Structural restrictions in local causal discovery: identifying direct causes of a target variable

Juraj Bodik[1*], Valérie Chavez-Demoulin[1]

[1] HEC, Université de Lausanne, Switzerland

Abstract

We consider the problem of learning a set of direct causes of a target variable from an observational joint distribution. Learning directed acyclic graphs (DAGs) that represent the causal structure is a fundamental problem in science. Several results are known when the full DAG is identifiable from the distribution, such as assuming a nonlinear Gaussian data-generating process. Often, we are only interested in identifying the direct causes of one target variable (local causal structure), not the full DAG. In this paper, we discuss different assumptions for the data-generating process of the target variable under which the set of direct causes is identifiable from the distribution. While doing so, we put essentially no assumptions on the variables other than the target variable. In addition to the novel identifiability results, we provide two practical algorithms for estimating the direct causes from a finite random sample and show their effectiveness on several benchmark data-sets. We apply this framework to learn direct causes of the reduction in the fertility rate in different countries.

Keywords— Causal inference, Identifiability, Target variable, Local causal discovery, Structural causal models

[*]Email of the corresponding author: Juraj.Bodik@unil.ch

1 Introduction

Causal reasoning holds great significance in numerous fields, such as public policy and decision-making or in medicine (Pearl, 2009). Randomized control experiments are widely accepted as the gold standard method for determining causal relationships (Imbens and Rubin, 2015). However, the feasibility of such experiments is often hindered by their high costs and ethical concerns. Therefore, it becomes crucial to estimate causal relations from observational data, that are obtained by observing a system without any interventions (Peters et al., 2017).

In this paper, we deal with the problem of estimating a set of direct causes of a target variable from a random sample. Typical research focus lies in estimating the full causal structure, while we are interested only in a local causal structure around one variable of interest. The main issue arises when several causal structures produce the same observed distribution; the set of direct causes can be unidentifiable. We can generally estimate only the Markov equivalence class (MEC).

A variety of research papers propose a methodology to deal with unidentifiable structures. These methods are either structural-restriction-based, that is, we add some additional assumptions about the functional relations between the variables, such as assuming nonlinear Gaussian data-generating process (Hoyer et al., 2008; Peters et al., 2014; Mooij et al., 2016; Peters et al., 2017; Immer et al., 2022; Bodik and Chavez-Demoulin, 2023); or score-based, that is, we pick a causal structure with the best fit on the data according to some score function (Chickering, 2002; Nowzohour and Bühlmann, 2016); or information-theory-based, using mutual information and approximations of Kolmogorov complexity (Janzing and Schölkopf, 2010; Marx and Vreeken, 2017; Tagasovska et al., 2020). However, these methods were designed either for a bivariate case or to infer the entire causal structure of the system.

Several methodologies have been proposed to infer a local structure around a target variable. These methods are typically divided into three categories: learning a local skeleton (unoriented graph), learning a minimal Markov blanket (a sufficient set), or learning a set of direct causes of the target variable (goal of this paper). Under causal sufficiency and faithfulness (Pearl, 2009), the PC algorithm (Spirtes et al., 2001) can identify the MEC and consistently learn the skeleton of the full structure. Yin et al. (2008), Aliferis et al. (2010), Wang et al. (2014) discuss modifications of the PC algorithm focusing only on the local structure. Gao and Aragam (2021) suggest a methodology for estimating the minimal Markov blanket based on comparing entropies of the variables. Azadkia et al. (2021) propose a method to learn the direct causes of the target variable under the assumption that the causes are identifiable (specifically, assuming that the underlying structure is a polytree). In contrast, our work aims to distinguish between different local causal structures (local MEC) using an structural-restrictions-based approach, where we take the ideas from classical approaches and use them locally.

In the following, we present the main ideas of the paper. The theory is based on a structural causal model (SCM Verma and Pearl (2013)) where a target variable Y is structurally generated as $Y = f_Y(\mathbf{X}_{pa_Y}, \varepsilon_Y)$, where $\mathbf{X} = (X_1, \dots, X_p)^\top$ are other variables in the system and $pa_Y \subseteq \{1, \dots, p\}$ are called parents (or direct causes) of Y, and $\varepsilon_Y \perp\!\!\!\perp \mathbf{X}_{pa_Y}$ is a noise variable. Our goal is to estimate the set pa_Y from a random sample from $(Y, \mathbf{X})$. Following the structural-restrictions-based ideology, we assume $f_Y \in \mathcal{F}$, where $\mathcal{F}$ is a subset of all measurable functions (for example, all linear functions).

Throughout the paper, we restrict the class of functions $\mathcal{F}$ in the following way. We assume that f_Y is invertible (notation $f_Y \in \mathcal{I}$) in the sense that

$$\text{there exists a function } f_Y^{\leftarrow} \text{ such that } \varepsilon_Y = f_Y^{\leftarrow}(\mathbf{X}_{pa_Y}, Y).$$

In other words, the noise variables can be recovered from the observed variables. Moreover, we assume that $f_Y \in \mathcal{I}_m$, where

$$\mathcal{I}_m = \{f \in \mathcal{I} : \ f \text{ is not constant in any of its arguments}\}.$$

This assumption is closely related to causal minimality (the subscript m in $\mathcal{I}_m$ represents the word "minimality"). For more details and a rigorous definition of the class $\mathcal{I}_m$, see Appendix A.1. Overall, we assume that $\mathcal{F} \subseteq \mathcal{I}_m$.

Our framework is built on a notion of $\mathcal{F}$-identifiability. Without loss of generality, we assume $\varepsilon_Y \sim U(0,1)$[1].

Definition 1. *A non-empty set $S \subseteq \{1,\dots,p\}$ is called an $\mathcal{F}$-**plausible** set of parents of Y, if*

$$\text{there exists } f \in \mathcal{F} \text{ such that for } \varepsilon_S := f^{\leftarrow}(\boldsymbol{X}_S, Y) \text{ holds } \varepsilon_S \perp\!\!\!\perp \boldsymbol{X}_S,\ \ \varepsilon_S \sim U(0,1). \tag{1}$$

*We define a set of $\mathcal{F}$-**identifiable** parents of Y as follows:*

$$S_{\mathcal{F}}(Y) := \bigcap_{S \subseteq \{1,\dots,p\}} \{S :\ S \text{ is } \mathcal{F}\text{-plausible set of parents of } Y\}.$$

The functional space $\mathcal{F}$ corresponds to our assumptions that we are willing to make about the data-generating process of Y. If we assume linearity of the covariates, this represents the assumption $f_Y \in \mathcal{F}_L$, where (recall that without loss of generality, we assume $\varepsilon_Y \sim U(0,1)$)

$$\mathcal{F}_L = \{f \in \mathcal{I}_m : f(\mathbf{x}, \varepsilon) = \beta^T \mathbf{x} + q^{-1}(\varepsilon) \text{ for some quantile function } q^{-1} \text{ and } \beta \in \mathbb{R}^{|\mathbf{x}|}\}. \tag{2}$$

Note that the restriction $f \in \mathcal{I}_m$ in (2) implies that the arguments $\beta_i \neq 0$. On the other hand, if we assume a Conditionally Parametric Causal Model ($CPCM(F)$, see (5) in Section 2), this corresponds to assuming $f_Y \in \mathcal{F}_F$ where

$$\mathcal{F}_F := \{f \in \mathcal{I}_m : f(\mathbf{x}, \varepsilon) = F^{-1}(\varepsilon; \theta(\mathbf{x})) \text{ for some function } \theta\}. \tag{3}$$

Table 1 describes all functional spaces considered in this paper.

The concept of $\mathcal{F}$-identifiability provides theoretical limitations for the causal estimates under assumption $f_Y \in \mathcal{F}$. If $S_{\mathcal{F}}(Y)$ contains one element, we can only identify one cause of Y, even if we observe an infinite number of observations. The main part of the paper consists of inferring which elements belong to $S_{\mathcal{F}}(Y)$: When does it hold that $S_{\mathcal{F}}(Y) = pa_Y$?

From a practical point of view, we propose two algorithms for estimating the direct causes of a target variable from a random sample. One provides an estimate of $S_{\mathcal{F}}(Y)$, that is, it tests the $\mathcal{F}$-plausibility of several sets and outputs their intersection. This shields us against the mistake of including a non-parent in the output. However, the output does not have to contain all direct causes. The second is a score-based algorithm estimating pa_Y based on a goodness-of-fit; even if several sets are $\mathcal{F}$-plausible, the output is the set with the best score. The first algorithm has strong theoretical guarantees for containing only the direct causes of Y. In contrast, the score-based algorithm outputs the "best looking" set of direct causes without theoretical guarantees for the output.

We primarily focus on the case when $\mathbf{X} = (X_1, \dots, X_p)$ are neighbors (either direct causes or direct effects) of Y in the corresponding SCM. Using the classical conditional independence approach and d-separation (see Section 2), we can eliminate other variables from being potential parents of Y. Nevertheless, the theory can be extended to non-neighbors as well.

In the following example, we provide an initial assessment of the findings presented in Section 3.

Example 1 (Teaser example with Gaussian assumptions)**.** *Consider an SCM defined as follows (see Figure 1): let (X_1, X_2, X_3) be normally distributed, and X_5 be non-degenerate. Let*

$$Y = \mu_Y(X_1, X_2, X_3) + \sigma_Y(X_1, X_2, X_3)\varepsilon_Y,\ \ \varepsilon_Y \text{ is Gaussian}, \quad \varepsilon_Y \perp\!\!\!\perp (X_1, X_2, X_3),$$

$$X_4 = \mu_4(X_2, Y, X_5) + \sigma_4(X_2, Y, X_5)\varepsilon_4,\ \ \varepsilon_4 \text{ is Gaussian}, \qquad \varepsilon_4 \perp\!\!\!\perp (X_1, X_2, X_3, Y),$$

where $(\mu_4, \sigma_4)^\top$ and $(\mu_Y, \sigma_Y)^\top$ are real functions that satisfy some (weak) assumptions presented in Section 3. In particular, $(\mu_4, \sigma_4)^\top$ are not in the form (6) and μ_Y is not additive (or σ_Y is not multiplicative) in its arguments.

Then, $S_{\mathcal{F}_F}(Y) = S_{\mathcal{F}_{LS}}(Y) = pa_Y = \{1,2,3\}$, where F is a Gaussian distribution function. This result follows from the theory presented in Section 3; in particular, it is a consequence of Lemma 4 in combination with Lemma 3, Consequence 1 and results in Example 2 combined with Proposition 1.

[1]To understand why this assumption is without loss of generality, consider the equality $f_Y(\mathbf{X}_{pa_Y}, \varepsilon) = f_Y(\mathbf{X}_{pa_Y}, q^{-1}(\varepsilon_Y))$, where $\varepsilon_Y \sim U(0,1)$ and q is a distribution function of ε. We define $\tilde{f}_Y(\mathbf{X}_{pa_Y}, \varepsilon_Y) = f_Y(\mathbf{X}_{pa_Y}, q^{-1}(\varepsilon_Y))$ and only work with $\tilde{f}_Y$.

Summary of different $\mathcal{F} \subset \mathcal{I}_m$ used in the paper
$\mathcal{F}_L = \{f \in \mathcal{I}_m : f(\mathbf{x}, \varepsilon) = \beta^T\mathbf{x} + q^{-1}(\varepsilon)$ for some quantile function q^{-1} and $\beta \neq 0\}$
$\mathcal{F}_A = \{f \in \mathcal{I}_m : f(\mathbf{x}, \varepsilon) = \mu(\mathbf{x}) + g^{-1}(\varepsilon)$ for some $\mu(\cdot)$ and quantile function $q^{-1}\}$
$\mathcal{F}_{LS} = \{f \in \mathcal{I}_m : f(\mathbf{x}, \varepsilon) = \mu(\mathbf{x}) + \sigma(\mathbf{x})q^{-1}(\varepsilon)$ for some functions $\mu, \sigma > 0$ and for some quantile function $q^{-1}\}$
$\mathcal{F}_F := \{f \in \mathcal{I}_m : f(\mathbf{x}, \varepsilon) = F^{-1}(\varepsilon; \theta(\mathbf{x}))$ for some function $\theta : \mathbb{R}^{\lvert\mathbf{x}\rvert} \to \mathbb{R}^q\}$

Table 1: The table summarizes different functional spaces $\mathcal{F}$ used in the paper. $\mathcal{F}_L$, $\mathcal{F}_A$, $\mathcal{F}_{LS}$ and $\mathcal{F}_F$ correspond to linearity assumption, additivity assumption, location-scale assumption and $CPCM(F)$ assumption, respectively.

This paper is structured as follows: Section 2 provides a standard notation and classical preliminary results from a theory of graphical models. Moreover, we briefly summarize causal discovery methods from the literature that are based on structural restrictions. In Section 3, we dive deeper into mathematical properties of $S_{\mathcal{F}}(Y)$, where the aim is to find conditions under which $S_{\mathcal{F}}(Y) = pa_Y$. In Section 4, we describe our proposed algorithms for estimating $S_{\mathcal{F}}(Y)$ and pa_Y from a random sample. Section 5 contains a short simulation study followed by an application on a real dataset. Finally, the paper has four appendices: Appendix A contains some detailed notions omitted from the main text for clarity, Appendix C includes the proofs of the theorems, Appendix B provides some auxiliary results needed for the proofs (especially Lemma B.2 is the core mathematical result of the paper) and Appendix D contains some details about the simulations and the application.

2 Preliminaries, notation and existing structural-restrictions-based methods

First, we introduce the graphical causal notation (e.g., Spirtes et al. (2001)). A DAG (directed acyclic graph) $\mathcal{G} = (V, E)$ contains a finite set of vertices (nodes) V and a set of directed edges E between distinct vertices, such that there exists no directed cycle or multiple edges. For a distinct pair $i, j \in V$ with an edge $i \to j$, we say that i is a **parent** of j and j is a **child** of i, notation $i \in pa_j(\mathcal{G})$ and $j \in ch_j(\mathcal{G})$, respectively. If there is a directed path from i to j, we say that i is an **ancestor** of j and j is a **descendant** of i, notation $i \in an_j(\mathcal{G})$ and $j \in de_i(\mathcal{G})$, respectively. The skeleton of $\mathcal{G}$ is the undirected graph obtained from $\mathcal{G}$ by replacing directed edges with undirected edges. Nodes $i, j, k \in V$ form a **v-structure** if i and j point to k, and there is no direct edge between i and j. In that case, k is called an unshielded **collider** and i, j are spouses, notation $i \in sp_j(\mathcal{G})$ (and $j \in sp_i(\mathcal{G})$). Even if i and j have an edge connecting them, node k is called a collider. We say that the node $i \in V$ is a **source** node if $pa_j(\mathcal{G}) = \emptyset$, notation $i \in Source(\mathcal{G})$. We omit the argument $\mathcal{G}$ if evident from the context.

Consider a random vector $(X_i)_{i \in V}$ over a probability space $(\Omega, \mathcal{A}, P)$. With a slight abuse of notation, we identify the vertices $j \in V$ with the variables X_j. We denote $\mathbf{X}_S = \{X_s : s \in S\}$ for $S \subseteq V$. Typically, we denote $V = \{0, \dots, p\}$, where X_0 (usually denoted by Y) is a target variable and $\mathbf{X} = (X_1, \dots, X_p)^\top$ are other variables (covariates). We assume that $(Y, \mathbf{X})$ follow a

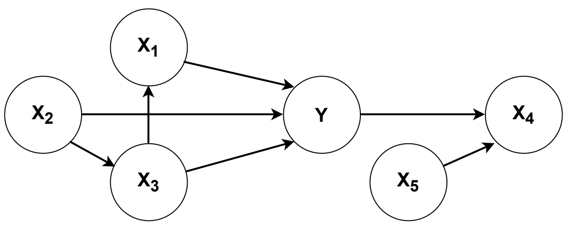

Figure 1: DAG corresponding to Example 1.

structural causal model (SCM) with a DAG $\mathcal{G}$ (Pearl, 2009), that is, each variable arises from a structural equation

$$X_i = f_i(\mathbf{X}_{pa_i}, \eta_i), \quad i = 0, 1, \dots, p, \tag{4}$$

where η_i are jointly independent random variables [2]. The measurable functions f_i are the assignments (or link functions), and we say that X_j is a direct cause of X_i if $j \in pa_i$.

A joint distribution $P_{\mathbf{X}}$ over $\mathbf{X}$ satisfies a **Markov property** with respect to $\mathcal{G}$ if $A \perp\!\!\!\perp_{\mathcal{G}} B \mid C \implies A \perp\!\!\!\perp B \mid C$ for $A, B, C \subseteq V$ are disjoint subsets of the vertices and $\perp\!\!\!\perp_{\mathcal{G}}$ represent a d-separation in $\mathcal{G}$ (Verma and Pearl, 2013). On the other hand, if for all $A, B, C \subseteq V$ disjoint subsets of the vertices hold $A \perp\!\!\!\perp_{\mathcal{G}} B \mid C \impliedby A \perp\!\!\!\perp B \mid C$, we say that the distribution is **faithful** with respect to $\mathcal{G}$. $P_{\mathbf{X}}$ satisfies **causal minimality** with respect to $\mathcal{G}$ if it is Markov with respect to $\mathcal{G}$, but not to any proper subgraph of $\mathcal{G}$. It can be shown that faithfulness implies causal minimality (Peters et al., 2017, Proposition 6.35). Two graphs $\mathcal{G}_1, \mathcal{G}_2$ are Markov equivalent if the set of distributions that are Markov with respect to $\mathcal{G}_1, \mathcal{G}_2$ is the same. We denote the **Markov equivalency class** of $\mathcal{G}$ as the set $MEC(\mathcal{G}) := \{\mathcal{G}' : \mathcal{G} \text{ and } \mathcal{G}' \text{ are Markov equivalent}\}$. Two graphs are Markov equivalent if they have the same skeleton and the same v-structures (Verma and Pearl, 2013).

A set of variables $\mathbf{X}$ is said to be causally **sufficient** if there is no hidden common cause that causes more than one variable in $\mathbf{X}$ (Spirtes, 2010). A set $S \subseteq V$ is called **Markov blanket** for X_i if $X_i \perp\!\!\!\perp X_{V \setminus (S \cup \{i\})} \mid \mathbf{X}_S$. A **Markov boundary** is a minimal Markov blanket, denoted by $MB_i(\mathcal{G})$. Note that $MB_i(\mathcal{G}) = pa_i(\mathcal{G}) \cup ch_i(\mathcal{G}) \cup sp_i(\mathcal{G})$.

For a $S \subseteq V$ we define a **projection** of graph $\mathcal{G}$ on S, denoted as $\mathcal{G}[S]$, as a graph with vertices S and the following edges: for distinct $i, j \in S$, there is an edge $i \to j$ in $\mathcal{G}[S]$ if and only if there is a directed path from i to j in $\mathcal{G}$ such that all vertices on this path except i, j do not belong to S ($i \to \cdots \to j$). Moreover, there is a bidirected edge $i - j$ if there exists a path between i, j in $\mathcal{G}$ that does not contain a collider and the first edge points towards i and the last edge points towards j ($i \leftarrow \cdots \to j$).

Let $(X_0, \mathbf{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an SCM (4) with DAG $\mathcal{G}_0$. Let $\mathcal{F}$ be a subset of all measurable functions. We say that the SCM follows an **$\mathcal{F}$-model**, if each structural equation in the SCM satisfies $f_i \in \mathcal{F}, i = 0, \dots, p$. We say that $\mathcal{G}_0$ is **identifiable** from the joint distribution under the $\mathcal{F}$-model (we also say that the causal model is identifiable, or that there does not exist a backward model) if there does not exist a graph $\mathcal{G}' \neq \mathcal{G}_0$ and functions $f_i' \in \mathcal{F}, i = 0, \dots, p$ generating the same joint distribution.

In the following, we discuss some well-known results from the literature that address the problem of identifiability and estimation of the causal structure; for a review see Glymour et al. (2019).

Shimizu et al. (2006) show that $\mathcal{G}_0$ is identifiable under the LiNGaM model (Linear Non-Gaussian additive Models where $\mathcal{F}$ consists of all linear functions and the noise variables are non-Gaussian). Hoyer et al. (2008) and Peters et al. (2014) develop a framework for ANM (additive noise models where $\mathcal{F}$ consists of functions additive in the last input, that is, $X_i = g(\mathbf{X}_{pa_i}) + \eta_i$). Under some weak assumptions on g, $\mathcal{G}_0$ is identifiable (Peters et al., 2014, Corollary 31) and the authors propose an algorithm for estimating $\mathcal{G}_0$ (for a review on ANM, see Mooij et al. (2016)). Zhang and Hyvärinen (2009) consider the PNL model (Post-Nonlinear Causal Model) where $\mathcal{F}$ consists of post-additive functions, that is,

$$X_i = g_1\big(g_2(\mathbf{X}_{pa_i}) + \eta_i\big)$$

with an invertible link function g_1. Note that the former two are special cases of the PNL model. However, the PNL model is identifiable under some technical assumptions and with several exceptions (Zhang and Hyvärinen, 2010). Park and Raskutti (2017) show identifiability in models where $var[X_i \mid \mathbf{X}_{pa_i}]$ is a quadratic function of $\mathbb{E}[X_i \mid \mathbf{X}_{pa_i}]$. Galanti et al. (2020) consider the neural SCM with representation $X_i = g_1\big(g_2(\mathbf{X}_{pa_i}), \eta_i\big)$, where g_1, g_2 are assumed to be neural networks. Khemakhem et al. (2021), Immer et al. (2022) and Strobl and Lasko (2022) propose several

[2] We distinguish between uniformly distributed noise variables (denoted by ε_i) and arbitrarily distributed noise variables (denoted by η_i). Under the relation $\varepsilon_i = q(\eta_i)$ where q is the distribution of η_i, these cases are equivalent and we use the former notation in the paper. However in Section 2, in order to be consistent with the notation used in the literature, we use the latter notation.

methods and identifiability results for location-scale models, where $\mathcal{F}$ consists of location-scale functions, that is, $X_i = g_1(\mathbf{X}_{pa_i}) + g_2(\mathbf{X}_{pa_i})\eta_i$.

Bodik and Chavez-Demoulin (2023) consider a class of $CPCM(F)$ (conditionally parametric causal models), where $X_i \mid \mathbf{X}_{pa_i}$ has the conditional distribution F with parameters $\theta_i(\mathbf{X}_{pa_i}) \in \mathbb{R}^q$ for some $q \in \mathbb{N}$. That is, the data-generating process is of the form

$$X_i = f_i(\mathbf{X}_{pa_i}, \varepsilon_i) = F^{-1}\big(\varepsilon_i; \theta_i(\mathbf{X}_{pa_i})\big), \quad \text{equivalently } X_i \mid \mathbf{X}_{pa_i} \sim F\big(\theta_i(\mathbf{X}_{pa_i})\big), \tag{5}$$

where $\varepsilon_i \sim U(0,1)$, and F is a known distribution function with parameters θ_i being functions of the direct causes of X_i. This model represents a system where the direct causes only affect some characteristics of the distribution but not the class itself. For example, if F is Gaussian, then we are in the Gaussian Location-scale models framework.

Example 2 (Gaussian case)**.** *Suppose that $(X_0, \boldsymbol{X})$ follow $CPCM(F)$ with a Gaussian distribution function F. This corresponds to $X_i \mid \boldsymbol{X}_{pa_i} \sim N\big(\mu_i(\boldsymbol{X}_{pa_i}), \sigma_i^2(\boldsymbol{X}_{pa_i})\big)$ for all $i = 0, \dots, d$ and for some functions $\theta_i = (\mu_i, \sigma_i)^\top : \mathbb{R}^{|pa_i|} \to \mathbb{R} \times \mathbb{R}^+$. Using linearity of mean and variance in a Gaussian distribution, this is equivalent to assuming that the data-generating process has a form*

$$X_i = \mu_i(\boldsymbol{X}_{pa_i}) + \sigma_i(\boldsymbol{X}_{pa_i})\eta_i, \qquad \eta_i \textit{ is Gaussian.}$$

Potentially, source nodes can have arbitrary distributions.

Consequence 3 in Bodik and Chavez-Demoulin (2023) shows that $\mathcal{G}_0$ is identifiable unless μ_i, σ_i have a specific functional form. More precisely, $\mathcal{G}_0$ is identifiable if the arguments of μ_i, σ_i (denoted by $\mu_{i,j}, \sigma_{i,j}$, $j = 1, \dots, |pa_i|$) do not satisfy:

$$\frac{1}{\sigma_{i,j}^2(x)} = ax^2 + c, \qquad \frac{\mu_{i,j}(x)}{\sigma_{i,j}^2(x)} = d + ex, \qquad x \in \mathbb{R}, \tag{6}$$

where $a, d, e \in \mathbb{R}, c > 0$ are some constants.

Bodik and Chavez-Demoulin (2023) show that $\mathcal{G}_0$ is generally identifiable in $CPCM(F)$ for any F belonging to the exponential family of distributions, with an exception when θ is equal to the sufficient statistic of F.

Example 3 (Pareto case)**.** *Suppose that $(X_0, \boldsymbol{X})$ follow $CPCM(F)$ where F is the Pareto distribution function, that is, the corresponding density function has a form $p(y) = \frac{\theta}{y^{\theta+1}}, \theta > 0, y \geq 1$. This model corresponds to a data generating process where*

$$X_i \mid \boldsymbol{X}_{pa_i} \sim Pareto\big(\theta_i(\boldsymbol{X}_{pa_i})\big).$$

If $\theta(\boldsymbol{X}_{pa_i})$ is small, then the tail of X_i is large and extremes occur more frequently. Consequence 1 in Bodik and Chavez-Demoulin (2023) shows that $\mathcal{G}_0$ is identifiable if $\theta(\boldsymbol{X}_{pa_i})$ do not have a logarithmic form in any of its arguments.

3 Properties of $\mathcal{F}$-identifiable parents

Recall that we assume the data-generating process of Y in the form

$$Y = f_Y(\mathbf{X}_{pa_Y}, \varepsilon_Y), \quad f_Y \in \mathcal{F}, \quad \varepsilon_Y \perp\!\!\!\perp \mathbf{X}_{pa_Y}, \quad \varepsilon_Y \sim U(0,1). \tag{$\heartsuit$}$$

The principle of independence of the cause and the mechanism directly implies that the set $S = pa_Y$ is always $\mathcal{F}$-plausible; under ($\heartsuit$) it always holds that

$$S_{\mathcal{F}}(Y) \subseteq pa_Y. \tag{7}$$

However, the equality $S_{\mathcal{F}}(Y) = pa_Y$ does not need to hold. Observe that

$$\text{if } \mathcal{F}_1 \subseteq \mathcal{F}_2, \text{ then } S_{\mathcal{F}_1}(Y) \supseteq S_{\mathcal{F}_2}(Y).$$

This is not surprising as the more restrictions we put on the data-generating process, the larger the set of identifiable parents.

In this section, we discuss which elements belong to $S_{\mathcal{F}}(Y)$. We will find out that under linearity, we typically get $S_{\mathcal{F}_L}(Y) = \emptyset$, that is, if the link function f_Y is linear, we can not identify any parents of Y. However, if the link function f_Y is "ugly", it typically holds that $S_{\mathcal{F}}(Y) = pa_Y$ unless $\mathcal{F}$ contains too many functions.

In Section 3.1, our focus will be on the case $\mathcal{F} = \mathcal{F}_L$. In Section 3.2, we focus mostly on the case where $\mathcal{F} = \mathcal{F}_F$ for different distribution functions F. Consequence 1 and Proposition 2 in Section 3.2.1 discuss the modifications of these results for $\mathcal{F} = \mathcal{F}_A$ and $\mathcal{F} = \mathcal{F}_{LS}$.

Section 3.3 contains a discussion with several examples illustrating the results presented in Sections 3.1 and 3.2.

3.1 Case $S_{\mathcal{F}}(Y) = \emptyset$ and linear models

The following remark shows why the concept of $\mathcal{F}$-identifiable parents is usually not very interesting for linear SCM. Recall that we interchangeably use the notation $X_0 = Y$ for the target variable, and recall that an SCM follows an $\mathcal{F}$-model if each structural equation in the SCM satisfies $f_i \in \mathcal{F}, i = 0, \dots, p$. Following this definition, note that a statement "$(Y, \mathbf{X})$ follow a linear SCM" is equivalent to a statement "$(Y, \mathbf{X})$ follow an $\mathcal{F}_L$-model".

Remark 1. *Consider $\mathcal{F} = \mathcal{F}_L$, where $(Y, \boldsymbol{X})$ follow an $\mathcal{F}_L$-model with DAG drawn in Figure 2A and*

$$Y = \beta_1 X_1 + \beta_2 X_2 + q^{-1}(\varepsilon_Y), \qquad \varepsilon_Y \perp\!\!\!\perp (X_1, X_2),\ \varepsilon_Y \sim U(0,1),$$

for some $\beta_1 \neq 0, \beta_2 \neq 0$ and a quantile function q^{-1}. Then $S_{\mathcal{F}_L}(Y) = \emptyset$.

Proof. We show that the set $S = \{1\}$ is $\mathcal{F}_L$-plausible. Intuitively, this holds since $Y - \beta_1 X_1 \perp\!\!\!\perp X_1$, and we do not put any restrictions on the noise variable. More precisely, we find $f \in \mathcal{F}_L$ such that (1) holds. Such f can be defined as $f(x, \varepsilon) = \beta_1 x + \tilde{q}^{-1}(\varepsilon)$ for $x \in \mathbb{R}, \varepsilon \in (0,1)$, where $\tilde{q}$ is the distribution function of $[\beta_2 X_2 + q^{-1}(\varepsilon_Y)]$. Then trivially $f \in \mathcal{F}_L$ and its inverse satisfies $f^{\leftarrow}(x, y) = \tilde{q}(y - \beta_1 x)$ for $x, y \in \mathbb{R}$. Hence, $\varepsilon_S = f^{\leftarrow}(X_1, Y) = \tilde{q}(Y - \beta_1 X_1) = \tilde{q}[\beta_2 X_2 + q^{-1}(\varepsilon_Y)] \perp\!\!\!\perp X_1$ and $\varepsilon_S \sim U(0,1)$. We have shown that f satisfies (1).

With a similar reasoning we get that $S = \{2\}$ is also $\mathcal{F}_L$-plausible. Therefore, $S_{\mathcal{F}_L}(Y) \subseteq \{1\} \cap \{2\} = \emptyset$. $\square$

A similar argument can be used in a more general case. The following lemma describes that in the linear structural causal models, $S_{\mathcal{F}_L}(Y)$ is usually empty, even if $pa_Y(\mathcal{G}_0) \neq \emptyset$.

Lemma 1. *Let $(Y, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an $\mathcal{F}_L$-model with DAG $\mathcal{G}_0$ and $pa_Y(\mathcal{G}_0) \neq \emptyset$. Then, $|S_{\mathcal{F}_L}(Y)| \leq 1$. Moreover, if there exist $a, b \in an_Y(\mathcal{G}_0)$ that are d-separated in $\mathcal{G}_0$, then $S_{\mathcal{F}_L}(Y) = \emptyset$.*

The proof is in Appendix C. Lemma 1 shows a more general principle that goes beyond the linear models. If we can *marginalize* a causal model to a smaller submodel without breaking $f_Y \in \mathcal{F}$, then only the submodel is relevant for inference about $S_{\mathcal{F}}(Y)$. We provide a more rigorous explanation of this.

Definition 2. *Let $\mathcal{F} \subseteq \mathcal{I}_m$ and let $(X_0, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an $\mathcal{F}$-model with DAG $\mathcal{G}_0$. For a non-empty set $S \subseteq \{0, \dots, p\}$, we say that the $\mathcal{F}$-model is* ***marginalizable*** *to S if $\boldsymbol{X}_S$ can also be written as an $\mathcal{F}$-model with some underlying DAG $\mathcal{G}_S$.*

We provide some intuition about the marginalizability on a specific example.

Remark 2. *Consider the following $\mathcal{F}_L$-model with DAG drawn in Figure 2B: $X_1 = \eta_1, X_2 = X_1 + \eta_2, X_0 = X_1 + X_2 + \eta_0$, where the noise variables (η_1, η_2, η_0 are not necessary uniformly distributed) are jointly independent.*

- *If $\eta_1, \eta_2, \eta_0 \overset{iid}{\sim} N(0, \sigma^2)$, then the $\mathcal{F}_L$-model is marginalizable to $S = \{0,1\}$ and to $S = \{0,2\}$.*
- *If η_1, η_2, η_0 are not Gaussian, then the $\mathcal{F}_L$-model is marginalizable to $S = \{0,1\}$, but not to $S = \{0,2\}$.*

Proof. The marginalizability with respect to $S = \{0, 1\}$ follows from the fact that we can rewrite $X_1 = \eta_1, X_0 = 2X_1 + \tilde{\eta}_0$, where $\tilde{\eta}_Y = \eta_2 + \eta_0 \perp\!\!\!\perp \eta_1$. For $S = \{0, 2\}$ and Gaussian noise, we can rewrite $X_2 = \tilde{\eta}_2, X_0 = \frac{3}{2}X_2 + \tilde{\eta}_0$, where $\tilde{\eta}_2 = \eta_1 + \eta_2$ and $\tilde{\eta}_0 = \frac{1}{2}\eta_1 - \frac{1}{2}\eta_2 + \eta_0 \perp\!\!\!\perp \tilde{\eta}_2$.

For $S = \{0, 2\}$ and non-Gaussian noise, we can never find $X_2 = \tilde{\eta}_2, X_0 = \beta X_2 + \tilde{\eta}_0$, where $\tilde{\eta}_2 \perp\!\!\!\perp \tilde{\eta}_0$ and $\beta \neq 0$, since for non-Gaussian variables $a\eta_1 + b\eta_2 \not\perp\!\!\!\perp c\eta_1 + d\eta_2$ always holds, for $a, b, c, d \in \mathbb{R} \setminus \{0\}$ (Darmois-Skitovič theorem (Darmois, 1953)). □

The following lemma states that marginalizable models typically have a small number of identifiable parents.

Lemma 2. *Let $\mathcal{F} \subseteq \mathcal{I}_m$. Let $(X_0, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an $\mathcal{F}$-model with DAG $\mathcal{G}_0$ and $pa_{X_0}(\mathcal{G}_0) \neq \emptyset$. Let $S \subseteq \{1, \dots, p\}$ be a non-empty set. If $(X_0, \boldsymbol{X})$ is marginalizable to $S \cup \{0\}$, then $S_{\mathcal{F}}(X_0) \subseteq S$.*

Proof. Since $(X_0, \mathbf{X})$ is marginalizable to $S \cup \{0\}$, $(X_0, \mathbf{X}_S)$ follows an $\mathcal{F}$-model. Therefore, there exist $f_0 \in \mathcal{F}$, such that $X_0 = f_0(X_{\tilde{S}}, \varepsilon_0)$ for some $\tilde{S} \subseteq S$, $\varepsilon_0 \perp\!\!\!\perp X_{\tilde{S}}$, $\varepsilon_0 \sim U(0, 1)$. In other words, $f_0^{\leftarrow}(X_{\tilde{S}}, X_0) \perp\!\!\!\perp X_{\tilde{S}}$, $f_0^{\leftarrow}(X_{\tilde{S}}, X_0) \sim U(0, 1)$, which is exactly the definition of $\mathcal{F}$-plausibility. Hence, $\tilde{S}$ is $\mathcal{F}$-plausible and consequently $S_{\mathcal{F}}(X_0) \subseteq \tilde{S} \subseteq S$. □

Continuing with Example 2, Lemma 2 gives us that $S_{\mathcal{F}_L}(X_0) = \emptyset$ in the Gaussian case, and $S_{\mathcal{F}}(X_0) = \{1\}$ in the non-Gaussian case. Merely assuming linearity is insufficient to determine whether X_2 is a parent of X_0.

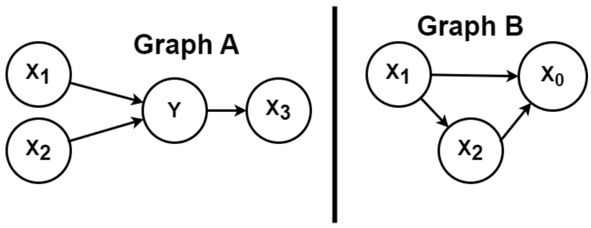

Figure 2: DAGs corresponding to Remark 1 (Graph A) and Remark 2 (Graph B)

3.2 Deriving assumptions under which $S_{\mathcal{F}}(Y) = pa_Y$

We give conditions under which all sets $S \subseteq \{1, \dots, p\}, S \neq pa_Y$ are not $\mathcal{F}$-plausible. We will focus on two main cases: when $S \cap ch_Y \neq \emptyset$ and $S \subset pa_Y$. We start with the case when $S \cap ch_Y \neq \emptyset$.

3.2.1 Case $S \cap ch_Y \neq \emptyset$

In the following, we show that we can use classical identifiability results from the literature for assessing $\mathcal{F}$-plausibility of a set S. Informally, if all variables in the SCM follow an identifiable $\mathcal{F}$-model, then any S containing a child of Y can not be $\mathcal{F}$-plausible. We use a notion of pairwise identifiability of an $\mathcal{F}$-model, defined in Appendix A.2. Pairwise identifiability describes the identifiability of the causal relation between each pair of random variables, conditioned on any other variables.

Proposition 1. *Let $(X_0, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an SCM with DAG $\mathcal{G}_0$. Let $S \subseteq \{1, \dots, p\}$ and denote $S_0 = S \cup \{0\}$. Assume that $\mathcal{G} = \mathcal{G}_0[S_0]$, the projection of $\mathcal{G}_0$ on S defined in Section 2, is a DAG. Let S contain a childless child of X_0, that is, $\exists j \in ch_0(\mathcal{G})$ such that $ch_j(\mathcal{G}) = \emptyset$.*

Let $\mathcal{F} \subseteq \mathcal{I}_m$ and let $(X_0, \boldsymbol{X}_S)$ follow an $\mathcal{F}$-model with graph $\mathcal{G}$, that is pairwise identifiable. Then, S is not $\mathcal{F}$-plausible.

The proof is in Appendix C. To provide an example of the usage of Proposition 1, consider $\mathcal{F} = \mathcal{F}_A$ and $X_1 \to Y \to X_2$. Let $X_2 = f_2(Y) + \eta_2$, where $Y \perp\!\!\!\perp \eta_2$ and η_2 has the Gaussian distribution and f_2 is non-linear. Combining Proposition 1 with Lemma 6 in Zhang and Hyvärinen (2009), we obtain that $S = \{2\}$ is not $\mathcal{F}_A$-plausible. Since $S = \{1\} = pa_Y$ is $\mathcal{F}_A$-plausible as long as $f_Y \in \mathcal{F}_A$, we get $S_{\mathcal{F}_A}(Y) = pa_Y = \{1\}$.

The following proposition discusses a different case when $\mathcal{F}$-unplausibility results from restricting support of Y by conditioning on the child of Y. This result is specific for a location-scale space of functions $\mathcal{F}_{LS}$, but can be modified for other types of $\mathcal{F}$.

Proposition 2 (Assuming bounded support)**.** *Let $(Y, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow a SCM with DAG $\mathcal{G}_0$. Let $S \subseteq \{1, \dots, p\}$ be a non-empty set. Let $\underline{\Psi}, \overline{\Psi} : \mathbb{R}^{|S|} \to \mathbb{R}$ be real functions such that*

$$supp(Y \mid \boldsymbol{X}_S = \boldsymbol{x}) = \big(\underline{\Psi}(\boldsymbol{x}), \overline{\Psi}(\boldsymbol{x})\big), \quad \forall \boldsymbol{x} \in supp(\boldsymbol{X}_S).$$

Moreover, let

$$\frac{Y - \underline{\Psi}(\boldsymbol{X}_S)}{\overline{\Psi}(\boldsymbol{X}_S) - \underline{\Psi}(\boldsymbol{X}_S)} \not\perp\!\!\!\perp \boldsymbol{X}_S. \tag{8}$$

Then, S is not $\mathcal{F}_{LS}$-plausible.

The proof can be found in Appendix C. Proposition 2 can be expressed as follows. If the support of Y given $\mathbf{X}_S = \mathbf{x}_S$ is bounded, then S can be $\mathcal{F}_{LS}$-plausible only in a very specific case when (8) does not hold. Typically, (8) holds if S contains a child of Y.

Example 4. *Consider SCM where Y is a parent of X_1 and $X_1 = Y + \eta$, where $Y \perp\!\!\!\perp \eta$. Assume that Y, η are non-negative ($supp(Y) = supp(\eta) = (0, \infty)$). Then, $\underline{\Psi}(x) = 0$ and $\overline{\Psi}(x) = x$, since the support of $[Y \mid Y + \eta = x]$ is $(0, x)$. Hence, (8) reduces to $\frac{Y}{X_1} \not\perp\!\!\!\perp X_1$. If $\frac{Y}{X_1} \not\perp\!\!\!\perp X_1$, then $S = \{1\}$ is not $\mathcal{F}_{LS}$-plausible.*

How strong is the assumption $\frac{Y}{X_1} \not\perp\!\!\!\perp X_1$? We claim that it holds in typical situations. A notable exception when $\frac{Y}{X_1} \perp\!\!\!\perp X_1$ holds is when Y, η have Gamma distributions with equal scales.

Proposition 2 is applicable only when S contains a child of Y. Generally, if $S \subseteq pa_Y$, then (8) typically does not hold, as the following example illustrates.

Example 5. *Consider a bivariate SCM with $X_1 \to Y$. Let $Y = X_1 + \eta$, where $X_1 \perp\!\!\!\perp \eta$. Assume that $supp(X) = supp(\eta) = (0, 1)$. Then, $\underline{\Psi}(x) = x$ and $\overline{\Psi}(x) = 1 + x$. Hence, (8) reduces to $Y - X_1 \not\perp\!\!\!\perp X_1$, which is not satisfied. Hence, Proposition 2 is not applicable.*

Proposition 2 can be also stated for a case when $\overline{\Psi}(x) = \infty$. In that case, we require stronger assumptions; we replace assumption (8) by $Y - \underline{\Psi}(\mathbf{X}_S) \not\perp\!\!\!\perp \mathbf{X}_S$ and replace $\mathcal{F}_{LS}$ by $\mathcal{F}_A$ (more restricted set where the scale is fixed).

3.2.2 Case $S \subset pa_Y$

In the following, we discuss the case when $S \subset pa_Y$. First, we introduce the notion of the unseparability of a real function. This is a fundamental notion since we will show that if f_Y is unseparable, then every $S \subset pa_Y$ is not $\mathcal{F}$-plausible. Later, we provide a characterization of unseparable functions, which leaves us with a powerful tool for inferring $\mathcal{F}$-plausible sets. For the characterization of unseparable functions, we mainly restrict our attention to sets $\mathcal{F} = \mathcal{F}_F$ for some distribution function F.

Definition 3. *Let $\boldsymbol{X} = (X_1, \dots, X_k)$ be a random vector and $\mathcal{F} \subseteq \mathcal{I}_m$. A function $f \in \mathcal{I}_m : \mathbb{R}^{k+1} \to \mathbb{R}$ is called $\mathcal{F}-$**unseparable wrt. $\boldsymbol{X}$**, if for all $S \subset \{1, \dots, k\}$ there exists $z \in \mathbb{R}$ such that*

$$\text{for all } g \in \mathcal{F} \text{ holds } g^{\leftarrow}\big(\boldsymbol{X}_S, f(\boldsymbol{X}, z)\big) \not\perp\!\!\!\perp \boldsymbol{X}_S. \tag{9}$$

The notion of unseparability of a function describes that we are not able to "erase" the effect of $\mathbf{X}_S$ on $Y = f(\mathbf{X}, z)$ without considering other variables. Note that the notion of $\mathcal{F}$-unseparability is a property of a real function; it does not depend on causal relations (only on the distribution of $\mathbf{X}$). We provide some intuition about the notion of unseparability on the following example.

Remark 3. *Let $k = 2$ and $\boldsymbol{X} = (X_1, X_2)$ be continuous with independent components. Consider the function $f(x_1, x_2, z) = x_1 x_2 z$ and $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{I}_m$ such that $\mathcal{F}_1$ contains only linear functions and $\mathcal{F}_2$ contains only multiplicative functions, that is, $\mathcal{F}_1 = \mathcal{F}_L$ and*

$$\mathcal{F}_2 = \{f \in \mathcal{I}_m : f(\boldsymbol{x}, \varepsilon) = g_1(x_1) \dots g_k(x_k) q^{-1}(\varepsilon) \\ \text{for some measurable functions } g_1, \dots, g_k \text{ and a quantile function } q^{-1}\}.$$

Then, f is $\mathcal{F}_1-$unseparable wrt. $\boldsymbol{X}$ but is not $\mathcal{F}_2-$unseparable wrt. $\boldsymbol{X}$.

Proof. Intuitively, f is $\mathcal{F}_1-$unseparable wrt. $\mathbf{X}$ because we can not find $\beta \in \mathbb{R}$ such that $X_1X_2\varepsilon - \beta X_1 \perp\!\!\!\perp X_1$. However, f is not $\mathcal{F}_2-$unseparable wrt. $\mathbf{X}$, since we can find g such that $X_1X_2\varepsilon \cdot g(X_1) \perp\!\!\!\perp X_1$. We give a more rigorous explanation of this.

First, we show that f is not $\mathcal{F}_2-$unseparable wrt. $\mathbf{X}$. Take $S = \{1\}$. Choose a function $g \in \mathcal{F}_2$ such that $g(x, z) = xz$. Its inverse has a form $g^{\leftarrow}(x, xz) = z$. Hence, $g^{\leftarrow}(X_S, f(\mathbf{X}, z)) = g^{\leftarrow}(X_1, X_1X_2z) = X_2z \perp\!\!\!\perp X_1$. Hence, f is not $\mathcal{F}_2-$unseparable wrt. $\mathbf{X}$, since we found a $g \in \mathcal{F}_2$ such that (9) is violated.

Second, we show that f is $\mathcal{F}_1-$unseparable wrt. $\mathbf{X}$. Consider $S = \{1\}$ and consider any function $g \in \mathcal{F}_1$. Let us write g in a form $g(x, z) = \beta x + q^{-1}(z)$ for some quantile function q^{-1} and $\beta \neq 0$. The inverse of g satisfies $g^{\leftarrow}(x, xz) = q(xz - \beta x)$. In order to show (9), we need to show that $q(X_1X_2z - \beta X_1) \not\perp\!\!\!\perp X_1$. In the rest of the section (and Appendix B), we develop a framework for proving statements such as this one. In particular, $X_1X_2z - \beta X_1 \not\perp\!\!\!\perp X_1$ holds due to Lemma B.2 part 2. Hence, for $S = \{1\}$ (9) is satisfied. For $S = \{2\}$ the proof follows analogously. □

The following lemma shows the connection between the $\mathcal{F}$-unseparability and $\mathcal{F}$-plausibility. It shows that the $\mathcal{F}$-unseparability implies the $\mathcal{F}$-unplausibility of all subsets of the parents.

Lemma 3. *Let $(Y, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an SCM with DAG $\mathcal{G}_0$ and $pa_Y(\mathcal{G}_0) \neq \emptyset$. Let $f_Y \in \mathcal{F} \subseteq \mathcal{I}_m$. If f_Y is $\mathcal{F}-$unseparable wrt. $\boldsymbol{X}_{pa_Y}$, then every $S \subset pa_Y(\mathcal{G}_0)$ is not an $\mathcal{F}$-plausible set of parents of Y.*

Proof. For a contradiction, let $S \subset pa_Y(\mathcal{G}_0)$ be an $\mathcal{F}$-plausible set of parents of Y. Then, we can find $f \in \mathcal{F}$ such that $f^{\leftarrow}(\mathbf{X}_S, Y) \perp\!\!\!\perp \mathbf{X}_S$. Since we have $Y = f_Y(\mathbf{X}_{pa_Y}, \varepsilon_Y)$, we can rewrite $f^{\leftarrow}\big(\mathbf{X}_S, f_Y(\mathbf{X}_{pa_Y}, \varepsilon_Y)\big) \perp\!\!\!\perp \mathbf{X}_S$. Since $\varepsilon_Y \perp\!\!\!\perp \mathbf{X}_{pa_Y}$, conditioning on $[\varepsilon_Y = z]$ for arbitrary $z \in (0, 1)$ will give us $f^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}_{pa_Y}, z)\big) \perp\!\!\!\perp \mathbf{X}_S$, which is a contradiction with unseparability. □

In the following, we focus on characterizing the $\mathcal{F}_F$-unseparability for different distributions F. We show that some large classes of $f \in \mathcal{F}_F$ are indeed $\mathcal{F}_F$-unseparable. We will restrict our attention to specific types of F, where the parameters act **additively**/**multiplicatively**/**location-scale**. Rigorous definitions of these types can be found in Appendix A.3. The Gaussian distribution with fixed variance is an example of F whose parameter acts additively. The Gaussian distribution with the fixed expectation or the Pareto distribution are examples of F whose parameter acts multiplicatively. Examples of Location-Scale types of distributions include Gaussian distribution, logistic distribution, or Cauchy distribution, among many others.

First, we consider one parameter case ($q = 1$ in (5)). The following two propositions are the main results of this subsection. They characterize $\mathcal{F}_F-$unseparability (under some weak assumptions on the distribution function F). Combining these results with Lemma 3 gives us a powerful tool for inferring $\mathcal{F}_F-$plausible sets.

Proposition 3. *Let F be a distribution function whose parameter acts post-multiplicatively. Let $\boldsymbol{X} = (X_1, \dots, X_k)$ be a continuous random vector with independent components.*

- *Consider $f \in \mathcal{F}_F$ in the form $f(\boldsymbol{x}, \varepsilon) = F^{-1}\big(\varepsilon, \theta(\boldsymbol{x})\big)$ with additive function $\theta(x_1, \dots, x_k) = h_1(x_1) + \dots + h_k(x_k)$, where h_i are continuous non-constant real functions. Then, f is $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*
- *Consider $f \in \mathcal{F}_F$ in the form $f(\boldsymbol{x}, \varepsilon) = F^{-1}\big(\varepsilon, \theta(\boldsymbol{x})\big)$ with multiplicative function $\theta(x_1, \dots, x_k) = h_1(\boldsymbol{x}_S) \cdot h_2(\boldsymbol{x}_{\{1,\dots,k\}\setminus S})$ for some $S \subsetneq \{1, \dots, k\}$, where h_1, h_2 are continuous non-constant non-zero real functions. Then, f is not $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*

Idea of the proof. Full proof is in Appendix C. Here we show the main steps of the first bullet-point in the case when $k = 2$, $h_1(x) = h_2(x) = x$ and F is the Pareto distribution function. Consider $S = \{1\}$ (the case $S = \{2\}$ follows similarly). We show that for $S = \{1\}$ and any $z \in (0, 1)$ there does not exist $g \in \mathcal{F}_F$ such that $g^{\leftarrow}\big(X_1, f(\mathbf{X}, z)\big) \perp\!\!\!\perp X_1$.

For a contradiction, assume that such g exists and write $g^{\leftarrow}\big(x, \cdot\big) = F\big(\cdot, \theta_g(x)\big)$ for some non-constant positive function θ_g. Using the form of the Pareto distribution, rewrite $X_1 \perp\!\!\!\perp$ $g^{\leftarrow}\big(X_1, f(\mathbf{X}, z)\big) = F\Big(f(\mathbf{X}, z), \theta_g(X_1)\Big) = F\Big(F^{-1}[z, \theta(\mathbf{X})], \theta_g(X_1)\Big) = z^{-\frac{\theta(\mathbf{X})}{\theta_g(X_1)}} = z^{-\frac{(X_1+X_2)}{\theta_g(X_1)}}$.

Using the identity $Z_1 \perp\!\!\!\perp Z_2 \implies f(Z_1) \perp\!\!\!\perp Z_2$ for any measurable function f and random variables Z_1, Z_2, we get that

$$X_1 \perp\!\!\!\perp \theta_g(X_1)(X_1 + X_2).$$

We show that this is not possible. Denote $\xi = \theta_g(X_1)(X_1 + X_2)$ and choose *distinct* a, b, c in the support of X_1 such that $\theta_g(b) \neq 0$ (since θ_g is non-constant and X_1 is non-binary, this is possible).

Since $\xi \perp\!\!\!\perp X_1$, then $\xi \mid [X_1 = a] \overset{D}{=} \xi \mid [X_1 = b] \overset{D}{=} \xi \mid [X_1 = c]$. Hence,

$$\theta_g(a)(a + X_2) \overset{D}{=} \theta_g(b)(b + X_2) \overset{D}{=} \theta_g(c)(c + X_2). \tag{10}$$

By dividing by a nonzero constant $\theta_g(b)$ and subtracting b we get

$$\{\frac{\theta_g(a)}{\theta_g(b)}a - b\} + \frac{\theta_g(a)}{\theta_g(b)}X_2 \overset{D}{=} X_2 \overset{D}{=} \{\frac{\theta_g(c)}{\theta_g(b)}c - b\} + \frac{\theta_g(c)}{\theta_g(b)}X_2.$$

It holds that (see lemma B.1) if $z_1 + z_2X_2 \overset{D}{=} X_2$ for some constants z_1, z_2, then $z_2 = \pm 1$. Hence in our case, it leads to $\frac{\theta_g(a)}{\theta_g(b)} = \pm 1$ and $\frac{\theta_g(c)}{\theta_g(b)} = \pm 1$. Therefore, at least two values of $\theta_g(a), \theta_g(b), \theta_g(c)$ have to be equal, and neither of them is zero. Without loss of generality $\theta_g(a) = \theta_g(b)$. Plugging this into equation (10), we get $a = b$ which is a contradiction since we chose them to be distinct. Therefore, the non-constant function θ_g does not exist, and hence also g does not exist. □

Proposition 4. *Let F be a distribution function whose parameter acts post-additively. Let $\boldsymbol{X} = (X_1, \dots, X_k)$ be a continuous random vector with independent components.*

- *Consider $f \in \mathcal{F}_F$ in the $f(\boldsymbol{x}, \varepsilon) = F^{-1}\big(\varepsilon, \theta(\boldsymbol{x})\big)$ with an additive function $\theta(x_1, \dots, x_k) = h_1(\boldsymbol{x}_S) + h_2(\boldsymbol{x}_{\{1,\dots,k\}\setminus S})$ for some non-empty $S \subset \{1, \dots, k\}$, where h_1, h_2 are continuous non-constant non-zero real functions. Then, f is not $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*
- *Consider $f \in \mathcal{F}_F$ in the form $f(\boldsymbol{x}, \varepsilon) = F^{-1}\big(\varepsilon, \theta(\boldsymbol{x})\big)$ with multiplicative function $\theta(x_1, \dots, x_k) = h_1(x_1).h_2(x_2)\dots h_k(x_k)$ where h_i are continuous non-constant non-zero real functions. Then, f is $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*

The proof follows similar steps as the proof of Proposition 3 and can be found in Appendix C. Proposition 3 and Proposition 4 reveal that to obtain $\mathcal{F}_F$-unseparability, the form of the parameter can not match the way how the parameter affects the distribution. If the parameter acts additively resp. multiplicatively, the parameter can not have an additive resp. multiplicative form. However, $\mathcal{F}_F$-unseparability typically holds if the forms do not match. Note that the multiplicative assumption of f_Y in the second bullet-point of Proposition 4 is not necessary and is only technical for proving Lemma B.2. As long as f_Y is not additive in any of its arguments, Lemma B.2 can be modified for much larger space of functions.

Proposition 3 and Proposition 4 are formulated only for one-parameter case (case $q = 1$). The following proposition discusses the case when $q = 2$, where we restrict ourselves to a location-scale family of distributions.

Proposition 5. *Let F have a Location-Scale type with $q = 2$ parameters. Let $\boldsymbol{X} = (X_1, \dots, X_k)$ be a continuous random vector with independent components. Consider $f \in \mathcal{F}_F$ in the form $f(\boldsymbol{x}, \varepsilon) = F^{-1}\big(\varepsilon, \theta(\boldsymbol{x})\big)$, where $\theta(\boldsymbol{x}) = \big(\mu(\boldsymbol{x}), \sigma(\boldsymbol{x})\big)^\top$ is additive in both components, that is, $\mu(\boldsymbol{x}) = h_{1,\mu}(x_1) + \dots + h_{k,\mu}(x_k)$ and $\sigma(\boldsymbol{x}) = h_{1,\sigma}(x_1) + \dots + h_{k,\sigma}(x_k)$ for some continuous non-constant non-zero functions $h_{i,\cdot}$, where moreover we assume $h_{i,\sigma} > 0$, $i = 1, \dots, k$. Then, f is $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*

The proof is in the Appendix C. The crucial assumption lies in the additivity of $\sigma(\mathbf{x})$; since this parameter acts multiplicatively in F and has an additive form, a similar reasoning as in Proposition 3 can be used (Proposition 5 shows that even if another parameter μ depends on $\mathbf{X}$, it does not ruin the results). Proposition 5 can also be reformulated such that both parameters have a multiplicative form.

The biggest drawback of Propositions 3, 4 and 5 lies in the fact that we assume independent components of $\mathbf{X}$, which is rarely the case. This assumption is only a technical assumption that simplifies the proof and is not necessary. It allows us to explicitly express the conditional distributions $h(X_i) \mid \mathbf{X}_S$, that we used in the proof. However, an explicit form of a conditional distribution can also be found in other cases, such as when $\mathbf{X}$ is Gaussian.

Lemma 4. *Let the assumptions from Proposition 5 hold, and let $\boldsymbol{X} = (X_1, \dots, X_k)$ be a non-degenerate Gaussian random vector (possibly with dependent components) and $h_{i,\sigma}$ be linear functions. Then, f is $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*

Proof. The proof follows the same steps as the proof of Proposition 5, with the only difference that we use Lemma B.2 part 4 instead of Lemma B.2 part 3 in the last step. □

The results of Propositions 3, 4 and 5 are not restricted for $\mathcal{F}_F$, but can be easily modified for the case $\mathcal{F} = \mathcal{F}_A$.

Consequence 1. *Consider $f_Y \in \mathcal{F}_A$ and let $(Y, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an SCM with DAG $\mathcal{G}_0$ where pa_Y are d-separated.*

- *If f_Y has a form $f_Y(\boldsymbol{x}, e) = h_1(\boldsymbol{x}_S) + h_2(\boldsymbol{x}_{pa_Y \setminus S}) + q^{-1}(e)$, $\quad \boldsymbol{x} \in \mathbb{R}^{|pa_Y|}, e \in (0,1)$, for some non-empty $S \subset pa_Y$, where h_1, h_2 are continuous non-constant real functions and q^{-1} is a quantile function. Then $S_{\mathcal{F}_A}(Y) = \emptyset$.*
- *If f_Y has a form $f_Y(\boldsymbol{x}, e) = h_1(x_1) \dots h_{|pa_Y|}(x_{|pa_Y|}) + q^{-1}(e)$, $\quad \boldsymbol{x} \in \mathbb{R}^{|pa_Y|}, e \in (0,1)$, where h_i are continuous non-constant non-zero real functions and q^{-1} is a quantile function. Then every $S \subset pa_Y$ is not $\mathcal{F}_A$-plausible.*

The proof is in Appendix C. As before, the assumption of d-separability can be replaced by assuming the normality of $\mathbf{X}_{pa_Y}$, the assumption of multiplicative form of f_Y in the second bullet-point can be weakened and the statement can be modified for $\mathcal{F} = \mathcal{F}_{LS}$.

3.3 Some examples

Example 6 (Extending Example 3)**.** *Consider an SCM with DAG drawn in Figure 2A, where the structural equation corresponding to Y is in the form (5) with F being the Pareto distribution with $\theta(X_1, X_2) = h_1(X_1) + h_2(X_2)$ for some continuous non-constant functions h_1, h_2. Moreover, let $X_3 = F^{-1}\big(\varepsilon_3, \theta_3(Y)\big)$, where $\theta_3(x) \neq a\log(x) + b$ for any $a, b \in \mathbb{R}$, $\varepsilon_3 \perp\!\!\!\perp Y$. Then $S_{\mathcal{F}_F}(Y) = \{1, 2\} = pa_Y$.*

The reason is as follows: if $3 \in S$, then S is not $\mathcal{F}_F$-plausible (Proposition 1 combined with results presented in Example 3). Cases $S = \{1\}$ or $S = \{2\}$ are also not $\mathcal{F}_F$-plausible, since the function $\theta(x_1, x_2) = h_1(x_1) + h_2(x_2)$ is $\mathcal{F}_F$-unseparable wrt (X_1, X_2) from Proposition 3. Since $S = \{1, 2\}$ is trivially $\mathcal{F}_F$-plausible from (7), we get $S_{\mathcal{F}_F}(Y) = \{1, 2\} = pa_Y$.

Example 7 (Additive noise models)**.** *Consider an SCM with DAG drawn in Figure 1 and $\mathcal{F} = \mathcal{F}_A$. Let (X_1, X_2, X_3) be normally distributed,*

$$Y = \mu_Y(X_1, X_2, X_3) + \eta_Y, \quad X_4 = \mu_1(Y) + \mu_2(X_5, \eta_4), \quad X_5 = \eta_5$$

for some independent (not necessarily uniformly distributed) noise variables η_Y, η_4, η_5, where μ_1, μ_2 are continuous non-zero functions. Let (Y, X_4) follow an identifiable bivariate additive noise model (that is, μ_1 does not satisfy a differential equation described in Section 3.1 in Peters et al. (2014)).

$S = \{5\}$ is not $\mathcal{F}_A$-plausible, since $Y - f(X_3) \perp\!\!\!\perp X_3$ can happen only if f is a constant function (and constant functions do not belong to $\mathcal{I}_m$). The same argument can be used whenever $5 \in S \not\ni 4$.

$S = \{4\}$ is not $\mathcal{F}_A$-plausible since (Y, X_4) follow an identifiable (additive) model, and Proposition 1 can be used. Sets $S \supseteq \{4, 5\}$ can be dealt with analogously.

If μ_Y is additive in the first component, that is, $\mu_Y(X_1, X_2, X_3) = h_1(X_1) + h(X_2, X_3)$ for some continuous non-constant functions h_1, h, then $S = \{1\}$ is $\mathcal{F}_A$-plausible and hence $S_{\mathcal{F}_A}(Y) \subseteq \{1\}$. However if μ_Y is not additive in any of its components, we get that any $S \subset \{1, 2, 3\}$ is not $\mathcal{F}_A$-plausible.

If μ_Y does not satisfy the differential equation described in Section 3.1 in Peters et al. (2014), we also get that every S with $4 \in S$ is not $\mathcal{F}_A$-plausible since (Y, X_4) follow an identifiable (additive) model and Proposition 1 can be applied.

We do not have to pay attention to the sets $S = \{1,2,3,4\}, \{1,2,3,5\}, \{1,2,3,4,5\}$ since $pa_Y \subset S$ and the set of $\mathcal{F}_A$-identifiable parents remain the same regardless of the plausibility of these sets.

In Example 7, we considered a variable X_5, which is not a neighbor of Y. In general, we can directly discard this variable as a potential parent using the classical graphical argument of d-separation. However, if $X_i \perp\!\!\!\perp Y$ for some $i \in \{1,\dots,p\}$, typically $i \notin S_{\mathcal{F}}(Y)$. In particular, it is easy to see that if $X_i \perp\!\!\!\perp Y$, then a set $S = \{i\}$ is not $\mathcal{F}$-plausible.

4 Estimation

In this section, we introduce two algorithms for estimating the set of direct causes of a target variable. One is based on estimating $S_{\mathcal{F}}(Y)$, and the second is a score-based algorithm that finds the best-fitting set for pa_Y.

We consider a random sample of size $n \in \mathbb{N}$ from $(Y, \mathbf{X}) \in \mathbb{R} \times \mathbb{R}^p$, where Y is the target variable, $\mathbf{X}$ are the covariates and $\mathcal{F} \subseteq \mathcal{I}_m$.

4.1 ISD algorithm for estimating $S_{\mathcal{F}}(Y)$

To determine the $\mathcal{F}$-plausibility of a non-empty set $S \subseteq \{1,\dots,p\}$, we follow the procedure below: First,

$$\text{estimate } \hat{\varepsilon}_S = \hat{f}^{\leftarrow}(Y, \mathbf{X}_S) \text{ under constraint } \hat{f} \in \mathcal{F}. \tag{11}$$

Then, answer the following questions:

1. (**I**ndependence) Is $\hat{\varepsilon}_S \perp\!\!\!\perp \mathbf{X}_S$?
2. (**S**ignificance) Is every X_i significant, $i \in S$?
3. (**D**istribution) Is $\hat{\varepsilon}_S \sim U(0,1)$?

We conclude that S is $\mathcal{F}$-plausible if and only if all questions 1,2 and 3 have positive answers. Finally, we

$$\text{define } \hat{S}_{\mathcal{F}}(Y) \text{ as an intersection of all sets that we marked as } \mathcal{F}\text{-plausible.}$$

Estimating (11) is a classical problem in machine learning.

- If $\mathcal{F} = \mathcal{F}_L$ we can use linear regression; regressing Y on $\mathbf{X}_S$ in a linear model $Y = \beta_S \mathbf{X}_S + \eta$ and output the residuals $\hat{\eta} := Y - \hat{\beta}_S \mathbf{X}_S$. Possibly, re-scale the residuals $\hat{\varepsilon} := \hat{q}(\hat{\eta})$ where $\hat{q}$ is the empirical distribution function of $\hat{\eta}$. See the discussion about question 3 below.
- If $\mathcal{F} = \mathcal{F}_A$, then we can apply random forest, neural networks, GAM, or other classical methods (Ho, 1995; Green and Silverman, 1994; Ripley, 1996). Using one of these methods, we estimate the conditional mean $\hat{\mu}$ and output the residuals $\hat{\eta} := Y - \hat{\mu}(\mathbf{X}_S)$. Possibly, re-scale the residuals $\hat{\varepsilon} := \hat{q}(\hat{\eta})$.
- If $\mathcal{F} = \mathcal{F}_{LS}$, we can apply GAMLSS (Stasinopoulos and Rigby, 2007), neural networks or similar methods (Immer et al., 2022) to estimate the conditional mean and variance $\hat{\mu}, \hat{\sigma}$ and output $\hat{\eta} := \frac{Y - \hat{\mu}(\mathbf{X}_S)}{\hat{\sigma}(\mathbf{X}_S)}$. Possibly, re-scale $\hat{\varepsilon} := \hat{q}(\hat{\eta})$.
- If $\mathcal{F} = \mathcal{F}_F$ for some distribution function F, we can use GAMLSS or GAM algorithms for estimating θ (the method depends on F. Often, we can also apply random forest). Then, we define $\hat{\varepsilon} := F\big(Y, \hat{\theta}(\mathbf{X}_S)\big)$.

For the test of independence (question 1), we can use a Kernel-based HSIC test (Pfister et al., 2018) or a copula-based test (Genest et al., 2019).

The assessment of significance (question 2) corresponds to the chosen regression method. Especially in the case of linear regression, we test if $\beta_i \neq 0$ for every $i \in S$. For GAM or GAMLSS, we analogously use the corresponding p-values. If we use random forest, we can use a permutation test for assessing the significance of the covariates (Paul and Dupont, 2014).

For the test of $\hat{\varepsilon} \sim U(0,1)$ (question 3), we can use a Kolmogorov-Smirnov or Anderson-Darling test (Razali and Yap, 2011). However, we typically skip this step. If $\mathcal{F} = \mathcal{F}_L, \mathcal{F}_A$ or $\mathcal{F}_{LS}$, the answer to question 3 is always positive since we can use a probability integral transform of the estimated noise to obtain $\hat{\varepsilon} \sim U(0,1)$. Note that $f \in \mathcal{F}$ if and only if $q \circ f \in \mathcal{F}$ where q is a distribution function (which is not longer true in general if $\mathcal{F} = \mathcal{F}_F$). Applying a monotonic transformation on $\hat{\varepsilon}$ does not affect questions 1 and 2. Therefore, this question can be omitted if $\mathcal{F} = \mathcal{F}_L, \mathcal{F}_A$ or $\mathcal{F}_{LS}$.

In our implementation, we opt for HSIC test, GAM estimation and the Anderson-Darling test with level $\alpha = 0.05$. Some comparisons of the performance of GAM and a random forest can be found in the supplementary package.

4.2 Score-based estimation of pa_Y

In the following, we propose a score-based algorithm for estimating the set of direct causes of Y. It is a counterpart of score-based algorithms for estimating the full DAG $\mathcal{G}_0$, following the ideas from Nowzohour and Bühlmann (2016); Peters et al. (2014); Bodik and Chavez-Demoulin (2023). Recall that under (♥), the set $S = pa_Y$ should satisfy that $\varepsilon_S \perp\!\!\!\perp \mathbf{X}_S$, every $X_i, i \in S$ is significant and $\varepsilon_S \sim U(0,1)$. Therefore, we use the following score function

$$\widehat{pa}_Y = \underset{\substack{S \subseteq \{1,\dots,p\} \\ S \neq \emptyset}}{\arg\max}\ score(S) = \underset{\substack{S \subseteq \{1,\dots,p\} \\ S \neq \emptyset}}{\arg\max}\ \lambda_1(Independence) + \lambda_2(Significance) + \lambda_3(Distribution),$$

where $\lambda_1, \lambda_2, \lambda_3 \in [0, \infty)$, "*Independence*" is a measure of independence between $(\hat{\varepsilon}_S, \mathbf{X}_S)$, "*Significance*" is a measure of significance of covariates $\mathbf{X}_S$ and "*Distribution*" is a distance between the distribution of $\hat{\varepsilon}_S$ and $U(0,1)$, where $\hat{\varepsilon}_S$ is the noise estimate obtained from (11).

The measure of independence can be chosen as the p-value of an independence test (such as the Kernel-based HSIC test or the copula-based test). *The measure of significance* corresponds to the estimation method analogously to the ISD case. For linear regression (similarly for GAM or GAMLSS), we compute the corresponding p-values for the hypotheses $\beta_i = 0, i \in S$. Then, *"Significance"* is the minus of the maximum of the corresponding p-values (worst case option). We can also use a permutation test to assess the covariate's significance in terms of the predictability power and choose the largest p-value. *The distance between the distribution of* $\hat{\varepsilon}_S$ *and* $U(0,1)$ can be chosen as the p-value of the Anderson-Darling test.

The choice of $\lambda_1, \lambda_2, \lambda_3$ describes weights we put on each of the three properties: if $\lambda_1 > \lambda_2, \lambda_3$, then our estimate will be very sensitive against the violation of the independence $\varepsilon_S \perp\!\!\!\perp \mathbf{X}_S$, but not as sensitive against the violation of the other two properties.

In our implementation, we opt for the following choices. The *Independence* term is the logarithm of the p-value of the Kernel-based HSIC test, and the *Distribution* term is the logarithm of the p-value of the Anderson-Darling test. We use GAM for the estimation in (11) and minus the logarithm of the maximum of the corresponding p-values for the *Significance* term. The logarithmic transformation of the three p-values is used to re-scale the values from $[0,1]$ to $(-\infty, 0]$. The practical choice for the weights is $\lambda_1 = \lambda_2 = \lambda_3 = 1$ (unless $\mathcal{F} = \mathcal{F}_L, \mathcal{F}_A$ or $\mathcal{F}_{LS}$ when we put $\lambda_3 = 0$).

4.3 Consistency

Consistency of the proposed algorithm follows directly from the results presented in Mooij et al. (2016), who showed consistency of the score-based DAG estimation for additive noise models. In the following, we assume $\mathcal{F} = \mathcal{F}_A$ although it is straightforward to generalize these results for other types of $\mathcal{F}$ (for a discussion about $\mathcal{F} = \mathcal{F}_{LS}$ see Sun and Schulte (2023) and for $\mathcal{F} = \mathcal{F}_F$ see Bodik and Chavez-Demoulin (2023)). For simplicity, we assume that the measure of independence is the negative value of HSIC test itself (not the p-value as we use in our implementation) and the noise estimate $\hat{\varepsilon}$ is *suitable* in the sense that

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \hat{\varepsilon}_i)^2\right) = 0,$$

where the expectation is taken with respect to the distribution of the random sample (see Appendix A.2 in Mooij et al. (2016)).

Proposition 6. *Consider $\mathcal{F} = \mathcal{F}_A$ and let $(Y, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an SCM with DAG $\mathcal{G}_0$ satisfying (♥). Let $pa_Y \neq \emptyset$ and assume that every $S \neq pa_Y$ is not $\mathcal{F}$-plausible. Then,*

$$\lim_{n\to\infty} \mathbb{P}(\widehat{pa}_Y \neq pa_Y) = 0, \tag{12}$$

where n is the size of the random sample and $\widehat{pa}_Y$ is our score-based estimate from Section 4.2 with $\lambda_1, \lambda_2 > 0, \lambda_3 = 0$, suitable estimation procedure and HSIC independence measure.

The proof is in Appendix C. If several sets are $\mathcal{F}$-plausible, the score-based algorithm provides no guarantees that pa_Y will have the best score among them (as opposed to the ISD algorithm that outputs their intersection).

On the population level, similar results can be easily stated for ISD algorithm. Given that we have infinite data, a consistent estimation method for (11) and a perfect independence test ("independence oracle"), ISD is correct in a sense that it outputs $\hat{S}_{\mathcal{F}_A}(Y) = S_{\mathcal{F}_A}(Y)$. This statement is made more precise in Lemma C.1 in Appendix C.

4.4 What if $pa_Y = \emptyset$?

The case $pa_Y = \emptyset$ has to be dealt with separately since an empty set can not be (by definition) $\mathcal{F}$-plausible. Moreover, the score-based algorithm always returns a non-empty set.

Some choices of $\mathcal{F}$ lead to a fully unrestricted marginal distribution of Y in the case of $pa_Y = \emptyset$. For example, assuming $\mathcal{F} = \mathcal{F}_L$ or $\mathcal{F}_A$ gives us $Y = f_Y(\varepsilon_Y) = g^{-1}(\varepsilon_Y)$ for an unrestricted quantile function g^{-1}, which fits into every situation. One possibility is to define a subset $\mathcal{P}_\emptyset \subset \{P : P \text{ is a distribution function}\}$ and say that $pa_Y = \emptyset$ is plausible if and only if $P_Y \in \mathcal{P}_\emptyset$, where P_Y is the distribution of Y. However, defining $\mathcal{P}_\emptyset$ would be a troublesome step in most applications.

The case $\mathcal{F} = \mathcal{F}_F$ gives us an elegant option for assessing the validity of $pa_Y = \emptyset$. If the marginal distribution of Y is F with some (unknown, but constant) parameters, we say that $pa_Y = \emptyset$ is plausible. In other words, we assess whether $P_Y \in \mathcal{P}_{F,\emptyset} = \{F_\theta : \theta \in \mathbb{R}^q\}$, which is equivalent to assessing $f_Y \in \{f \in \mathcal{F}_F : f \text{ is a univariate}\} = \{f : f(\varepsilon) = F^{-1}(\varepsilon, \theta), \text{ for some constant } \theta\}$. Even though a marginal assessment $P_Y \in \mathcal{P}_{F,\emptyset}$ elegantly corresponds to the previous $\mathcal{F}_F$ framework, it does not mean that it is a reasonable approach.

Sometimes, we encounter situations where we have a subset of covariates, and we know that one of the covariates acts as a parent of Y, although we are uncertain which one. Alternatively, we may have a covariate for which we know that it is a parent of Y. These scenarios can arise, for example, when conducting a do-intervention or when determining the orientation of certain edges using Meek rules (Meek, 1995). Consequently, assuming that $pa_Y \neq \emptyset$ can often be justified either by expert knowledge about the problem or by employing other causal inference methods.

4.5 How to choose $\mathcal{F}$

The choice of the class $\mathcal{F}$ is a crucial step in the algorithm. The choice of an appropriate model is a common problem in classical statistics, however, it is more subtle in causal discovery. It has been shown (Peters et al., 2014) that methods based on restricted structural equation models can outperform traditional constraint-based methods (when the problem is estimating the entire DAG). Even if assumption such as additive noise or Gaussian distribution of the effect given the causes can appear as very strong, such methods turned out to be very useful, and small violations of the model still lead to a good estimation procedure.

The size of $\mathcal{F}$ is the most important part. If $\mathcal{F}$ contains too many functions, we obtain that most sets are $\mathcal{F}$-plausible. On the other hand, too restrictive $\mathcal{F}$ can lead to rejecting all sets as potential causes. If we have knowledge about the data-generating process (such as when Y is a sum of many small events), choosing $\mathcal{F}_F$ for a distribution function F (such as Gaussian) is reasonable. If not, marginal distributions can be helpful for choosing an appropriate F. For a similar discussion about the choice of an appropriate model in causal discovery, see Section 5.4 in Bodik and Chavez-Demoulin (2023).

Finding an automatic, data-driven choice of $\mathcal{F}$ can lead to new and interesting results and broaden the applicability of this approach. In practice, we can try several different choices of $\mathcal{F}$ (choices based on, for example, marginal distributions) and observe the results. For the choices $\mathcal{F} = \mathcal{F}_A$ or $\mathcal{F}_{LS}$, there exist a numerous papers justifying such assumptions (when estimating the full DAG Peters et al. (2014), Peters et al. (2017), Immer et al. (2022), Khemakhem et al. (2021), Sun and Schulte (2023)).

5 Simulations and application

In this section, we highlight some of the theoretical results on simulated data (Section 5.1), we assess the performance of the algorithms on three benchmark datasets (Section 5.2), and we apply our methodology on real data (Section 5.3).

To randomly generate a d-dimensional function, we use the concept of the Perlin noise generator (Perlin, 1985). Examples of such generated functions can be found in Appendix D.1. The two algorithms presented in Section 4, all simulations, and the Perlin noise generator are coded in the programming language `R` (R Core Team, 2022) and can be found in the supplementary package or `https://github.com/jurobodik/Structural-restrictions-in-local-causal-discovery.git`.

5.1 Additive case under Proposition 4

We conduct a simulation study highlighting the results from Proposition 4. Consider a target variable Y with two parents X_1, X_2, where $\mathbf{X} = (X_1, X_2)$ is a centered normal random vector with correlation $c \in \mathbb{R}$. The generating process of Y is the following:

$$Y = g_1(X_1) + g_2(X_2) + \gamma \cdot g_{1,2}(X_1, X_2) + \eta, \quad \text{with } \eta \sim N(0,1) \text{ and } \gamma \in \mathbb{R}, \tag{13}$$

where $g_1, g_2, g_{1,2}$ are fixed functions generated using the Perlin noise approach.

Proposition 4 suggests that if $c = \gamma = 0$, then $S_{\mathcal{F}}(Y) = \emptyset$. Lemma 4 suggests that if $c \in \mathbb{R}, \gamma \neq 0$ then we should find that $S_{\mathcal{F}}(Y) = pa_Y = \{1, 2\}$. The case $c \in \mathbb{R}, \gamma = 0$ is unclear and it depends on g_1, g_2; see Remark 2. Moreover, the choice of c can affect the finite sample properties.

Figure 3 confirms these results. For a range of parameters $c \in [0, 0.9], \gamma \in [0, 1]$ we generate 50 times such a random dataset of size $n = 500$ and estimate $S_{\mathcal{F}}(Y)$ using the ISD algorithm. If γ is small, we discover direct causes of Y only in a small number of cases. However, the larger the γ, the larger the number of discovered parents. Figure 3 also suggests that the correlation between the parents can be, in fact, beneficial. The reason is that even if (13) is additive in each component, the correlation between the components can create a bias in estimating g_1 (resp. g_2). This results in a dependency between the residuals and X_1 (resp. X_2) in the model where we regress Y on X_1 (or on X_2), and we are more likely to reject the plausibility of $S = \{1\}$ (or $S = \{2\}$).

5.2 Benchmarks

We create three benchmark datasets to assess the performance of our methodology. Two of them correspond to additive noise models ($\mathcal{F} = \mathcal{F}_A$), the third to $\mathcal{F} = \mathcal{F}_F$ with the Pareto distribution F.

The first benchmark dataset consists of $\mathbf{X} = (X_1, X_2, X_3, X_4)^\top$ and the response variable Y, where $pa_Y = \{1\}$ with the corresponding graph drawn in Figure 4A. The data-generating process is as follows:

$$X_1 = \eta_1, \;\; Y = g_Y(X_1) + \eta_Y, \;\; X_i = g_i(Y, \eta_i), \;\; i = 2, 3, 4,$$

where g_Y, g_2, g_3, g_4 are fixed functions generated using the Perlin noise approach, $\eta_1, \eta_2, \eta_3, \eta_4$ are correlated uniformly distributed noise variables and $\eta_Y \sim N(0, 1)$ is independent of $\eta_1, \dots, \eta_4$. The challenge is to find the (one) direct cause among all variables.

The second benchmark dataset consists of $\mathbf{X} = (X_1, X_2, X_3)^\top$ and the response variable Y, where $pa_Y = \{1, 2, 3\}$ with the corresponding graph drawn in Figure 4B. The data-generating

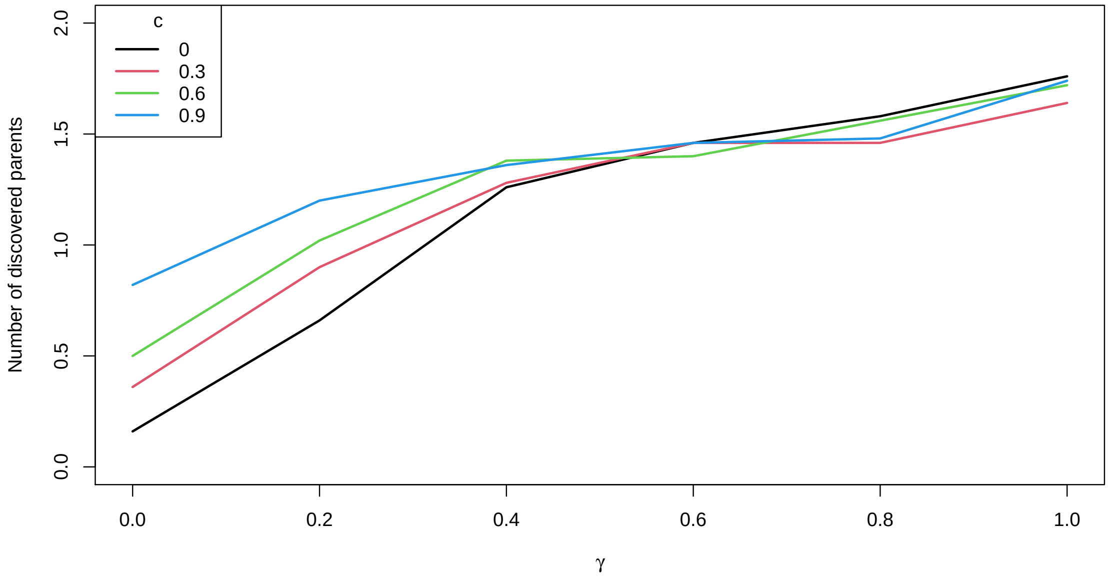

Figure 3: Results of the simulation study corresponding to the additive case of Section 5.1.

process is as follows:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1, c, c \\ c, 1, c \\ c, c, 1 \end{pmatrix} \right), \quad Y = g_Y(X_1, X_2, X_3) + \eta_Y, \quad \text{where } \eta_Y \sim N(0,1),$$

for $c = 0.5$ and a fixed function g_Y generated using the Perlin noise approach. The challenge is to estimate as many direct causes of Y as possible.

The third benchmark dataset consists of $\mathbf{X} = (X_1, X_2, X_3)^\top$ and the response variable Y corresponding to the DAG C of Figure 4. Here, every edge is randomly oriented; either $\to$ or $\leftarrow$ with probability $\frac{1}{2}$. The source variables (variables without parents) are generated following the standard Gaussian distribution. Y is generated as (5) with the Pareto distribution F with a fixed function $\theta(\mathbf{X}_{pa_Y})$ generated using the Perlin noise approach. Finally, if X_i is the effect of Y, it is generated as $X_i = f_i(Y, \eta_i)$ where $\eta_i \sim U(0,1)$, $\eta_i \perp\!\!\!\perp Y$ and f_i is a fixed function generated as a combination of functions generated using the Perlin noise approach.

In all data-sets, we consider a sample size of $n = 500$.

We compare our proposed algorithms with specific methods for causal discovery, which are: RESIT (Peters et al., 2014), CAM-UV (Maeda and Shimizu, 2021), pairwise bQCD (Tagasovska et al., 2020), pairwise IGCI with the Gaussian reference measure (Janzing and Schölkopf, 2010), pairwise Slope (Marx and Vreeken, 2017). When we use the term "pairwise", we are referring to orienting each edge between (X_i, Y) separately, $i = 1, \dots, p$.

For evaluating the performance, we simulate 100 repetitions of each of the three benchmark datasets and use two metrics: "percentage of discovered correct causes" and "percentage of no false positives" which measures the percentage of cases with no incorrect variable in the set of estimated causes. As an example, consider $pa_Y = \{1,2,3\}$. If we estimate $\widehat{pa}_Y = \{1,2\}$ in 80% of cases and $\widehat{pa}_Y = \{1,4,5\}$ in 20% of cases, the percentage of discovered correct causes is $\frac{2}{3}\frac{8}{10} + \frac{1}{3}\frac{2}{10} \approx 60\%$ and the percentage of no false positives is 80%.

Table 2 shows the performance of all methodologies. We observe that our two algorithms outperform the other approaches by a significant margin. The IDE algorithm never includes a wrong covariate in the set of causes. On the other hand, although the scoring algorithm demonstrates better overall performance and power, it tends to include non-causal variables more frequently.

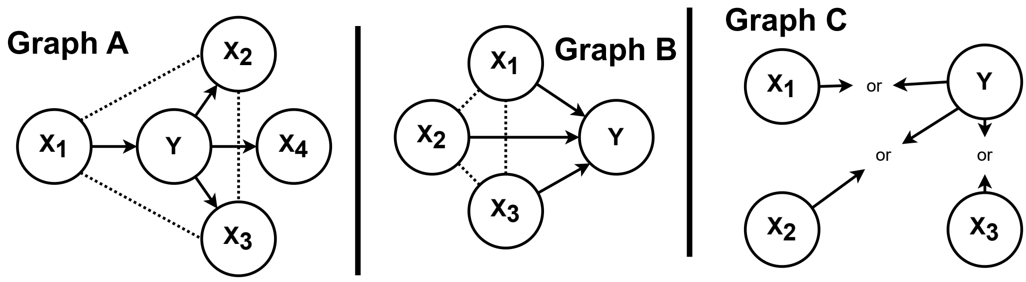

Figure 4: Graph A corresponds to the first benchmark dataset, where the noises of X_1, X_2, X_3 are correlated (denoted by dashed lines). Graph B corresponds to the second benchmark dataset, where (X_1, X_2, X_3) is generated as a correlated multivariate normally distributed random vector. Graph C corresponds to the third benchmark dataset, where each edge is randomly oriented, either $\rightarrow$ or $\leftarrow$ with probability $\frac{1}{2}$.

	First Benchmark	**Second Benchmark**	**Third Benchmark**	**Total (Mean)**
IDE algorithm	98%/ 100%	82%/ 100%	72%/ 100%	84%/ 100%
Score algorithm	100%/ 100%	98%/ 100%	92%/ 88%	93%/ 96%
RESIT	52%/ 18%	36%/ 100%	2%/ 94%	30%/ 71%
CAM-UV	96%/ 40%	2%/ 100%	0%/ 100%	32%/ 80%
Pairwise bQCD	100%/ 0%	56%/ 100%	80%/ 26%	78%/ 42%
Pairwise IGCI	0%/ 48%	0%/ 100%	70%/ 34%	17%/ 70%
Pairwise Slope	100%/ 2%	100%/ 100%	100%/ 22%	100%/ 41%

Table 2: Performance of different algorithms on the three benchmark datasets of Section 5.2. The first number represents the "percentage of discovered correct causes", and the second is the "percentage of no false positives".

5.3 Real-world example

To illustrate our methodology on a real-world example, we consider data on the fertility rate like in Heinze-Deml et al. (2018). The target variable of interest is Y = 'Fertility rate' measured yearly in more than 200 countries. Developing countries exhibit a significantly higher fertility rate when compared to Western countries (Cheng et al., 2022). The fertility rate can be predicted by considering covariates such as the 'infant mortality rate' or 'GDP'. However, if one wants to explore the potential effect of a particular law or a policy change, it becomes necessary to leverage the causal knowledge of the underlying system.

Randomized studies are not possible to design in this context since factors like 'infant mortality rate' cannot be isolated for manipulation. Even so, understanding the impact of policies to reduce infant mortality rates within a country remains an important question, even if randomized studies are unfeasible.

Here, we consider covariates $\mathbf{X} = (X_1, X_2, X_3, X_4)^\top$, where X_1 ='GDP (in US dollars)', X_2 ='Education expenditure (% of GDP)', X_3 ='Infant mortality rate (infant deaths per 1,000 live births)', X_4 ='Continent'. The data come from World Bank (a,b); United Nations. Visualizations of the data can be found in Appendix D.2.

Explaining changes in fertility rate is still a topical issue for which no apparent rational explanation exists. In our study, we focus on using our developed framework to provide data-driven answers about the potential causes of changes in fertility rates.

In the study by Heinze-Deml et al. (2018), the authors employed a method called invariant causal prediction (ICP, Peters et al. (2016)) to determine the possible causes of Y. The ICP approach relies on observing an environmental variable (closely related to a context variable or an

$\mathcal{F}$	$\mathcal{F}$-plausible sets	ISD estimate of the $\mathcal{F}$-identifiable set $\hat{S}_{\mathcal{F}}(Y)$	Score-based estimate of $\widehat{pa}_Y$
$\mathcal{F}_{LS}$	{2,3}, {3,4}, {1,2,3}, {1,3,4}	{3}	{2,3}
$\mathcal{F}_{F_1}$	{2,3}, {2,3,4}, {1,2,3}, {1,3,4}	{3}	{1,2,3}
$\mathcal{F}_{F_2}$	{3}, {2,3}, {3,4}, {1,2,3}, {1,3,4}, {2,3,4},{1,2,3,4}	{3}	{1, 2,3,4} (with almost equal scores for {1,2,3}, {2.3.4})

Table 3: Resulting estimates of causes of changes in fertility rates from real data of Section 5.3. Here, F_1 is a Gaussian distribution, and F_2 is a Gamma distribution.

instrumental variable Mooij et al. (2020)). However, selecting a suitable environmental variable can be a subject of debate, and in many cases, it may be challenging to identify one. Nevertheless, we acknowledge the potential benefits of utilizing such a variable, as it can enhance the reliability of the results.

We apply the methodology developed in this paper to estimate the causes of Y. We tried several choices of $\mathcal{F}$; in particular $\mathcal{F}_A, \mathcal{F}_{LS}, \mathcal{F}_{F_1}, \mathcal{F}_{F_2}, \mathcal{F}_{F_3}$ for the choices $F_1, F_2, F_3 = Gaussian$, $Gamma$, $Pareto$, respectively. These candidate choices come from a preliminary inference on the marginal distributions of the data. For the choice $\mathcal{F} = \mathcal{F}_A$, or $\mathcal{F}_{F_3}$, we observe that all sets $S \subseteq \{1, \dots, 4\}$ are strongly rejected as $\mathcal{F}$-plausible and our estimate is an empty set. One reason for this is the restricted aspects provided by these specific choices of $\mathcal{F}$; our data show much more complex relations than those that can be described by just one parameter (the mean in the case $\mathcal{F}_A$ and the tail index in the case $\mathcal{F}_{F_3}$).

Applying our methodology with the choices $\mathcal{F}_{LS}, \mathcal{F}_{F_1}, \mathcal{F}_{F_2}$, we obtain the results described in Table 3. The results suggest that X_3 is the identifiable cause of Y. This is in line with findings from Heinze-Deml et al. (2018) (backed up by research from sociology Hirschman (1994)), who also discover the variable X_3 to be causal. Furthermore, the score-based estimate indicates that X_2 is a member of $\widehat{pa}_Y$ across all three selections of $\mathcal{F}$. This suggests that X_2 is a cause of Y as well, even though the score-based estimate does not have the same guarantees as the set $\hat{S}_{\mathcal{F}}(Y)$. Note that sets $\{2, 3\}, \{1, 2, 3\}$ are $\mathcal{F}$-plausible for all three choices of $\mathcal{F}$.

We emphasize that the findings rely on the causal sufficiency of the variables employed; an assumption that can surely be questioned. For instance, other variables such as 'religious beliefs' or a 'political situation' may explain the fertility rate, but are hard to measure.

6 Conclusion and future work

In this work, we have discussed the problem of estimating direct causes of a target variable Y from an observational joint distribution. We introduced the concept of $\mathcal{F}$-identifiability, a crucial concept that describes the variables that can be inferred as causal under the assumption $f_Y \in \mathcal{F}$. Our theory mainly focuses on describing the set of the $\mathcal{F}$-identifiable parents of Y. We have explored various choices of $\mathcal{F}$ and develop a formal theory to determine when it is possible to identify all causes of Y. Additionally, we have presented two algorithms for estimating the set of parents of Y from a random sample. To evaluate the performance of our algorithms, we have created several benchmark datasets and demonstrated their effectiveness. Furthermore, we applied our methodology to a real-world problem related to the causes of the fertility rate.

In this study, we attempted to leverage established results from the theory of identifiability of the entire graph $\mathcal{G}$, but numerous research directions still need to be explored. Our theoretical results apply only to covariates that are neighbours of the target variable. Do our findings encompass non-adjacent variables as well? One limitation of our framework is the necessity of a choice of $\mathcal{F}$. Employing an automatic model selection approach could enhance performance and expand the applicability of this approach. Modifying other causal discovery methodologies on a local scale can also be an interesting future research direction.

Ultimately, although our findings can provide valuable insights into the causal structure, further research is necessary to thoroughly assess the practical applicability of our approach in real-world settings. We believe that the developed theory in this work can be beneficial for a deeper understanding of causal structure and the theoretical limitations of purely data-driven methodologies for causal inference.

Conflict of interest and data availability

The code and data are available in the repository, or upon request from the authors.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work was supported by the Swiss National Science Foundation.

A Appendix

A.1 Class of invertible functions, invertible causal model and causal minimality

First, let us formally introduce a class of measurable functions $\mathcal{I}_m$ and a subclass of SCM called invertible causal models.

Definition 4. *Let $\mathcal{X}_x \subseteq \mathbb{R}^p, \mathcal{X}_y \subseteq \mathbb{R}, \mathcal{X}_z \subseteq \mathbb{R}$ be measurable sets. A measurable function $f : \mathcal{X}_x \times \mathcal{X}_y \to \mathcal{X}_z$ is called* ***invertible for the last element****, notation $f \in \mathcal{I}$, if there exists a function $f^{\leftarrow} : \mathcal{X}_x \times \mathcal{X}_z \to \mathcal{X}_y$ that fulfills the following: $\forall \boldsymbol{x} \in \mathcal{X}_x, \forall y \in \mathcal{X}_y, z \in \mathcal{X}_z$ such that $y = f(\boldsymbol{x}, z)$, then $z = f^{\leftarrow}(\boldsymbol{x}, y)$. Moreover, denote*

$$\mathcal{I}_m = \{f \in \mathcal{I} : \ f \textit{ is not constant in any of its arguments}\}.$$

We define the ***ICM*** *(invertible causal model) as a SCM (4) with structural equations satisfying $f_i \in \mathcal{I}_m$ for all $i = 0, \dots, p$.*

The previous definition describes that the element z in a relation $y = f(\mathbf{x}, z)$ can be uniquely recovered from $(\mathbf{x}, y)$. To provide an example, for the function $f(x, z) = x + z$ holds $f^{\leftarrow}(x, y) = y - x$, since $f^{\leftarrow}(x, f(x, z)) = f(x, z) - x = z$. More generally, for the additive function defined as $f(\mathbf{x}, z) = g_1(\mathbf{x}) + g_2(z)$, holds $f^{\leftarrow}(\mathbf{x}, y) = g_2^{-1}(y - g_1(\mathbf{x}))$, if the inverse g^{-1} exists, $\mathbf{x} \in \mathbb{R}^d, y, z \in \mathbb{R}$. Overall, if f is differentiable and the partial derivative of $f(\mathbf{x}, z)$ with respect to z is monotonic, then $f \in \mathcal{I}$ (follows from Inverse function theorem Baxandall and Liebeck (1986)). Note that $f_i \in \mathcal{I}_m$ implies causal minimality of the ICM model, as the following lemma suggests.

Lemma A.1. *Consider a distribution generated by ICM with graph $\mathcal{G}_0$ (see Definition 4). Let all structural equations $f_j \in \mathcal{I}_m, \forall j = 0, \dots, p$. Then, the distribution is causally minimal with respect to $\mathcal{G}_0$. Conversely, if there exists $f_j \in \mathcal{I} \setminus \mathcal{I}_m$, then the causal minimality is violated.*

Proof. The second claim follows directly from Proposition 4 in Peters et al. (2014). As for the first claim, we use a similar approach as in Proposition 17 from Peters et al. (2014).

Let $f_j \in \mathcal{I}_m$ for all $j = 0, \dots, p$ and let the causal minimality be violated, i.e. let $\tilde{\mathcal{G}}$ be a subgraph of $\mathcal{G}_0$ such that the distribution is Markov wrt. $\tilde{\mathcal{G}}$. Find $i, j \in \mathcal{G}_0$ such that $i \to j$ in $\mathcal{G}_0$ but $i \not\to j$ in $\tilde{\mathcal{G}}$.

In graph $\mathcal{G}_0$ we have a structural equation $X_j = f_j(\mathbf{X}_{pa_j(\mathcal{G}_0)}, \varepsilon_j) = f_j(X_i, \mathbf{X}_{pa_j(\mathcal{G}_0)\setminus\{i\}}, \varepsilon_j)$ but in $\tilde{\mathcal{G}}$ we have $X_j = \tilde{f}_j(\mathbf{X}_{pa_j(\mathcal{G}_0)\setminus\{i\}}, \varepsilon_j)$. Hence, functions $f_j(X_i, \mathbf{X}_{pa_j(\mathcal{G}_0)\setminus\{i\}}, \varepsilon_j)$ and $\tilde{f}_j(\mathbf{X}_{pa_j(\mathcal{G}_0)\setminus\{i\}}, \varepsilon_j)$ has to be equal when we condition on $(X_i, \mathbf{X}_{pa_j(\mathcal{G}_0)\setminus\{i\}}) = (x_i, \mathbf{x})$. Then, in order for functions $f_j(x_i, \mathbf{x}, \varepsilon_j)$ and $\tilde{f}_j(\mathbf{x}, \varepsilon_j)$ to be equal (stressing out that $\varepsilon_j \perp\!\!\!\perp \mathbf{X}_{pa_j}$), f_j can not depend on its first argument. This contradicts $f \in \mathcal{I}_m$. $\square$

A.2 Pairwise identifiability

In the following, we define the concept of pairwise identifiability of an $\mathcal{F}$-model.

Definition 5. *Let $(X_0, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow a SCM (4) with DAG $\mathcal{G}_0$. Let $\mathcal{F}$ be a subset of all measurable functions. We say that the $\mathcal{F}$-model is **identifiable**, if there does not exist a graph $\mathcal{G}' \neq \mathcal{G}_0$ and functions $f_i' \in \mathcal{F}, i = 0, \dots, p$ generating the same joint distribution.*

*We say that the $\mathcal{F}$-model is **pairwise identifiable**, if for all $i, j \in \mathcal{G}_0, i \in pa_j$ hold the following: $\forall S \subseteq V$ such that $pa_j \setminus \{i\} \subseteq S \subseteq nd_j \setminus \{i, j\}$ there exist $\boldsymbol{x}_S : p_S(\boldsymbol{x}_S) > 0$ satisfying that a bivariate model defined as $Z_1 = \tilde{\varepsilon}_1, Z_2 = \tilde{f}(Z_1, \varepsilon_j)$ is identifiable, where $P_{\tilde{\varepsilon}_1} = P_{X_i | \boldsymbol{X}_S = \boldsymbol{x}_S}$, $\tilde{f}(x, \varepsilon) = f(\boldsymbol{x}_{pa_j \setminus \{i\}}, x, \varepsilon)$, $\tilde{\varepsilon}_1 \perp\!\!\!\perp \varepsilon_j$.*

A similar notion was mentioned in the concept of additive noise models (Peters et al., 2014, Definition 27). Here, we assume an arbitrary $\mathcal{F}$-model. In the bivariate case, the notion of identifiability and pairwise identifiability trivially coincides. Note the following observation.

Lemma A.2. *Pairwise identifiable $\mathcal{F}$-model is identifiable.*

Proof. For a contradiction, let there be two $\mathcal{F}$-models with causal graphs $\mathcal{G} \neq \mathcal{G}'$ that both generate the same joint distribution $P_{(X_0, \mathbf{X})}$. Using Proposition 29 in Peters et al. (2014), there exists variables $L, K \in \{X_0, \dots, X_p\}$ such that

- $K \to L$ in $\mathcal{G}$ and $L \to K$ in $\mathcal{G}'$,
- $S := \underbrace{\{pa_L(\mathcal{G}) \setminus \{K\}\}}_{\mathbf{Q}} \cup \underbrace{\{pa_K(\mathcal{G}') \setminus \{L\}\}}_{\mathbf{R}} \subseteq \{nd_L(\mathcal{G}) \cap nd_K(\mathcal{G}') \setminus \{K, L\}\}$.

For this S, choose x_S according to the condition in the definition of pairwise identifiability. We use the notation $x_S = (x_q, x_r)$ where $q \in \mathbf{Q}, r \in \mathbf{R}$ and we define $K^\star := K \mid \{X_S = x_S\}$ and $L^\star := L \mid \{X_S = x_S\}$. Now we use Lemma 36 and Lemma 37 from Peters et al. (2014). Since $K \to L$ in $\mathcal{G}$, we get

$$K^\star = \tilde{\varepsilon}_{K^\star}, \qquad L^\star = f_{L^\star}(K^\star, \varepsilon_L),$$

where $\tilde{\varepsilon}_{K^\star} = K \mid \{X_S = x_S\}$ and $\varepsilon_L \perp\!\!\!\perp K^\star$. We obtained a bivariate $\mathcal{F}$-model with $K^\star \to L^\star$. However, the same holds for the other direction; From $L \to K$ in $\mathcal{G}'$, we get

$$L^\star = \tilde{\varepsilon}_{L^\star}, \qquad K^\star = f_{K^\star}(L^\star, \varepsilon_K),$$

where $\tilde{\varepsilon}_{L^\star} = L \mid \{X_S = x_S\}$ and $\varepsilon_K \perp\!\!\!\perp L^\star$. We obtained a bivariate $\mathcal{F}$-model with $L^\star \to K^\star$, which is a contradiction. □

A.3 Post-additive/post-multiplicative/location-scale distributions

Consider a distribution function F with $q \in \mathbb{N}$ parameters $\theta = (\theta_1, \dots, \theta_q)^\top$.

Definition 6. *Let F be a distribution function with one ($q = 1$) parameter θ. We say that the **parameter acts post-additively** in F, if there exists an invertible real function f_2 and a function $f_1 \in \mathcal{I}_m$ such that for all θ_1, θ_2 holds* [3]

$$F_{\theta_1}\big(F_{\theta_2}^{-1}(z)\big) = f_1\big(z, f_2(\theta_1) + \theta_2\big), \ \ \forall z \in (0, 1). \tag{14}$$

*We say that the **parameter acts post-multiplicatively** in F, if there exist an invertible real function f_2 and a function $f_1 \in \mathcal{I}_m$ such that for all θ_1, θ_2 holds*

$$F_{\theta_1}\big(F_{\theta_2}^{-1}(z)\big) = f_1\big(z, f_2(\theta_1) \cdot \theta_2\big), \ \ \forall z \in (0, 1). \tag{15}$$

*Let F be a distribution function with two ($q = 2$) parameters $\theta = (\mu, \sigma)^\top \in \mathbb{R} \times \mathbb{R}_+$. We say that F is a **Location-Scale** distribution, if for all θ holds*

$$F_\theta\left(\frac{x - \mu}{\sigma}\right) = F_{\theta_0}(x), \quad \forall x \in \mathbb{R},$$

where F_{θ_0} is called standard distribution and corresponds to a parameter $\theta_0 = (0, 1)^\top$.

[3] Notation $F_{\theta_1}\big(F_{\theta_2}^{-1}(z)\big)$ is equivalent to $F(F^{-1}(z, \theta_2), \theta_1)$. We believe that it improves the readability.

Examples of F whose parameter acts post-additively include a Gaussian distribution with fixed variance or a Logistic distribution/Gumbel distribution with fixed scales. Note that typically $f_2(x) = -x$ since $F_{\theta_1}\big(F_{\theta_1}^{-1}(z)\big) = z$ needs to hold.

Examples of F whose parameter acts post-multiplicatively include a Gaussian distribution with the fixed expectation or a Pareto distribution (where $F_{\theta_1}\big(F_{\theta_2}^{-1}(z)\big) = z^{\frac{\theta_1}{\theta_2}} = f_1\big(z, f_2(\theta_1)\cdot\theta_2\big)$ for $f_1(z,x) = z^{-1/x}$ and $f_2(x) = -1/x$). Functions f_1, f_2 are not necessarily uniquely defined.

Examples of Location-Scale types of distributions include Gaussian distribution, logistic distribution, or Cauchy distribution, among many others.

B Auxiliary results

The median $med(X)$ is defined as $m \in \mathbb{R}$ such that $P(X \leq m) \geq \frac{1}{2} \leq P(X \geq m)$. It always exists, and is unique for continuous random variables.

Lemma B.1. *Let X be a non-degenerate continuous real random variable. Let $a, b \in \mathbb{R}$ such that*

$$a + bX \overset{D}{=} X. \tag{16}$$

Then, either $(a,b) = (0,1)$ or $(a,b) = (2med(X), -1)$.

Proof. Idea of the proof assuming a finite variance of X: If X has finite variance, then (16) implies $var(a+bX) = var(X)$, rewriting gives us $b^2 var(X) = var(X)$ and hence, $b = \pm 1$. Now, (16) also implies $\mathbb{E}(a+bX) = \mathbb{E}(X)$ hence $a = (1-b)\mathbb{E}(X)$. Therefore if $b = 1$ then $a = 0$, and if $b = -1$ then $a = 2\mathbb{E}(X)$.

Proof without the moment assumption: (16) implies that for any $q \in (0.5, 1)$, the difference between the q quantile and $(1-q)$ quantile should be the same on both sides of (16). Denote $F_X^{-1}(q)$ a $q-$quantile of X and assume that $F_X^{-1}(q) \neq F_X^{-1}(1-q)$ (since X is non-degenerate, there exist such q). We get

$$F_{a+bX}^{-1}(q) - F_{a+bX}^{-1}(1-q) = F_X^{-1}(q) - F_X^{-1}(1-q) =: D.$$

Consider $b \geq 0$. Using linearity of the quantile function, we obtain $a + bF_X^{-1}(q) - \big(a + bF_X^{-1}(1-q)\big) = D$ and hence, $bD = D$ which gives us $b = 1$. If $b < 0$ then an identity $F_{a+bX}^{-1}(q) = a + \big(1 - F_{-bX}^{-1}(1-q)\big) = a + \big(1 + bF_X^{-1}(1-q)\big)$ hold. Hence, we get $a + [1 + bF_X^{-1}(1-q)] - [a + \big(1 + bF_X^{-1}(1-q)\big)] = D$. Rewriting the left side we get $-bD = D$ which gives us $b = -1$.

In the case when $b = 1$, trivially $a = 0$ since otherwise $med(a+X) \neq med(X)$. If $b = -1$, then applying median on both sides of (16) gives us $med(a - X) = med(X)$ and hence, $a = 2med(X)$ as we wanted to show. ☐

Lemma B.2. *Let $\boldsymbol{X} = (X_1, \dots, X_k)$ be a continuous random vector with independent components and $s < k$. Let $h_1, \dots, h_k$ be continuous non-constant real functions.*

1. *There does not exist a non-zero function f such that*

$$f(X_1, \dots, X_s)\big(h_1(X_1) + \dots + h_k(X_k)\big) \perp\!\!\!\perp (X_1, \dots, X_s). \tag{17}$$

2. *Moreover, let $h_1, \dots, h_k$ be non-zero. Then, there does not exist a non-zero function f such that*

$$f(X_1, \dots, X_s) + h_1(X_1)h_2(X_2)\dots h_k(X_k) \perp\!\!\!\perp (X_1, \dots, X_s). \tag{18}$$

3. *Let $h : \mathbb{R}^{k-s} \to \mathbb{R}$ be measurable function such that $h(X_{s+1}, \dots, X_k)$ is non-degenerate continuous random variable. There do not exist functions f_1, f_2, where f_2 is positive non-constant, such that*

$$f_1(X_1, \dots, X_s) + f_2(X_1, \dots, X_s)h(X_{s+1}, \dots, X_k) \perp\!\!\!\perp (X_1, \dots, X_s) \tag{19}$$

4. *Let $\boldsymbol{X} = (X_1, \dots, X_k)$ be non-degenerate Gaussian random vector (possibly with dependent components) and $\beta = (\beta_{s+1}, \dots, \beta_k)^\top \in \mathbb{R}^{k-s}$. There do not exist functions f_1, f_2, where f_2 is positive non-constant, such that*

$$f_1(X_1, \dots, X_s) + f_2(X_1, \dots, X_s)(\beta_{s+1}X_{s+1} + \dots + \beta_k X_k) \perp\!\!\!\perp (X_1, \dots, X_s) \tag{20}$$

Proof. Let us introduce functionals (not a norm) $||\cdot||_{plus}$ and $||\cdot||_{times}$, defined by $||\mathbf{a}||_{plus} = a_1 + \dots + a_d$, $||\mathbf{a}||_{times} = a_1 a_2 \dots a_d$, for $\mathbf{a} = (a_1, \dots, a_d)^\top \in \mathbb{R}^d$.

We use notation $\mathbf{X}_S = (X_1, \dots, X_s)^\top$ and denote a function $h_S : \mathbb{R}^s \to \mathbb{R}^s : h_S(\mathbf{x}) = (h_1(x_1), \dots, h_s(x_s))^\top$.

Part 1: For a contradiction, let such f exist. First, some notation: Let $Y = h_{s+1}(X_{s+1}) + \dots + h_k(X_k)$ and define $\xi := f(\mathbf{X}_S)(||h_S(\mathbf{X}_S)||_{plus} + Y)$, which is the left hand side of (17).

Choose $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^s$ in the support of $\mathbf{X}_S$ such that $||h_S(\mathbf{a})||_{plus}, ||h_S(\mathbf{b})||_{plus}, ||h_S(\mathbf{c})||_{plus}$ are distinct and $f(\mathbf{b}) \neq 0$ (it is possible since h_i are non-constant).

Since $\xi \perp\!\!\!\perp \mathbf{X}_S$, then $\xi \mid [\mathbf{X}_S = \mathbf{a}] \overset{D}{=} \xi \mid [\mathbf{X}_S = \mathbf{b}] \overset{D}{=} \xi \mid [\mathbf{X}_S = \mathbf{c}]$. Hence,

$$f(\mathbf{a})(||h_S(\mathbf{a})||_{plus} + Y) \overset{D}{=} f(\mathbf{b})(||h_S(\mathbf{b})||_{plus} + Y) \overset{D}{=} f(\mathbf{c})(||h_S(\mathbf{c})||_{plus} + Y). \tag{21}$$

By dividing by a nonzero constant $f(\mathbf{b})$ and subtracting a constant $||h_S(\mathbf{b})||_{plus}$, we get

$$\frac{f(\mathbf{a})}{f(\mathbf{b})}||h_S(\mathbf{a})||_{plus} - ||h_S(\mathbf{b})||_{plus} + \frac{f(\mathbf{a})}{f(\mathbf{b})}Y \overset{D}{=} Y \overset{D}{=} \frac{f(\mathbf{c})}{f(\mathbf{b})}||h_S(\mathbf{c})||_{plus} - ||h_S(\mathbf{b})||_{plus} + \frac{f(\mathbf{c})}{f(\mathbf{b})}Y.$$

Now we use lemma B.1. It gives us that $\frac{f(\mathbf{a})}{f(\mathbf{b})} = \pm 1$ and also $\frac{f(\mathbf{c})}{f(\mathbf{b})} = \pm 1$. Therefore, at least two values of $f(\mathbf{a}), f(\mathbf{b}), f(\mathbf{c})$ has to be equal (and neither of them are zero). WLOG $f(\mathbf{a}) = f(\mathbf{c})$. Plugging this into equation (21), we get $||h_S(\mathbf{a})||_{plus} = ||h_S(\mathbf{c})||_{plus}$ which is a contradiction since we chose them distinct.

Part 2: We proceed very similarly as in the previous part. For a contradiction, let such f exist. First, some notation: let $Y = h_{s+1}(X_{s+1}) \dots h_k(X_k)$ and define $\xi := f(\mathbf{X}_S) + (||h_S(\mathbf{X}_S)||_{times} \cdot Y)$, which is the left hand side of (18).

Choose $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^s$ in the support of $\mathbf{X}_S$ such that $||h_S(\mathbf{a})||_{times}, ||h_S(\mathbf{b})||_{times}, ||h_S(\mathbf{c})||_{times}$ are distinct and $f(\mathbf{b}) \neq 0$.

Since $\xi \perp\!\!\!\perp \mathbf{X}_S$, then $\xi \mid [\mathbf{X}_S = \mathbf{a}] \overset{D}{=} \xi \mid [\mathbf{X}_S = \mathbf{b}] \overset{D}{=} \xi \mid [\mathbf{X}_S = \mathbf{c}]$. Hence,

$$f(\mathbf{a}) + ||h_S(\mathbf{a})||_{times} \cdot Y \overset{D}{=} f(\mathbf{b}) + ||h_S(\mathbf{b})||_{times} \cdot Y \overset{D}{=} f(\mathbf{c}) + ||h_S(\mathbf{c})||_{times} \cdot Y. \tag{22}$$

By dividing by a nonzero constant $f(\mathbf{b})$ and subtracting a constant $||h_S(\mathbf{b})||_{times}$, we get

$$\frac{f(\mathbf{a})}{f(\mathbf{b})}||h_S(\mathbf{a})||_{times} - ||h_S(\mathbf{b})||_{times} + \frac{f(\mathbf{a})}{f(\mathbf{b})}Y \overset{D}{=} Y \overset{D}{=} \frac{f(\mathbf{c})}{f(\mathbf{b})}||h_S(\mathbf{c})||_{times} - ||h_S(\mathbf{b})||_{times} + \frac{f(\mathbf{c})}{f(\mathbf{b})}Y.$$

Now we use lemma B.1. It gives us that $\frac{f(\mathbf{a})}{f(\mathbf{b})} = \pm 1$ and also $\frac{f(\mathbf{c})}{f(\mathbf{b})} = \pm 1$. Therefore, at least two values of $f(\mathbf{a}), f(\mathbf{b}), f(\mathbf{c})$ has to be equal (and neither of them are zero). WLOG $f(\mathbf{a}) = f(\mathbf{b})$. Plugging this into equation (22), we get $||h_S(\mathbf{a})||_{times} = ||h_S(\mathbf{b})||_{times}$ which is a contradiction since we chose them distinct.

Part 3: For a contradiction, let f_1, f_2 exist. Denote $Y = h(X_{s+1}, \dots, X_k)$. Choose $\mathbf{a}, \mathbf{b} \in \mathbb{R}^s$ in the support of $\mathbf{X}_S$ such that $f_2(\mathbf{a}) \neq f_2(\mathbf{b}) \neq 0$. From (19), we get $f_1(\mathbf{a}) + f_2(\mathbf{a})Y \overset{D}{=} f_1(\mathbf{b}) + f_2(\mathbf{b})Y$. By rewriting we get $\frac{f_1(\mathbf{a}) - f_1(\mathbf{b})}{f_2(\mathbf{b})} + \frac{f_2(\mathbf{a})}{f_2(\mathbf{b})}Y \overset{D}{=} Y$. Applying Lemma B.1 we obtain $\frac{f_2(\mathbf{a})}{f_2(\mathbf{b})} = \pm 1$. Since f_2 is positive, we get $f_2(\mathbf{a}) = f_2(\mathbf{b})$. That is a contradiction.

Part 4: We use the following well-known result: For a multivariate normal vector

$$\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)^\top \sim N\Bigg(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{1,1}, \Sigma_{1,2} \\ \Sigma_{2,1}, \Sigma_{2,2} \end{pmatrix} \Bigg),$$

where $\mathbf{Z}_1, \mathbf{Z}_2$ is a partition of $\mathbf{Z}$ into smaller sub-vectors, holds that $(\mathbf{Z}_1 | \mathbf{Z}_2 = a)$, the conditional distribution of the first partition given the second, has distribution equal to $N(\mu_a, \tilde{\Sigma})$, where

$\mu_a = \mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(a - \mu_2)$, $\tilde{\Sigma} = \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}$. Specially, the covariance structure does not depend on a.

Proof of Part 4: For a contradiction, let f_1, f_2 exist. Choose $\mathbf{a}, \mathbf{b} \in \mathbb{R}^s$ in the support of $\mathbf{X}_S$ such that $f_2(\mathbf{a}) \neq f_2(\mathbf{b}) \neq 0$. Denote $Y = \beta_{s+1}X_{s+1} + \cdots + \beta_k X_k$ and $Y_a := (Y \mid \mathbf{X}_S = a)$ and $Y_b := (Y \mid \mathbf{X}_S = b)$ (do not mistake the notation with a do-intervention; Y_a is simply a conditional distribution of Y given $\mathbf{X}_S = a$). Notice that Y is also Gaussian.

The well-known result gives us that $\mu_a + Y_a \overset{D}{=} \mu_b + Y_b$, where μ_a, μ_b are constants depending on the mean and covariance structure of $\mathbf{X}$, and on a, b.

From (20), we get

$$f_1(\mathbf{a}) + f_2(\mathbf{a})Y_a \overset{D}{=} f_1(\mathbf{b}) + f_2(\mathbf{b})Y_b.$$

By rewriting we get $\frac{f_1(\mathbf{a}) - f_1(\mathbf{b}) + f_2(a)(\mu_a - \mu_b)}{f_2(\mathbf{b})} + \frac{f_2(\mathbf{a})}{f_2(\mathbf{b})}Y_b \overset{D}{=} Y_b$. Applying Lemma B.1 we obtain $\frac{f_2(\mathbf{a})}{f_2(\mathbf{b})} = \pm 1$. Since f_2 is positive, we get $f_2(\mathbf{a}) = f_2(\mathbf{b})$. That is a contradiction.

□

Lemma B.3. *Let X, Y be continuous random variables and f is a (non-random) strictly increasing function. Then,*

$$X \perp\!\!\!\perp Y \iff f(X) \perp\!\!\!\perp Y. \tag{23}$$

Proof. This statement is trivial. □

C Proofs for Section 3

Lemma 1. *Let $(Y, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an $\mathcal{F}_L$-model with DAG $\mathcal{G}_0$ and $pa_Y(\mathcal{G}_0) \neq \emptyset$. Then, $|S_{\mathcal{F}_L}(Y)| \leq 1$ ($|S|$ represent the number of elements of the set S). Moreover, if there exist $a, b \in an_Y(\mathcal{G}_0)$ that are d-separated in $\mathcal{G}_0$, then $S_{\mathcal{F}_L}(Y) = \emptyset$.*

Proof. First, we will show $|S_{\mathcal{F}_L}(Y)| \leq 1$. Let $a \in an_Y(\mathcal{G}_0) \cap Source(\mathcal{G}_0)$. Such a exists since $pa_Y(\mathcal{G}_0) \neq \emptyset$. We will show that $S = \{a\}$ is $\mathcal{F}_L-$plausible set.

Denote $X_0 := Y$. Since $\mathcal{G}_0$ is acyclic, it is possible express recursively each variable $X_j, j = 0, \ldots, p$, as a weighted sum of the noise terms $\varepsilon_0, \ldots, \varepsilon_p$ that belong to the ancestors of X_j. Let us write the Linear SCM with notation

$$X_i = \sum_{j \in pa_i} \beta_{j,i}X_j + \varepsilon_i = \sum_{j \in an_i} \beta_{j \to i}\varepsilon_j,$$

where $\beta_{j,i}$ are non-zero constants and $\beta_{j \to i}$ is the sum of distinct weighted directed paths from node j to node i, with a convention $\beta_{j \to j} := 1$. [4]

Using this notation, note that

$$X_0 = \sum_{j \in an_0} \beta_{j \to i}\varepsilon_j = \beta_{a \to 0}\varepsilon_a + \sum_{j \in an_0 \setminus \{a\}} \beta_{j \to i}\varepsilon_j = \beta_{a \to 0}X_a + \sum_{j \in an_0 \setminus \{a\}} \beta_{j \to i}\varepsilon_j,$$

where $X_a \perp\!\!\!\perp \sum_{j \in an_i \setminus \{a\}} \beta_{j \to i}\varepsilon_j$ since $a \in Source(\mathcal{G}_0)$. Hence, $Y - \beta_{a \to 0}X_a \perp\!\!\!\perp X_a$, which is almost the definition of $\mathcal{F}_L$-plausibility of set $S = \{a\}$. More rigorously, for $S = \{a\}$ we can find $f \in \mathcal{F}_L$ such that $f_Y^{\leftarrow}(X_S, Y) \perp\!\!\!\perp X_S$ and $f_Y^{\leftarrow}(X_S, Y) \sim U(0, 1)$. This function can be defined as

$$f(x, \varepsilon) = \beta_{a \to 0}x + g^{-1}(\varepsilon), \ \ x \in \mathbb{R}, \varepsilon \in (0, 1),$$

where g is the distribution function of $(Y - \beta_{a \to 0}X_S)$. This function obviously satisfy $f \in \mathcal{F}_L$. Moreover, since $f_Y^{\leftarrow}(X_S, Y) = g(Y - \beta_{a \to 0}X_S)$, it holds that $f_Y^{\leftarrow}(X_S, Y) \perp\!\!\!\perp X_S$ and $f_Y^{\leftarrow}(X_S, Y) \sim U(0, 1)$, what we wanted to show. Hence $|S_{\mathcal{F}_L}(Y)| \leq 1$, since $S_{\mathcal{F}_L}(Y) \subseteq S = \{a\}$.

Now, let $a, b \in an_Y(\mathcal{G}_0)$ that are d-separated in $\mathcal{G}_0$. Let $a', b' \in \mathcal{G}_0$ such that $a' \in \big\{an_a(\mathcal{G}_0) \cup \{a\}\big\} \cap Source(\mathcal{G}_0)$, $b' \in \big\{an_b(\mathcal{G}_0) \cup \{b\}\big\} \cap Source(\mathcal{G}_0)$. They are well defined since the sets $an_a(\mathcal{G}_0) \cup$

[4] To provide an example of the notation, if $X_1 = \varepsilon_1, X_2 = 2X_1 + \varepsilon_2, X_3 = 3X_1 + 4X_2 + \varepsilon_3$, then $X_3 = 11\varepsilon_1 + 4\varepsilon_2 + 1\varepsilon_3 = \beta_{1 \to 3}\varepsilon_1 + \beta_{2 \to 3}\varepsilon_2 + \beta_{3 \to 3}\varepsilon_1$.

$\{a\}$, $an_b(\mathcal{G}_0) \cup \{b\}$ must contain some source node. Since a, b are d-separated, $\{an_a(\mathcal{G}_0) \cup \{a\}\}$ and $\{an_b(\mathcal{G}_0) \cup \{b\}\}$ are disjoint sets and hence, $a' \neq b'$ (they are even d-separated in $\mathcal{G}_0$(Verma and Pearl, 2013)).

Using the same argument as in the first part of the proof, since $a' \in an_Y(\mathcal{G}_0) \cap Source(\mathcal{G}_0)$, it holds that $S = \{a\}$ is $\mathcal{F}_L-$plausible set. Also $S = \{b\}$ is $\mathcal{F}_L-$plausible set since $b' \in an_Y(\mathcal{G}_0) \cap Source(\mathcal{G}_0)$. Together $S_{\mathcal{F}_L}(Y) \subseteq \{a\}$ and $S_{\mathcal{F}_L}(Y) \subseteq \{b\}$. We showed that $S_{\mathcal{F}_L}(Y) = \emptyset$. □

Proposition 1. *Let $(X_0, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an SCM with DAG $\mathcal{G}_0$. Let $S \subseteq \{1, \dots, p\}$ and denote $S_0 = S \cup \{0\}$. Assume that $\mathcal{G} = \mathcal{G}_0[S_0]$, the projection of $\mathcal{G}_0$ on S defined in Section 2, is a DAG. Let S contain a childless child of X_0, i.e. $\exists j \in ch_0(\mathcal{G})$ such that $ch_j(\mathcal{G}) = \emptyset$.*

Let $\mathcal{F} \subseteq \mathcal{I}_m$ and let $(X_0, \boldsymbol{X}_S)$ follow an $\mathcal{F}$-model with graph $\mathcal{G}$, that is pairwise identifiable. Then, S is not $\mathcal{F}$-plausible.

Proof. Recall the definition of pairwise identifiability from Appendix A.2. For a contradiction, let S be $\mathcal{F}$-plausible. The idea of the proof is that we define two bivariate $\mathcal{F}$-models; one with $X_0 \to X_j$ and one with $X_j \to X_0$, which will lead to a contradiction with the pairwise identifiability.

Since $(X_0, \mathbf{X}_S)$ follow an $\mathcal{F}$-model, we can write $X_i = f_i(\mathbf{X}_{pa_i(\mathcal{G})}, \varepsilon_i)$, where $f_i \in \mathcal{F}$ and ε_i are jointly independent, $i \in S_0$. We use the pairwise identifiability condition. For a specific choice (X_0, X_j) (where X_j is a child of X_0), and $\tilde{S} = nd_j(\mathcal{G}) \setminus \{0, j\} = S \setminus \{j\}$ (the second equality holds since j is a childless child), there exist $\mathbf{x}_{\tilde{S}} : p_{\tilde{S}}(\mathbf{x}_{\tilde{S}}) > 0$ satisfying that a bivariate $\mathcal{F}$-model defined as

$$\tilde{X}_0 = \tilde{\varepsilon}_0, \tilde{X}_j = \tilde{f}_j(\tilde{X}_0, \tilde{\varepsilon}_j) \tag{24}$$

is identifiable, where $P_{\tilde{\varepsilon}_0} = P_{X_0 | \mathbf{X}_{\tilde{S}} = x_{\tilde{S}}}$ and $\tilde{f}_j(x, \varepsilon) = f(\mathbf{x}_{pa_j \setminus \{0\}}, x, \varepsilon)$, $\tilde{\varepsilon}_j \perp\!\!\!\perp \tilde{\varepsilon}_0$.

From the fact that S is $\mathcal{F}$-plausible, there exist $f \in \mathcal{F}$ such that $\varepsilon_S := f^{\leftarrow}(\mathbf{X}_S, X_0)$ satisfy $\varepsilon_S \perp\!\!\!\perp \mathbf{X}_S, \varepsilon_S \sim U(0,1)$. Hence, we can define a model

$$\tilde{\tilde{X}}_j = \tilde{\tilde{\varepsilon}}_j, \tilde{\tilde{X}}_0 = \tilde{\tilde{f}}(\tilde{\tilde{X}}_j, \varepsilon_S),$$

where $P_{\tilde{\tilde{\varepsilon}}_j} = P_{X_j | \mathbf{X}_{\tilde{S}} = x_{\tilde{S}}}$ and $\tilde{\tilde{f}}(\mathbf{x}, \varepsilon) = f(\mathbf{x}_{\tilde{S}}, x, \varepsilon)$. In this model, $\varepsilon_S \perp\!\!\!\perp \tilde{\tilde{\varepsilon}}_j$.

Now, note that $(\tilde{X}_0, \tilde{X}_j) \overset{D}{=} (\tilde{\tilde{X}}_0, \tilde{\tilde{X}}_j)$, since both sides are distributed as $[(X_0, X_j) \mid X_{\tilde{S}}]$. This is a contradiction with the identifiability of (24). Therefore, S is not $\mathcal{F}_F$-plausible. □

Proposition 2. *Let $(Y, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow a SCM with DAG $\mathcal{G}_0$. Let $S \subseteq \{1, \dots, p\}$ be a non-empty set. Let $\underline{\Psi}, \overline{\Psi} : \mathbb{R}^{|S|} \to \mathbb{R}$ be real functions such that*

$$supp(Y \mid \boldsymbol{X}_S = \boldsymbol{x}) = (\underline{\Psi}(\boldsymbol{x}), \overline{\Psi}(\boldsymbol{x})), \quad \forall \boldsymbol{x} \in supp(\boldsymbol{X}_S).$$

Moreover, let

$$\frac{Y - \underline{\Psi}(\boldsymbol{X}_S)}{\overline{\Psi}(\boldsymbol{X}_S) - \underline{\Psi}(\boldsymbol{X}_S)} \not\perp\!\!\!\perp \boldsymbol{X}_S. \tag{8}$$

Then, S is not $\mathcal{F}_{LS}$-plausible.

Proof. For a contradiction, let S be $\mathcal{F}_{LS}$-plausible. Hence, there exists $f \in \mathcal{F}_{LS}$ such that

$$f^{\leftarrow}(\mathbf{X}_S, Y) \perp\!\!\!\perp \mathbf{X}_S \tag{25}$$

Since $f \in \mathcal{F}_{LS}$, we can write $f^{\leftarrow}(\mathbf{x}, y) = q\big(\frac{y - \mu(\mathbf{x})}{\sigma(\mathbf{x})}\big)$ for some functions $\mu(\cdot), \sigma(\cdot) > 0$ and for some (continuous) distribution function $q(\cdot)$. Using this notation, (25) is equivalent to

$$\frac{Y - \mu(\mathbf{X}_S)}{\sigma(\mathbf{X}_S)} \perp\!\!\!\perp \mathbf{X}_S. \tag{26}$$

Denote $W_{\mathbf{x}} := (Y \mid \mathbf{X}_S = \mathbf{x})$. From (26) we get that for all $\mathbf{x}, \mathbf{y}$ in the support of $\mathbf{X}_S$ must hold

$$\frac{W_{\mathbf{x}} - \mu(\mathbf{x})}{\sigma(\mathbf{x})} \overset{D}{=} \frac{W_{\mathbf{y}} - \mu(\mathbf{y})}{\sigma(\mathbf{y})}. \tag{27}$$

Hence, also supports must match, i.e. (27) implies

$$\frac{\underline{\Psi}(\mathbf{x}) - \mu(\mathbf{x})}{\sigma(\mathbf{x})} = \frac{\underline{\Psi}(\mathbf{y}) - \mu(\mathbf{y})}{\sigma(\mathbf{y})}, \qquad \frac{\overline{\Psi}(\mathbf{x}) - \mu(\mathbf{x})}{\sigma(\mathbf{x})} = \frac{\overline{\Psi}(\mathbf{y}) - \mu(\mathbf{y})}{\sigma(\mathbf{y})},$$

for all $\mathbf{x}, \mathbf{y}$ in the support of $\mathbf{X}_S$. Solving for μ, σ gives us

$$\mu(\mathbf{x}) = c_1 + \underline{\Psi}(\mathbf{x}), \qquad \sigma(\mathbf{x}) = c_2 \cdot [\overline{\Psi}(\mathbf{x}) - \underline{\Psi}(\mathbf{x})],$$

where $c_1 \in \mathbb{R}, c_2 \in \mathbb{R}_+$ are some constants. Plugging this into (26) gives us a contradiction with (8). □

Proposition 3. *Let F be a distribution function whose parameter acts post-multiplicatively. Let $\boldsymbol{X} = (X_1, \dots, X_k)$ be a continuous random vector with independent components.*

- *Consider $f \in \mathcal{F}_F$ in the form $f(\boldsymbol{x}, \varepsilon) = F^{-1}(\varepsilon, \theta(\boldsymbol{x}))$ with an additive function $\theta(x_1, \dots, x_k) = h_1(x_1) + \dots + h_k(x_k)$, where h_i are continuous non-constant real functions. Then, f is $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*
- *Consider $f \in \mathcal{F}_F$ in the form $f(\boldsymbol{x}, \varepsilon) = F^{-1}(\varepsilon, \theta(\boldsymbol{x}))$ with a multiplicative function $\theta(x_1, \dots, x_k) = h_1(\boldsymbol{x}_S) \cdot h_2(\boldsymbol{x}_{\{1,\dots,k\}\setminus S})$ for some $S \subsetneq \{1, \dots, k\}$, where h_i are continuous non-constant non-zero real functions. Then, f is not $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*

Proof. **The first bullet-point**: For a contradiction, consider that there exist $S \subset \{1, \dots, k\}$, $z \in (0, 1)$, $g \in \mathcal{F}_F$ such that $g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) \perp\!\!\!\perp \mathbf{X}_S$. Since $g \in \mathcal{F}_F$, we can write $g^{\leftarrow}(\mathbf{x}_S, \cdot) = F\big(\cdot, \theta_g(\mathbf{x}_S)\big)$ for some non-constant function θ_g. Hence, simply rewriting

$$\mathbf{X}_S \perp\!\!\!\perp g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) = F[F^{-1}\big(z, \theta(\mathbf{X})\big), \theta_g(\mathbf{X}_S)] = f_1[z, f_2\big(\theta_g(\mathbf{X}_S)\big) \cdot \theta(\mathbf{X})].$$

We use an identity, that if $X \perp\!\!\!\perp Y \implies f(X) \perp\!\!\!\perp Y$ for any measurable function f. Since f_1 is invertible, we obtain

$$\mathbf{X}_S \perp\!\!\!\perp f_2\big(\theta_g(\mathbf{X}_S)\big) \cdot \theta(\mathbf{X}). \tag{28}$$

Define $\tilde{\theta}_g(\mathbf{X}_S) := f_2\big(\theta_g(\mathbf{X}_S)\big)$. Finally, since $\theta(\mathbf{X})$ is an additive function from the assumptions, (28) is equivalent to

$$\tilde{\theta}_g(\mathbf{X}_S)[h_1(X_1) + \dots + h_k(X_k)] \perp\!\!\!\perp \mathbf{X}_S.$$

However, that is a contradiction with Lemma B.2 part 1.

The second bullet-point: We find an appropriate function $g \in \mathcal{F}_F$ such that for any $z \in (0, 1)$ holds $g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) \perp\!\!\!\perp \mathbf{X}_S$. Since $g \in \mathcal{F}_F$ we write $g^{\leftarrow}(\mathbf{x}_S, \cdot) = F\big(\cdot, \theta_g(\mathbf{x}_S)\big)$ for some θ_g.

Rewrite

$$g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) = F[F^{-1}\big(z, \theta(\mathbf{X})\big), \theta_g(\mathbf{X}_S)] = f_1[z, f_2\big(\theta_g(\mathbf{X}_S)\big) \cdot \theta(\mathbf{X})],$$

where f_1, f_2 are from (15). We choose θ_g such that $f_2\big(\theta_g(\mathbf{x}_S)\big) = \frac{1}{h_1(\mathbf{x}_S)}$. Obviously $g \in \mathcal{F}_F$. Then, by extending θ to its multiplicative form, we get

$$f_1[z, f_2\big(\theta_g(\mathbf{X}_S)\big) \cdot \theta(\mathbf{X})] = f_1[z, h_2(\mathbf{X}_{\{1,\dots,k\}\setminus S})] \perp\!\!\!\perp \mathbf{X}_S.$$

Together, we found $g \in \mathcal{F}_F$ defined by $g^{\leftarrow}(\mathbf{x}_S, \cdot) = F\big(\cdot, f_2^{-1}(\frac{1}{h_1(\mathbf{x}_S)})\big)$ that satisfy $g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) \perp\!\!\!\perp \mathbf{X}_S$. Hence, f is not $\mathcal{F}_F-$unseparable wrt $\mathbf{X}$. □

Proposition 4. *Let F be a distribution function whose parameter acts post-additively. Let $\boldsymbol{X} = (X_1, \dots, X_k)$ be a continuous random vector with independent components.*

- *Consider $f \in \mathcal{F}_F$ in the form $f(\boldsymbol{x}, \varepsilon) = F^{-1}(\varepsilon, \theta(\boldsymbol{x}))$ with an additive function $\theta(x_1, \dots, x_k) = h_1(\boldsymbol{x}_S) + h_2(\boldsymbol{x}_{\{1,\dots,k\}\setminus S})$ for some non-empty $S \subset \{1, \dots, k\}$, where h_i are continuous non-constant non-zero real functions. Then, f is not $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*
- *Consider $f \in \mathcal{F}_F$ in the form $f(\boldsymbol{x}, \varepsilon) = F^{-1}\big(\varepsilon, \theta(\boldsymbol{x})\big)$ with a multiplicative function $\theta(x_1, \dots, x_k) = h_1(x_1) \cdot h_2(x_2) \dots h_k(x_k)$ where h_i are continuous non-constant non-zero real functions. Then, f is $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*

Proof. **The first bullet-point**: We find an appropriate function $g \in \mathcal{F}_F$ such that for any $z \in (0,1)$ holds $g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) \perp\!\!\!\perp \mathbf{X}_S$. Since $g \in \mathcal{F}_F$ we write $g^{\leftarrow}(\mathbf{x}_S, \cdot) = F\big(\cdot, \theta_g(\mathbf{x}_S)\big)$ for some θ_g.

Rewrite

$$g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) = F\Bigg(F^{-1}[z, \theta(\mathbf{X})], \theta_g(\mathbf{X}_S)\Bigg) = f_1[z, f_2\big(\theta_g(\mathbf{X}_S)\big) + \theta(\mathbf{X})],$$

where f_1, f_2 are from (14). We choose θ_g such that $f_2(\theta_g(\mathbf{x}_S)) = -h_1(\mathbf{x}_S)$. Obviously $g \in \mathcal{F}_F$. Then, by extending θ to its additive form, we get

$$f_1[z, f_2\big(\theta_g(\mathbf{X}_S)\big) + \theta(\mathbf{X})] = f_1[z, h_2(\mathbf{X}_{\{1,\dots,k\}\setminus S})] \perp\!\!\!\perp \mathbf{X}_S.$$

Together, we found $g \in \mathcal{F}_F$ defined by $g^{\leftarrow}(\mathbf{x}_S, \cdot) = F\big(\cdot, f_2^{-1}(\frac{1}{h_1(\mathbf{x}_S)})\big)$ that satisfy $g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) \perp\!\!\!\perp \mathbf{X}_S$. Hence, f is not $\mathcal{F}_F-$unseparable wrt. $\mathbf{X}$.

The second bullet-point: For a contradiction, consider that there exist $S \subset \{1, \dots, k\}$, $z \in (0,1)$, $g \in \mathcal{F}_F$ such that $g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) \perp\!\!\!\perp \mathbf{X}_S$. Since $g \in \mathcal{F}_F$, we can write $g^{\leftarrow}(\mathbf{x}_S, \cdot) = F\big(\cdot, \theta_g(\mathbf{x}_S)\big)$ for some non-constant function θ_g. Hence, simply rewriting

$$\mathbf{X}_S \perp\!\!\!\perp g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) = F[F^{-1}\big(z, \theta(\mathbf{X})\big), \theta_g(\mathbf{X}_S)] = f_1[z, f_2\big(\theta_g(\mathbf{X}_S)\big) + \theta(\mathbf{X})].$$

We use an identity, that if $X \perp\!\!\!\perp Y \implies f(X) \perp\!\!\!\perp Y$ for any measurable function f. Since f_1 is invertible, we obtain

$$\mathbf{X}_S \perp\!\!\!\perp f_2\big(\theta_g(\mathbf{X}_S)\big) + \theta(\mathbf{X}) \tag{29}$$

Define $\tilde{\theta}_g(\mathbf{X}_S) := f_2\big(\theta_g(\mathbf{X}_S)\big)$. Finally, since $\theta(\mathbf{X})$ is an multiplicative function from the assumptions, (29) is equivalent to

$$\tilde{\theta}_g(\mathbf{X}_S) + \big(h_1(X_1) \dots h_k(X_k)\big) \perp\!\!\!\perp \mathbf{X}_S.$$

However, that is a contradiction with Lemma B.2 part 2. ☐

Proposition 5. *Let F has a Location-Scale type with $q = 2$ parameters. Let $\boldsymbol{X} = (X_1, \dots, X_k)$ be a continuous random vector with independent components. Consider $f \in \mathcal{F}_F$ in the form $f(\boldsymbol{x}, \varepsilon) = F^{-1}\big(\varepsilon, \theta(\boldsymbol{x})\big)$, where $\theta(\boldsymbol{x}) = \big(\mu(\boldsymbol{x}), \sigma(\boldsymbol{x})\big)^\top$ is additive in both components, i.e. $\mu(\boldsymbol{x}) = h_{1,\mu}(x_1) + \dots + h_{k,\mu}(x_k)$ and $\sigma(\boldsymbol{x}) = h_{1,\sigma}(x_1) + \dots + h_{k,\sigma}(x_k)$ for some continuous non-constant non-zero functions $h_{i,\cdot}$, where moreover we assume $h_{i,\sigma} > 0$, $i = 1, \dots, k$. Then, f is $\mathcal{F}_F-$unseparable wrt $\boldsymbol{X}$.*

Proof. For a contradiction, consider that there exist $S \subsetneq \{1, \dots, k\}$, $z \in (0,1)$, $g \in \mathcal{F}_F$ such that $g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) \perp\!\!\!\perp \mathbf{X}_S$. Since $g \in \mathcal{F}_F$, we can write $g^{\leftarrow}(\mathbf{x}_S, \cdot) = F\big(\cdot, \theta_g(\mathbf{x}_S)\big)$ for some function $\theta_g = (\mu_g, \sigma_g)$ that is non-constant in neither of the arguments. Hence, simply rewriting

$$\mathbf{X}_S \perp\!\!\!\perp g^{\leftarrow}\big(\mathbf{X}_S, f(\mathbf{X}, z)\big) = F[F^{-1}\big(z, \theta(\mathbf{X})\big), \theta_g(\mathbf{X}_S)] = F_{\theta_0}\Bigg(\frac{[\mu(\mathbf{X}) + \sigma(\mathbf{X})F_{\theta_0}^{-1}(z)] - \mu_g(\mathbf{X}_S)}{\sigma_g(\mathbf{X}_S)}\Bigg),$$

where F_{θ_0} is a distribution function of a standardised random variable (e.g. if F is Gaussian, then $F_{\theta_0}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} dy$).

We use an identity $X \perp\!\!\!\perp Y \implies f(X) \perp\!\!\!\perp Y$ for any measurable function f to obtain

$$\mathbf{X}_S \perp\!\!\!\perp \frac{[\mu(\mathbf{X}) + \sigma(\mathbf{X})F_{\theta_0}^{-1}(z)] - \mu_g(\mathbf{X}_S)}{\sigma_g(\mathbf{X}_S)}.$$

$$\mathbf{X}_S \perp\!\!\!\perp \frac{[\mu(\mathbf{X}) + \sigma(\mathbf{X})F_{\theta_0}^{-1}(z)] - \mu_g(\mathbf{X}_S)}{\sigma_g(\mathbf{X}_S)}. \tag{30}$$

Equation (30) can be equivalently rewritten into

$$\mathbf{X}_S \perp\!\!\!\perp f_1(\mathbf{X}_S) + f_2(\mathbf{X}_S)h(\mathbf{X}_{S^c}), \tag{31}$$

where $S^c = \{1,\dots,k\}\setminus S$, $S = (S[1],\dots,S[s])$,

$$f_1(\mathbf{x}) = \frac{h_{S[1],\mu}(x_1)+\cdots+h_{S[s],\mu}(x_s)+[h_{S[1],\sigma}(x_1)+\cdots+h_{S[s],\sigma}(x_s)]F_{\theta_0}^{-1}(z)-\mu_g(\mathbf{x})}{\sigma_g(\mathbf{x})},$$

$$f_2(\mathbf{x}) = \frac{1}{\sigma_g(\mathbf{x})}$$

$$h(\mathbf{x}) = [h_{S^c[1],\sigma}(x_1)+\cdots+h_{S^c[k-s],\sigma}(x_s)]F_{\theta_0}^{-1}(z)$$

However, independence (31) is a contradiction with Lemma B.2 part 3. □

Consequence 1. *Consider $f_Y \in \mathcal{F}_A$ and let $(Y,\boldsymbol{X}) \in \mathbb{R}\times\mathbb{R}^p$ follow a SCM with DAG $\mathcal{G}_0$ where pa_Y are d-separated.*

- *If f_Y has a form*

$$f_Y(\boldsymbol{x},e) = h_1(\boldsymbol{x}_S)+h_2(\boldsymbol{x}_{pa_Y\setminus S})+q^{-1}(e), \quad \boldsymbol{x}\in\mathbb{R}^{|pa_Y|}, e\in(0,1),$$

  for some non-empty $S\subset pa_Y$, where h_1,h_2 are continuous non-constant real functions and q^{-1} is a quantile function. Then $S_{\mathcal{F}_A}(Y)=\emptyset$.

- *If f_Y has a form*

$$f_Y(\boldsymbol{x},e) = h_1(x_1)\dots h_{|pa_Y|}(x_{|pa_Y|})+q^{-1}(e), \quad \boldsymbol{x}\in\mathbb{R}^{|pa_Y|}, e\in(0,1),$$

  where h_i are continuous non-constant non-zero real functions and q^{-1} is a quantile function. Then every $S\subset pa_Y$ is not $\mathcal{F}_A$-plausible.

Proof. **The first bullet-point**: We show that sets S and $S^c := pa_Y\setminus S$ are both $\mathcal{F}_A$-plausible. Consider a function $f\in\mathcal{F}_A$ satisfying $f^{\leftarrow}(\mathbf{X}_S,Y) := \tilde{q}\big(Y-h_1(\mathbf{X}_S)\big)$, where $\tilde{q}$ is a distribution function of $[h_2(\mathbf{X}_{pa_Y\setminus S})+q^{-1}(\varepsilon_Y)]$. Then, $f^{\leftarrow}(\mathbf{X}_S,Y)\perp\!\!\!\perp\mathbf{X}_S$ since $Y-h_1(\mathbf{X}_S)=h_2(\mathbf{X}_{pa_Y\setminus S})+q^{-1}(\varepsilon_Y)\perp\!\!\!\perp\mathbf{X}_S$ and we apply identity B.3. Moreover, $f^{\leftarrow}(\mathbf{X}_S,Y)\sim U(0,1)$ trivially. We showed that set S satisfy every property for being $\mathcal{F}_A$-plausible. Set S^c can be analogously shown to be $\mathcal{F}_A$-plausible as well. Therefore, $S_{\mathcal{F}_A}(Y)\subseteq S\cap S^c=\emptyset$.

The second bullet-point: For a contradiction, consider that there exist a non-empty set $S\subset pa_Y$ that is $\mathcal{F}_A$-plausible. That means, there exist $f\in\mathcal{F}_A$ such that (1) holds. Since $f\in\mathcal{F}_A$, we can write $f(\mathbf{x},e)=\mu(\mathbf{x})+\tilde{q}^{-1}(e)$, $\mathbf{x}\in\mathbb{R}^{|S|}, e\in(0,1)$ for some function μ and quantile function $\tilde{q}^{-1}$. Also, we can write $f^{\leftarrow}(\mathbf{x},y)=\tilde{q}\big(y-\mu(\mathbf{x})\big)$, $\mathbf{x}\in\mathbb{R}^{|S|}, y\in\mathbb{R}$ (see discussion in Appendix A.1). Using (1) and identity B.3, we have that $Y-\mu(\mathbf{X}_S)\perp\!\!\!\perp\mathbf{X}_S$.

From the definition of f_Y we have that $Y=h_1(X_1)\dots h_{|pa_Y|}(X_{|pa_Y|})+q^{-1}(\varepsilon_Y)$, where $\varepsilon_Y\perp\!\!\!\perp\mathbf{X}_{pa_Y}$. Hence, we have

$$\begin{aligned} Y-\mu(\mathbf{X}_S)&\perp\!\!\!\perp\mathbf{X}_S\\ h_1(X_1)\dots h_{|pa_Y|}(X_{|pa_Y|})+q^{-1}(\varepsilon_Y)-\mu(\mathbf{X}_S)&\perp\!\!\!\perp\mathbf{X}_S\\ h_1(X_1)\dots h_{|pa_Y|}(X_{|pa_Y|})-\mu(\mathbf{X}_S)&\perp\!\!\!\perp\mathbf{X}_S \end{aligned}$$

That is a contradiction with Lemma B.2 part 2. □

Proposition 6. *Consider $\mathcal{F}=\mathcal{F}_A$ and let $(Y,\boldsymbol{X})\in\mathbb{R}\times\mathbb{R}^p$ follow an SCM with DAG $\mathcal{G}_0$ satisfying (♥). Assume that every $S\neq pa_Y$ is not $\mathcal{F}$-plausible. Then,*

$$\lim_{n\to\infty}\mathbb{P}(\widehat{pa}_Y\neq pa_Y)=0, \tag{32}$$

where n is the size of the random sample and $\widehat{pa}_Y$ is our score-based estimate from Section 4.2 with $\lambda_1,\lambda_2>0,\lambda_3=0$, suitable estimation procedure and HSIC independence measure.

Proof. This result is a trivial consequence of Theorem 20 in Mooij et al. (2016). We use the same notation. For a rigorous definition of $HSIC$ and $\widehat{HSIC}$, see Appendix A.1 in Mooij et al. (2016).

We show that $score(S) > score(pa_Y)$ as $n \to \infty$ for any $S \neq pa_Y$. The $score(S)$ is defined as the weighted sum of *Independence* and *Significance* terms. Let us first concentrate on the former. By definition, we write $Independence = -\widehat{HSIC}(\mathbf{X}_S, \hat{\varepsilon}_S)$. On a population level, it holds (Lemma 12 in Mooij et al. (2016)) that $HSIC(\mathbf{X}_S, \varepsilon_S) > 0$ and $HSIC(\mathbf{X}_{pa_Y}, \varepsilon_{pa_Y}) = 0$, since $\mathbf{X}_S$ and ε_S are not independent (because S is not $\mathcal{F}$-plausible) and $\mathbf{X}_{pa_Y}$ and ε_{pa_Y} are independent (by definition of the SCM). By Theorem 20 in Mooij et al. (2016), we obtain that $\widehat{HSIC}(\mathbf{X}_{pa_Y}, \hat{\varepsilon}_{pa_Y}) \to HSIC(\mathbf{X}_{pa_Y}, \varepsilon_{pa_Y}) = 0$ and $\widehat{HSIC}(\mathbf{X}_S, \hat{\varepsilon}_S) \to HSIC(\mathbf{X}_S, \varepsilon_S) > 0$, as $n \to \infty$. Therefore, the independence term is strictly smaller (for some large n) for S than for pa_Y.

Let us focus on *Significance* term [5]. Since all $\mathbf{X}_{pa_Y}$ are significant (otherwise $f_Y \notin \mathcal{I}_m$) we get that $Significance \to 0$ as $n \to \infty$ for pa_Y. Moreover, by definition $Significance \geq 0$ for S.

Together, we get that $score(pa_Y) > score(S)$ for large n, since the *Independence* term is strictly smaller (for large n) for S than for pa_Y and *Significance* term converges to 0 for pa_Y and is non-negative. We showed that pa_Y has the largest score among all $S \subseteq \{1, \dots, p\}$ (for n large enough). □

Lemma C.1. *Consider $\mathcal{F} = \mathcal{F}_A$ and let $(Y, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^p$ follow an SCM with DAG $\mathcal{G}_0$ satisfying (♥), with $S_{\mathcal{F}_A}(Y) \neq \emptyset$. Then, ISD algorithm used with a consistent estimation (11) and an independence-oracle is guaranteed to estimate $\hat{S}_{\mathcal{F}_A}(Y) = S_{\mathcal{F}_A}(Y)$.*

Proof. Similar statement can be found about RESIT algorithm (Peters et al., 2014). However here, the proof is trivial. For every $\mathcal{F}_A$-plausible set S, we have that $\varepsilon_S \perp\!\!\!\perp \mathbf{X}_S$. Since our estimation is consistent, we also have (in the limit) that $\hat{\varepsilon}_S \perp\!\!\!\perp \mathbf{X}_S$. Therefore, our independence oracle will output YES on question 1. Trivially, all $\mathbf{X}_S$ are significant since S is $\mathcal{F}_A$-plausible set, and $\hat{\varepsilon}_S \sim U(0,1)$ is trivially satisfied since for $\mathcal{F} = \mathcal{F}_A$ is the third question redundant. Therefore, ISD algorithm will mark the set S as $\mathcal{F}_A$-plausible and the ISD estimation is correct. □

D Appendix: Simulations and application

D.1 Functions generated using Perlin noise approach

In the following, we provide examples of functions generated using Perlin noise approach. For one dimensional case, let $X_1, \eta_Y \overset{iid}{\sim} N(0,1)$ and $Y = g(X_1) + \eta_Y$, where g is generated using Perlin noise approach. Such (typical) datasets are plotted in Figure 5.

For two dimensional case, let $X_1, X_2, \eta_Y \overset{iid}{\sim} N(0,1)$ and $Y = g(X_1, X_2) + \eta_Y$, where g is generated using Perlin noise approach. Such (typical) dataset is plotted in Figure 6.

For three dimensional case, let $X_1, X_2, X_3, \eta_Y \overset{iid}{\sim} N(0,1)$ and $Y = g(X_1, X_2, X_3) + \eta_Y$, where g is generated using Perlin noise approach. Visualisation of a four dimensional dataset is a bit tricky; Figure 7 represents the 3-dimensional slices of the function.

D.2 Application

[5] We work with the *Significance* term somehow vaguely in this proof. However, we only need the property that $Significance \to 0$ as $n \to \infty$ for pa_Y, which is satisfied for any reasonable method for assessing significance of covariates.

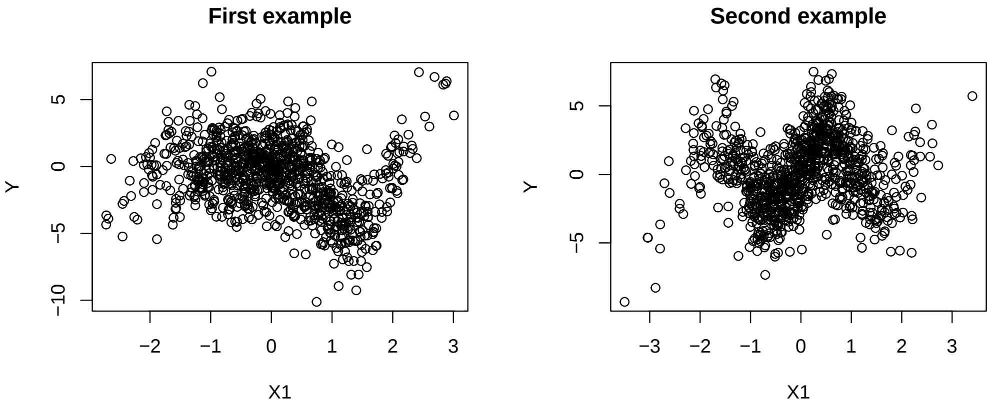

Figure 5: Two examples of one dimensional functions generated using Perlin noise approach.

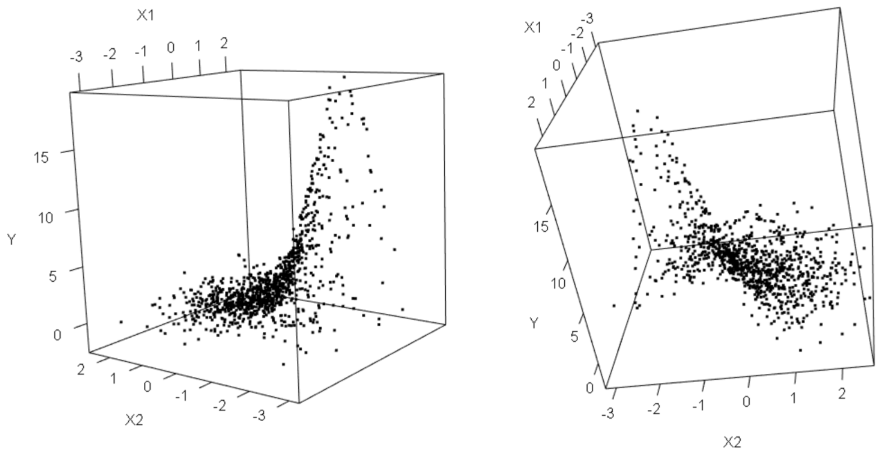

Figure 6: A typical two-dimensional function generated using Perlin noise approach. Figure shows this function from two different angles.

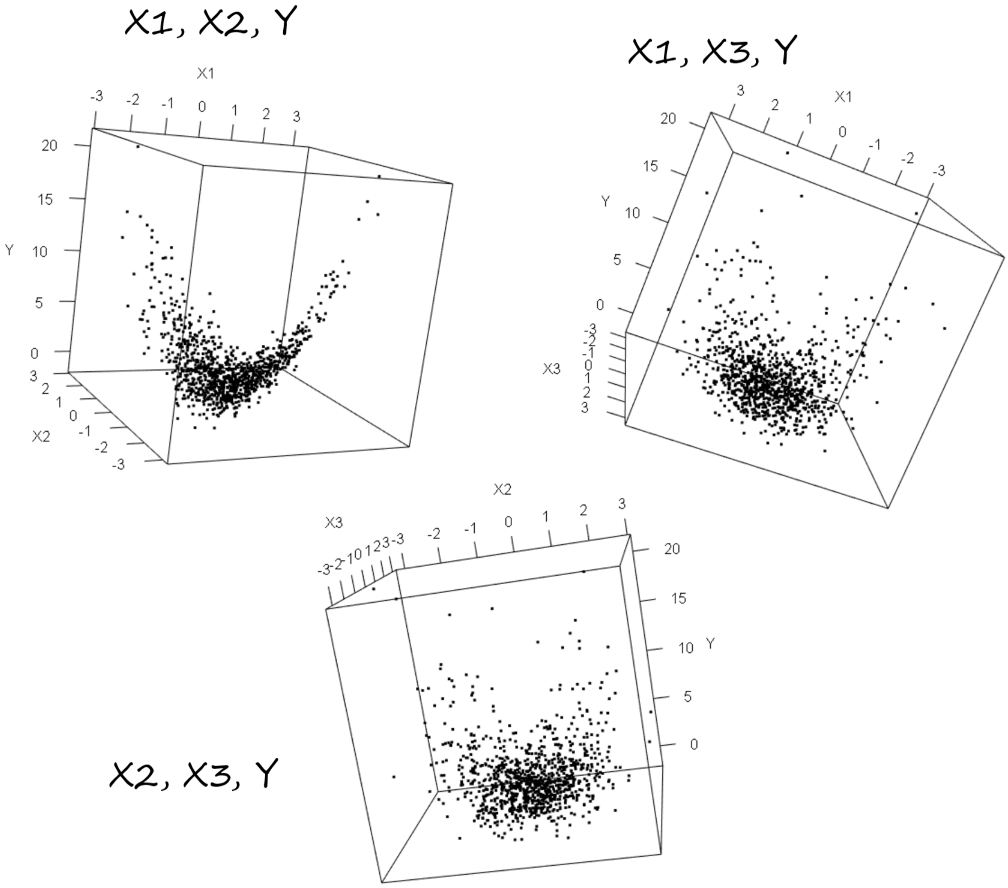

Figure 7: A typical three-dimensional function generated using Perlin noise approach. Figure shows projections of this function.

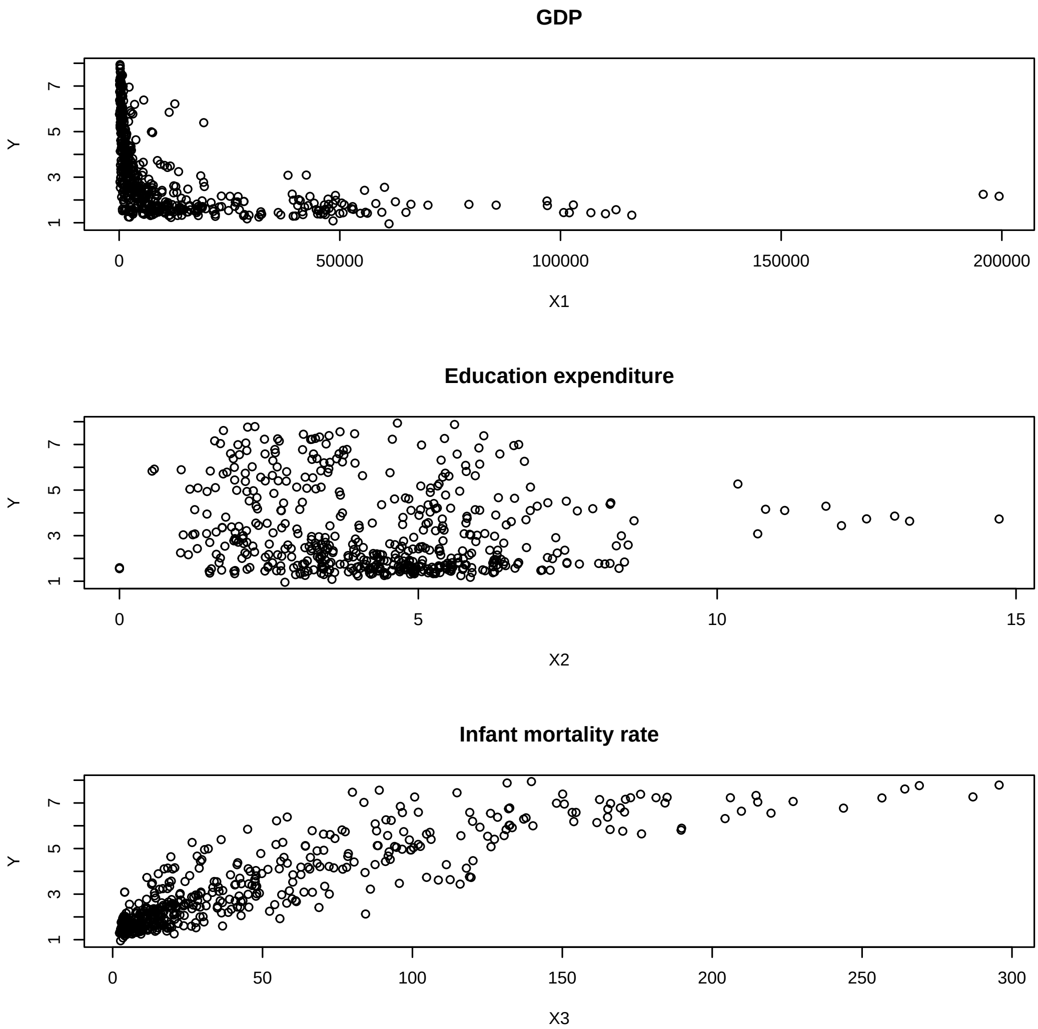

Figure 8: Visualization of marginal relations between the target variable (fertility rate) and the covariates from Section 5.3.

References

C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(7):171–234, 2010. URL http://jmlr.org/papers/v11/aliferis10a.html.

M. Azadkia, A. Taeb, and P. Bühlmann. A fast non-parametric approach for causal structure learning in polytrees, 2021. URL https://arxiv.org/abs/2111.14969.

P. Baxandall and H. Liebeck. *The Inverse Function Theorem: Vector Calculus*. New York: Oxford University Press, 1986. ISBN 0-19-859652-9.

J. Bodik and V. Chavez-Demoulin. Identifiability of causal graphs under nonadditive conditionally parametric causal models, 2023. URL https://arxiv.org/abs/2303.15376.

H. Cheng, W. Luo, S. Si, X. Xin, Z. Peng, H. Zhou, H. Liu, and Y. Yu. Global trends in total fertility rate and its relation to national wealth, life expectancy and female education. *BMC Public Health*, 22(1):1346, 2022. doi: 10.1186/s12889-022-13656-1.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002. doi: 10.1162/153244303321897717.

G. Darmois. Analyse générale des liaisons stochastiques: Etude particulière de l'analyse factorielle linéaire. *Review of the International Statistical Institute*, 21(1/2):2–8, 1953. ISSN 1532-4435. doi: 10.2307/1401511.

T. Galanti, O. Nabati, and L. Wolf. A critical view of the structural causal model. *Preprint*, 2020. URL https://arxiv.org/abs/2002.10007.

M. Gao and B. Aragam. Efficient bayesian network structure learning via local markov boundary search. *Advances in Neural Information Processing Systems*, 34, 2021. URL https://arxiv.org/abs/2110.06082.

C. Genest, J.G. Nešlehová, B. Rémillard, and O.A. Murphy. Testing for independence in arbitrary distributions. *Biometrika*, 106(1):47–68, 2019. doi: 10.1093/biomet/asy059.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019. doi: 10.3389/fgene.2019.00524.

P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall/CRC, 1994. ISBN 9780412300400. doi: 10.1201/b15710.

Ch. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018. doi: 10.1515/jci-2017-0016.

C. Hirschman. Why fertility changes. *Annual Review of Sociology*, 20:203–233, 1994. doi: 10.1146/annurev.so.20.080194.001223.

T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995. doi: 10.1109/ICDAR.1995.598994.

P. Hoyer, D. Janzing, J.M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, Cambridge, 2015. doi: 10.1017/CBO9781139025751.

A. Immer, Ch. Schultheiss, J. E. Vogt, B. Schölkopf, and P. Bühlmann. On the identifiability and estimation of causal location-scale noise models. *arXiv preprint arXiv:2210.09054*, 2022.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010. doi: 10.1109/TIT.2010.2060095.

I. Khemakhem, R. Monti, R. Leech, and A.Hyvarinen. Causal autoregressive flows. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3520–3528. PMLR, 2021. URL https://proceedings.mlr.press/v130/khemakhem21a.html.

T. N. Maeda and S. Shimizu. Causal additive models with unobserved variables. In *Proc. 27th Conference on Uncertainty in Artificial Intelligence (UAI2021)*, volume 161 of *Proceedings of Machine Learning Research*, pages 97–106. PMLR, 2021.

A. Marx and J. Vreeken. Telling cause from effect using mdl-based local and global regression. *Knowledge and Information Systems*, 2017. doi: 10.1109/ICDM.2017.40.

C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 403–410, 1995. ISBN 1558603859.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020. URL http://jmlr.org/papers/v21/17-123.html.

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(1):1103–1204, 2016.

Ch. Nowzohour and P. Bühlmann. Score-based causal learning in additive noise models. *Statistics: A Journal of Theoretical and Applied Statistics*, 50(3):471–485, 2016. doi: 10.1080/02331888.2015.1060237.

G. Park and G. Raskutti. Learning quadratic variance function (qvf) dag models via overdispersion scoring (ods). *Journal of Machine Learning Research*, 18(1):8300–8342, 2017. URL https://proceedings.neurips.cc/paper/2015/file/fccb60fb512d13df5083790d64c4d5dd-Paper.pdf.

J. Paul and P. Dupont. Statistically interpretable importance indices for random forests. *23rd Annual Machine Learning Conference of Belgium and the Netherlands (BENELEARN)*, 2014. URL http://hdl.handle.net/2078.1/147350.

J. Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2009. ISBN 978-0521895606.

K. Perlin. An image synthesizer. *SIGGRAPH Comput. Graph.*, 19(0097-8930):287–296, 1985. doi: 10.1145/325165.325247.

J. Peters, J.M. Mooij, 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. URL https://doi.org/10.1111/rssb.12167.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319. URL https://library.oapen.org/bitstream/id/056a11be-ce3a-44b9-8987-a6c68fce8d9b/11283.pdf.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B*, 80(1):5–31, 2018. doi: 10.1111/rssb.12235.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

N. M. Razali and B. W. Yap. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model. Analytics*, 2, 2011.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996. doi: 10.1017/CBO9780511812651.

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

P. Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(54):1643–1662, 2010. URL http://jmlr.org/papers/v11/spirtes10a.html.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search, 2nd Edition*, volume 1. The MIT Press, 1 edition, 2001. URL https://EconPapers.repec.org/RePEc:mtp:titles:0262194406.

D.M. Stasinopoulos and R.A. Rigby. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7), 2007. doi: 10.18637/jss.v023.i07.

E.V. Strobl and T. A. Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *arXiv preprint arXiv:2205.13085*, 2022.

X. Sun and O. Schulte. Cause-effect inference in location-scale noise models: Maximum likelihood vs. independence testing, 2023.

N. Tagasovska, V. Chavez-Demoulin, and T. Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9311–9323, 2020. URL https://proceedings.mlr.press/v119/tagasovska20a.html.

United Nations. United Nations Statistics Division: Development Data. URL https://www.un.org/development/desa/pd/data-landing-page. Accessed: June 13, 2023.

T. S. Verma and J. Pearl. On the equivalence of causal models. *CoRR*, 2013. URL http://arxiv.org/abs/1304.1108.

Ch. Wang, Y. Zhou, Q. Zhao, and Z. Geng. Discovering and orienting the edges connected to a target variable in a dag via a sequential local learning approach. *Computational Statistics and Data Analysis*, 77:252–266, 2014. ISSN 0167-9473. doi: 10.1016/j.csda.2014.03.003.

World Bank. World Bank Open Data: Government expenditure on education, a. URL https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS. Accessed: June 13, 2023.

World Bank. World Bank Open Data: GDP per capita, b. URL https://data.worldbank.org/indicator/NY.GDP.PCAP.CD. Accessed: June 13, 2023.

J. Yin, Y. Zhou, C. Wang, P. He, C. Zheng, and Z. Geng. Partial orientation and local structural learning of causal networks for prediction. In *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, volume 3 of *Proceedings of Machine Learning Research*, pages 93–105, Hong Kong, 03–04 Jun 2008. PMLR.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 647–655. AUAI Press, 2009. ISBN 9780974903958.

K. Zhang and A. Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, volume 6 of *Proceedings of Machine Learning Research*, pages 157–164. PMLR, 2010. URL https://proceedings.mlr.press/v6/zhang10a.html.