeneity, where different regional dynamics may influence the causal relationship between conflict and forest loss, we further conduct an analysis at the department level (see Figure 4.5). In fact, there is considerable spatial variation in the estimated causal effects, with significant positive as well as negative effects (Figure 4.5 middle). From a modeling perspective, this variation may be seen as evidence for an interaction effect between conflict and the assumed hidden confounders. In most departments, the estimated causal effect is negative (although mostly insignificant), meaning that conflict tends to decrease forest loss. The strongest positive and significant causal influence is identified in the La Guajira department ($\hat{T} = 0.398$, $P = 0.047$). Although this region is commonly associated with semi-arid to very dry conditions, most conflicts occured in the South-Western areas, at the beginning of Caribbean tropical forests (see Figure 4.1). In fact, these zones have also been identified by Negret et al. [2019] as having been strongly affected by deforestation pressure in the wake of conflict. Interestingly, the neighboring Magdalena department shows the opposite effect ($\hat{T} = -0.218$, $P = 0.004$), which might point to a different socio-political reality. It may also reflect the fact that this department experienced high forest loss after the ceasefire in the entire Colombia [Prem et al., 2020]. The positive effect in the department of Huila ($\hat{T} = 0.095$, $P = 0.023$) is again in line with the findings by Negret et al. [2019] (based on a visual inspection of their attribution maps). Out of the 8 departments that are mostly controlled by FARC (Figure 4.5 right), 6 have a negative test statistic, meaning that conflict reduces forest loss. This can be explained in part by the internal governance of this group, where forest cover was a strategic advantage for both their own protection as well as for cocaine production. Overall, of course, the peace-induced acceleration of forest loss has to be discussed with caution, and should not be interpreted reversely as if conflict per se is a measure of environmental protection as it has been discussed, for instance, by Clerici et al. [2020].

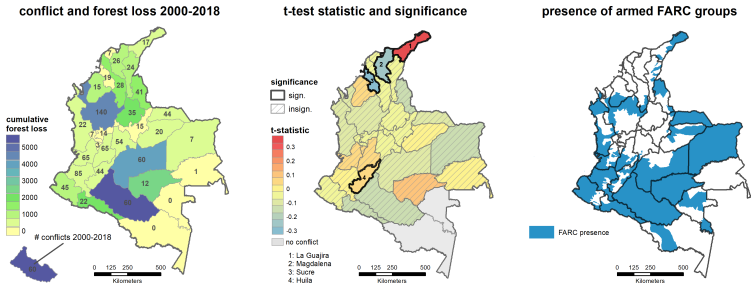


FIGURE 4.5. Regional analysis of conflict and forest loss in Colombia. The left panel shows the total forest loss and the total number of conflict events from 2000 to 2018 aggregated at department level. The most severe incidences of forest loss occur in the Northern Andean forests and on the northern borders of the Amazon region. In the middle panel, we report for each department estimates $\hat{T}$ and test results for $H_0 : T = 0$, using the methodology described in Sections 4.2 and 4.3. We used a test level of $\alpha = 0.05$, and report significances without multiple-testing adjustment. In most departments, the estimated causal effect is negative (blue, conflict reduces forest loss), although mainly insignificant. We identify four departments with statistically significant results, hereof two with a positive causal effect (La Guarija and Huila) and two with a negative causal effect (Magdalena and Sucre). In total, there are 8 departments that are mostly controlled by FARC (above 75% FARC presence, right panel). Out of these, 6 departments have a negative test statistic (conflict reduces forest loss).

4.4.5. Interpretation of our results in light of the Colombian peace process

In late 2012, negotiations that later would be known as "Colombian peace process" started between the then president of Colombia and the strongest group of rebels, the FARC, and lasted until 2016. Controversies in the country culminated in the rejection of the agreement in a national referendum. Despite this failure, peace was declared by both parties upon a revised agreement in October 2016, and became effective in the subsequent year.[5] While severe incidences continued to occur in the year leading up to the final agreement, the negotiations marked a steadily decreasing number of conflicts, see Figure 4.6 (left). Since this decrease of conflicts is the consequence of governmental intervention, rather than a natural resolution of local tensions, the peace process provides an opportunity to verify the intervention effects estimated in Section 4.4.4. As can be seen in Figure 4.6 (right), Colombia experienced a steep increase in the total forest loss in the final phase of the peace negotiations. Although there may be several other factors which have contributed to this development, we observe that these results align with our previous finding of an overall negative causal effect of conflict on forest loss ($\hat{T} < 0$).

4.5. Conclusions and future work

4.5.1. Methodology

This paper introduces ways to discuss causal inference for spatio-temporal data. From a methodological perspective, it contains three main contributions: the definition of a class of causal models for multivariate spatio-temporal stochastic processes, a procedure for estimating causal effects within this model class, and a non-parametric hypothesis test for the overall existence of such effects. Our method allows for the influence of arbitrarily many latent confounders, as long as these confounders do not vary across time (or space). We

[5] The agreement was signed only by the FARC, while other guerilla groups remain active.

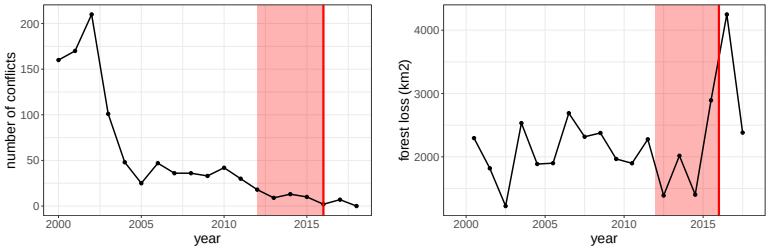


FIGURE 4.6. Total number of conflicts (left) and total forest loss (right) in the years 2000–2018 in Colombia. In the right panel, the height of the curve at $x + 0.5$ corresponds to the total forest loss between years x and $x + 1$. The shaded area marks the Colombian peace process, which began in September 2012. A final agreement was reached in October 2016 (vertical red line).

prove asymptotic consistency of our estimator, and verify this finding empirically using simulated data. Our results hold under weak assumptions on the data generating process, and do not rely on any particular distributional properties of the data. We prove sufficient conditions under which these rather general assumptions hold true. The proposed testing procedure is based on data resampling and provably obtains valid level in finite samples.

Our work can be extended into several directions. We proved that Assumption 4.1 holds for regularly sampled stationary Gaussian processes. Such settings allow for the application of well-known theorems about stationary and ergodic sequences. We hypothesize that the assumption also holds in the more general case, where the marginalized process does not resemble a (collection of) stationary sequence(s), as long as certain mixing conditions are satisfied. For example, we believe that if the original spatial process $\mathbf{H}^1$ is weakly stationary and mixing, Assumption 4.1 holds under any spatial sampling scheme $(s_n)_{n\in\mathbb{N}}$ with $\|s_n\| \to \infty$ as $n \to \infty$.

In our method for estimating causal effects, we allow the regression model for (X_s^t, Y_s^t) to change arbitrarily (within the specified function class) between neighboring locations. It may be worthwhile exploring how smoothness assumptions on the hidden variables can

be incorporated in the modeling process to gain statistical efficiency of the proposed estimator. Likewise, such smoothness assumptions would allow for alternative permutation schemes (data can be permuted spatially wherever **H** is assumed to be constant) which could lead to increased power of our hypothesis tests.

4.5.2. Case study

We have applied our methodology to the problem of quantifying the causal influence of conflict on forest loss in Colombia. Conflict events are predictive of exceedances in forest loss, but we find no evidence of a causal relation when analyzing this problem on country level: once all (time-invariant) confounders are adjusted for, there is a negative but insignificant correlation between conflict and forest loss. Our analysis on department level suggests that this non-finding could be due to locally varying effects of opposite directionality, which would approximately cancel out in our final estimate. In most departments, we find negative (mostly insignificant) effect of conflict on forest loss, although we also identify a few departments where conflict seems to increase forest loss. The overall negative influence of conflict on forest loss estimated by our method is in line with the observation that in the final phase of the peace process, which stopped many of the existing conflicts, the total forest loss in Colombia has increased. However, these results should be interpreted with caution. Overall, we find that, once all time-invariant confounders are adjusted for, conflicts have only weak explanatory power for predicting forest loss, and the potential causal effect is therefore likely to be small, compared to other drivers of forest loss. The chain of events which link the occurrence of an armed conflict with the clearing of local forests is rather complex, and we hope that future research will be able to shed further light on the underlying causal relationship.

Acknowledgments

This work contributes to the Global Land Programme, and was supported by a research grant (18968) from VILLUM FONDEN. M.D.M. thanks the European Space Agency for funding the Earth

System Data Lab. We are grateful to Steffen Lauritzen for helpful comments on ergodicity, and Niels Richard Hansen and Niklas Pfister for insightful discussions on the sufficiency conditions in Section 4.2.3.2. We further thank Lina Estupinan-Suarez and Alejandro Salazar for helpful discussions on the environmental impacts of the Colombian peace process.

Appendices

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

A.1. Structural causal models

Below, we formally define structural causal models [Pearl, 2009, Bollen, 1989], and use a presentation similar to Peters et al. [2017, Chapter 6].

Definition A.1 (Structural causal model)**.** *A* structural causal model *(SCM) over variables* $(Z_1, \dots, Z_p)$ *consists of a family of structural assignments*

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

where for each $j \in \{1, \dots, p\}$, $\mathbf{PA}_j \subseteq \{Z_1, \dots, Z_p\} \setminus \{Z_j\}$ *denotes the* parent set *of variable* Z_j*, and a product distribution over the noise variables* $(N_1, \dots, N_p)$*. Every SCM induces a graph over the nodes in* $\{Z_1, \dots, Z_p\}$*: 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 $Z_1, \dots, Z_p$. 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.

A.2. Parametrizations of the models IID and HMM

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

$$\mathbf{\Gamma}^{\mathrm{IID}}_{ij}(\gamma) = \begin{cases} \gamma_j & j < \ell \\ 1 - \sum_{k=1}^{\ell-1} \gamma_k & j = \ell \end{cases}$$

and

$$\mathbf{\Gamma}_{ij}^{\mathrm{HMM}}(\gamma) = \begin{cases} \gamma_{i\ell+j} & j < \ell \\ 1 - \sum_{k=1}^{\ell-1} \gamma_{i\ell+k} & j = \ell. \end{cases}$$

For the regression matrix Θ, we consider the two types of parameter constraints discussed in Section 2.3.2. 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 $\mathbf{\Theta}^c : \mathcal{T}^c \to \mathbb{R}^{p \times \ell}$ and $\mathbf{\Theta}^= : \mathcal{T}^= \to \mathbb{R}^{p \times \ell}$, for all $i \in \{1, \dots, p+1\}$ and $j \in \{1, \dots, \ell\}$ given by

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

Both of the parameter constraints induced by $(\mathbf{\Theta}^c, \mathcal{T}^c)$ and $(\mathbf{\Theta}^=, \mathcal{T}^=)$ ensure the existence of the maximum likelihood estimator, see Theorem 2.1. Since all of the above coordinate mappings are linear in θ and γ, Assumption (A4) in Section 2.3.5 is satisfied for any pair $(\mathbf{\Theta}, \mathbf{\Gamma})$ with $\mathbf{\Theta} \in \{\mathbf{\Theta}^c, \mathbf{\Theta}^=\}$ and $\mathbf{\Gamma} \in \{\mathbf{\Gamma}^{\mathrm{IID}}, \mathbf{\Gamma}^{\mathrm{HMM}}\}$.

A.3. Proofs

A.3.1. Proof of Proposition 2.2

Recall that by Definition A.1, we require the underlying causal graph to be acyclic. For every $t \in \{1, \dots, n\}$, 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 A.1), it follows that $(X_t^{\mathrm{PA}^0(Y)}, H_t^*) \perp\!\!\!\perp N_t$. Using the i.i.d. assumption on $(N_t)_{t \in \{1,\dots,n\}}$, we have that for all t and for all x, h, the distribution of $Y_t \,|\, (X_t^{\mathrm{PA}^0(Y)} = x, H_t^* = h) \stackrel{d}{=} f(x, h, N_t)$ does not depend on t, which shows that $S^* = \mathrm{PA}^0(Y)$ satisfies (2.2.1). 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 $(\mathbf{Y}, \mathbf{X})$. □

A.3.2. Proof of Theorem 2.1

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 $\beta = (\beta_1, \dots, \beta_\ell) \in \mathcal{B} := \mathbb{R}^{p \times \ell}$ 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 $h \in \{1, \dots, \ell\}^m$ and every $j \in \{1, \dots, \ell\}$ define $T_{h=j} := \{t \in \{1, \dots, m\} : h_t = j\}$ and write the likelihood function as $G = \sum_{h \in \{1,\dots,\ell\}^m} g_h$, where

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

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 s.t. $\lim_{n\to\infty} G(\phi^{*n}) = G^*$. 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 A.1 and Lemma A.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 2.1.

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

Lemma A.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 $(\boldsymbol{\Theta}^c, \mathcal{T}^c)$ and $(\boldsymbol{\Theta}^=, \mathcal{T}^=)$ separately.

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

$$g_h(\phi^n) \begin{cases} \to 0 \text{ as } n\to\infty & \text{if } T_{h=j} \neq \emptyset \\ \text{does not depend on } \delta_j^n & \text{otherwise,} \end{cases} \tag{A.3.1}$$

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 $(\tilde{\phi}^n)_{n\in\mathbb{N}}$ 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 $\{1,\ldots,m\}$ 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

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

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 A.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 $\lim_{n\to\infty} G(\tilde{\phi}^n) \geq \limsup_{n\to\infty} G(\phi^n)$.*

Proof. The idea of the proof is as follows. We construct a bounded sequence $(\tilde{\beta}^n)_{n\in\mathbb{N}}$, s.t. 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 $\lim_{n\to\infty} G(\tilde{\phi}^n) \geq \limsup_{n\to\infty} G(\phi^n)$. 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 A.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 $(\phi^n)_{n\in\mathbb{N}} = (\beta_1^n, \ldots, \beta_\ell^n, \delta^n, \gamma^n)$ be as stated in Lemma A.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 $(\tilde{\phi}_{(k)}^n)_{n\in\mathbb{N}}$ 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 $\limsup_{n\to\infty} G(\tilde{\phi}_{(k)}^n) \geq \limsup_{n\to\infty} G(\phi^n)$. We then take $k' \in K^\infty \setminus \{k\}$ and similarly construct $(\phi_{(k,k')}^n)_{n\in\mathbb{N}}$ from $(\tilde{\phi}_{(k)}^n)_{n\in\mathbb{N}}$ such that $\limsup_{n\to\infty} G(\tilde{\phi}_{(k,k')}^n) \geq \limsup_{n\to\infty} G(\tilde{\phi}_{(k)}^n)$. 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 $\limsup_{n\to\infty} G(\tilde{\phi}^n) \geq \lim_{n\to\infty} G(\phi^n)$. 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 $k_1, \ldots, k_j \in K^\infty$ (if $j = 0$, this corresponds to the base case). For simplicity write $(\check{\phi}^n)_{n\in\mathbb{N}} = (\tilde{\phi}_{(k_1,\ldots,k_j)}^n)_{n\in\mathbb{N}}$. Pick an arbitrary $k \in K^\infty \setminus \{k_1, \ldots, k_j\}$. If for all $t \in \{1, \ldots, m\}$, $|x_t \beta_k^n| \to \infty$ as $n \to \infty$, we could (similar to the proof of Lemma A.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 $\mathcal{S}_1 := \{s \in \{1,\dots,m\} : |x_s\beta_k^n| \to \infty \text{ as } n \to \infty\}$, $\mathcal{S}_2 := \{1,\dots,m\} \setminus \mathcal{S}_1$, $\mathcal{H}_1 := \{h \in \{1,\dots,\ell\}^m : T_{h=k} \cap \mathcal{S}_1 \neq \emptyset\}$ 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 $h \in \{1,\dots,\ell\}^m$ 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 $\{\sigma_j^2(\delta^n)\}_{n\in\mathbb{N}}$, $j \in \{1,\dots,\ell\}$, 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

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

We now show that the above defines a bounded sequence.

($\circ$) 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 $\{u_1, \dots, u_d\}$ be an orthonormal basis for $\mathrm{span}_{\mathbb{R}}(\{x_s : s \in \mathcal{S}\})$. 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 $(|u_1\tilde{\beta}_k^n|)_{n\in I_{\mathcal{S}}^b}, \dots, (|u_d\tilde{\beta}_k^n|)_{n\in I_{\mathcal{S}}^b}$ (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.

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

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 $(\tilde{\beta}_k^n)_{n\in\mathbb{N}}$, we thus have that for all $n \notin I_h^\infty$ and for all $s \in T_{h=k}$, the function values $\mathcal{N}(y_s \,|\, x_s\tilde{\beta}_k^n, \sigma_k^2(\delta^n))$ and $\mathcal{N}(y_s \,|\, x_s\beta_k^n, \sigma_k^2(\delta^n))$ coincide. Consequently, we have that for all $n \notin I_h^\infty$, $g_h(\tilde{\phi}_k^n) = g_h(\check{\phi}^n)$. 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

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

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

This completes the proof of Lemma A.2. □

A.3.3. Proof of Theorem 4.2

We start by introducing some notation to be used in the proofs of Theorem 4.2 and Theorem 2.3. 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 4.2 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), $\kappa \mapsto f(x, y \,|\, \kappa)$ 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 $\gamma \mapsto \mathbf{\Gamma}_{jk}(\gamma)$ are continuous, (L4) for all $j \in \{0, \ldots, \ell\}$, $\mathbb{E}_0[|\log f(X_1, Y_1 \,|\, \kappa_j(\theta^0))|] < \infty$, (L5) for all $\kappa \in \mathcal{K}$, there exists a $\delta > 0$ such that $\mathbb{E}_0[\sup_{\kappa' : \|\kappa - \kappa'\| < \delta} (\log f(X_1, Y_1 \,|\, \kappa'))^+] < \infty$, and (L6) for every $k \in \{1, \ldots, \ell\}$, the class $\mathcal{SR}^k(\cdot \,|\, X_1)$ satisfies the following identifiability property. Define

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

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

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

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

Then, for every $k \in \{1, \ldots, \ell\}$, φ^k is a one-to-one map of $\mathcal{Q}^k$ onto $\mathcal{SR}^k(\cdot \,|\, X_1)$. 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 $\kappa \mapsto \mathcal{N}(y \,|\, x, \kappa)$ and (L3) is implied by (A4). For (L4), we see that for all $j \in \{0, \ldots, \ell\}$,

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

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

$$\begin{aligned}
&\mathbb{E}_0 \left[\sup_{\kappa' : \|\kappa' - \kappa\| < \delta} (\log f(X_1, Y_1 | \kappa'))^+ \right] \\
&\leq \mathbb{E}_0 \left[\sup_{\kappa' : \|\kappa' - \kappa\| < \delta} (\log f(Y_1 | X_1, \kappa'))^+ + |\log f(X_1)| \right] \\
&\leq \mathbb{E}_0 \left[\sup_{\sigma' : \|\sigma'^2 - \sigma^2\| < \delta} (-\frac{1}{2} \log(2\pi\sigma'^2))^+ + |\log f(X_1)| \right] \\
&\leq \mathbb{E}_0 \left[\frac{1}{2} |\log(\pi\sigma^2)| + |\log f(X_1)| \right] < \infty.
\end{aligned}$$

It is left to prove (L6), the identifiability of the classes $\mathcal{SR}^k(\cdot \,|\, X_1)$. 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. Let $k \in \{1, \ldots, \ell\}$ and $q = \{(\lambda_1, \beta_1, \sigma_1^2), \ldots, (\lambda_k, \beta_k, \sigma_k^2)\}, q' = \{(\lambda_1', \beta_1', {\sigma_1'}^2), \ldots, (\lambda_k', \beta_k', {\sigma_k'}^2)\} \in \mathcal{Q}^k$, 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 $Q'_{\cdot j} = (\lambda_j', {\sigma_j'}^2, {\beta_j'}^T)^T$ 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 $x \in \text{int}(\text{supp}(X_1))$,

$$f_q(y \,|\, x) = \sum_{j=0}^{k} \lambda_j f(y \,|\, x, \kappa_j) = \sum_{j=0}^{k} \lambda_j' f(y \,|\, x, \kappa_j') = f_{q'}(y \,|\, x)$$

for almost all y. It now follows from Teicher [1963, Proposition 1] that, for all $x \in \text{int}(\text{supp}(X_1))$,

$$\{(\lambda_1, \sigma_1^2, x\beta_1), \ldots, (\lambda_k, \sigma_k^2, x\beta_k)\} = \{(\lambda_1', {\sigma'}_1^2, x\beta_1'), \ldots, (\lambda_k', {\sigma'}_k^2, x\beta_k')\}. \tag{A.3.2}$$

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 (A.3.2) means that for every

$z \in \mathcal{Z} := \mathbb{R}^2 \times \text{int}(\text{supp}(X_1))$, 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

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

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 $Q = Q'M_0^T$. □

A.3.4. Proof of Theorem 2.3

Throughout the proof, we make use of the notation introduced in the first paragraph of Appendix A.3.3. Theorem 2.3 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 $\phi^0 \in \text{int}(\mathcal{P})$ 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 $j \in \{1,\dots,\ell\}$, $\mathbb{P}_0(\rho(X_1,Y_1) = \infty \,|\, H_1 = j) < 1$.

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 $j \in \{1,\dots,\ell\}$ and $k \in \{1,\dots,d\}$, 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 $bx_i^r x_k^s y^t$ for some $i,k \in \{1,\dots,p\}$ and $r,s,t \in \{0,\dots,4\}$ with $r+s+t \le 4$. Then, for every $\kappa \in \mathcal{K}$, $\int g(|x|,|y|) f(x,y\,|\,\kappa) d(x,y) < \infty$, where $|x| = (|x_1|,\dots,|x_p|)$. This result follows from the fact that for every x, $\int |y|^t f(y\,|\,x,\kappa) dy$ is a polynomial of $|x|$ of degree t, and the assumption that, for all $j \in \{1,\dots,p\}$, $\mathbb{E}[|X_1^j|^4] < \infty$.

For (B3)(i), we treat all derivatives simultaneously. Let $n \in \{1,2\}$, $i_1,\dots,i_n \in \{1,\dots,K\}$ and $j \in \{1,\dots,\ell\}$ 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

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

(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), $\sup_{\theta\in\mathcal{T}_0} |g_\theta(x, y)| \leq g(|x|, |y|)$.

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 $x_k(\beta_{jk}(\theta^0) - \text{sign}(x_k)\varepsilon) \leq x_k\beta_{jk}(\theta) \leq x_k(\beta_{jk}(\theta^0)+\text{sign}(x_k)\varepsilon)$. With $s(x) = (\text{sign}(x_1), \ldots, \text{sign}(x_p))$ it follows that for all (x, y) and all $\theta \in \mathcal{T}_0$, $y-x(\beta_j(\theta^0)-\text{diag}(s(x))\varepsilon) \leq y - x\beta_j(\theta) \leq y - x(\beta_j(\theta^0) + \text{diag}(s(x))\varepsilon)$. 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 $\theta \in \mathcal{T}_0$,

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

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

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

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

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

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

A.3.5. Proof of Corollary 2.1

Let (A1)–(A7) hold true. By Theorem 4.2, 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 $(\pi^m_{\mathcal{P}})_{m\in\mathbb{N}} = ((\pi^m_{\mathcal{T}}, \pi^m_{\mathcal{G}}))_{m\in\mathbb{N}}$ of permutations on $\mathcal{P}$, such that, $\mathbb{P}_0$-almost surely, the sequence of MLEs $(\pi^m_{\mathcal{P}}(\hat{\phi}_m))_{m\in\mathbb{N}}$ converges to ϕ^0 as $m \to \infty$. For $\alpha \in (0, 1)$ and for every $m \in \mathbb{N}$, we then have

$$\begin{aligned}\mathbb{P}^m_0(\theta^0 \in C^\alpha_{\text{adjusted}}(\hat{\theta}_m)) &\geq \mathbb{P}^m_0(\theta^0 \in C^\alpha(\pi^m_{\mathcal{T}}(\hat{\theta}_m))) \\ &= \mathbb{P}^m_0(\phi^0 \in C^\alpha(\pi^m_{\mathcal{T}}(\hat{\theta}_m)) \times \mathcal{G}).\end{aligned}$$

By Theorem 2.3, the right hand side converges to $1-\alpha$ as $m \to \infty$.□

A.3.6. Proof of Theorem 2.4

By Corollary 2.1, 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, $\mathbb{P}^n_0(\hat{S}_n \subseteq S^*) \geq \mathbb{P}^n_0(\varphi^n_{S^*} \text{ accepts } H^n_{0,S^*})$, the result follows. □

A.4. Further details on likelihood optimization

Below, we describe the two optimization methods NLM and EM. Since the loglikelihood function (2.3.1) 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`).

A.4.1. Method I ("NLM"): Non-Linear Maximization

This method maximizes the loglikelihood function (2.3.1) 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 (2.3.2). An equality constraint on the error variances can be enforced directly by using the parametrization $(\boldsymbol{\Theta}^{=}, \mathcal{T}^{=})$ described in Appendix A.2. 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].

A.4.2. 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 $P^{(t)}_{(\mathbf{y},\mathbf{x})}$ 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

$$Q(\phi \,|\, \phi^{(t)}) := \mathbb{E}_{P^{(t)}_{(\mathbf{y},\mathbf{x})}} \left[\ell_{\text{complete}}(\mathbf{y}, \mathbf{x}, \mathbf{H} \,|\, \phi)\right] \tag{A.4.1}$$

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 $P^{(t)}_{(\mathbf{y},\mathbf{x})}$ and Q depend on the choice of model. In model IID, $P^{(t)}_{(\mathbf{y},\mathbf{x})}$ 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, (A.4.1) 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 (A.4.1), 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 $(\boldsymbol{\Theta}^{=}, \mathcal{T}^{=})$ from Appendix A.2. A lower bound on the error variances is enforced by restarting the algorithm whenever

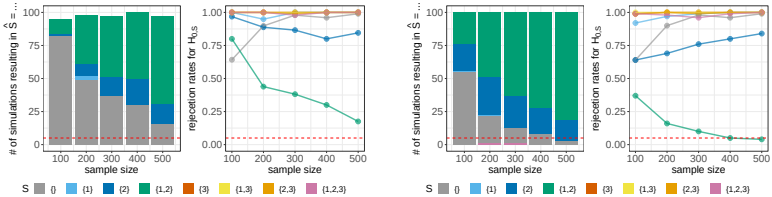


FIGURE A.1. Output of ICPH (bar plots) and rejection rates for individual hypotheses (curve plots) for the experiment in Section 2.5.1.4 with parameter constraint $\sigma^2_{Y1}, \sigma^2_{Y2} \geq 10^{-16}$ (left) and $\sigma^2_{Y1} = \sigma^2_{Y2}$ (right), using the EM-algorithm as optimization routine. The results are very similar to those presented in Figure 2.8, 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.

an update $\phi^{(t)}$ contains a variance component that deceeds the lower bound (`mixreg` uses the threshold 10^{-16}).

Figure A.1 shows numerical results for ICPH when using the EM-algorithm as optimization routine. The results should be compared to Figure 2.8, 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.

A.5. Additional numerical experiments

In this section, we present additional experimental results. In all simulations, we use slight adaptations of the SCM in Section 2.5.1.2, and measure the performance of ICPH using rejection rates for non-causality (similar to Figure 2.9). All results are summarized in Figure A.2.

A.5.1. 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 (2.2.6). 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 2.5.1.2 with parameters sampled as in Section 2.5.1.3. 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 A.2 (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.

A.5.2. 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 2.5.1.2. 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

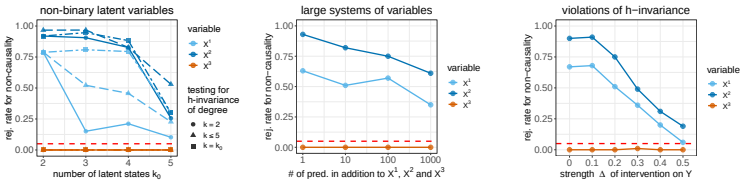


FIGURE A.2. Rejection rates for non-causality of the variables X^1, X^2 and X^3 for the experiments described in Appendix A.5. We investigate the performance of ICPH for non-binary variables (left), for large numbers of predictors (middle), and under violations of the h-invariance assumption (right). By simultaneously testing for h-invariance of different degrees (see Appendix A.5.1 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 2.1, 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).

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 A.2 (middle) suggest that even for a large number of predictors, ICPH is generally able to infer S^* (provided that S^* contains only few variables).

A.5.3. Violations of the h-invariance assumption

The theoretical guarantees of our method rely on the existence of an h-invariant set (Assumption 2.1). 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 2.5.1.2, but include direct interventions on Y. For increasing values of $\Delta \in \{0, 0.1, \ldots, 0.5\}$, we change the coefficients $(\beta^Y_{11}, \beta^Y_{21})$ 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 A.2 right).

B | The difficult task of distribution generalization in nonlinear models

B.1. Transforming causal models

As illustrated in Remark 3.1, our framework is able to model cases where causal relations between the observed variables are given explicitly, e.g., by an SCM. The key insight is that most of these causal relations can be absorbed by the hidden confounding H on which we make few restrictions. To show how this can be done in a general setting, let us consider the following SCM

$$\begin{aligned} A &\coloneqq \varepsilon_A & X &\coloneqq w(X,Y) + g(A) + h_2(H, \varepsilon_X) \\ H &\coloneqq \varepsilon_H & Y &\coloneqq f(X) + h_1(H, \varepsilon_Y). \end{aligned} \tag{B.1.1}$$

Assume that this SCM is uniquely solvable in the sense that there exists a unique function F such that $(A, H, X, Y) = F(\varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y)$ almost surely, see Bongers et al. [2016] for more details. Denote by F_X the coordinates of F that correspond to the X variable (i.e., the coordinates from $r+q+1$ to $r+q+d$). Assume further that there exist functions $\tilde{g}$ and $\tilde{h}_2$ such that

$$F_X(\varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y) = \tilde{g}(\varepsilon_A) + \tilde{h}_2((\varepsilon_H, \varepsilon_Y), \varepsilon_X). \tag{B.1.2}$$

This decomposition is not always possible, but it exists in the following settings, for example: (i) *There are no A variables.* As discussed in Section 3.2 our framework also works if no A variables exist. In these cases, the additive decomposition (B.1.2) becomes trivial. (ii) *There are further constraints on the full SCM.* The additive decomposition (B.1.2) holds if, for example, w is a linear function or A only enters the structural assignments of covariates X which have at most Y as a descendant.

Using the decomposition in (B.1.2), we can define the following SCM

$$\begin{aligned} A &\coloneqq \varepsilon_A & X &\coloneqq \tilde{g}(A) + \tilde{h}_2(\tilde{H}, \varepsilon_X) \\ \tilde{H} &\coloneqq \varepsilon_{\tilde{H}} & Y &\coloneqq f(X) + h_1(\tilde{H}), \end{aligned} \tag{B.1.3}$$

where $\varepsilon_{\tilde{H}}$ has the same distribution as $(\varepsilon_H, \varepsilon_Y)$ in the previous model. This model fits the framework described in Section 3.2, where the noise term in Y is now taken to be constantly zero. Both

SCMs (B.1.1) and (B.1.3) induce the same observational distribution and the same function f appears in the assignments of Y.

It is further possible to express the set of interventions on the covariates X in the original SCM (B.1.1) as a set of interventions on the covariates in the reduced SCM (B.1.3). The description of a class of interventions in the full SCM (B.1.1) may, however, become more complex if we consider them in the reduced SCM (B.1.3). In particular, to apply the developed methodology, one needs to check whether the interventions in the reduced SCM is a well-behaved set of interventions (this is not necessarily the case) and how the support of all X variables behaves under that specific intervention. We now discuss the case that the causal graph induced by the full SCM is a directed acyclic graph (DAG).

Intervention type. First, we consider which types of interventions in (B.1.1) translate to well-behaved interventions in (B.1.3). Importantly, interventions on A in the full SCM reduce to regular interventions A also in the reduced SCM. Similarly, performing hard interventions on all components of X in the full SCM leads to the same intervention in the reduced SCM, which is in particular both confounding-removing and confounding-preserving. For interventions on subsets of the X, this is not always the case. To see that, consider the following example.

$$\begin{aligned} A &:= \varepsilon_A \\ X_1 &:= \varepsilon_1,\ X_2 := Y + \varepsilon_2 \\ Y &:= X_1 + \varepsilon_Y \end{aligned} \quad \text{(B.1.4)} \qquad \begin{aligned} A &:= \varepsilon_A, \quad H := \varepsilon_Y \\ X &:= (\varepsilon_1, H + \varepsilon_1 + \varepsilon_2) \\ Y &:= X_1 + H \end{aligned} \quad \text{(B.1.5)}$$

with $\varepsilon_A, \varepsilon_1, \varepsilon_2, \varepsilon_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, where (B.1.4) represents the full SCM and (B.1.5) corresponds to the reduced SCM using our framework. Consider now, in the full SCM, the intervention $X_1 := i$, for some $i \in \mathbb{R}$. In the reduced SCM, this intervention corresponds to the intervention $X = (X_1, X_2) := (i, H + i + \varepsilon_2)$, which is neither confounding-preserving nor confounding-removing.[1] On the other

[1] This may not come as a surprise since without the help of an instrument, it is impossible to distinguish whether a covariate is an ancestor or a descendant of Y.

hand, any intervention on X_2 or A in the full SCM model corresponds to the same intervention in the reduced SCM. We can generalize these observations to the following statements:

- *Interventions on A:* If we intervene on A in the full SCM (B.1.1) (i.e., by replacing the structural assignment of A with $\psi^i(I^i, \varepsilon_A^i)$), then this translates to an equivalent intervention in the reduced SCM (B.1.3).

- *Hard interventions on all X:* If we intervene on all X in the full SCM (B.1.1) by replacing the structural assignment of X with an independent random variable $I \in \mathbb{R}^d$, then this translates to the same intervention in the reduced SCM (B.1.3) which is confounding-removing.

- *No X is a descendant of Y and there is no unobserved confounding H:* If we intervene on X in the full SCM (B.1.1) (i.e., by replacing the structural assignment of X with $\psi^i(g, A^i, \varepsilon_X^i, I^i)$), then this translates to a potentially different but confounding-removing intervention in the reduced SCM (B.1.3). This is because the reduced SCM (B.1.3) does not include unobserved variables H in this case.

- *Hard interventions on a variable X_j which has at most Y as a descendant:* If we intervene on X_j in the full SCM (B.1.1) by replacing the structural assignment of X_j with an independent random variable I, then this intervention translates to a potentially different but confounding-preserving intervention.

Other settings may yield well-behaved interventions, too, but may require more assumptions on the full SCM model (B.1.1) or further restrictions on the intervention classes.

Intervention support. A support-reducing intervention in the full SCM can translate to a support-extending intervention in the reduced SCM. Consider the following example.

$$\begin{aligned} X_1 &:= \varepsilon_1 \\ X_2 &:= X_1 + \mathbf{1}\{X_1 = 0.5\} \\ Y &:= X_2 + \varepsilon_Y \end{aligned} \tag{B.1.6}$$

$$\begin{aligned} X &:= (\varepsilon_1, \varepsilon_1 + \mathbf{1}\{\varepsilon_1 = 0.5\}) \\ Y &:= X_2 + \varepsilon_Y, \end{aligned} \tag{B.1.7}$$

with $\varepsilon_1, \varepsilon_Y \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$. As before, (B.1.6) represents the full SCM, whereas (B.1.7) corresponds to the reduced SCM converted to fit our framework. Under the observational distribution, the support of X_1 and X_2 is equal to the open interval $(0,1)$. Consider now the support-reducing intervention $X_1 := 0.5$ in (B.1.6). Within our framework, such an intervention would correspond to the intervention $X = (X_1, X_2) := (0.5, 1.5)$, which is support-extending. This example is rather special in that the SCM consists of a function that changes on a null set of the observational distribution. With appropriate assumptions to exclude similar degenerate cases, it is possible to show that support-reducing interventions in (B.1.1) correspond to support-reducing interventions within our framework (B.1.3).

B.2. Sufficient conditions for Assumption 1 in IV settings

Assumption 3.1 states that f is identified on the support of X from the observational distribution of (Y, X, A). Whether this assumption is satisfied depends on the structure of $\mathcal{F}$ but also on the other function classes $\mathcal{G}, \mathcal{H}_1, \mathcal{H}_2$ and $\mathcal{Q}$ that make up the model class $\mathcal{M}$ from which we assume that the distribution of (Y, X, A) is generated.

Identifiability of the causal function in the presence of instrumental variables is a well-studied problem in econometrics literature. Most prominent is the literature on identification in linear SCMs [e.g., Fisher, 1966, Greene, 2003]. However, identification has also been studied for various other parametric function classes. We say that $\mathcal{F}$ is a parametric function class if it can be parametrized by some finite dimensional parameter set $\Theta \subseteq \mathbb{R}^p$. We here consider classes of the form

$$\mathcal{F} := \{f(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R} \,|\, \theta : \Theta \to \mathbb{R}, \theta \mapsto f(x, \theta) \text{ is } C^2 \text{ for all } x \in \mathbb{R}^d\}.$$

Consistent estimation of the parameter θ_0 using instrumental variables in such function classes has been studied extensively in the econometric literature [e.g., Amemiya, 1974, Jorgenson and Laffont, 1974, Kelejian, 1971]. These works also contain rigorous results on how instrumental variable estimators of θ_0 are constructed and under which conditions consistency (and thus identifiability) holds. Here, we give an argument on why the presence of the exogenous variables A yields identifiability under certain regularity conditions. Assume that $\mathbb{E}[h_1(H, \varepsilon_Y)|A] = 0$, which implies that the true causal function $f(\cdot, \theta_0)$ satisfies the population orthogonality condition

$$\mathbb{E}[l(A)^\top (Y - f(X, \theta_0))] = \mathbb{E}\big[l(A)^\top \mathbb{E}[h_1(H, \varepsilon_Y)|A]\big] = 0, \quad \text{(B.2.1)}$$

for some measurable mapping $l : \mathbb{R}^q \to \mathbb{R}^g$, for some $g \in \mathbb{N}_{>0}$. Clearly, θ_0 is identified from the observational distribution if the map $\theta \mapsto \mathbb{E}[l(A)^\top (Y - f(X, \theta))]$ is zero if and only if $\theta = \theta_0$. Furthermore, since $\theta \mapsto f(x, \theta)$ is differentiable for all $x \in \mathbb{R}^d$, the mean value theorem yields that, for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, there exists an intermediate point $\tilde{\theta}(x, \theta, \theta_0)$ on the line segment between θ and θ_0 such that

$$f(x, \theta) - f(x, \theta_0) = D_\theta f(x, \tilde{\theta}(x, \theta, \theta_0))(\theta - \theta_0),$$

where, for each $x \in \mathbb{R}^d$, $D_\theta f(x, \theta) \in \mathbb{R}^{1 \times p}$ is the derivative of $\theta \mapsto f(x, \theta)$ evaluated in θ. Composing the above expression with the random vector X, multiplying with $l(A)$ and taking expectations yields that

$$\begin{aligned} &\mathbb{E}[l(A)(Y - f(X, \theta_0))] - \mathbb{E}[l(A)(Y - f(X, \theta))] \\ &\quad = \mathbb{E}[l(A) D_\theta f(X, \tilde{\theta}(X, \theta, \theta_0))](\theta_0 - \theta). \end{aligned}$$

Hence, if $\mathbb{E}[l(A) D_\theta f(X, \tilde{\theta}(X, \theta, \theta_0))] \in \mathbb{R}^{g \times p}$ is of rank p for all $\theta \in \Theta$ (which implies $g \geq p$), then θ_0 is identifiable as it is the only parameter that satisfies the population orthogonality condition of (B.2.1). As θ_0 uniquely determines the entire function, we get identifiability of $f \equiv f(\cdot, \theta_0)$, not only on the support of X but the entire domain $\mathbb{R}^d$, i.e., both Assumptions 3.1 and 3.2 are satisfied. In the case that $\theta \mapsto f(x, \theta)$ is linear, i.e. $f(x, \theta) = f(x)^T \theta$ for all $x \in \mathbb{R}^d$, the above rank condition reduces to $\mathbb{E}[l(A) f(X)^T] \in \mathbb{R}^{g \times p}$ having rank

p (again, implying that $g \geq p$). Furthermore, when $(x, \theta) \mapsto f(x, \theta)$ is bilinear, a reparametrization of the parameter space ensures that $f(x, \theta) = x^T\theta$ for $\theta \in \Theta \subseteq \mathbb{R}^d$. In this case, the rank condition can be reduced to the well-known rank condition for identification in a linear SCM, namely that $\mathbb{E}[AX^T] \in \mathbb{R}^{q\times p}$ is of rank p.

Finally, identifiability and methods of consistent estimation of the causal function have also been studied for non-parametric function classes. The conditions for identification are rather technical, however, and we refer the reader to Newey [2013], Newey and Powell [2003] for further details.

B.3. Choice of test statistic

By considering the variables $B(X) = (B_1(X), \ldots, B_k(X))$ and $C(A) = (C_1(A), \ldots, C_k(A))$ as vectors of covariates and instruments, respectively, our setting in Section 3.5.2 reduces to the classical (just-identified) linear IV setting. We could therefore use a test statistics similar to the one propsed by the PULSE [Jakobsen and Peters, 2020]. With a notation that is slightly adapted to our setting, this estimator tests $\tilde{H}_0(\theta)$ using the test statistic

$$T_n^1(\theta) = c(n)\frac{\|\mathbf{P}(\mathbf{Y} - \mathbf{B}\theta)\|_2^2}{\|\mathbf{Y} - \mathbf{B}\theta\|_2^2},$$

where $\mathbf{P}$ is the projection onto the columns of $\mathbf{C}$, and $c(n)$ is some function with $c(n) \sim n$ as $n \to \infty$. Under the null hypothesis, T_n^1 converges in distribution to the χ_k^2 distribution, and diverges to infinity in probability under the general alternative. Using this test statistic, $\tilde{H}_0(\theta)$ is rejected if and only if $T_n^1(\theta) > q(\alpha)$, where $q(\alpha)$ is the $(1-\alpha)$-quantile of the χ_k^2 distribution. The acceptance region of this test statistic is asymptotically equivalent with the confidence region of the Anderson-Rubin test [Anderson and Rubin, 1949] for the causal parameter θ^0. Using the above test results in a consistent estimator for θ^0 [Jakobsen and Peters, 2020, Theorem 3.12]; the proof exploits the particular form of T_n^1 without explicitly imposing that assumptions (C1) and (C2) hold.

If the number k of basis functions is large, however, numerical experiments suggest that the above test has low power in finite sample

settings. As default, our algorithm therefore uses a different test based on a penalized regression approach. This test has been proposed in Chen et al. [2014] for inference in nonparametric regression models. We now introduce this procedure with a notation that is adapted to our setting. For every $\theta \in \mathbb{R}^k$, let $R_\theta = Y - B(X)^\top \theta$ be the residual associated with θ. We then test the slightly stronger hypothesis

$$\bar{H}_0(\theta): \text{ there exists } \sigma_\theta^2 > 0 \text{ such that } \mathbb{E}[R_\theta \mid A] \overset{\text{a.s.}}{=} 0 \text{ and } \mathrm{Var}[R_\theta \mid A] = \sigma_\theta^2$$

against the alternative that $\mathbb{E}[R_\theta \mid A] = m(A)$ for some smooth function m. To see that the above hypothesis implies $\tilde{H}_0(\theta)$ (and therefore $H_0(\theta)$, see Section 3.5.2.1), let $\theta \in \mathbb{R}^k$ be such that $\bar{H}_0(\theta)$ holds true. Then,

$$\begin{aligned}\mathbb{E}[C(A)(Y - B(X)^\top \theta)] &= \mathbb{E}[C(A)R_\theta] = \mathbb{E}[\mathbb{E}[C(A)R_\theta \mid A]] \\ &= \mathbb{E}[C(A)\mathbb{E}[R_\theta \mid A]] = 0,\end{aligned}$$

showing that also $\tilde{H}_0(\theta)$ holds true. Thus, if $\tilde{H}_0(\theta)$ is false, then also $\bar{H}_0(\theta)$ is false. As a test statistic $T_n^2(\theta)$ for $\bar{H}_0(\theta)$, we use (up to a normalization) the squared norm of a penalized regression estimate of m, evaluated at the data $\mathbf{A}$, i.e., the TSLS loss $\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$. In the fixed design case, where $\mathbf{A}$ is non-random, it has been shown that, under $\bar{H}_0(\theta)$ and certain additional regularity conditions, it holds that

$$\frac{\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 - \sigma_\theta^2 c_n}{\sigma_\theta^2 d_n} \xrightarrow{\text{d}} \mathcal{N}(0,1),$$

where c_n and d_n are known functions of $\mathbf{C}$, $\mathbf{M}$ and δ [Chen et al., 2014, Theorem 1]. The authors further state that the above convergence is unaffected by exchanging σ_θ^2 with a consistent estimator $\hat{\sigma}_\theta^2$, which motivates our use of the test statistic

$$T_n^2(\theta) := \frac{\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 - \hat{\sigma}_{\theta,n}^2 c_n}{\hat{\sigma}_{\theta,n}^2 d_n},$$

where $\hat{\sigma}_{\theta,n}^2 := \frac{1}{n-1}\sum_{i=1}^n \|(\mathbf{I}_n - \mathbf{P}_\delta)(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$. As a rejection threshold $q(\alpha)$ we use the $1-\alpha$ quantile of a standard normal distribution. For results on the asymptotic power of the test defined by T^2, we refer to Section 2.3 in Chen et al. [2014].

In our software package, both of the above tests are available options.

B.4. Addition to experiments

B.4.1. Sampling of the causal function

To ensure linear extrapolation of the causal function, we have chosen a function class consisting of natural cubic splines, which, by construction, extrapolate linearly outside the boundary knots. We now describe in detail how we sample functions from this class for the experiments in Section 3.5.2.4. Let $q_{\min}$ and $q_{\max}$ be the respective 5%- and 95% quantiles of X, and let $B_1, \ldots, B_4$ be a basis of natural cubic splines corresponding to 5 knots placed equidistantly between $q_{\min}$ and $q_{\max}$. We then sample coefficients $\beta_i \overset{\text{iid}}{\sim} \mathcal{U}(-1, 1)$, $i = 1, \ldots, 4$, and construct f as $f = \sum_{i=1}^{4} \beta_i B_i$. For illustration, we have included 18 realizations in Figure B.1.

B.4.2. Violations of the linear extrapolation assumption

We have assumed that the true causal function extrapolates linearly outside the 90% quantile range of X. We now investigate the performance of our method for violations of this assumption. To do so, we again sample from the model (3.5.4), with $\alpha_A = \alpha_H = \alpha_\varepsilon = 1/\sqrt{3}$. For each data set, the causal function is sampled as follows. Let $q_{\min}$ and $q_{\max}$ be the 5%- and 95% quantiles of X. We first generate a function $\tilde{f}$ that linearly extrapolates outside $[q_{\min}, q_{\max}]$ as described in Section B.4.1. For a given threshold κ, we then draw $k_1, k_2 \overset{\text{iid}}{\sim} \mathcal{U}(-\kappa, \kappa)$ and construct f for every $x \in \mathbb{R}$ by

$$f(x) = \tilde{f}(x) + \frac{1}{2} k_1 ((x - q_{\min})_-)^2 + \frac{1}{2} k_2 ((x - q_{\max})_+)^2,$$

such that the curvature of f on $(-\infty, q_{\min}]$ and $[q_{\max}, \infty)$ is k_1 and k_2, respectively. Figure B.2 shows results for $\kappa = 0, 1, 2, 3, 4$. As the curvature increases, the ability to generalize decreases.

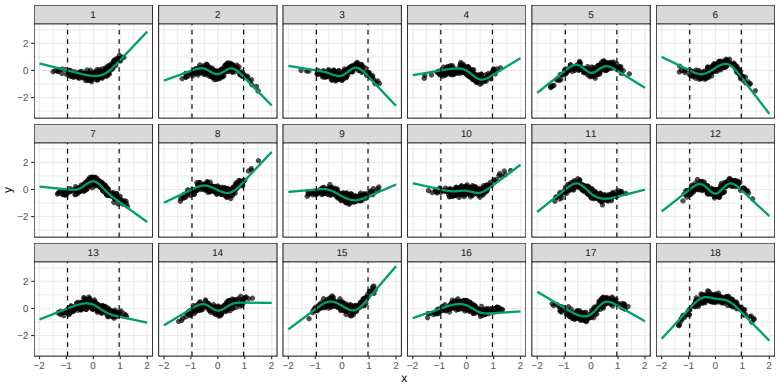


FIGURE B.1. The plots show independent realizations of the causal function that is used in all our experiments. These are sampled from a linear space of natural cubic splines, as described in Appendix B.4.1. To ensure a fair comparison with the alternative method, NPREGIV, the true causal function is chosen from a model class different from the one assumed by the NILE.

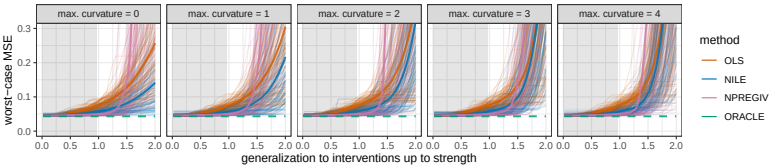


FIGURE B.2. Worst-case mean squared error for increasingly strong violations of the linear extrapolation assumption. The grey area marks the inner 90 % quantile range of X in the training distribution. As the curvature of f outside the domain of the observed data increases, it becomes difficult to predict the interventional behavior of Y for strong interventions. However, even in situations where the linear extrapolation assumption is strongly violated, it remains beneficial to extrapolate linearly.

B.5. Proofs

B.5.1. Proof of Proposition 3.1

Proof. Assume that $\mathcal{I}$ is a set of interventions on X with at least one confounding-removing intervention. Let $i \in \mathcal{I}$ and $f_\diamond \in \mathcal{F}$, then we have the following expansion

$$\begin{aligned}\mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] = \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_{M(i)}[\xi_Y^2] \\ + 2\mathbb{E}_{M(i)}[\xi_Y(f(X) - f_\diamond(X))],\end{aligned} \tag{B.5.1}$$

where $\xi_Y = h_1(H, \varepsilon_Y)$. For any intervention $i \in \mathcal{I}$ the causal function always yields an identical loss. In particular, it holds that

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2] = \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[\xi_Y^2] = \mathbb{E}_M[\xi_Y^2], \tag{B.5.2}$$

where we used that the distribution of ξ_Y is not affected by an intervention on X. The loss of the causal function can never be better than the minimax loss, that is,

$$\inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2] = \mathbb{E}_M[\xi_Y^2]. \tag{B.5.3}$$

In other words, the minimax solution (if it exists) is always better than or equal to the causal function. We will now show that when $\mathcal{I}$ contains at least one confounding-removing intervention, then the minimax loss is dominated by any such intervention.

Fix $i_0 \in \mathcal{I}$ to be a confounding-removing intervention and let (X, Y, H, A) be generated by the SCM $M(i_0)$. Recall that there exists a map ψ^{i_0} such that $X := \psi^{i_0}(g, h_2, A, H, \varepsilon_X, I^{i_0})$ and that $X \perp\!\!\!\perp H$ as i_0 is a confounding-removing intervention. Furthermore, since the vectors A, H, ε_X, ε_Y and I^{i_0} are mutually independent, we have that $(X, H) \perp\!\!\!\perp \varepsilon_Y$ which together with $X \perp\!\!\!\perp H$ implies X, H and ε_Y are mutually independent, and hence $X \perp\!\!\!\perp h_1(H, \varepsilon_Y)$. Using this independence we get that $\mathbb{E}_{M(i_0)}[\xi_Y(f(X) - f_\diamond(X))] = \mathbb{E}_M[\xi_Y]\mathbb{E}_{M(i_0)}[(f(X) - f_\diamond(X))]$. Hence, (B.5.1) for the intervention i_0 together with the modeling assumption $\mathbb{E}_M[\xi_Y] = 0$ implies that for all $f_\diamond \in \mathcal{F}$,

$$\mathbb{E}_{M(i_0)}[(Y - f_\diamond(X))^2] = \mathbb{E}_{M(i_0)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] \geq \mathbb{E}_M[\xi_Y^2].$$

This proves that the smallest loss at a confounding-removing intervention is achieved by the causal function. Denoting the non-empty subset of confounding-removing interventions by $\mathcal{I}_{\text{cr}} \subseteq \mathcal{I}$, this implies

$$\begin{aligned} \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] &\geq \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}_{\text{cr}}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \\ &\geq \inf_{f_\diamond \in \mathcal{F}} \mathbb{E}_{M(i_0)}[(Y - f_\diamond(X))^2] \\ &= \mathbb{E}_M[\xi_Y^2]. \end{aligned} \tag{B.5.4}$$

Combining (B.5.3) and (B.5.4) it immediately follows that

$$\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] = \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2],$$

and hence

$$f \in \arg\min_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2],$$

which completes the proof of Proposition 3.1. □

B.5.2. Proof of Proposition 3.2

Proof. Let $\mathcal{F}$ be the class of all linear functions and let $\mathcal{I}$ denote the set of interventions on X that satisfy

$$\sup_{i\in\mathcal{I}} \lambda_{\min}\big(\mathbb{E}_{M(i)}\big[XX^\top\big]\big) = \infty.$$

We claim that the causal function $f(x) = b^\top x$ is the unique minimax solution of (3.3.1). We prove the result by contradiction. Let $\bar{f} \in \mathcal{F}$ (with $\bar{f}(x) = \bar{b}^\top x$) be such that

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - \bar{b}^\top X)^2] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - b^\top X)^2],$$

and assume that $\|\bar{b} - b\|_2 > 0$. For a fixed $i \in \mathcal{I}$, we get the following bound

$$\begin{aligned}\mathbb{E}_{M(i)}[(b^\top X - \bar{b}^\top X)^2] &= (b-\bar{b})^\top \mathbb{E}_{M(i)}[XX^\top](b-\bar{b}) \\ &\geq \lambda_{\min}(\mathbb{E}_{M(i)}[XX^\top])\|b-\bar{b}\|_2^2.\end{aligned}$$

Since we assumed that the minimal eigenvalue is unbounded, this means that we can choose $i \in \mathcal{I}$ such that $\mathbb{E}_{M(i)}[(b^\top X - \bar{b}^\top X)^2]$ can be arbitrarily large. However, applying Proposition 3.3, this leads to a contradiction since $\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(b^\top X - \bar{b}^\top X)^2] \leq 4\,\mathrm{Var}_M(\xi_Y)$ cannot be satisfied. Therefore, it must holds that $\bar{b} = b$, which moreover implies that f is indeed a solution to the minimax problem $\arg\min_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2]$, as it achieves the lowest possible objective value. This completes the proof of Proposition 3.2. □

B.5.3. Proof of Proposition 3.3

Proof. Let $\mathcal{I}$ be a set of interventions on X or A and let $f_\diamond \in \mathcal{F}$ with

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2]. \tag{B.5.5}$$

For any $i \in \mathcal{I}$, the Cauchy-Schwartz inequality implies that

$$\begin{aligned}
&\mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \\
&= \mathbb{E}_{M(i)}[(f(X) + \xi_Y - f_\diamond(X))^2] \\
&= \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_{M(i)}[\xi_Y^2] + 2\mathbb{E}_{M(i)}[\xi_Y(f(X) - f_\diamond(X))] \\
&\geq \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] - 2\left(\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2]\mathbb{E}_M[\xi_Y^2]\right)^{\frac{1}{2}}.
\end{aligned}$$

A similar computation shows that the causal function f satisfies

$$\mathbb{E}_{M(i)}[(Y - f(X))^2] = \mathbb{E}_M[\xi_Y^2].$$

So by condition (B.5.5) this implies for any $i \in \mathcal{I}$ that

$$\begin{aligned}
&\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] \\
&\qquad - 2\left(\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2]\mathbb{E}_M[\xi_Y^2]\right)^{\frac{1}{2}} \leq \mathbb{E}_M[\xi_Y^2],
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] \leq 2\sqrt{\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2]\mathbb{E}_M[\xi_Y^2]} \\
\iff \quad &\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] \leq 4\mathbb{E}_M[\xi_Y^2].
\end{aligned}$$

As this inequality holds for all $i \in \mathcal{I}$, we can take the supremum over all $i \in \mathcal{I}$, which completes the proof of Proposition 3.3. □

B.5.4. Proof of Proposition 3.4

Proof. As argued before, we have that for all $i \in \mathcal{I}_1$,

$$\mathbb{E}_{M(i)}\left[(Y - f(X))^2\right] = \mathbb{E}_{M(i)}\left[\xi_Y^2\right] = \mathbb{E}_M\left[\xi_Y^2\right].$$

Let now $f_1^* \in \mathcal{F}$ be a minimax solution w.r.t. $\mathcal{I}_1$. Then, using that the causal function f lies in $\mathcal{F}$, it holds that

$$\sup_{i \in \mathcal{I}_1} \mathbb{E}_{M(i)}\left[(Y - f_1^*(X))^2\right] \leq \sup_{i \in \mathcal{I}_1} \mathbb{E}_{M(i)}\left[(Y - f(X))^2\right] = \mathbb{E}_M\left[\xi_Y^2\right].$$

Moreover, if $\mathcal{I}_2 \subseteq \mathcal{I}_1$, then it must also hold that

$$\sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\left[(Y - f_1^*(X))^2\right] \leq \mathbb{E}_M\left[\xi_Y^2\right] = \sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\left[(Y - f(X))^2\right].$$

To prove the second part, we give a one-dimensional example. Let $\mathcal{F}$ be linear (i.e., $f(x) = bx$) and let $\mathcal{I}_1$ consist of shift interventions on X of the form

$$X^i := g(A^i) + h_2(H^i, \varepsilon_X^i) + c,$$

with $c \in [0, K]$. Then, the minimax solution f_1^* (where $f_1^*(x) = b_1^* x$) with respect to $\mathcal{I}_1$ is not equal to the causal function f as long as $\mathrm{Cov}(X, \xi_Y)$ is strictly positive. This can be seen by explicitly computing the OLS estimator for a fixed shift c and observing that the worst-case loss is attained at $c = K$. Now let $\mathcal{I}_2$ be a set of interventions of the same form as $\mathcal{I}_1$ but including shifts with $c > K$ such that $\mathcal{I}_2 \not\subseteq \mathcal{I}_1$. Since $\mathcal{F}$ consists of linear functions, we know that the loss $\mathbb{E}_{M(i)}\big[(Y - f_1^*(X))^2\big]$ can become arbitrarily large, since

$$\begin{aligned}
&\mathbb{E}_{M(i)}\big[(Y - f_1^*(X))^2\big] \\
&\quad = (b - b_1^*)^2 \mathbb{E}_{M(i)}[X^2] + \mathbb{E}_M[\xi_Y^2] + 2(b - b_1^*)\mathbb{E}_{M(i)}[\xi_Y X] \\
&\quad = (b - b_1^*)^2 (c^2 + \mathbb{E}_M[X^2] + 2c\mathbb{E}_M[X]) + \mathbb{E}_M[\xi_Y^2] \\
&\qquad + 2(b - b_1^*)(\mathbb{E}_M[\xi_Y X] + \mathbb{E}_M[\xi_Y]c),
\end{aligned}$$

and $(b - b^*)^2 > 0$. In contrast, the loss for the causal function is always $\mathbb{E}_M[\xi_Y^2]$, so the worst-case loss of f_1^* becomes arbitrarily worse than that of f. This completes the proof of Proposition 3.4. □

B.5.5. Proof of Proposition 3.5

Proof. Let $\varepsilon > 0$. By definition of the infimum, we can find $f^* \in \mathcal{F}$ such that

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] \right| \leq \varepsilon.$$

Let now $\tilde{M} \in \mathcal{M}$ be s.t. $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. By assumption, the left-hand side of the above inequality is unaffected by substituting M for $\tilde{M}$, and the result thus follows. □

B.5.6. Proof of Proposition 3.6

Proof. Let $\mathcal{I}$ be a well-behaved set of interventions on X. We consider two cases; (A) all interventions in $\mathcal{I}$ are confounding-preserving

and (B) there is at least one intervention in $\mathcal{I}$ that is confounding-removing.

Case (A): In this case, we prove the result in two steps: (i) We show that (A, ξ_X, ξ_Y) is identified from the observational distribution $\mathbb{P}_M$. (ii) We show that this implies that the intervention distributions (X^i, Y^i), $i \in \mathcal{I}$, are also identified from the observational distribution, and conclude by using Proposition 3.5. Some of the details will be slightly technical because we allow for a large class of distributions (e.g., there is no assumption on the existence of densities).

We begin with step (i). In this case, $\mathcal{I}$ is a set of confounding-preserving interventions on X, and we have that $\mathrm{supp}_{\mathcal{I}}(X) \subseteq \mathrm{supp}(X)$. Fix $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$ and let $(\tilde{X}, \tilde{Y}, \tilde{H}, \tilde{A})$ be generated by the SCM of $\tilde{M}$. We have that $(X, Y, A) \stackrel{\mathrm{d}}{=} (\tilde{X}, \tilde{Y}, \tilde{A})$ and by Assumption 3.1, we have that $f \equiv \tilde{f}$ on $\mathrm{supp}(X)$, hence $f(X) \stackrel{\text{a.s.}}{=} \tilde{f}(X)$. Further, fix any $B \in \mathcal{B}(\mathbb{R}^p)$ (i.e., in the Borel sigma-algebra on $\mathbb{R}^p$) and note that

$$\begin{aligned}\mathbb{E}_M[\mathbb{1}_B(A)X|A] &= \mathbb{E}_M[\mathbb{1}_B(A)g(A) + \mathbb{1}_B(A)h_2(H, \varepsilon_X)|A] \\ &= \mathbb{E}_M[\mathbb{1}_B(A)g(A)|A] + \mathbb{1}_B(A)\mathbb{E}[h_2(H, \varepsilon_X)] \\ &= \mathbb{1}_B(A)g(A),\end{aligned}$$

almost surely. Here, we have used our modeling assumption $\mathbb{E}[h_2(H, \varepsilon_X)] = 0$. Hence, by similar arguments for $\mathbb{E}_{\tilde{M}}(\mathbb{1}_B(\tilde{A})\tilde{X}|\tilde{A})$ and the fact that $(X, Y, A) \stackrel{\mathrm{d}}{=} (\tilde{X}, \tilde{Y}, \tilde{A})$ we have that

$$\mathbb{1}_B(A)g(A) \stackrel{\text{a.s.}}{=} \mathbb{E}_M(\mathbb{1}_B(A)X|A) \stackrel{\mathrm{d}}{=} \mathbb{E}_{\tilde{M}}(\mathbb{1}_B(\tilde{A})\tilde{X}|\tilde{A}) \stackrel{\text{a.s.}}{=} \mathbb{1}_B(\tilde{A})\tilde{g}(\tilde{A}).$$

We conclude that $\mathbb{1}_B(A)g(A) \stackrel{\mathrm{d}}{=} \mathbb{1}_B(\tilde{A})\tilde{g}(\tilde{A})$ for any $B \in \mathcal{B}(\mathbb{R}^p)$. Let $\mathbb{P}$ and $\tilde{\mathbb{P}}$ denote the respective background probability measures on which the random elements (X, Y, H, A) and $(\tilde{X}, \tilde{Y}, \tilde{H}, \tilde{A})$ are defined. Fix any $F \in \sigma(A)$ (i.e., in the sigma-algebra generated by A) and note that there exists a $B \in \mathcal{B}(\mathbb{R}^p)$ such that $F = \{A \in B\}$.

Since $A \overset{\mathrm{d}}{=} \tilde{A}$, we have that,

$$\begin{aligned}\int_F g(A)\,\mathrm{d}\mathbb{P} &= \int \mathbb{1}_B(A) g(A)\,\mathrm{d}\mathbb{P} = \int \mathbb{1}_B(\tilde{A})\tilde{g}(\tilde{A})\,\mathrm{d}\tilde{\mathbb{P}} \\ &= \int \mathbb{1}_B(A)\tilde{g}(A)\,\mathrm{d}\mathbb{P} = \int_F \tilde{g}(A)\,\mathrm{d}\mathbb{P}.\end{aligned}$$

Both $g(A)$ and $\tilde{g}(A)$ are $\sigma(A)$-measurable and they agree integral-wise over every set $F \in \sigma(A)$, so we must have that $g(A) \overset{\text{a.s.}}{=} \tilde{g}(A)$. With $\eta(a,b,c) = (a, c - \tilde{f}(b), b - \tilde{g}(a))$ we have that

$$\begin{aligned}(A, \xi_Y, \xi_X) &\overset{\text{a.s.}}{=} (A, Y - \tilde{f}(X), X - \tilde{g}(A)) = \eta(A, X, Y) \\ &\overset{\mathrm{d}}{=} \eta(\tilde{A}, \tilde{X}, \tilde{Y}) = (\tilde{A}, \tilde{\xi}_Y, \tilde{\xi}_X),\end{aligned}$$

so $(A, \xi_Y, \xi_X) \overset{\mathrm{d}}{=} (\tilde{A}, \tilde{\xi}_Y, \tilde{\xi}_X)$. This completes step (i).

Next, we proceed with step (ii). Take an arbitrary intervention $i \in \mathcal{I}$ and let $\phi^i, I^i, \tilde{I}^i$ with $I^i \overset{\mathrm{d}}{=} \tilde{I}^i$, $I^i \perp\!\!\!\perp (\varepsilon^i_X, \varepsilon^i_Y, \varepsilon^i_H, \varepsilon^i_A) \sim Q$ and $\tilde{I}^i \perp\!\!\!\perp (\tilde{\varepsilon}^i_X, \tilde{\varepsilon}^i_Y, \tilde{\varepsilon}^i_H, \tilde{\varepsilon}^i_A) \sim \tilde{Q}$ be such that the structural assignments for X^i and $\tilde{X}^i$ in $M(i)$ and $\tilde{M}(i)$, respectively, are given as

$$\begin{aligned}X^i &:= \phi^i(A^i, g(A^i), h_2(H^i, \varepsilon^i_X), I^i) \quad \text{and} \\ \tilde{X}^i &:= \phi^i(\tilde{A}^i, \tilde{g}(\tilde{A}^i), \tilde{h}_2(\tilde{H}^i, \tilde{\varepsilon}^i_X), \tilde{I}^i).\end{aligned}$$

Define $\xi^i_X := h_2(H^i, \varepsilon^i_X)$, $\xi^i_Y := h_1(H^i, \varepsilon^i_Y)$, $\tilde{\xi}^i_X := \tilde{h}_2(\tilde{H}^i, \tilde{\varepsilon}^i_X)$ and $\tilde{\xi}^i_Y := \tilde{h}_1(\tilde{H}^i, \tilde{\varepsilon}^i_Y)$. Then, it holds that

$$(A^i, \xi^i_X, \xi^i_Y) \overset{\mathrm{d}}{=} (A, \xi_X, \xi_Y) \overset{\mathrm{d}}{=} (\tilde{A}, \tilde{\xi}_X, \tilde{\xi}_Y) \overset{\mathrm{d}}{=} (\tilde{A}^i, \tilde{\xi}^i_X, \tilde{\xi}^i_Y),$$

where we used step (i), that (A^i, ξ^i_X, ξ^i_Y) and (A, ξ_X, ξ_Y) are generated by identical functions of the noises and that $(\varepsilon_X, \varepsilon_Y, \varepsilon_H, \varepsilon_A)$ and $(\varepsilon^i_X, \varepsilon^i_Y, \varepsilon^i_H, \varepsilon^i_A)$ have identical distributions. Adding a random variable with the same distribution, that is mutually independent with all other variables, on both sides does not change the distribution of the bundle, hence

$$(A^i, \xi^i_X, \xi^i_Y, I^i) \overset{\mathrm{d}}{=} (\tilde{A}^i, \tilde{\xi}^i_X, \tilde{\xi}^i_Y, \tilde{I}^i).$$

Define $\kappa(a,b,c,d) := (\phi^i(a,\tilde{g}(a),b,d), \tilde{f}(\phi^i(a,\tilde{g}(a),b,d)) + c)$. As shown in step (i) above, we have that $g(A^i) \overset{\text{a.s.}}{=} \tilde{g}(A^i)$. Furthermore, since $\text{supp}(X^i) \subseteq \text{supp}(X)$ we have that $f(X^i) \overset{\text{a.s.}}{=} \tilde{f}(X^i)$, and hence

$$\begin{aligned}(X^i, Y^i) &\overset{\text{a.s.}}{=} (X^i, \tilde{f}(X^i) + \xi_Y^i) \\ &= (\phi^i(A^i, g(A^i), \xi_X^i, I^i), \tilde{f}(\phi^i(A^i, g(A^i), \xi_X^i, I^i)) + \xi_Y^i) \\ &\overset{\text{a.s.}}{=} (\phi^i(A^i, \tilde{g}(A^i), \xi_X^i, I^i), \tilde{f}(\phi^i(A^i, \tilde{g}(A^i), \xi_X^i, I^i)) + \xi_Y^i) \\ &= \kappa(A^i, \xi_X^i, \xi_Y^i, I^i) \overset{\text{d}}{=} \kappa(\tilde{A}^i, \tilde{\xi}_X^i, \tilde{\xi}_Y^i, \tilde{I}^i) = (\tilde{X}^i, \tilde{Y}^i).\end{aligned}$$

Thus, $\mathbb{P}^{(X,Y)}_{M(i)} = \mathbb{P}^{(X,Y)}_{\tilde{M}(i)}$, which completes step (ii). Since $i \in \mathcal{I}$ was arbitrary, the result now follows from Proposition 3.5.

Case (B): Assume that the set of interventions $\mathcal{I}$ contains at least one confounding-removing intervention. Let $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ be such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. Then, by Proposition 3.1, it follows that the causal function $\tilde{f}$ is a minimax solution w.r.t. $(\tilde{M}, \mathcal{I})$. By Assumption 3.1, we further have that $\tilde{f}$ and f coincide on $\text{supp}(X) \supseteq \text{supp}_{\mathcal{I}}(X)$. Hence, it follows that

$$\begin{aligned}\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f_\diamond(X))^2] &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - \tilde{f}(X))^2] \\ &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f(X))^2],\end{aligned}$$

showing that also f is a minimax solution w.r.t. $(\tilde{M}, \mathcal{I})$. This completes the proof of Proposition 3.6. □

B.5.7. Proof of Proposition 3.7

Proof. We first show that the causal parameter β is not a minimax solution. Let $u := \sup \mathcal{I} < \infty$, since $\mathcal{I}$ is bounded, and take $b =$

$\beta + 1/(\sigma u)$. By an explicit computation we get that

$$\begin{aligned}\inf_{b_\diamond \in \mathbb{R}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - b_\diamond X)^2\right] &\leq \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - bX)^2\right] \\ &= \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(\varepsilon_Y + \tfrac{1}{\sigma} H - \tfrac{1}{\sigma u} i H)^2\right] \\ &= \sup_{i \in \mathcal{I}} \left[1 + \left(1 - \tfrac{i}{u}\right)^2\right] \\ &< 2 = \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - \beta X)^2\right],\end{aligned}$$

where the last inequality holds because $0 < 1 + (1 - i/u)^2 < 2$ for all $i \in \mathcal{I}$, and since $\mathcal{I} \subseteq \mathbb{R}_{>0}$ is compact with upper bound u. Hence,

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - \beta X)^2\right] - \inf_{b_\diamond \in \mathbb{R}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - b_\diamond X)^2\right] > 0,$$

proving that the causal parameter is not a minimax solution for model M w.r.t. $(\mathcal{F}, \mathcal{I})$. Recall that in order to prove that $(\mathbb{P}_M, \mathcal{M})$ does not generalize with respect to $\mathcal{I}$ we have to show that there exists an $\varepsilon > 0$ such that for all $b \in \mathbb{R}$ it holds that

$$\sup_{\tilde{M} : \mathbb{P}_{\tilde{M}} = \mathbb{P}_M} \left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - bX)^2\right] - \inf_{b_\diamond \in \mathbb{R}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - b_\diamond X)^2\right] \right| \geq \varepsilon.$$

Thus, it remains to show that for all $b \neq \beta$ there exists a model $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$ such that the generalization loss is bounded below uniformly by a positive constant. We will show the stronger statement that for any $b \neq \beta$, there exists a model $\tilde{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, such that under $\tilde{M}$, b results in arbitrarily large generalization error. Let $c > 0$ and $i_0 \in \mathcal{I}$. Define

$$\tilde{\sigma} \coloneqq \frac{\operatorname{sign}\left((\beta - b) i_0\right) \sqrt{1 + c} - 1}{(\beta - b) i_0} > 0,$$

and let $\tilde{M} \coloneqq M(\gamma, \beta, \tilde{\sigma}, Q)$. By construction of the model class $\mathcal{M}$, it holds that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. Furthermore, by an explicit computation

we get that

$$\begin{aligned}
&\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y-bX)^2\big] \\
&\quad \geq \mathbb{E}_{\tilde{M}(i_0)}\big[(Y-bX)^2\big] = \mathbb{E}_{\tilde{M}(i_0)}\big[((\beta-b)i_0 H + \varepsilon_Y + \tfrac{1}{\tilde{\sigma}}H)^2\big] \\
&\quad = \mathbb{E}_{\tilde{M}(i_0)}\big[([(\beta-b)i_0\tilde{\sigma}+1]\varepsilon_H + \varepsilon_Y)^2\big] = [(\beta-b)i_0\tilde{\sigma}+1]^2 + 1 \\
&\quad = ((\beta-b)i_0\tilde{\sigma})^2 + 2(\beta-b)i_0\tilde{\sigma} + 2 \\
&\quad = (\operatorname{sign}((\beta-b)i_0)\sqrt{1+c} - 1)^2 + 2\operatorname{sign}((\beta-b)i_0)\sqrt{1+c} \\
&\quad = c+2. \qquad \text{(B.5.6)}
\end{aligned}$$

Finally, by definition of the infimum, it holds that

$$\inf_{b_\diamond\in\mathbb{R}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y-b_\diamond X)^2\big] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y-\beta X)^2\big] = 2. \quad \text{(B.5.7)}$$

Combining (B.5.6) and (B.5.7) yields that the generalization error is bounded below by c. That is,

$$\big|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y-bX)^2\big] - \inf_{b_\diamond\in\mathbb{R}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y-b_\diamond X)^2\big]\big| \geq c.$$

The above results make no assumptions on γ, and hold true, in particular, if $\gamma \neq 0$ (in which case Assumption 3.1 is satisfied, see Appendix B.2). This completes the proof of Proposition 3.7. □

B.5.8. Proof of Proposition 3.8

Proof. Let $\tilde{M} \in \mathcal{M}$ be such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. By Assumptions 3.1 and 3.2, it holds that $f \equiv \tilde{f}$. The proof now proceeds analogously to that of Proposition 3.6. □

B.5.9. Proof of Proposition 3.9

Proof. By Assumption 3.1, f is identified on $\operatorname{supp}^M(X)$ by the observational distribution $\mathbb{P}_M$. Let $\mathcal{I}$ be a set of interventions containing at least one confounding-removing intervention. For any

$\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$, Proposition 3.1 yields that the causal function is a minimax solution. That is,

$$\begin{aligned}\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big] &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - \tilde{f}(X))^2\big] \\ &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[\xi_Y^2] = \mathbb{E}_{\tilde{M}}[\xi_Y^2], \quad \text{(B.5.8)}\end{aligned}$$

where we used that any intervention $i \in \mathcal{I}$ does not affect the distribution of $\xi_Y = \tilde{h}_2(H, \varepsilon_Y)$. Now, assume that $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ satisfies $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. Since $(\mathbb{P}_M, \mathcal{M})$ satisfies Assumption 3.1, we have that $f \equiv \tilde{f}$ on $\text{supp}^M(X) = \text{supp}^{\tilde{M}}(X)$. Let f^* be any function in $\mathcal{F}$ such that $f^* = f$ on $\text{supp}^M(X)$. We first show that $\|\tilde{f} - f^*\|_{\mathcal{I},\infty} \leq 2\delta K$, where $\|f\|_{\mathcal{I},\infty} := \sup_{x \in \text{supp}^M_{\mathcal{I}}(X)} \|f(x)\|$. By the mean value theorem, for all $f_\diamond \in \mathcal{F}$ it holds that $|f_\diamond(x) - f_\diamond(y)| \leq K\|x - y\|$, for all $x, y \in \mathcal{D}$. For any $x \in \text{supp}^M_{\mathcal{I}}(X)$ and $y \in \text{supp}^M(X)$ we have

$$\begin{aligned}\big|\tilde{f}(x) - f^*(x)\big| &= \big|\tilde{f}(x) - \tilde{f}(y) + f^*(y) - f^*(x)\big| \\ &\leq \big|\tilde{f}(x) - \tilde{f}(y)\big| + \big|f^*(y) - f^*(x)\big| \\ &\leq 2K\|x - y\|,\end{aligned}$$

where we used the fact that $\tilde{f}(y) = f(y) = f^*(y)$, for all $y \in \text{supp}^M(X)$. In particular, it holds that

$$\begin{aligned}\|\tilde{f} - f^*\|_{\mathcal{I},\infty} &= \sup_{x \in \text{supp}^M_{\mathcal{I}}(X)} \big|\tilde{f}(x) - f^*(x)\big| \\ &\leq 2K \sup_{x \in \text{supp}^M_{\mathcal{I}}(X)} \inf_{y \in \text{supp}^M(X)} \|x - y\| \qquad \text{(B.5.9)} \\ &= 2\delta K.\end{aligned}$$

For any $i \in \mathcal{I}$ we have that

$$\begin{aligned}\mathbb{E}_{\tilde{M}(i)}\big[(Y - f^*(X))^2\big] &= \mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) + \xi_Y - f^*(X))^2\big] \qquad \text{(B.5.10)} \\ &= \mathbb{E}_{\tilde{M}}\big[\xi_Y^2\big] + \mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f^*(X))^2\big] \\ &\qquad + 2\mathbb{E}_{\tilde{M}(i)}\big[\xi_Y(\tilde{f}(X) - f^*(X))\big].\end{aligned}$$

Next, we can use Cauchy-Schwarz, (B.5.8) and (B.5.9) in (B.5.10) to get that

$$\begin{aligned}
&\left|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f^*(X))^2\right] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right]\right| \\
&\quad = \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f^*(X))^2\right] - \mathbb{E}_{\tilde{M}}[\xi_Y^2] \\
&\quad = \sup_{i\in\mathcal{I}} \left(\mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) - f^*(X))^2\right] + 2\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y(\tilde{f}(X) - f^*(X))\right]\right) \\
&\quad \leq 4\delta^2 K^2 + 4\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}, \qquad \text{(B.5.11)}
\end{aligned}$$

proving the first statement. Finally, if $\mathcal{I}$ consists only of confounding-removing interventions, then the bound in (B.5.11) can be improved by using that $\mathbb{E}[\xi_Y] = 0$ together with $H \perp\!\!\!\perp X$. In that case, we get that $\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y(\tilde{f}(X) - f(X))\right] = 0$ and hence the bound becomes $4\delta^2 K^2$. This completes the proof of Proposition 3.9. □

B.5.10. Proof of Proposition 3.10

Proof. By Assumption 3.1, f is identified on $\mathrm{supp}^M(X)$ by the observational distribution $\mathbb{P}_M$. Let $\mathcal{I}$ be a set of confounding-preserving interventions. For a fixed $\varepsilon > 0$, let $f^* \in \mathcal{F}$ be a function satisfying

$$|\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f^*(X))^2)\right] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2)\right]| \leq \varepsilon. \qquad \text{(B.5.12)}$$

Fix any secondary model $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. The general idea is to derive an upper bound for $\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f^*(X))^2]$ and a lower bound for $\inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f_\diamond(X))^2]$ which will allow us to bound the absolute difference of interest.

Since $(\mathbb{P}_M, \mathcal{M})$ satisfies Assumption 3.1, we have that $f \equiv \tilde{f}$ on $\mathrm{supp}^M(X) = \mathrm{supp}^{\tilde{M}}(X)$. We first show that $\|\tilde{f} - f\|_{\mathcal{I},\infty} \leq 2\delta K$, where $\|f\|_{\mathcal{I},\infty} := \sup_{x\in\mathrm{supp}^M_{\mathcal{I}}(X)} \|f(x)\|$. By the mean value theorem, for all $f_\diamond \in \mathcal{F}$ it holds that $|f_\diamond(x) - f_\diamond(y)| \leq K\|x - y\|$, for all

$x, y \in \mathcal{D}$. For any $x \in \operatorname{supp}^M_{\mathcal{I}}(X)$ and $y \in \operatorname{supp}^M(X)$ we have

$$\begin{aligned}\big|\tilde{f}(x) - f(x)\big| &= \big|\tilde{f}(x) - \tilde{f}(y) + f(y) - f(x)\big| \\ &\leq \big|\tilde{f}(x) - \tilde{f}(y)\big| + \big|f(y) - f(x)\big| \\ &\leq 2K\|x - y\|,\end{aligned}$$

where we used the fact that $\tilde{f}(y) = f(y)$, for all $y \in \operatorname{supp}_M(X)$. In particular, it holds that

$$\begin{aligned}\|\tilde{f} - f\|_{\mathcal{I},\infty} &= \sup_{x \in \operatorname{supp}^M_{\mathcal{I}}(X)} \big|\tilde{f}(x) - f(x)\big| \\ &\leq 2K \sup_{x \in \operatorname{supp}^M_{\mathcal{I}}(X)} \inf_{y \in \operatorname{supp}^M(X)} \|x - y\| \\ &= 2\delta K.\end{aligned} \tag{B.5.13}$$

Let now $i \in \mathcal{I}$ be fixed. The term $\xi_Y = h_1(H, \varepsilon_Y)$ is not affected by the intervention i. Furthermore, $\mathbb{P}^{(X,\xi_Y)}_{M(i)} = \mathbb{P}^{(X,\xi_Y)}_{\tilde{M}(i)}$ since i is confounding-preserving (this can be seen by a slight modification to the arguments from case (A) in the proof of Proposition 3.6). Thus, for any $f_\diamond \in \mathcal{F}$ we have that

$$\begin{aligned}&\mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big] \\ &\quad= \mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) + \xi_Y - f_\diamond(X) + f(X) - f(X))^2\big] \\ &\quad= \mathbb{E}_{\tilde{M}(i)}\big[\xi_Y^2\big] + \mathbb{E}_{\tilde{M}(i)}\big[(f(X) - f_\diamond(X))^2\big] + \mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f(X))^2\big] \\ &\qquad\quad + 2\mathbb{E}_{\tilde{M}(i)}\big[\xi_Y(f(X) - f_\diamond(X))\big] \\ &\qquad\quad + 2\mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\big] \\ &\qquad\quad + 2\mathbb{E}_{\tilde{M}(i)}\big[\xi_Y(\tilde{f}(X) - f(X))\big] \\ &\quad= \mathbb{E}_{M(i)}\big[\xi_Y^2\big] + \mathbb{E}_{M(i)}\big[(f(X) - f_\diamond(X))^2\big] + \mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))^2\big] \\ &\qquad\quad + 2\mathbb{E}_{M(i)}\big[\xi_Y(f(X) - f_\diamond(X))\big] \\ &\qquad\quad + 2\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\big] \\ &\qquad\quad + 2\mathbb{E}_{M(i)}\big[\xi_Y(\tilde{f}(X) - f(X))\big] \\ &\quad= \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] + L^i_1(\tilde{f}) + L^i_2(\tilde{f}, f_\diamond) + L^i_3(\tilde{f}),\end{aligned} \tag{B.5.14}$$

where, we have made the following definitions

$$\begin{aligned} L_1^i(\tilde{f}) &:= \mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))^2\big], \\ L_2^i(\tilde{f}, f_\diamond) &:= 2\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\big], \\ L_3^i(\tilde{f}) &:= 2\mathbb{E}_{M(i)}\big[\xi_Y(\tilde{f}(X) - f(X))\big]. \end{aligned}$$

Using (B.5.13) it follows that

$$0 \leq L_1^i(\tilde{f}) \leq 4\delta^2 K^2, \tag{B.5.15}$$

and by the Cauchy-Schwarz inequality it follows that

$$\big|L_3^i(\tilde{f})\big| \leq 2\sqrt{\mathrm{Var}_M(\xi_Y) 4\delta^2 K^2} = 4\delta K \sqrt{\mathrm{Var}_M(\xi_Y)}. \tag{B.5.16}$$

Let now $f_\diamond \in \mathcal{F}$ be any function such that

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2)\big] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - \tilde{f}(X))^2)\big], \tag{B.5.17}$$

then by (B.5.13), the Cauchy-Schwarz inequality and Proposition 3.3, it holds for all $i \in \mathcal{I}$ that

$$\begin{aligned} L_2^i(\tilde{f}, f_\diamond) &= 2\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\big] \\ &= 2\mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\big] \\ &= -2\mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f(X))^2\big] \\ &\quad + 2\mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f(X))(\tilde{f}(X) - f_\diamond(X))\big] \\ &\geq -8\delta^2 K^2 - 2\sqrt{4\delta^2 K^2}\sqrt{4\,\mathrm{Var}_M(\xi_Y)} \\ &= -8\delta^2 K^2 - 8\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}, \end{aligned} \tag{B.5.18}$$

where, in the third equality, we have added and subtracted the term $2\mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f(X))\tilde{f}(X)\big]$. Now let $\mathcal{S} := \{f_\diamond \in \mathcal{F} : \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - \tilde{f}(X))^2\big]\}$ be the set of all functions satisfying (B.5.17). Due to (B.5.14), (B.5.15), (B.5.16) and (B.5.18)

we have the following lower bound of interest

$$\begin{aligned}
&\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big] \\
&= \inf_{f_\diamond \in \mathcal{S}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big] \qquad \text{(B.5.19)} \\
&= \inf_{f_\diamond \in \mathcal{S}} \sup_{i \in \mathcal{I}} \big\{\mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] + L_1^i(\tilde{f}) + L_2^i(\tilde{f}, f_\diamond) + L_3^i(\tilde{f})\big\} \\
&\geq \inf_{f_\diamond \in \mathcal{S}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] - 8\delta^2 K^2 - 8\delta K \sqrt{\mathrm{Var}_M(\xi_Y)} \\
&\qquad - 4\delta K \sqrt{\mathrm{Var}_M(\xi_Y)} \\
&\geq \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] - 8\delta^2 K^2 - 12\delta K \sqrt{\mathrm{Var}_M(\xi_Y)}.
\end{aligned}$$

Next, we construct the aforementioned upper bound of interest. To that end, note that

$$\begin{aligned}
&\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f^*(X))^2\big] \qquad \text{(B.5.20)} \\
&\quad = \sup_{i \in \mathcal{I}} \Big\{\mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] + L_1^i(\tilde{f}) + L_2^i(\tilde{f}, f^*) + L_3^i(\tilde{f})\Big\},
\end{aligned}$$

by (B.5.14). We have already established upper bounds for $L_1^i(\tilde{f})$ and $L_3^i(\tilde{f})$ in (B.5.15) and (B.5.16), respectively. In order to control $L_2^i(\tilde{f}, f^*)$ we introduce an auxiliary function. Let $\bar{f}^* \in \mathcal{F}$ satisfy

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2)\big] \leq \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f(X))^2)\big], \quad \text{(B.5.21)}$$

and

$$\Big|\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2\big] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big]\Big| \leq \varepsilon. \qquad \text{(B.5.22)}$$

Choosing such a $\bar{f}^* \in \mathcal{F}$ is always possible. If f is an ε-minimax solution, i.e., it satisfies (B.5.22), then choose $\bar{f}^* = f$. Otherwise, if f is not a ε-minimax solution, then choose any $\bar{f}^* \in \mathcal{F}$ that is an ε-minimax solution (which is always possible). In this case we have that

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2\big] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] \leq \varepsilon,$$

and

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f(X))^2\big] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] \geq \varepsilon,$$

which implies that (B.5.21) is satisfied. We can now construct an upper bound on $L_2^i(\tilde{f}, f^*)$ in terms of $L_2^i(\tilde{f}, \bar{f}^*)$ by noting that for all $i \in \mathcal{I}$

$$\begin{aligned}
&\big|L_2^i(\tilde{f}, f^*)\big| \\
&= 2\big|\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - f^*(X))\big]\big| \\
&\leq 2\big|\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - \bar{f}^*(X))\big]\big| \qquad \text{(B.5.23)} \\
&\qquad + 2\mathbb{E}_{M(i)}\big|(\tilde{f}(X) - f(X))(\bar{f}^*(X) - f^*(X))\big| \\
&= \big|L_2^i(\tilde{f}, \bar{f}^*)\big| + 2\mathbb{E}_{M(i)}\big|(\tilde{f}(X) - f(X))(\bar{f}^*(X) - f^*(X))\big| \\
&\leq \big|L_2^i(\tilde{f}, \bar{f}^*)\big| + 2\sqrt{\mathbb{E}_{M(i)}\left[(\tilde{f}(X) - f(X))^2\right]\mathbb{E}_{M(i)}\left[(\bar{f}^*(X) - f^*(X))^2\right]} \\
&\leq \big|L_2^i(\tilde{f}, \bar{f}^*)\big| + 4\delta K\sqrt{\mathbb{E}_{M(i)}\left[(\bar{f}^*(X) - f^*(X))^2\right]},
\end{aligned}$$

where we used the triangle inequality, Cauchy-Schwarz inequality and (B.5.13) . Furthermore, (B.5.13) and (B.5.21) together with Proposition 3.3 yield the following bound

$$\begin{aligned}
|L_2^i(\tilde{f}, \bar{f}^*)| &= 2\big|\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - \bar{f}^*(X))\big]\big| \\
&= 2\sqrt{\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))^2\big]\mathbb{E}_{M(i)}\big[(f(X) - \bar{f}^*(X))^2\big]} \\
&\leq 2\sqrt{4\delta^2K^2}\sqrt{4\operatorname{Var}_M(\xi_Y)} \\
&= 8\delta K\sqrt{\operatorname{Var}_M(\xi_Y)}, \qquad \text{(B.5.24)}
\end{aligned}$$

for any $i \in \mathcal{I}$. Thus, it suffices to construct an upper bound on the second term in the final expression in (B.5.23). Direct computation leads to

$$\begin{aligned}
\mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] = \mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2\big] \\
+ \mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big] \\
+ 2\mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))(\bar{f}^*(X) - f^*(X))\big].
\end{aligned}$$

Rearranging the terms and applying the triangle inequality and Cauchy-Schwarz results in

$$\begin{aligned}
&\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big] \\
&\quad = \mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] - \mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2\big] \\
&\qquad - 2\mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))(\bar{f}^*(X) - f^*(X))\big] \\
&\quad \leq \big|\mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big]\big| \\
&\qquad + \big|\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] - \mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2\big]\big| \\
&\qquad + 2\mathbb{E}_{M(i)}\big|(Y - \bar{f}^*(X))(\bar{f}^*(X) - f^*(X))\big| \\
&\quad \leq 2\varepsilon + 2\sqrt{\mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2\big]}\sqrt{\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]} \\
&\quad \leq 2\varepsilon + 2\sqrt{\mathrm{Var}_M(\xi_Y)}\sqrt{\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]},
\end{aligned}$$

for any $i \in \mathcal{I}$. Here, we used that both f^* and $\bar{f}^*$ are ε-minimax solutions with respect to M and that $\bar{f}^*$ satisfies (B.5.21) which implies that

$$\begin{aligned}
\mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2)\big] &\leq \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f(X))^2)\big] \\
&= \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[\xi_Y^2\big] = \mathrm{Var}_M[\xi_Y],
\end{aligned}$$

for any $i \in \mathcal{I}$, as ξ_Y is unaffected by an intervention on X. Thus, $\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]$ must satisfy $\ell(\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]) \leq 0$, where $\ell : [0, \infty) \to \mathbb{R}$ is given by $\ell(z) = z - 2\varepsilon - 2\sqrt{\mathrm{Var}_M(\xi_Y)}\sqrt{z}$. The linear term of ℓ grows faster than the square root term, so the largest allowed value of $\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]$ coincides with the largest root of $\ell(z)$. The largest root is given by

$$C^2 := 2\varepsilon + 2\,\mathrm{Var}_M(\xi_Y) + 2\sqrt{\mathrm{Var}_M(\xi_Y)^2 + 2\varepsilon\,\mathrm{Var}_M(\xi_Y)},$$

where $(\cdot)^2$ refers to the square of C. Hence, for any $i \in \mathcal{I}$ it holds that

$$\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big] \leq C^2. \tag{B.5.25}$$

Hence by (B.5.23), (B.5.24) and (B.5.25) we have that the following upper bound is valid for any $i \in \mathcal{I}$.

$$\big|L_2^i(\tilde{f}, f^*)\big| \leq 8\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} + 4\delta KC. \tag{B.5.26}$$

Thus, using (B.5.20) with (B.5.15), (B.5.16) and (B.5.26), we get the following upper bound

$$\begin{aligned}&\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y-f^*(X))^2\big] \qquad\qquad (B.5.27)\\ &\quad\leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y-f^*(X))^2\big] + 4\delta^2K^2 + 4\delta KC + 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}.\end{aligned}$$

Finally, by combining the bounds (B.5.19) and (B.5.27) together with (B.5.12) we get that

$$\begin{aligned}&\Big|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y-f^*(X))^2\big] - \inf_{f_\diamond\in\mathcal{F}}\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y-f_\diamond(X))^2\big]\Big|\\ &\quad\leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y-f^*(X))^2\big] - \inf_{f_\diamond\in\mathcal{F}}\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y-f_\diamond(X))^2\big]\\ &\qquad\quad + 4\delta^2K^2 + 4\delta KC + 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}\\ &\qquad\quad + 8\delta^2K^2 + 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}\\ &\quad\leq \varepsilon + 12\delta^2K^2 + 24\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} + 4\delta KC. \qquad\qquad (B.5.28)\end{aligned}$$

Using that all terms are positive, we get that

$$C = \sqrt{\mathrm{Var}_M(\xi_Y)} + \sqrt{\mathrm{Var}_M(\xi_Y)+2\varepsilon} \leq 2\sqrt{\mathrm{Var}_M(\xi_Y)} + \sqrt{2\varepsilon}$$

Hence, (B.5.28) is bounded above by

$$\varepsilon + 12\delta^2K^2 + 32\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} + 4\sqrt{2}\delta K\sqrt{\varepsilon}.$$

This completes the proof of Proposition 3.10. □

B.5.11. Proof of Proposition 3.11

Proof. Let $\bar{f} \in \mathcal{F}$ and $c > 0$. By assumption, $\mathcal{I}$ is a well-behaved set of support-extending interventions on X. Since $\mathrm{supp}^M_{\mathcal{I}}(X) \setminus \mathrm{supp}^M(X)$ has non-empty interior, there exists an intervention $i_0 \in \mathcal{I}$ and $\varepsilon > 0$ such that $\mathbb{P}_{M(i_0)}(X \in B) \geq \varepsilon$, for some open subset $B \subsetneq \bar{B}$, such that $\mathrm{dist}(B, \mathbb{R}^d \setminus \bar{B}) > 0$, where $\bar{B} := \mathrm{supp}^M_{\mathcal{I}}(X) \setminus \mathrm{supp}^M(X)$. Let $\tilde{f}$ be any continuous function satisfying that, for all $x \in B \cup (\mathbb{R}^d \setminus \bar{B})$,

$$\tilde{f}(x) = \begin{cases} \bar{f}(x) + \gamma, & x \in B \\ f(x), & x \in \mathbb{R}^d \setminus \bar{B}, \end{cases}$$

where $\gamma := \varepsilon^{-1/2}\left\{(2\mathbb{E}_{\tilde{M}}[\xi_Y^2]+c)^{1/2}+(\mathbb{E}_{\tilde{M}}[\xi_Y^2])^{1/2}\right\}$.

Consider a secondary model $\tilde{M}=(\tilde{f},g,h_1,h_2,Q)\in\mathcal{M}$. Then, by Assumption 3.1, it holds that $\mathbb{P}_M=\mathbb{P}_{\tilde{M}}$. Since $\mathcal{I}$ only consists of interventions on X, it holds that $\mathbb{P}_{M(i_0)}(X\in B)=\mathbb{P}_{\tilde{M}(i_0)}(X\in B)$ (this holds since all components of $\tilde{M}$ and M are equal, except for the function f, which is not allowed to enter in the intervention on X). Therefore,

$$\begin{aligned}\mathbb{E}_{\tilde{M}(i_0)}\big[(Y-\bar{f}(X))^2\big] &\geq \mathbb{E}_{\tilde{M}(i_0)}\big[(Y-\bar{f}(X))^2\mathbb{1}_B(X)\big]\\ &= \mathbb{E}_{\tilde{M}(i_0)}\big[(\gamma+\xi_Y)^2\mathbb{1}_B(X)\big]\\ &\geq \gamma^2\varepsilon+2\gamma\mathbb{E}_{\tilde{M}(i_0)}\big[\xi_Y\mathbb{1}_B(X)\big]\\ &\geq \gamma^2\varepsilon-2\gamma\left(\mathbb{E}_{\tilde{M}}\big[\xi_Y^2\big]\varepsilon\right)^{1/2}\\ &= c+\mathbb{E}_{\tilde{M}}[\xi_Y^2], \end{aligned} \tag{B.5.29}$$

where the third inequality follows from Cauchy–Schwarz. Further, by the definition of the infimum it holds that

$$\inf_{f_\diamond\in\mathcal{F}}\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\big[(Y-f_\diamond(X))^2\big]\leq\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\big[(Y-\tilde{f}(X))^2\big]=\mathbb{E}_{\tilde{M}}[\xi_Y^2]. \tag{B.5.30}$$

Therefore, combining (B.5.29) and (B.5.30), the claim follows. □

B.5.12. Proof of Proposition 3.12

Proof. We prove the result by showing that under Assumption 3.3 it is possible to express interventions on A as confounding-preserving interventions on X and applying Propositions 3.6 and 3.8. To avoid confusion, we will throughout this proof denote the true model by $M^0=(f^0,g^0,h_1^0,h_2^0,Q^0)$. Fix an intervention $i\in\mathcal{I}$. Since it is an intervention on A, there exist ψ^i and I^i such that for any $M=(f,g,h_1,h_2,Q)\in\mathcal{M}$, the intervened SCM $M(i)$ is of the form

$$\begin{aligned}A^i &:= \psi^i(I^i,\varepsilon_A^i), \quad H^i := \varepsilon_H^i,\\ X^i &:= g(A^i)+h_2(H^i,\varepsilon_X^i),\\ Y^i &:= f(X^i)+h_1(H^i,\varepsilon_Y^i),\end{aligned}$$

where $(\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i) \sim Q$. We now define a confounding-preserving intervention j on X, such that, for all models $\tilde{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, the distribution of (X, Y) under $\tilde{M}(j)$ coincides with that under $\tilde{M}(i)$. To that end, define the intervention function

$$\bar{\psi}^j(h_2, A^j, H^j, \varepsilon_X^j, I^j) := g^0(\psi^i(I^j, A^j)) + h_2(H^j, \varepsilon_X^j),$$

where g^0 is the fixed function corresponding to model M, and therefore not an argument of $\bar{\psi}^j$. Let now j be the intervention on X satisfying that, for all $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$, the intervened model $M(j)$ is given as

$$\begin{aligned} A^j &:= \varepsilon_A^j, \quad H^j := \varepsilon_H^j, \\ X^j &:= \bar{\psi}^j(h_2, A^j, H^j, \varepsilon_X^j, I^j), \\ Y^j &:= f(X^j) + h_1(H^j, \varepsilon_Y^j), \end{aligned}$$

where $(\varepsilon_X^j, \varepsilon_Y^j, \varepsilon_A^j, \varepsilon_H^j) \sim Q$ and where I^j is chosen such that $I^j \overset{\mathrm{d}}{=} I^i$. By definition, j is a confounding-preserving intervention. Let now $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q})$ be such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, and let $(\tilde{X}^i, \tilde{Y}^i)$ and $(\tilde{X}^j, \tilde{Y}^j)$ be generated under $\tilde{M}(i)$ and $\tilde{M}(j)$, respectively. By Assumption 3.3, it holds for all $a \in \mathrm{supp}(A) \cup \mathrm{supp}_{\mathcal{I}}(A)$ that $\tilde{g}(a) = g^0(a)$. Hence, we get that

$$\begin{aligned} (\tilde{X}^i, \tilde{Y}^i) &\overset{\mathrm{d}}{=} (\tilde{g}(\psi^i(I^i, \tilde{\varepsilon}_A^i)) + \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i), \\ &\qquad \tilde{f}(\tilde{g}(\psi^i(I^i, \tilde{\varepsilon}_A^i)) + \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i)) + \tilde{h}_1(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_Y^i)) \\ &= (g^0(\psi^i(I^i, \tilde{\varepsilon}_A^i)) + \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i), \\ &\qquad \tilde{f}(g^0(\psi^i(I^i, \tilde{\varepsilon}_A^i)) + \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i)) + \tilde{h}_1(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_Y^i)) \\ &\overset{\mathrm{d}}{=} (g^0(\psi^i(I^j, \tilde{\varepsilon}_A^j)) + \tilde{h}_2(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j), \\ &\qquad \tilde{f}(g^0(\psi^i(I^j, \tilde{\varepsilon}_A^j)) + \tilde{h}_2(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j)) + \tilde{h}_1(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_Y^j)) \\ &\overset{\mathrm{d}}{=} (\bar{\psi}^j(\tilde{h}_2, \tilde{\varepsilon}_A^j, \tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j, I^j), \\ &\qquad \tilde{f}(\bar{\psi}^j(\tilde{h}_2, \tilde{\varepsilon}_A^j, \tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j, I^j)) + \tilde{h}_1(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_Y^j)) \\ &\overset{\mathrm{d}}{=} (\tilde{X}^j, \tilde{Y}^j), \end{aligned}$$

as desired. Since $i \in \mathcal{I}$ was arbitrary, we have now shown that there exists a mapping π from $\mathcal{I}$ into a set $\mathcal{J}$ of confounding-preserving

(and hence a well-behaved set) of interventions on X, such that for all $\tilde{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, $\mathbb{P}^{(X,Y)}_{\tilde{M}(i)} = \mathbb{P}^{(X,Y)}_{\tilde{M}(\pi(i))}$. Hence, we can rewrite Equation (3.4.1) in Definition 3.1 in terms of the set $\mathcal{J}$. The result now follows from Propositions 3.6 and 3.8. □

B.5.13. Proof of Proposition 3.13

Proof. Let $b \in \mathbb{R}^d$ be such that $f(x) = b^\top x$ for all $x \in \mathbb{R}^d$. We start by characterizing the error $\mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big]$. Let us consider models of the form $\tilde{M} = (f, \tilde{g}, h_1, h_2, Q) \in \mathcal{M}$ for some function $\tilde{g} \in \mathcal{G}$ with $\tilde{g}(a) = g(a)$ for all $a \in \mathrm{supp}_M(A)$. Clearly, any such model satisfies that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. For every $a \in \mathcal{A}$, let $i_a \in \mathcal{I}$ denote the corresponding hard intervention on A. For every $a \in \mathcal{A}$ and $b_\diamond \in \mathbb{R}^d$, we then have

$$
\begin{aligned}
&\mathbb{E}_{\tilde{M}(i_a)}\big[(Y - b_\diamond^\top X)^2\big] \\
&= \mathbb{E}_{\tilde{M}(i_a)}\big[(b^\top X + \xi_Y - b_\diamond^\top X)^2\big] \\
&= (b - b_\diamond)^\top \mathbb{E}_{\tilde{M}(i_a)}[XX^\top](b - b_\diamond) \\
&\qquad + 2(b - b_\diamond)^\top \mathbb{E}_{\tilde{M}(i_a)}[X\xi_Y] + \mathbb{E}_{\tilde{M}(i_a)}\big[\xi_Y^2\big] \\
&= (b - b_\diamond)^\top \underbrace{\big(\tilde{g}(a)\tilde{g}(a)^\top + \mathbb{E}_M[\xi_X \xi_X^\top]\big)}_{=: K_{\tilde{M}}(a)} (b - b_\diamond) \\
&\qquad + 2(b - b_\diamond)^\top \mathbb{E}_M[\xi_X \xi_Y] + \mathbb{E}_M\big[\xi_Y^2\big],
\end{aligned}
\tag{B.5.31}
$$

where we have used that, under i_a, the distribution of (ξ_X, ξ_Y) is unaffected. We now show that, for any $\tilde{M}$ with the above form, the causal function f does not minimize the worst-case mean squared error across interventions in $\mathcal{I}$. The idea is to show that the worst-case mean squared error (B.5.31) strictly decreases at $b_\diamond = b$ in the direction $u := \mathbb{E}_M[\xi_X \xi_Y] / \|\mathbb{E}_M[\xi_X \xi_Y]\|_2$. For every $a \in \mathcal{A}$ and $s \in \mathbb{R}$, define

$$
\begin{aligned}
\ell_{\tilde{M},a}(s) &:= \mathbb{E}_{\tilde{M}(i_a)}\big[(Y - (b + su)^\top X)^2\big] \\
&= u^\top K_{\tilde{M}}(a) u \cdot s^2 - 2u^\top \mathbb{E}_M[\xi_X \xi_Y] \cdot s + \mathbb{E}_M\big[\xi_Y^2\big].
\end{aligned}
$$

For every a, $\ell'_{\tilde{M},a}(0) = -2\|\mathbb{E}_M[\xi_X \xi_Y]\|_2 < 0$, showing that $\ell_{\tilde{M},a}$ is strictly decreasing at $s = 0$ (with a derivative that is bounded

away from 0 across all $a \in \mathcal{A}$). By boundedness of $\mathcal{A}$ and by the continuity of $a \mapsto \ell''_{\tilde{M},a}(0) = 2u^\top K_{\tilde{M}}(a)u$, it further follows that $\sup_{a\in\mathcal{A}}|\ell''_{\tilde{M},a}(0)| < \infty$. Hence, we can find $s_0 > 0$ such that for all $a \in \mathcal{A}$, $\ell_{\tilde{M},a}(0) > \ell_{\tilde{M},a}(s_0)$. It now follows by continuity of $(a, s) \mapsto \ell_{\tilde{M},a}(s)$ that

$$\begin{aligned}\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b^\top X)^2\big] &= \sup_{a\in\mathcal{A}} \ell_{\tilde{M},a}(0) \\ &> \sup_{a\in\mathcal{A}} \ell_{\tilde{M},a}(s_0) \\ &= \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - (b + s_0u)^\top X)^2\big],\end{aligned}$$

showing that $b + s_0u$ attains a lower worst-case mean squared error than b.

We now show that all functions other than f may result in an arbitrarily large error. Let $\bar{b} \in \mathbb{R}^d \setminus \{b\}$ be given, and let $j \in \{1, \dots, d\}$ be such that $b_j \neq \bar{b}_j$. The idea is to construct a function $\tilde{g} \in \mathcal{G}$ such that, under the corresponding model $\tilde{M} = (f, \tilde{g}, h_1, h_2, Q) \in \mathcal{M}$, some hard interventions on A result in strong shifts of the jth coordinate of X. Let $a \in \mathcal{A}$. Let $e_j \in \mathbb{R}^d$ denote the jth unit vector, and assume that $\tilde{g}(a) = ne_j$ for some $n \in \mathbb{N}$. Using (B.5.31), it follows that

$$\begin{aligned}\mathbb{E}_{\tilde{M}(i_a)}\big[(Y - \bar{b}^\top X)^2\big] = n^2(\bar{b}_j - b_j)^2 &+ (\bar{b} - b)^\top \mathbb{E}_M[\xi_X\xi_X^\top](\bar{b} - b) \\ &+ 2(\bar{b} - b)^\top \mathbb{E}_M[\xi_X\xi_Y] + \mathbb{E}_M\big[\xi_Y^2\big].\end{aligned}$$

By letting $n \to \infty$, we see that the above error may become arbitrarily large. Given any $c > 0$, we can therefore construct $\tilde{g}$ such that $\mathbb{E}_{\tilde{M}(i_a)}\big[(Y - \bar{b}^\top X)^2\big] \geq c + \mathbb{E}_M\big[\xi_Y^2\big]$. By carefully choosing $a \in \operatorname{int}(\mathcal{A} \setminus \operatorname{supp}_M(A))$, this can be done such that $\tilde{g}$ is continuous and $\tilde{g}(a) = g(a)$ for all $a \in \operatorname{supp}_M(A)$, ensuring that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. It

follows that

$$\begin{aligned} c &\leq \mathbb{E}_{\tilde{M}(i_a)}\left[(Y - \bar{b}^\top X)^2\right] - \mathbb{E}_M\left[\xi_Y^2\right] \\ &= \mathbb{E}_{\tilde{M}(i_a)}\left[(Y - \bar{b}^\top X)^2\right] - \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - b^\top X)^2\right] \\ &\leq \mathbb{E}_{\tilde{M}(i_a)}\left[(Y - \bar{b}^\top X)^2\right] - \inf_{b_\diamond \in \mathbb{R}^d} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - b_\diamond^\top X)^2\right] \\ &\leq \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - \bar{b}^\top X)^2\right] - \inf_{b_\diamond \in \mathbb{R}^d} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - b_\diamond^\top X)^2\right], \end{aligned}$$

which completes the proof of Proposition 3.13. □

B.5.14. Proof of Proposition 3.14

Proof. By assumption, $\mathcal{I}$ is a set of interventions on X or A of which at least one is confounding-removing. Now fix any

$$\tilde{M} = (f_{\eta_0}(x; \tilde{\theta}), \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M},$$

with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$. By Proposition 3.1, we have that a minimax solution is given by the causal function. That is,

$$\begin{aligned} \inf_{f_\diamond \in \mathcal{F}_{\eta_0}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_{\eta_0}(X; \tilde{\theta}))^2\right] \\ &= \mathbb{E}_M[\xi_Y^2], \end{aligned}$$

where we used that ξ_Y is unaffected by an intervention on X. By the support restriction $\operatorname{supp}^M(X) \subseteq (a, b)$ we know that

$$\begin{aligned} f_{\eta_0}(x; \theta^0) &= B(x)^\top \theta^0, \\ f_{\eta_0}(x; \tilde{\theta}) &= B(x)^\top \tilde{\theta}, \\ f_{\eta_0}(x; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}) &= B(x)^\top \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}, \end{aligned}$$

for all $x \in \operatorname{supp}^M(X)$. Furthermore, as $Y = B(X)^\top \theta^0 + \xi_Y$ $\mathbb{P}_M$-almost surely, we have that

$$\begin{aligned} \mathbb{E}_M\left[C(A)Y\right] &= \mathbb{E}_M\left[C(A)B(X)^\top \theta^0\right] + \mathbb{E}_M\left[C(A)\xi_Y\right] \\ &= \mathbb{E}_M\left[C(A)B(X)^\top\right]\theta^0, \end{aligned} \tag{B.5.32}$$

where we used the assumptions that $\mathbb{E}[\xi_Y] = 0$ and $A \perp\!\!\!\perp \xi_Y$ by the exogeneity of A. Similarly,

$$\mathbb{E}_{\tilde{M}}[C(A)Y] = \mathbb{E}_{\tilde{M}}\left[C(A)B(X)^\top\right]\tilde{\theta}.$$

As $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$, it holds $\mathbb{E}_M[C(A)Y] = \mathbb{E}_{\tilde{M}}[C(A)Y]$ and $\mathbb{E}_M[C(A)B(X)^\top] = \mathbb{E}_{\tilde{M}}[C(A)B(X)^\top]$, hence

$$\mathbb{E}_M\left[C(A)B(X)^\top\right]\tilde{\theta} = \mathbb{E}_M\left[C(A)B(X)^\top\right]\theta^0 \iff \tilde{\theta} = \theta^0,$$

by assumption (B2), which states that $\mathbb{E}[C(A)B(X)^\top]$ is of full rank (bijective). In other words, the causal function parameterized by θ^0 is identified from the observational distribution. Assumptions 3.1 and 3.2 are therefore satisfied. Furthermore, we also have that

$$\begin{aligned}
&\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\right] \\
&\quad= \sup_{i\in\mathcal{I}} \Big\{\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\right] + \mathbb{E}_{\tilde{M}(i)}\left[\xi_Y^2\right] \\
&\qquad\quad + 2\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))\right]\Big\} \\
&\quad\le \sup_{i\in\mathcal{I}} \Big\{\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\right] + \mathbb{E}_{\tilde{M}(i)}\left[\xi_Y^2\right] \\
&\qquad\quad + 2\sqrt{\mathbb{E}_{\tilde{M}(i)}\left[\xi_Y^2\right]\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\right]}\Big\} \\
&\quad\le \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\right] + \mathbb{E}_M\left[\xi_Y^2\right] \\
&\qquad\quad + 2\sqrt{\mathbb{E}_M\left[\xi_Y^2\right]\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\right]},
\end{aligned}$$

by Cauchy-Schwarz inequality, where we additionally used that $\mathbb{E}_{\tilde{M}(i)}[\xi_Y^2] = \mathbb{E}_M[\xi_Y^2]$ as ξ_Y is unaffected by interventions on X. Thus,

$$\begin{aligned}
&\Big|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\right] - \inf_{f_\diamond\in\mathcal{F}_{\eta_0}}\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right]\Big| \\
&\quad\le \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\right] \\
&\qquad + 2\sqrt{\mathbb{E}_M\left[\xi_Y^2\right]\sup_{i\in\mathcal{I}}\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\right]}.
\end{aligned}$$

For the next few derivations let $\hat{\theta} = \hat{\theta}^n_{\lambda^\star_n, \eta_0, \mu}$ for notational simplicity. Note that, for all $x \in \mathbb{R}$,

$$\begin{aligned}(f_{\eta_0}(x;\theta^0) - f_{\eta_0}(x;\hat{\theta}))^2 \leq\ & (\theta^0 - \hat{\theta})^\top B(x)B(x)^\top(\theta^0 - \hat{\theta}) \\ & + (B(a)^\top(\theta^0 - \hat{\theta}) + B'(a)^\top(\theta^0 - \hat{\theta})(x-a))^2 \\ & + (B(b)^\top(\theta^0 - \hat{\theta}) + B'(b)^\top(\theta^0 - \hat{\theta})(x-b))^2.\end{aligned}$$

The second term has the following upper bound

$$\begin{aligned}&(B(a)^\top(\theta^0 - \hat{\theta}) + B'(a)^\top(\theta^0 - \hat{\theta})(x-a))^2 \\ &\quad = (\theta^0 - \hat{\theta})^\top B(a)B(a)^\top(\theta^0 - \hat{\theta}) \\ &\qquad + (x-a)^2(\theta^0 - \hat{\theta})^\top B'(a)B'(a)^\top(\theta^0 - \hat{\theta}) \\ &\qquad + 2(x-a)(\theta^0 - \hat{\theta})^\top B'(a)B(a)^\top(\theta^0 - \hat{\theta}) \\ &\quad \leq \lambda_{\max}(B(a)B(a)^\top)\|\theta^0 - \hat{\theta}\|_2^2 \\ &\qquad + (x-a)^2\lambda_{\max}(B'(a)B'(a)^\top)\|\theta^0 - \hat{\theta}\|_2^2 \\ &\qquad + 2(x-a)\lambda_{\max}((B'(a)B(a)^\top + B(a)B'(a)^\top)/2)\|\theta^0 - \hat{\theta}\|_2^2,\end{aligned}$$

where $\lambda_{\max}$ denotes the maximum eigenvalue. An analogous upper bound can be constructed for the third term. Thus, by combining these two upper bounds with a similar upper bound for the first term, we arrive at

$$\begin{aligned}&\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}))^2\right] \\ &\quad \leq \lambda_{\max}(\mathbb{E}_{\tilde{M}(i)}[B(X)B(X)^\top])\|\theta^0 - \hat{\theta}\|_2^2 \\ &\qquad + \lambda_{\max}(B(a)B(a)^\top)\|\theta^0 - \hat{\theta}\|_2^2 \\ &\qquad + \mathbb{E}_{\tilde{M}(i)}[(X-a)^2]\lambda_{\max}(B'(a)B'(a)^\top)\|\theta^0 - \hat{\theta}\|_2^2 \\ &\qquad + 2\mathbb{E}_{\tilde{M}(i)}[X-a]\lambda_{\max}((B'(a)B(a)^\top + B(a)B'(a)^\top)/2)\|\theta^0 - \hat{\theta}\|_2^2 \\ &\qquad + \lambda_{\max}(B(b)B(b)^\top)\|\theta^0 - \hat{\theta}\|_2^2 \\ &\qquad + \mathbb{E}_{\tilde{M}(i)}[(X-b)^2]\lambda_{\max}(B'(b)B'(b)^\top)\|\theta^0 - \hat{\theta}\|_2^2 \\ &\qquad + 2\mathbb{E}_{\tilde{M}(i)}[X-b]\lambda_{\max}((B'(b)B(b)^\top + B(b)B'(b)^\top)/2)\|\theta^0 - \hat{\theta}\|_2^2.\end{aligned}$$

Ass. (B1) says that $\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[X^2]$ and $\sup_{i\in\mathcal{I}} \lambda_{\max}(\mathbb{E}_{\tilde{M}(i)}[B(X)B(X)^\top])$ are finite. Hence, the supremum of each of the above terms is finite. That is, there exists a constant $c > 0$ such that

$$\Big| \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\big] - \inf_{f_\diamond\in\mathcal{F}_{\eta_0}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big]\Big|$$
$$\leq c\|\theta^0 - \hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}\|_2^2 + 2\sqrt{\mathbb{E}_M\big[\xi_Y^2\big]c}\|\theta^0 - \hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}\|_2.$$

It therefore suffices to show that

$$\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} \xrightarrow[n\to\infty]{P} \theta^0,$$

with respect to the distribution induced by M. To simplify notation, we henceforth drop the M subscript in the expectations and probabilities. Note that by the rank conditions in (B2), and the law of large numbers, we may assume that the corresponding sample product moments satisfy the same conditions. That is, for the purpose of the following arguments, it suffices that the sample product moment only satisfies these rank conditions asymptotically with probability one.

Let $B := B(X)$, $C := C(A)$, let $\mathbf{B}$ and $\mathbf{C}$ be row-wise stacked i.i.d. copies of $B(X)^\top$ and $C(A)^\top$, and recall the definition $\mathbf{P}_\delta := \mathbf{C}\left(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M}\right)^{-1}\mathbf{C}^\top$. By convexity of the objective function we can find a closed form expression for our estimator of θ^0 by solving the corresponding normal equations. The closed form expression is given by

$$\hat{\theta}^n_{\lambda,\eta,\mu} := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^k} \|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 + \gamma\theta^\top\mathbf{K}\theta,$$
$$= \left(\frac{\mathbf{B}^\top\mathbf{B}}{n} + \lambda^\star_n\frac{\mathbf{B}^\top\mathbf{P}_\delta\mathbf{P}_\delta\mathbf{B}}{n} + \frac{\gamma\mathbf{K}}{n}\right)^{-1}\left(\frac{\mathbf{B}^\top\mathbf{Y}}{n} + \lambda^\star_n\frac{\mathbf{B}^\top\mathbf{P}_\delta\mathbf{P}_\delta\mathbf{Y}}{n}\right),$$

where we used that $\lambda^\star_n \in [0,\infty)$ almost surely by (C2). Consequently (using standard convergence arguments and that $n^{-1}\gamma\mathbf{K}$ and $n^{-1}\delta\mathbf{M}$ converges to zero in probability), if $\lambda^\star_n$ diverges to infinity in probability as n tends to infinity, then

$$\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} \xrightarrow{P} \Big(\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\Big)^{-1}\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CY\right]$$
$$= \theta^0.$$

Here, we also used that the terms multiplied by $\lambda_n^\star$ are the only asymptotically relevant terms. These are the standard arguments that the K-class estimator (with minor penalized regression modifications) is consistent as long as the parameter $\lambda_n^\star$ converges to infinity, or, equivalently, $\kappa_n^\star = \lambda_n^\star/(1+\lambda_n^\star)$ converges to one in probability.

We now consider two cases: *(i)* $\mathbb{E}[B\xi_Y] \neq 0$ and *(ii)* $\mathbb{E}[B\xi_Y] = 0$, corresponding to the case with unmeasured confounding and without, respectively. For *(i)* we show that $\lambda_n^\star$ converges to infinity in probability and for *(ii)* we show consistency by other means (as $\lambda_n^\star$ might not converge to infinity in this case).

Case (i): The confounded case $\mathbb{E}[B\xi_Y] \neq 0$. It suffices to show that

$$\lambda_n^\star := \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \xrightarrow[n\to\infty]{P} \infty.$$

To that end, note that for fixed $\lambda \geq 0$ we have that

$$\hat{\theta}^n_{\lambda,\eta_0,\mu} \xrightarrow[n\to\infty]{P} \theta_\lambda, \tag{B.5.33}$$

where

$$\begin{aligned}\theta_\lambda := &\left(\mathbb{E}\left[BB^\top\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)^{-1} \\ &\times\left(\mathbb{E}\left[BY\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CY\right]\right).\end{aligned} \tag{B.5.34}$$

Recall that (B.5.32) states that $\mathbb{E}\left[CY\right] = \mathbb{E}\left[CB^\top\right]\theta^0$. Using (B.5.32) and that $Y = B^\top\theta^0 + \xi_Y$ $\mathbb{P}_M$-almost surely, we have that the latter factor of (B.5.34) is given by

$$\begin{aligned}&\mathbb{E}\left[BY\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CY\right] \\ &\quad= \mathbb{E}\left[BB^\top\right]\theta^0 + \mathbb{E}\left[B\xi_Y\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\theta^0 \\ &\quad= \left(\mathbb{E}\left[BB^\top\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)\theta^0 + \mathbb{E}\left[B\xi_Y\right]\end{aligned}$$

Inserting this into (B.5.34) we arrive at the following representation of θ_λ

$$\theta_\lambda = \theta^0 + \left(\mathbb{E}\left[BB^\top\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)^{-1}\mathbb{E}\left[B\xi_Y\right]. \tag{B.5.35}$$

Since $\mathbb{E}[B\xi_Y] \neq 0$ by assumption, the above yields that

$$\forall \lambda \geq 0 : \qquad \theta^0 \neq \theta_\lambda. \tag{B.5.36}$$

Now we prove that $\lambda_n^\star$ diverges to infinity in probability as n tends to infinity. That is, for any $\lambda \geq 0$ we will prove that

$$\lim_{n\to\infty} \mathbb{P}(\lambda_n^\star \leq \lambda) = 0.$$

We fix an arbitrary $\lambda \geq 0$. By (B.5.36) we have that $\theta^0 \neq \theta_\lambda$. This implies that there exists an $\varepsilon > 0$ such that $\theta^0 \notin \overline{B(\theta_\lambda, \varepsilon)}$, where $\overline{B(\theta_\lambda, \varepsilon)}$ is the closed ball in $\mathbb{R}^k$ with center θ_λ and radius ε. By the consistency result (B.5.33), we know that the sequence of events $(A_n)_{n\in\mathbb{N}}$, for every $n \in \mathbb{N}$, given by

$$A_n := (|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon) = (\hat{\theta}^n_{\lambda,\eta_0,\mu} \in \overline{B(\theta_\lambda, \varepsilon)}),$$

satisfies $\mathbb{P}(A_n) \to 1$ as $n \to \infty$. By assumption (C3) we have that

$$\tilde{\lambda} \mapsto T_n(\theta^n_{\tilde{\lambda},\eta_0,\mu}), \qquad \text{and} \qquad \theta \mapsto T_n(\theta),$$

are weakly decreasing and continuous, respectively. Together with the continuity of $\tilde{\lambda} \mapsto \hat{\theta}^n_{\tilde{\lambda},\eta_0,\mu}$, this implies that also the mapping $\tilde{\lambda} \mapsto T_n(\hat{\theta}^n_{\tilde{\lambda},\eta_0,\mu})$ is continuous. It now follows from Ass. (C2) (stating that $\lambda_n^\star$ is almost surely finite) that for all $n \in \mathbb{N}$, $\mathbb{P}(T_n(\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}) \leq q(\alpha)) = 1$. Furthermore, since $\tilde{\lambda} \mapsto T_n(\theta^n_{\tilde{\lambda},\eta_0,\mu})$ is weakly decreasing, it follows that

$$\begin{aligned}
\mathbb{P}(\lambda_n^\star \leq \lambda) &= \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}) \leq q(\alpha)\}) \\
&\leq \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\}) \\
&= \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap A_n) \\
&\qquad + \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap A_n^c) \\
&\leq \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap \{|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon\}) \\
&\qquad + \mathbb{P}(A_n^c).
\end{aligned}$$

It now suffices to show that the first term converges to zero, since $\mathbb{P}(A_n^c) \to 0$ as $n \to \infty$. We have

$$\begin{aligned}
&\mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap \{|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon\}) \\
&\quad \leq \mathbb{P}\Big(\{\lambda_n^\star \leq \lambda\} \cap \Big\{ \inf_{\theta \in \overline{B(\theta_\lambda,\varepsilon)}} T_n(\theta) \leq q(\alpha)\Big\} \cap \{|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon\}\Big) \\
&\quad \leq \mathbb{P}\Big(\inf_{\theta \in \overline{B(\theta_\lambda,\varepsilon)}} T_n(\theta) \leq q(\alpha)\Big) \\
&\quad \xrightarrow{P} 0,
\end{aligned}$$

as $n \to \infty$, since $\overline{B(\theta_\lambda, \varepsilon)}$ is a compact set not containing θ^0. Here, we used that the test statistic (T_n) is assumed to have compact uniform power (C1). Hence, $\lim_{n\to\infty} \mathbb{P}(\lambda_n^\star \leq \lambda) = 0$ for any $\lambda \geq 0$, proving that $\lambda_n^\star$ diverges to infinity in probability, which ensures consistency.

Case (ii): the unconfounded case $\mathbb{E}[B(X)\xi_Y] = 0$. Recall that

$$\begin{aligned}
\hat{\theta}^n_{\lambda,\eta_0,\mu} &:= \arg\min_{\theta \in \mathbb{R}^k} \|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda \|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 + \gamma \theta^\top \mathbf{K} \theta \\
&= \arg\min_{\theta \in \mathbb{R}^k} l^n_{\mathrm{OLS}}(\theta) + \lambda l^n_{\mathrm{TSLS}}(\theta) + \gamma l_{\mathrm{PEN}}(\theta), \qquad \text{(B.5.37)}
\end{aligned}$$

where we defined $l^n_{\mathrm{OLS}}(\theta) := n^{-1}\|\mathbf{Y} - \mathbf{B}\theta\|_2^2$, $l^n_{\mathrm{TSLS}}(\theta) := n^{-1}\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$, and $l_{\mathrm{PEN}}(\theta) := n^{-1}\theta^\top \mathbf{K}\theta$. For any $0 \leq \lambda_1 < \lambda_2$ we have

$$\begin{aligned}
&l^n_{\mathrm{OLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \lambda_1 l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \gamma l_{\mathrm{PEN}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) \\
&\quad \leq l^n_{\mathrm{OLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \lambda_1 l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \gamma l_{\mathrm{PEN}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) \\
&\quad = l^n_{\mathrm{OLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \lambda_2 l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \gamma l_{\mathrm{PEN}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) \\
&\qquad + (\lambda_1 - \lambda_2) l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) \\
&\quad \leq l^n_{\mathrm{OLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \lambda_2 l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \gamma l_{\mathrm{PEN}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) \\
&\qquad + (\lambda_1 - \lambda_2) l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}),
\end{aligned}$$

where we used (B.5.37). Rearranging this inequality and dividing by $(\lambda_1 - \lambda_2)$ yields

$$l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) \geq l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}),$$

proving that $\lambda \mapsto l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda,\eta_0,\mu})$ is weakly decreasing. Thus, since $\lambda^\star_n \geq 0$ almost surely, we have that

$$\begin{aligned} l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}) &\leq l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{0,\eta_0,\mu}) \\ &= n^{-1}(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu})^\top \mathbf{P}_\delta \mathbf{P}_\delta (\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu}). \end{aligned} \tag{B.5.38}$$

Furthermore, recall from (B.5.33) that

$$\hat{\theta}^n_{0,\eta_0,\mu} \xrightarrow[n\to\infty]{P} \theta_0 = \theta^0, \tag{B.5.39}$$

where the last equality follows from (B.5.35) using that we are in the unconfounded case $\mathbb{E}[B(X)\xi_Y] = 0$. By expanding and deriving convergence statements for each term, we get

$$\begin{aligned} &(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu})^\top \mathbf{P}_\delta \mathbf{P}_\delta (\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu}) \\ &\quad \xrightarrow[n\to\infty]{P} (\mathbb{E}[YC^\top] - \theta_0 \mathbb{E}[BC^\top])\mathbb{E}[C^\top C]^{-1}(\mathbb{E}[CY] - \mathbb{E}[CB^\top]\theta_0) \\ &\quad = 0, \end{aligned} \tag{B.5.40}$$

where we used Slutsky's theorem, the weak law of large numbers, (B.5.39) and (B.5.32). Thus, by (B.5.38) and (B.5.40) it holds that

$$l^n_{\mathrm{TSLS}}(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}) = n^{-1}\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu})\|_2^2 \xrightarrow[n\to\infty]{P} 0.$$

For any $z \in \mathbb{R}^n$ we have that

$$\begin{aligned} &\|\mathbf{P}_\delta z\|_2^2 \\ &= z^\top \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top\mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top z \\ &= z^\top \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top z \\ &= \|(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top z\|_2^2, \end{aligned}$$

hence

$$\begin{aligned} &\|H_n - G_n\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}\|_2^2 \\ &\quad = \|n^{-1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu})\|_2^2 \\ &\quad \xrightarrow{P} 0, \end{aligned} \tag{B.5.41}$$

where for each $n \in \mathbb{N}$, $G_n \in \mathbb{R}^{k\times k}$ and $H_n \in \mathbb{R}^{k\times 1}$ are defined as

$$\begin{aligned} G_n &:= n^{-1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top\mathbf{B}, \text{ and} \\ H_n &:= n^{-1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top\mathbf{Y}. \end{aligned}$$

Using the weak law of large numbers, the continuous mapping theorem and Slutsky's theorem, it follows that, as $n \to \infty$,

$$\begin{aligned} G_n &\xrightarrow{P} G := E[CC^\top]^{1/2}E[CC^\top]^{-1}E[CB^\top], \text{ and} \\ H_n &\xrightarrow{P} H := E[CC^\top]^{1/2}E[CC^\top]^{-1}E[CY] \\ &\qquad = E[CC^\top]^{1/2}E[CC^\top]^{-1}E[CB^\top]\theta^0 \\ &\qquad = G\theta^0, \end{aligned}$$

where the second to last equality follows from (B.5.32). Together with (B.5.41), we now have that

$$\|G_n\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - G\theta^0\|_2^2 \le \|G_n\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - H_n\|_2^2 + \|H_n - G\theta^0\|_2^2 \xrightarrow[n\to\infty]{P} 0.$$

Furthermore, by the rank assumptions in (B2) we have that $G_n \in \mathbb{R}^{k\times k}$ is of full rank (with probability tending to one), hence

$$\begin{aligned} \|\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - \theta^0\|_2^2 &= \|G_n^{-1}G_n(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - \theta^0)\|_2^2 \\ &\le \|G_n^{-1}\|_{\text{op}}^2\|G_n(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - \theta^0)\|_2^2 \\ &\xrightarrow{P} \|G^{-1}\|_{\text{op}}^2 \cdot 0 \\ &= 0, \end{aligned}$$

as $n \to \infty$, proving the proposition. □

C | Towards Causal Inference for Spatio-Temporal Data: Conflict and Forest Loss in Colombia

C.1. Examples

Let $(\mathbf{X}, \mathbf{Y}, \mathbf{H})$ come from an LSCM satisfying condition (L1) described in Section 4.2.3.2. Below, we give two examples of distributions over $(\mathbf{X}, \mathbf{H})$ for which also conditions (L2) and (L3) hold true. In both cases, $\frac{1}{m}(\mathbf{\Phi}_s^m)^\top \mathbf{\Phi}_s^m$ converges in probability to some limit matrix of the form $\mathbb{E}_\nu[\varphi(X)\varphi(X)^\top]$ for some measure ν with full support on $\mathbb{R}^d$. To see that $\mathbb{E}_\nu[\varphi(X)\varphi(X)^\top]$ is strictly positive definite, let $v \in \mathbb{R}^p$ be such that $0 = v^\top \mathbb{E}_\nu[\varphi(X)\varphi(X)^\top]v = \mathbb{E}_\nu[\|\varphi(X)^\top v\|_2^2]$. By continuity of φ, it follows that $\varphi^\top v \equiv 0$, and the linear independence of $\varphi_1, \ldots, \varphi_p$ implies that $v = 0$.

Example C.1 (Temporally ergodic $\mathbf{X}$)**.** *Let* $(\mathbf{X}, \mathbf{Y}, \mathbf{H})$ *come from an LSCM satisfying Assumption (L1). Assume that for every* $\mathbf{h} \in \mathcal{H}$ *and* $s \in \mathbb{R}^2$*, it holds that under* $\mathbb{P}_\mathbf{h}$*,* $\mathbf{X}_s$ *is a stationary and mixing process with a marginal distribution that has full support on* $\mathbb{R}^d$ *(e.g., a vector autoregessive process with additive Gaussian noise). Assume further that* $\mathbb{E}_\mathbf{h}[|\xi_s^1|^2] < \infty$ *and* $\mathbb{E}_\mathbf{h}[|\varphi_i(X_s^1)|^2 < \infty]$ *for all* $i \in \{1, \ldots, p\}$*. Analogously to the proof of Proposition 4.3, we can then show that for each* $i, j \in \{1, \ldots, p\}$*, the sequences* $(\varphi_i(X_s^t)\xi_s^t)_{t\in\mathbb{N}}$ *and* $(\varphi_i(X_s^t)\varphi_j(X_s^t))_{t\in\mathbb{N}}$ *are ergodic under* $\mathbb{P}_\mathbf{h}$*, and it follows that*

$$(\tfrac{1}{m}(\mathbf{\Phi}_s^m)^\top \boldsymbol{\xi}_s^m)_i = \frac{1}{m}\sum_{t=1}^m \varphi_i(X_s^t)\xi_s^t \to \mathbb{E}_\mathbf{h}[\varphi_i(X_s^1)\xi_s^1]$$

and

$$(\tfrac{1}{m}(\mathbf{\Phi}_s^m)^\top \mathbf{\Phi}_s^m)_{ij} = \frac{1}{m}\sum_{t=1}^m \varphi_i(X_s^t)\varphi_j(X_s^t) \to \mathbb{E}_\mathbf{h}[\varphi_i(X_s^1)\varphi_j(X_s^1)]$$

as $m \to \infty$ *in probability under* $\mathbb{P}_\mathbf{h}$*. Since for all* $s \in \mathbb{R}^2$*,* $\mathbb{E}_\mathbf{h}[\varphi(X_s^1)\xi_s^1] = \mathbb{E}_\mathbf{h}[\varphi(X_s^1)] \cdot \mathbb{E}_\mathbf{h}[\xi_s^1] = 0$ *and* $\mathbb{E}_\mathbf{h}[\varphi(X_s^1)\varphi(X_s^1)^\top] \succ 0$*, the above implies (L2) and (L3).*

Example C.2 (Temporally independent $\mathbf{X}$ with convergent mixture distributions)**.** *Let* $(\mathbf{X}, \mathbf{Y}, \mathbf{H})$ *come from an LSCM satisfying Assumption (L1) for some bounded functions* $\varphi_1, \ldots, \varphi_p$*. Assume that for every* $s \in \mathbb{R}^2$*, the variables* $X_s^1, X_s^2, \ldots$ *are conditionally independent given* $\mathbf{H}$ *(they are not required to be identically distributed),*

and that for every $\mathbf{h} \in \mathcal{H}$*, the sequence of mixture distributions*

$$\mathbb{P}^m_{s,\mathbf{h}} := \frac{1}{m}\sum_{t=1}^{m} \mathbb{P}_{X^t_s|\mathbf{H}=\mathbf{h}}, \quad m \in \mathbb{N}, \tag{C.1.1}$$

converges, for $m \to \infty$*, weakly towards some limit measure* $\mathbb{P}^\infty_{s,\mathbf{h}}$ *with full support on* $\mathbb{R}^d$*. Then, conditions (L1) and (L2) are satisfied. To see this, let* $\mathbf{h} \in \mathcal{H}$ *and* $s \in \mathbb{R}^2$ *be fixed for the rest of this example. Let* $m \in \mathbb{N}$ *and* $\delta > 0$*. Since* $\mathbb{E}_{\mathbf{h}}[(\mathbf{\Phi}^m_s)^\top \boldsymbol{\xi}^m_s] = 0$*, it follows from Chebychev's inequality that for all* $i \in \{1, \dots, p\}$*,*

$$\begin{aligned}\mathbb{P}_{\mathbf{h}}(|\tfrac{1}{m}((\mathbf{\Phi}^m_s)^\top \boldsymbol{\xi}^m_s)_i| > \delta) &\leq \frac{1}{\delta^2}\mathrm{Var}_{\mathbf{h}}(\frac{1}{m}((\mathbf{\Phi}^m_s)^\top \boldsymbol{\xi}^m_s)_i) \\ &= \frac{\mathbb{E}_{\mathbf{h}}((\xi^1_0)^2)}{\delta^2 m} \underbrace{\mathbb{E}^m_{s,\mathbf{h}}[\varphi_i(X)^2]}_{\text{unif. bounded}} \to 0,\end{aligned}$$

as $m \to \infty$*, showing that (L2) is satisfied. To prove (L3), let* $M^m := \mathbb{E}^m_{s,\mathbf{h}}[\varphi(X)\varphi(X)^\top]$*,* $m \in \mathbb{N}$*, and* $M^\infty := \mathbb{E}^\infty_{s,\mathbf{h}}[\varphi(X)\varphi(X)^\top]$ *(to simplify notation, we here omit the implicit dependence on* $\mathbf{h}$ *and* s*). By assumption on* $(\mathbb{P}^m_{s,\mathbf{h}})_{m\in\mathbb{N}}$*,* M^m *converges entrywise to* M^∞ *as* $m \to \infty$*. Together with another application of Chebychev's inequality, it follows that for all* $i, j \in \{1, \dots, p\}$*,*

$$\begin{aligned}&\mathbb{P}_{\mathbf{h}}(|\tfrac{1}{m}((\mathbf{\Phi}^m_s)^\top \mathbf{\Phi}^m_s)_{ij} - M^\infty_{ij}| > 2\delta) \\ &\quad\leq \mathbb{P}_{\mathbf{h}}(|\tfrac{1}{m}((\mathbf{\Phi}^m_s)^\top \mathbf{\Phi}^m_s)_{ij} - M^m_{ij}| > \delta) + \mathbb{P}_{\mathbf{h}}(|M^m_{ij} - M^\infty_{ij}| > \delta) \\ &\quad\leq \frac{1}{\delta^2 m}\underbrace{\mathbb{E}^m_{s,\mathbf{h}}[\varphi_i(X)^2\varphi_j(X)^2]}_{\text{unif. bounded}} + \underbrace{\mathbb{P}_{\mathbf{h}}(|M^m_{ij} - M^\infty_{ij}| > \delta)}_{=0 \text{ for } m \text{ large}} \to 0,\end{aligned}$$

as $m \to \infty$*, showing that* $\frac{1}{m}((\mathbf{\Phi}^m_s)^\top \mathbf{\Phi}^m_s)$ *converges entrywise to* $M^\infty \succ 0$ *in probability under* $\mathbb{P}_{\mathbf{h}}$*, and (L3) follows.*

Remark C.1 (Necessity of the convergence of mixtures)**.** *The convergence assumption on* $\mathbb{P}^m_{s,\mathbf{h}}$ *is crucial for obtaining the above consistency result. It is easy to construct examples of* $(\mathbb{P}_{X^t_s|\mathbf{H}=\mathbf{h}})_{t\in\mathbb{N}}$ *where this assumption fails to hold. For example, let* $\mathbb{P}_{\mathbf{h}}(X^t_s \in (-\infty, -1]^d) = 1$ *whenever* $\lfloor \log_2 t \rfloor$ *is even, and* $\mathbb{P}_{\mathbf{h}}(X^t_s \in [1, \infty)^d) = 1$

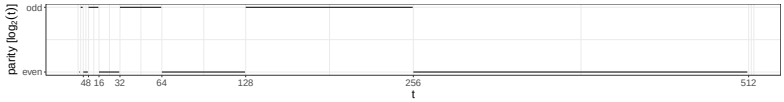


FIGURE C.1. Visualization of the example in Remark C.1. Whenever the parity of $\lfloor \log_2 t \rfloor$ changes from even to odd, the entire mass of $\mathbb{P}_{X_s^t|\mathbf{H}=\mathbf{h}}$ moves from $(-\infty,-1]^d$ to $[1,\infty)^d$, and vice versa. In this case, the mixture $\mathbb{P}^m_{s,\mathbf{h}}$ in (C.1.1) does not converge, and the consistency in Proposition 4.4 does not hold in general.

whenever $\lfloor \log_2 t \rfloor$ is odd. This construction is visualized in Figure C.1. Then, both sequences $(\mathbb{P}_{\mathbf{h}}(X_s^t \in (-\infty,-1]^d))_{t\in\mathbb{N}}$ and $(\mathbb{P}_{\mathbf{h}}(X_s^t \in [1,\infty)^d))_{t\in\mathbb{N}}$ alternate between zero and one, with a frequency chosen such that for all $k \geq 2$ even, $\mathbb{P}^{2^k-1}_{s,\mathbf{h}}([1,\infty)^d) = 2/3$, and for all $k \geq 3$ odd, $\mathbb{P}^{2^k-2}_{s,\mathbf{h}}((-\infty,-1]^d) = 2/3$, showing that $\mathbb{P}^m_{s,\mathbf{h}}$ does not converge. In this case, the dataset $\{(X_s^t, Y_s^t) : t \in \{1,\dots,m\}\}$ alternates between mostly containing pairs (X_s^t, Y_s^t) with $X_s^t \in (-\infty,-1]^d$ and mostly containing pairs (X_s^t, Y_s^t) with $X_s^t \in [1,\infty)^d$. If the functional dependence of Y_s^t on X_s^t differs between these two domains, the estimator $\hat{f}^m_{Y|X}$ does therefore not converge in general.

Remark C.2 (Conditions implying the convergence of mixtures)**.** *We can make the convergence assumption on $\mathbb{P}^m_{s,\mathbf{h}}$ more concrete in the case where the distributions in $(\mathbb{P}_{X_s^t|\mathbf{H}=\mathbf{h}})_{t\in\mathbb{N}}$ differ only in their respective mean vectors. Assume there exist functions $\mu_s^t : \mathcal{Z}_\ell \to \mathbb{R}^d$ and $g_s : \mathcal{Z}_\ell \times \mathbb{R}^d \to \mathbb{R}^d$, $(s,t) \in \mathbb{R}^2 \times \mathbb{N}$, and a d-dimensional error process $\boldsymbol{\zeta} \perp\!\!\!\perp \mathbf{H}$, such that for each $s \in \mathbb{R}^2$, $\zeta_s^1, \zeta_s^2, \dots$ are i.i.d., and such that for all $(s,t) \in \mathbb{R}^2 \times \mathbb{N}$ it holds that $X_s^t = \mu_s^t(\mathbf{H}) + g_s(\mathbf{H}, \zeta_s^t)$. Assume further that for each $\mathbf{h} \in \mathcal{H}$ and $s \in \mathbb{R}^2$, $g_s(\mathbf{h}, \zeta_s^0)$ has strictly positive density $f_{s,\mathbf{h}}$ w.r.t. the Lebesgue measure on $\mathbb{R}^d$. We can then ensure convergence of the mixture distributions $\mathbb{P}^m_{s,\mathbf{h}}$ by requiring that for each $\mathbf{h} \in \mathcal{H}$ and $s \in \mathbb{R}^2$ there exists some density function $f^{mix}_{s,\mathbf{h}}$*

on $\mathbb{R}^d$*, such that for all* $x \in \mathbb{R}^d$ *it holds that*[1]

$$\lim_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{1}_{(-\infty,x]}(\mu_s^t(\mathbf{h})) = \int_{(-\infty,x]} f_{s,\mathbf{h}}^{mix}(z)dz.$$

(Intuitively, this equation states that, in the limit $m \to \infty$*, the set* $\{\mu_s^t(\mathbf{h}) : t \in \{1,\dots,m\}\}$ *looks like an i.i.d. sample drawn from the distribution with density* $f_{s,\mathbf{h}}^{mix}$*.) For all* $\mathbf{h} \in \mathcal{H}$*,* $s \in \mathbb{R}^2$*,* $m \in \mathbb{N}$ *and* $x \in \mathbb{R}^d$*, we then have*

$$\begin{aligned}
\mathbb{P}_{s,\mathbf{h}}^m((-\infty,x]) &= \frac{1}{m}\sum_{t=1}^{m} \int_{(-\infty,x]} f_{s,\mathbf{h}}(v - \mu_s^t(\mathbf{h}))dv \\
&= \int_{\mathbb{R}^d} \frac{1}{m}\sum_{t=1}^{m} f_{s,\mathbf{h}}(v - \mu_s^t(\mathbf{h}))\,\mathbb{1}_{(-\infty,x]}(v)dv \\
&= \int_{\mathbb{R}^d} f_{s,\mathbf{h}}(v)\frac{1}{m}\sum_{t=1}^{m} \mathbb{1}_{(-\infty,x-v]}(\mu_s^t(\mathbf{h}))dv,
\end{aligned}$$

and it follows from the dominated convergence theorem that

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{P}_{s,\mathbf{h}}^m((-\infty,x]) &= \int_{\mathbb{R}^k} f_{s,\mathbf{h}}(v) \int_{(-\infty,x-v]} f_{s,\mathbf{h}}^{mix}(z)dzdv \\
&= \int_{(-\infty,x]} \int_{\mathbb{R}^k} f_{s,\mathbf{h}}(v) f_{s,\mathbf{h}}^{mix}(z-v)dvdz \\
&= \int_{(-\infty,x]} (f_{s,\mathbf{h}} * f_{s,\mathbf{h}}^{mix})(z)dz,
\end{aligned}$$

showing that $\mathbb{P}_{s,\mathbf{h}}^m$ *converges weakly to the measure with the convoluted density* $f_{s,\mathbf{h}} * f_{s,\mathbf{h}}^{mix}$*. Since* $f_{s,\mathbf{h}}$ *is strictly positive, this measure has full support on* $\mathbb{R}^d$*.*

[1] By slight abuse of notation, we use $(-\infty, x]$ to denote the product set $\bigtimes_{i=1}^{d}(-\infty, x_i]$.

C.2. Proofs

C.2.1. Proof of Proposition 4.1

By definition, intervening on $\mathbf{X}$ leaves the conditional distribution $\mathbf{Y} \mid (\mathbf{X}, \mathbf{H})$ unchanged. Under $\mathbb{P}_x$, the property (4.2.2) therefore still holds for the same error process $\boldsymbol{\varepsilon}$. Since also the marginal distribution of $\mathbf{H}$ is unaffected by the intervention, we have that

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_x}[Y_s^t] &= \mathbb{E}_{\mathbb{P}_x}[f(X_s^t, H_s^t, \varepsilon_s^t)] = \mathbb{E}_{\mathbb{P}_x}[f(x, H_s^t, \varepsilon_s^t)] \\ &= \mathbb{E}[f(x, H_s^t, \varepsilon_s^t)] = \mathbb{E}[f(x, H_0^1, \varepsilon_0^1)] = f_{\mathrm{AVE}(X \to Y)}(x),\end{aligned}$$

as desired. □

C.2.2. Proof of Theorem 4.2

Consider a fixed $x \in \mathcal{X}$. For every $n, m \in \mathbb{N}$ we have that

$$\begin{aligned}&\hat{f}^{nm}_{\mathrm{AVE}(X \to Y)}(\mathbf{X}_n^m, \mathbf{Y}_n^m)(x) - f_{\mathrm{AVE}(X \to Y)}(x) \\ &\quad= \frac{1}{n} \sum_{i=1}^{n} \hat{f}^m_{Y|X}(\mathbf{X}_{s_i}^m, \mathbf{Y}_{s_i}^m)(x) - \mathbb{E}\left[f_{Y|(X,H)}(x, H_0^1)\right] \\ &\quad= \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}^m_{Y|X}(\mathbf{X}_{s_i}^m, \mathbf{Y}_{s_i}^m)(x) - f_{Y|(X,H)}(x, H_{s_i}^1)\right) \\ &\qquad+ \frac{1}{n} \sum_{i=1}^{n} f_{Y|(X,H)}(x, H_{s_i}^1) - \mathbb{E}\left[f_{Y|(X,H)}(x, H_0^1)\right] \\ &\quad= r_1(\mathbf{X}_n^m, \mathbf{Y}_n^m, \mathbf{H}_n^1) + r_2(\mathbf{H}_n^1),\end{aligned}$$

where

$$r_1(\mathbf{X}_n^m, \mathbf{Y}_n^m, \mathbf{H}_n^1) := \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}^m_{Y|X}(\mathbf{X}_{s_i}^m, \mathbf{Y}_{s_i}^m)(x) - f_{Y|(X,H)}(x, H_{s_i}^1)\right) \text{ and}$$

$$r_2(\mathbf{H}_n^1) := \frac{1}{n} \sum_{i=1}^{n} f_{Y|(X,H)}(x, H_{s_i}^1) - \mathbb{E}\left[f_{Y|(X,H)}(x, H_0^1)\right].$$

It follows that for any $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(\mathbf{X}^m_n, \mathbf{Y}^m_n)(x) - f^0_{\mathrm{AVE}(X\to Y)}(x)\right| > \delta\right) \\
&\quad \leq \mathbb{P}\left(|r_1(\mathbf{X}^m_n, \mathbf{Y}^m_n, \mathbf{H}^1_n)| > \delta/2\right) + \mathbb{P}\left(|r_2(\mathbf{H}^1_n)| > \delta/2\right).
\end{aligned}$$

Let now $\alpha > 0$ be arbitrary. By Assumption 4.1, there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $\mathbb{P}(|r_2(\mathbf{H}^1_n)| \geq \delta/2) \leq \alpha/2$. By Assumption 4.2, we can for any such $n \geq N$ find $M_n \in \mathbb{N}$, such that for all $i = 1, \ldots, n$ and all $m \geq M_n$ it holds that $\mathbb{P}(|\hat{f}^m_{Y|X}(\mathbf{X}^m_{s_i}, \mathbf{Y}^m_{s_i})(x) - f_{Y|(X,H)}(x, H^1_{s_i})| > \delta/2) \leq \alpha/(2n)$. For all $m \geq M_n$ we then have

$$\begin{aligned}
&\mathbb{P}\left(|r_1(\mathbf{X}^m_n, \mathbf{Y}^m_n, \mathbf{H}^1_n)| > \delta/2\right) \\
&\quad \leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \left|\hat{f}^m_{Y|X}(\mathbf{X}^m_{s_i}, \mathbf{Y}^m_{s_i})(x) - f_{Y|(X,H)}(x, H^1_{s_i})\right| > \delta/2\right) \\
&\quad \leq \mathbb{P}\left(\bigcup_{i=1}^n \left\{\left|\hat{f}^m_{Y|X}(\mathbf{X}^m_{s_i}, \mathbf{Y}^m_{s_i})(x) - f_{Y|(X,H)}(x, H^1_{s_i})\right| > \delta/2\right\}\right) \\
&\quad \leq \sum_{i=1}^n \mathbb{P}\left(\left|\hat{f}^m_{Y|X}(\mathbf{X}^m_{s_i}, \mathbf{Y}^m_{s_i})(x) - f_{Y|(X,H)}(x, H^1_{s_i})\right| > \delta/2\right) \\
&\quad \leq \sum_{i=1}^n \alpha/(2n) = \alpha/2,
\end{aligned}$$

and the result follows. □

C.2.3. Proof of Proposition 4.3

By construction, $(H^1_{s_n})_{n\in\mathbb{N}}$ can be decomposed into m subsequences $(H^1_{s_{(n-1)m+j}})_{n\in\mathbb{N}}$, $j \in \{1, \ldots, m\}$, each of which corresponds to an equally spaced sampling of $\mathbf{H}^1$ along the first spatial axis. We fist prove that each of these subsequences satisfies Assumption 4.1, and then conclude that the same must hold for the original sequence $(H^1_{s_n})_{n\in\mathbb{N}}$. Let $j \in \{1, \ldots, m\}$ and let $\varphi : \mathbb{R}^\ell \to \mathbb{R}$ be a measurable function with $\mathbb{E}[|\varphi(H^1_0)|] < \infty$. For notational simplicity, let for each $n \in \mathbb{N}$, $Z_n := H^1_{s_{(n-1)m+j}}$. The idea is to apply an ergodic theorem for real-valued stationary and ergodic time series [e.g., Rønn-Nielsen

and Sokol, 2013, Corollary 2.3.13] to the sequence $(\varphi(Z_n))_{n\in\mathbb{N}}$. By stationarity of the process $\mathbf{H}^1$, and by choice of the sampling scheme, $(\varphi(Z_n))_{n\in\mathbb{N}}$ is indeed stationary. We need to show that $(\varphi(Z_n))_{n\in\mathbb{N}}$ is also ergodic. Using Rønn-Nielsen and Sokol [2013, Lemma 2.3.15], this follows by proving the following mixing condition: for all $p, q \geq 1$ and all $A_1, \ldots, A_p \in \mathcal{B}(\mathbb{R})$ and $B_1, \ldots, B_q \in \mathcal{B}(\mathbb{R})$, it holds that

$$\begin{aligned}&\mathbb{P}(\varphi(Z_1) \in A_1, \ldots, \varphi(Z_p) \in A_p, \varphi(Z_{n+1}) \in B_1, \ldots, \varphi(Z_{n+q}) \in B_q) \\ &\to \mathbb{P}(\varphi(Z_1) \in A_1, \ldots, \varphi(Z_p) \in A_p) \cdot \mathbb{P}(\varphi(Z_1) \in B_1, \ldots, \varphi(Z_q) \in B_q),\end{aligned} \tag{C.2.1}$$

as $n \to \infty$. Since the finite-dimensional distributions of $(Z_n)_{n\in\mathbb{N}}$ are Gaussian, this condition is easily verified. Let $p, q \geq 1$, and let $\mathbb{P}_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbb{P}_2 = \mathcal{N}(\mu_2, \Sigma_2)$ be the distributions of $(Z_1, \ldots, Z_p)$ and $(Z_1, \ldots, Z_q)$, respectively. Property (C.2.1) follows if we can show that $(Z_1, \ldots, Z_m, Z_{n+1}, \ldots, Z_{n+p})$ converges to $\mathbb{P}_1 \otimes \mathbb{P}_2 = \mathcal{N}((\mu_1, \mu_2), \mathrm{diag}(\Sigma_1, \Sigma_2))$ in distribution as $n \to \infty$. Convergence of the mean vector is trivial, and convergence of the covariance matrix follows by the assumption on C and our choice of spatial sampling (the distance between the respective locations at which $(Z_1, \ldots, Z_m)$ and $(Z_{n+1}, \ldots, Z_{n+p})$ are observed tends to infinity as n increases). To prove that the limit distribution is indeed Gaussian, one can then consider characteristic functions and apply a combination of Levy's Continuity Theorem [e.g., Williams, 1991, Theorem 18.1] and the Cramér-Wold Theorem [Cramér and Wold, 1936]. This proves that $\frac{1}{n}\sum_{i=1}^n \varphi(Z_i) \to \mathbb{E}[\varphi(Z_1)]$ in probability as $n \to \infty$, i.e., the subsequence $(H^1_{s_{(n-1)m+j}})_{n\in\mathbb{N}}$ satisfies Assumption 4.1. Since j was arbitrary, this holds true for all $j \in \{1, \ldots, m\}$. It remains to prove that also the original sequence $(H^1_{s_n})_{n\in\mathbb{N}}$ satisfies Assumption 4.1.

Let an integrable function $\varphi : \mathbb{R}^\ell \to \mathbb{R}$ be given, and assume first that $\mathbb{E}[\varphi(H^1_0)] = 0$. For every $j \in \{1, \ldots, m\}$ and $i \in \mathbb{N}$, define $S^j_i := \sum_{k=1}^i \varphi(H^1_{s_{(k-1)m+j}})$. By the first part of the proof, we have that for all j, $\frac{1}{i}S^j_i \to 0$ in probability as $i \to \infty$. We want to show that also $\frac{1}{n}\sum_{k=1}^n \varphi(H^1_{s_k}) \to 0$ in probability as $n \to \infty$. Let $\delta, \alpha > 0$ and choose $I \in \mathbb{N}$ such that for all $j \in \{1, \ldots, m\}$ and $i \geq I$, $\mathbb{P}(\frac{1}{i}|S^j_i| > \delta/m) \leq \alpha/m$. Define $N := mI + 1$ and pick an arbitrary $n \geq N$. We can then write $n = im + j$ for some $i \geq I$ and

$j \in \{1, \dots, m\}$. With $J_1 := \{1, \dots, j\}$ and $J_2 = \{1, \dots, m\} \setminus J_1$, we then have

$$\begin{aligned}
\mathbb{P}(\frac{1}{n}|\sum_{k=1}^{n} \varphi(H^1_{s_k})| > \delta) &= \mathbb{P}(\frac{1}{n}|\sum_{j' \in J_1} S^{j'}_{i+1} + \sum_{j' \in J_2} S^{j'}_i| > \delta) \\
&\leq \mathbb{P}(\sum_{j' \in J_1} \frac{1}{i+1}|S^{j'}_{i+1}| + \sum_{j' \in J_2} \frac{1}{i}|S^{j'}_i| > \delta) \\
&\leq \sum_{j' \in J_1} \mathbb{P}(\frac{1}{i+1}|S^{j'}_{i+1}| > \delta/m) \\
&\quad + \sum_{j' \in J_2} \mathbb{P}(\frac{1}{i}|S^{j'}_i| > \delta/m) \leq \alpha,
\end{aligned}$$

which completes the proof in the case where $\mathbb{E}[\varphi(H^1_0)] = 0$. The general case follows by applying the above result to the function $\tilde{\varphi} = \varphi - \mathbb{E}[\varphi(H^1_0)]$. □

C.2.4. Proof of Proposition 4.4

Let $s \in \mathbb{R}^2$ and $\mathbf{h} \in \mathcal{H}$. With $\gamma_s := f_1(h^1_s)$, it follows from (L1) that for all x and t, we have $\mathbb{E}[Y^t_s \,|\, X^t_s = x, H^t_s = h^t_s] = \varphi(x)^\top \gamma_s$. It therefore suffices to prove that $\hat{\gamma}^m_s \to \gamma_s$ in probability under $\mathbb{P}_{\mathbf{h}}$. For the ease of notation, we omit all sub- and superscripts from $\mathbf{Y}^m_s$, $\mathbf{\Phi}^m_s$ and $\boldsymbol{\xi}^m_s$ in the below calculations. Let $c > 0$ be such that (L3) holds

true and let $\delta > 0$ be arbitrary. For every $m \in \mathbb{N}$, we then have

$$
\begin{aligned}
&\mathbb{P}_{\mathbf{h}}(\|\gamma_s - \hat{\gamma}_s^m\|_2 > \delta) \\
&\quad = \mathbb{P}_{\mathbf{h}}(\|\gamma_s - (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathbf{Y}\|_2 > \delta) \\
&\quad = \mathbb{P}_{\mathbf{h}}(\|\gamma_s - (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\gamma_s + \boldsymbol{\xi})\|_2 > \delta) \\
&\quad = \mathbb{P}_{\mathbf{h}}(\|(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\xi}\|_2 > \delta) \\
&\quad \leq \mathbb{P}_{\mathbf{h}}(\|(\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\|_2\|\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\xi}\|_2 > \delta\,|\,) \\
&\quad = \mathbb{P}_{\mathbf{h}}((\lambda_{\min}(\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}))^{-1}\|\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\xi}\|_2 > \delta) \\
&\quad = \mathbb{P}_{\mathbf{h}}((\lambda_{\min}(\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}))^{-1}\|\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\xi}\|_2 > \delta \text{ and } \lambda_{\min}\left(\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right) > c) \\
&\qquad + \mathbb{P}_{\mathbf{h}}((\lambda_{\min}(\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}))^{-1}\|\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\xi}\|_2 > \delta \text{ and } \lambda_{\min}\left(\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right) \leq c) \\
&\quad \leq \mathbb{P}_{\mathbf{h}}(\|\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\xi}\|_2 > c\delta) + \mathbb{P}_{\mathbf{h}}(\lambda_{\min}\left(\tfrac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right) \leq c),
\end{aligned}
$$

which tends to zero as $m \to \infty$ by (L2) and (L3). □

C.2.5. Proof of Proposition 4.5

Recall our definition of $\mathcal{H} \subseteq \mathcal{Z}_\ell$ as the set of functions $\mathbf{h} : \mathbb{R}^2 \times \mathbb{N} \to \mathbb{R}^\ell$ that are constant in the time-argument. Since $\mathbf{H}$ is time-invariant, we have that $\mathbb{P}(\mathbf{H} \in \mathcal{H}) = 1$. It therefore suffices to prove that for all $\mathbf{h} \in \mathcal{H}$, $(\mathbf{X}_n^m, \sigma(\mathbf{Y}_n^m)) \overset{d}{=} (\mathbf{X}_n^m, \mathbf{Y}_n^m)$ under $\mathbb{P}(\,\cdot\,|\,\mathbf{H} = \mathbf{h})$. Assume that H_0 holds true, and let σ be a permutation of $\{1, \dots, m\}$. Then, there exists a function $\tilde{f} : \mathbb{R}^{\ell+1} \to \mathbb{R}$ and an error process $\boldsymbol{\varepsilon} \perp\!\!\!\perp (\mathbf{X}, \mathbf{H})$ such that for all $(s,t) \in \mathbb{R}^2 \times \mathbb{N}$, $Y_s^t = \tilde{f}(H_s, \varepsilon_s^t)$. It follows that the conditional distribution of $\mathbf{Y}\,|\,(\mathbf{X}, \mathbf{H})$ does not depend on $\mathbf{X}$, and hence that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{H}$. Further, since $\boldsymbol{\varepsilon}^1, \boldsymbol{\varepsilon}^2, \dots$ are i.i.d., we have that for all $\mathbf{h} \in \mathcal{H}$, $\mathbf{Y}^1, \dots, \mathbf{Y}^m$ are i.i.d. under $\mathbb{P}(\,\cdot\,|\,\mathbf{H} = \mathbf{h})$. For all $\mathbf{h} \in \mathcal{H}$, it therefore holds that $(\mathbf{X}_n^m, \sigma(\mathbf{Y}_n^m)) \overset{d}{=} (\mathbf{X}_n^m, \mathbf{Y}_n^m)$ under $\mathbb{P}(\,\cdot\,|\,\mathbf{H} = \mathbf{h})$, and the result follows. □

C.3. Further results on resampling tests

C.3.1. Temporal autocorrelation in the response variable

A central assumption of the LSCM model class is that the error process of $\mathbf{Y}$ is independent over time. This assumption says that all dependencies between different temporal instances of $\mathbf{Y}$ are induced via the covariates $\mathbf{X}$ or the time-invariant confounders $\mathbf{H}$. In practice, there may be other time-varying conditions influencing forest loss, thereby inducing a temporal dependence in $\mathbf{Y}$ which cannot be explained by $(\mathbf{X}, \mathbf{H})$. In this case, the exchangeability property in Proposition 4.5, and therefore the level of our resampling test, is violated. To incorporate temporal autocorrelation in the response variable, we adopt a block-permutation scheme: we divide the period 2000–2018 into 6 blocks (2000–2002, 2003–2005, ..., 2016–2018), and perform a block-wise permutation of the data from $\mathbf{Y}$. This procedure leaves the within-block dependence structure in $\mathbf{Y}$ intact. The results align with our previous findings: $P = 0.892$ for the test of an instantaneous effect, and $P = 0.498$ for the test of a temporally lagged effect (when using the same test statistics as in Section 4.4.4).

C.3.2. Spatial block-permutation scheme for Model 1

In Section 4.4.3, we describe alternative permutation schemes to test the null hypotheses in Models 1 and 2. Strictly speaking, we require additional exchangeability assumptions on $\mathbf{Y}$ to ensure the validity of the corresponding resampling tests. Here, we investigate an alternative permutation scheme for Model 1. To account for the spatial autocorrelation in $\mathbf{Y}$, we adopt a spatial block-permutation: for every year 2000–2018, observations are grouped into blocks of size 100km $\times$ 100km. To obtain resampled datasets, we then permute values of $\mathbf{Y}$ in these blocks of data, thereby leaving the spatial dependence within each block intact. Observations which do not fall in any of the blocks are permuted randomly. As seen in Figure C.2, this procedure slightly increases the p-value, but does not affect the significance of the test.

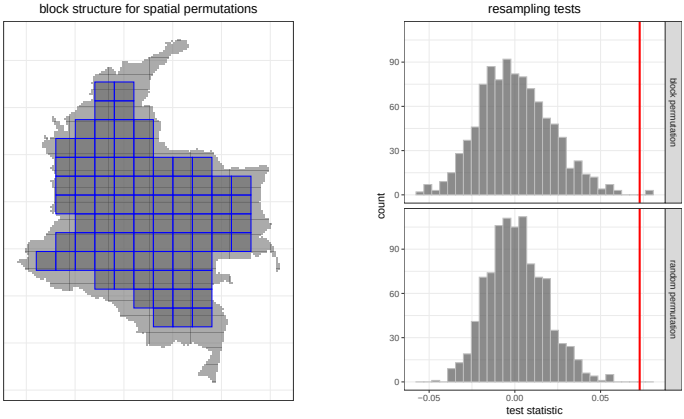


FIGURE C.2. Block structure for the spatial permutation scheme (left) and results of resampling tests (right) for the null hypothesis in Model 1 from Section 4.4.4. The test statistic $\hat{T} = \hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(1) - \hat{f}^{nm}_{\mathrm{AVE}(X\to Y)}(0)$ is indicated by a red vertical bar. The empirical distribution of the test statistic under this permutation scheme (top right) has a higher variance than under the permutation scheme used in Section 4.4.4 (bottom right), resulting in a slightly larger p-value of 0.008 compared with the p-value of 0.002 for the original test. The significance of the test does not change.

Bibliography

S. S. Abadeh, P. M. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

J. Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.

T. Amemiya. The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2:105–110, 1974.

T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63, 1949.

A. Angelsen and D. Kaimowitz. Rethinking the causes of deforestation: lessons from economic models. *The world bank research observer*, 14(1):73–98, 1999.

M. Appel and E. Pebesma. On-demand processing of data cubes from satellite image collections with the gdalcubes library. *Data*, 4(3):92, 2019.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *ArXiv e-prints (1907.02893)*, 2019.

D. Armenteras, E. Cabrera, N. Rodríguez, and J. Retana. National and regional determinants of tropical deforestation in Colombia. *Regional Environmental Change*, 13(6):1181–1193, 2013.

D. Armenteras, L. Schneider, and L. M. Dávalos. Fires in protected areas reveal unforeseen costs of Colombian peace. *Nature ecology & evolution*, 3(1):20–23, 2019.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

V. Avitabile, M. Herold, G. B. Heuvelink, S. L. Lewis, O. L. Phillips, G. P. Asner, J. Armston, P. S. Ashton, L. Banin, N. Bayol, et al. An integrated pan-tropical biomass map using multiple reference datasets. *Global change biology*, 22(4):1406–1420, 2016.

P. L. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In *21st Annual Conference on Learning Theory (COLT)*, 2008.

M. Baumann and T. Kuemmerle. The impacts of warfare and armed conflict on land systems. *Journal of land use science*, 11(6):672–688, 2016.

A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

M. Bertolacci, E. Cripps, O. Rosen, J. W. Lau, S. Cripps, et al. Climate inference on daily rainfall across the Australian continent, 1876–2015. *The Annals of Applied Statistics*, 13(2):683–712, 2019.

D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

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.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10 (Sep):2137–2155, 2009.

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

J. Blanchet, Y. Kang, K. Murthy, and F. Zhang. Data-driven optimal transport cost selection for distributionally robust optimization. In *2019 Winter Simulation Conference (WSC)*, pages 3740–3751. IEEE, 2019.

K. A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, New York, 1989.

S. Bongers, J. Peters, B. Schölkopf, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *ArXiv e-prints (1611.06221v3)*, 2016.

R. J. Bowden and D. A. Turkington. *Instrumental Variables*. Econometric Society Monographs. Cambridge University Press, 1985.

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

V. Butsic, M. Baumann, A. Shortland, S. Walker, and T. Kuemmerle. Conservation and conflict in the Democratic Republic of Congo: The impacts of warfare, mining, and protected areas on deforestation. *Biological conservation*, 191:266–273, 2015.

K. M. Carlson, L. M. Curran, G. P. Asner, A. M. Pittman, S. N. Trigg, and J. M. Adeney. Carbon emissions from forest conversion by Kalimantan oil palm plantations. *Nature Climate Change*, 3 (3):283–287, 2013.

A. Castro-Nunez, O. Mertz, A. Buritica, C. C. Sosa, and S. T. Lee. Land related grievances shape tropical forest-cover in areas affected by armed-conflict. *Applied Geography*, 85:39–50, 2017.

M. A. Chaplain, G. Singh, and J. C. McLachlan. *On growth and form: spatio-temporal pattern formation in biology*. Wiley, 1999.

H. Chen, Y. Wang, R. Li, and K. Shear. A note on a nonparametric regression test through penalized splines. *Statistica Sinica*, 24: 1143, 2014.

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.

R. Christiansen and J. Peters. Switching regression models and causal inference in the presence of discrete latent variables. *Journal of Machine Learning Research*, 21(41), 2020.

R. Christiansen, M. Baumann, T. Kuemmerle, M. D. Mahecha, and J. Peters. Towards causal inference for spatio-temporal data: Conflict and forest loss in Colombia. *arXiv preprint arXiv:2005.08639*, 2020a.

R. Christiansen, N. Pfister, M. E. Jakobsen, N. Gnecco, and J. Peters. The difficult task of distribution generalization in nonlinear models. *arXiv preprint arXiv:2006.07433*, 2020b.

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.

N. Clerici, D. Armenteras, P. Kareiva, R. Botero, J. Ramírez-Delgado, G. Forero-Medina, J. Ochoa, C. Pedraza, L. Schneider, C. Lora, et al. Deforestation in Colombian protected areas increased during post-conflict periods. *Scientific Reports*, 10(1):1–10, 2020.

P. Collier et al. Economic causes of civil conflict and their implications for policy. Technical Report 76632, The World Bank, Washington, D.C., United States, 2000.

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

H. Cramér and H. Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1(4):290–294, 1936.

N. Cressie and C. K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

M. Croicu and R. Sundberg. UCDP georeferenced event dataset codebook version 4.0. *Journal of Peace Research*, 50(4):523–532, 2015.

G. Csurka. Domain adaptation for visual applications: A comprehensive survey. In G. Csurka, editor, *Domain Adaptation in Computer Vision Applications*. Springer, 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.

S. Darolles, Y. Fan, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

L. M. Dávalos, K. M. Sanchez, and D. Armenteras. Deforestation and coca cultivation rooted in twentieth-century development projects. *BioScience*, 66(11):974–982, 2016.

S. B. David, T. Lu, T. Luu, and D. Pal. Impossibility theorems for domain adaptation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 129–136. PMLR, 2010.

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

R. S. DeFries, T. Rudel, M. Uriarte, and M. Hansen. Deforestation driven by urban population growth and agricultural trade in the twenty-first century. *Nature Geoscience*, 3(3):178–181, 2010.

J. Duchi, P. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song. Robust physical-world attacks on deep learning models. *ArXiv e-prints (1707.08945)*, 2017.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: models, methods and applications*. Springer Science & Business Media, 2013.

F. M. Fisher. *The identification problem in econometrics*. McGraw-Hill, New York, NY, USA, 1966.

W. A. Fuller. Some properties of a modification of the limited information estimator. *Econometrica: Journal of the Econometric Society*, pages 939–953, 1977.

K. M. Gaynor, K. J. Fiorella, G. H. Gregory, D. J. Kurz, K. L. Seto, L. S. Withey, and J. S. Brashares. War and wildlife: linking armed conflict to conservation. *Frontiers in Ecology and the Environment*, 14(10):533–542, 2016.

A. E. Gelfand, H.-J. Kim, C. Sirmans, and S. Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003.

E. Giorgi, P. J. Diggle, R. W. Snow, and A. M. Noor. Geostatistical methods for disease mapping and visualisation using data from spatio-temporally referenced prevalence surveys. *International Statistical Review*, 86(3):571–597, 2018.

G. Giuliani, G. Camara, B. Killough, and S. Minchin. Earth observation open science: Enhancing reproducible science using data cubes, 2019.

N. P. Gleditsch, H. Strand, M. Eriksson, M. Sollenberg, and P. Wallensteen. Armed conflict 1946–2001: A new dataset. *Journal of Peace Research*, 39:615–637, 2002.

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.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ArXiv e-prints (1412.6572)*, 2014.

V. Gorsevski, E. Kasischke, J. Dempewolf, T. Loboda, and F. Grossmann. Analysis of the impacts of armed conflict on the Eastern Afromontane forest region on the South Sudan—Uganda border using multitemporal Landsat imagery. *Remote Sensing of Environment*, 118:10–20, 2012.

C. W. Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.

W. H. Greene. *Econometric analysis*. Pearson Education, Upper Saddle River, New Jersey, USA, 2003.

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. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu. A survey of learning causality with data: Problems and methods. *arXiv preprint arXiv:1809.09337*, 2018.

T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.

A. R. Hall. *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press, 2005.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. Stehman, S. J. Goetz, T. R. Loveland, et al. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853, 2013.

A. Harrison. Blood timber: how Europe helped fund war in the Central African Republic: a report. `https://apo.org.au/node/55984`, 2015. Accessed: 2020-05-15.

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.

C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *ArXiv e-prints (1710.11469)*, 2017.

C. Hof, M. B. Araújo, W. Jetz, and C. Rahbek. Additive threats from pathogens, climate and land-use change for global amphibian diversity. *Nature*, 480(7378):516–519, 2011.

S. Holly, M. H. Pesaran, and T. Yamagata. A spatio-temporal model of house prices in the USA. *Journal of Econometrics*, 158(1):160–173, 2010.

J. L. Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394, 2011.

P. Hoyer, S. Shimizu, A. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008.

Z. Hu and L. J. Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

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.

L. C. Irland. State failure, corruption, and warfare: challenges for forest policy. *Journal of Sustainable Forestry*, 27(3):189–223, 2008.

M. E. Jakobsen and J. Peters. Distributional robustness of K-class estimators and the PULSE. *arXiv preprint arXiv:2005.03353*, 2020.

D. Janzing, J. Peters, J. M. Mooij, and B. Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 249–257. AUAI Press, 2009.

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.

D. W. Jorgenson and J.-J. Laffont. Efficient estimation of nonlinear simultaneous equations with additive disturbances. In *Annals of Economic and Social Measurement, Volume 3, number 4*, pages 615–640. National Bureau of Economic Research, Inc, 1974.

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.

H. H. Kelejian. Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association*, 66(334):373–374, 1971.

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.

H. Kreft and W. Jetz. Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences*, 104(14):5925–5930, 2007.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

E. F. Lambin and P. Meyfroidt. Global land use change, economic globalization, and the looming land scarcity. *Proceedings of the National Academy of Sciences*, 108(9):3465–3472, 2011.

D. M. Landholm, P. Pradhan, and J. P. Kropp. Diverging forest land use dynamics induced by armed conflict across the tropics. *Global Environmental Change*, 56:86–94, 2019.

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

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

P. Le Billon. The political ecology of transition in Cambodia 1989–1999: war, peace and forest exploitation. *Development and change*, 31(4):785–805, 2000.

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.

B. Liu, B. Huang, and W. Zhang. *Spatio-temporal analysis and optimization of land use/cover change: Shenzhen as a case study.* CRC Press, 2017.

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.

A. C. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 587–596, 2009.

Q. Luo, W. Lu, W. Cheng, P. A. Valdes-Sosa, X. Wen, M. Ding, and J. Feng. Spatio-temporal Granger causality: A new framework. *NeuroImage*, 79:241–263, 2013.

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.

S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 10846–10856. Curran Associates, Inc., 2018.

M. D. Mahecha, F. Gans, G. Brandt, R. Christiansen, S. E. Cornell, N. Fomferra, G. Kraemer, J. Peters, P. Bodesheim, G. Camps-Valls, et al. Earth system data cubes unravel global multivariate dynamics. *Earth System Dynamics Discussions*, 11:201–234, 2020.

R. S. Mariano. Simultaneous equation model estimators: Statistical properties and practical implications. In B. H. Baltagi, editor, *A*

Companion to Theoretical Econometrics, chapter 6, pages 122–43. Blackwell Publishing Ltd, NJ, USA, 2001.

D. Martini, J. Pacheco-Labrador, O. Perez-Priego, C. van der Tol, T. S. El-Madany, T. Julitta, M. Rossini, M. Reichstein, R. Christiansen, U. Rascher, et al. Nitrogen and phosphorus effect on sun-induced fluorescence and gross primary productivity in mediterranean grassland. *Remote Sensing*, 11(21):2562, 2019.

N. Meinshausen. Causality from a distributional robustness point of view. In *IEEE Data Science Workshop*, pages 6–10, 2018.

N. Meinshausen and P. Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

J.-M. Montero, G. Fernández-Avilés, and J. Mateu. *Spatial and spatio-temporal geostatistical modeling and kriging*, volume 998. John Wiley & Sons, 2015.

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.

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

S. Nativi, P. Mazzetti, and M. Craglia. A view-based model of data-cube to support big earth data systems interoperability. *Big Earth Data*, 1(1-2):75–99, 2017.

P. J. Negret, J. Allan, A. Braczkowski, M. Maron, and J. E. Watson. Need for conservation planning in postconflict Colombia. *Conservation Biology*, 31, 2017.

P. J. Negret, L. Sonter, J. E. Watson, H. P. Possingham, K. R. Jones, C. Suarez, J. M. Ochoa-Quintero, and M. Maron. Emerging evidence that armed conflict and coca cultivation influence deforestation patterns. *Biological Conservation*, 239:108176, 2019.

W. K. Newey. Nonparametric instrumental variables estimation. *American Economic Review*, 103(3):550–56, 2013.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111 – 2245. Elsevier, 1994.

W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

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.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345 – 1359, 2010.

J. M. Pavía, B. Larraz, and J. M. Montero. Election forecasts using spatiotemporal models. *Journal of the American Statistical Association*, 103(483):1050–1059, 2008.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, USA, 2nd edition, 2009.

T. R. Pearson, S. Brown, and F. M. Casarim. Carbon emissions from tropical forest degradation caused by logging. *Environmental Research Letters*, 9(3):034017, 2014.

C. S. Peirce. A theory of probable inference. In C. S. Peirce, editor, *Studies in logic by members of the Johns Hopkins University*, pages 126–181. Little, Brown and Co, 1883.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

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.

T. Pettersson and P. Wallensteen. Armed conflicts, 1946–2014. *Journal of peace research*, 52(4):536–550, 2015.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

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.

N. Pfister, E. G. Williams, J. Peters, R. Aebersold, and P. Bühlmann. Stabilizing variable selection and regression. *arXiv preprint arXiv:1911.01850*, 2019c.

M. Prem, S. Saavedra, and J. F. Vargas. End-of-conflict deforestation: Evidence from Colombia's peace agreement. *World Development*, 129:104852, 2020.

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

J. S. Racine and T. Hayfield. *np: Nonparametric Kernel Smoothing Methods for Mixed Data Types*, 2018. URL `https://CRAN.R-project.org/package=np`. R package version 0.60–10.

T. Richardson, P. Spirtes, et al. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

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.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Causal transfer in machine learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

A. Rønn-Nielsen and A. Sokol. Advanced probability. `http://web.math.ku.dk/noter/filer/vidsand12.pdf`, 2013. Accessed: 2020-05-15.

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

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

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

A. M. Sánchez-Cuervo and T. M. Aide. Consequences of the armed conflict, forced human displacement, and land abandonment on forest cover change in Colombia: A multi-scaled analysis. *Ecosystems*, 16(6):1052–1070, 2013.

A. M. Sánchez-Cuervo, T. M. Aide, M. L. Clark, and A. Etter. Land cover change in Colombia: surprising forest recovery trends between 2001 and 2010. *PloS one*, 7(8), 2012.

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

N. Sani, J. Lee, and I. Shpitser. Identification and estimation of causal effects defined by shift interventions. In *Proceedings of the 36th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2020.

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.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262. Omnipress, 2012.

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.

M. Sherman. *Spatial statistics and spatio-temporal data: covariance functions and directional properties*. John Wiley & Sons, 2011.

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.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

I. Shpitser, T. VanderWeele, and J. M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*, pages 527–536, Dec 2010. ISBN 9780974903965.

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 Learning Research*, 7:191–246, 2006.

A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. *ArXiv e-prints (1710.10571)*, 2017.

L. J. Sonter, D. Herrera, D. J. Barrett, G. L. Galford, C. J. Moran, and B. S. Soares-Filho. Mining drives extensive deforestation in the Brazilian Amazon. *Nature Communications*, 8(1):1–7, 2017.

P. Spirtes and K. Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, page 3. Springer, 2016.

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.

H. Theil. Repeated least squares applied to complete equation systems. *The Hague: central planning bureau*, 1953.

H. Theil. *Economic forecasts and policy*. North-Holland, Amsterdam, Netherlands, 1958.

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.

T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-88)*, pages 352–359, Corvallis, Oregon, 1988. AUAI Press.

R. Volpi, P. Morerio, S. Savarese, and V. Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5495–5504, 2018.

E. J. Ward, J. E. Jannot, Y.-W. Lee, K. Ono, A. O. Shelton, and J. T. Thorson. Using spatiotemporal species distribution models to identify temporally evolving hotspots of species co-occurrence. *Ecological Applications*, 25(8):2198–2209, 2015.

N. Wiener. The theory of prediction. In E. Beckenbach, editor, *Modern Mathematics for Engineers*. McGraw-Hill, New York, NY, 1956.

C. K. Wikle, A. Zammit-Mangion, and N. Cressie. *Spatio-temporal Statistics with R*. CRC Press, 2019.

D. Williams. *Probability with martingales*. Cambridge university press, 1991.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, Cambridge, MA, 2010.

A. Zammit-Mangion, J. Rougier, N. Schön, F. Lindgren, and J. Bamber. Multivariate spatio-temporal modelling for assessing Antarctica's present-day contribution to sea-level rise. *Environmetrics*, 26(3):159–177, 2015.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Corvallis, Oregon, USA, 2009. AUAI Press.

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.

J. Y. Zhu, C. Sun, and V. O. Li. An extended spatio-temporal Granger causality model for air quality estimation with heterogeneous urban big data. *IEEE Transactions on Big Data*, 3(3): 307–319, 2017.

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.