Journal of Machine Learning Research 26 (2025) 1-55 Submitted 10/24; Revised 11/25; Published 11/25

Identifiability of Causal Graphs under Non-Additive Conditionally Parametric Causal Models

Juraj Bodik JURAJ.BODIK@UNIL.CH
HEC Lausanne
University of Lausanne
CH-1015 Lausanne, Switzerland

Valérie Chavez-Demoulin VALERIE.CHAVEZ@UNIL.CH
HEC Lausanne
University of Lausanne
CH-1015 Lausanne, Switzerland

Editor: Jin Tian

Abstract

Existing approaches to causal discovery often rely on restrictive modeling assumptions that limit their applicability in real-world settings, particularly when data are heavy-tailed or contain a mixture of discrete and continuous variables. Identifiability of causal graphs has been established under several structural models, including linear non-Gaussian models, post-nonlinear models, and location-scale models. However, these frameworks may not capture the diversity of distributions observed in practice. To address this, we introduce Conditionally Parametric Causal Models (CPCM), a flexible class of models where the conditional distribution of the effect, given its cause, belongs to a known parametric family such as Gaussian, Poisson, Gamma, or Pareto. These models are adaptable to a wide range of practical situations, where the cause influences not only the mean but also the variance or tail behavior of the effect. We demonstrate the identifiability of CPCM by leveraging the concept of sufficient statistics. Furthermore, we propose an algorithm for estimating the causal structure from random samples drawn from CPCM. We evaluate the empirical properties of our methodology on various datasets, demonstrating state-of-the-art performance across multiple benchmarks.

Keywords: causal discovery, structural causal models, identifiability, higher moments, exponential family

1. Introduction

Understanding causal relations, as opposed to mere statistical associations, allows us to predict the effects of interventions that modify a system (Pearl and Mackenzie, 2019). Identifying such causal structures is central to many scientific disciplines (Imbens and Rubin, 2015). Yet different data-generating mechanisms can lead to the same observational distribution, making causal inference inherently challenging. While observing a system under interventions can reliably reveal its causal structure, performing interventions is often expensive, ethically problematic (Greenland et al., 1999), or simply unfeasible. This motivates the growing focus on estimating causal structure directly from observational data.

Over recent decades, extensive work has focused on building a rigorous mathematical framework for the "language" of causal inference, largely formalized through structural causal models (SCMs) (Pearl, 2009). Consider random variables $\mathbf{X} = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$. An SCM with an underlying acyclic graph $\mathcal{G}$ specifies a data-generating process through structural equations

$$X_i = f_i(\mathbf{X}_{pa_i}, \varepsilon_i), \qquad i = 1, \ldots, d,$$

where f_i are causal (link) functions, pa_i denotes the set of parents (direct causes) of X_i in $\mathcal{G}$, and ε_i are jointly independent noise variables. The central objective of causal discovery is to recover the causal structure, represented by the graph $\mathcal{G}$. If $\mathcal{G}$ and the conditional distributions are known, the joint distribution of $\mathbf{X}$ follows directly. Here we face the inverse task: given only the distribution of $\mathbf{X}$ (or a sample from it), we aim to infer $\mathcal{G}$. This is generally impossible without imposing additional assumptions on the underlying SCM (Peters et al., 2017).

The existing literature in the field presents numerous methods and corresponding results for causal discovery under various assumptions on the SCM (Glymour et al., 2019). When observing multiple environments following different interventions, the assumptions can be significantly less restrictive (Peters et al., 2016; Mooij et al., 2020; Wang et al., 2024). However, if the goal is to uncover causal relationships based solely on an observed random sample, the assumptions become more strict; typically assuming additive noise (Shimizu et al., 2006; Peters et al., 2014; Mooij et al., 2016; Montagna et al., 2023). This assumption of additivity $X_i = f_i(\mathbf{X}_{pa_i}) + \varepsilon_i$ suggests that $\mathbf{X}_{pa_i}$ influences only the mean of X_i, while the tail, variance, and higher moments remain fixed. This is a strong assumption, as the tail or other characteristics of the random variable can provide different information about the causal structure.

In this paper, we develop a framework where $\mathbf{X}_{pa_i}$ can arbitrarily affect the mean, variance, tail, or other characteristics of X_i. However, a caution has to be taken because if the model is too general, the causal structure will become unidentifiable, meaning that multiple causal structures could produce the same distribution of $\mathbf{X}$.

Example 1 *A useful model that allows the parent variables to influence both the mean and variance of X_i is given by the structural equation $X_i = \mu(\boldsymbol{X}_{pa_i}) + \sigma(\boldsymbol{X}_{pa_i})\,\varepsilon_i$, where ε_i is Gaussian. Equivalently, the model for the conditional distribution is*

$$X_i \mid \boldsymbol{X}_{pa_i} \sim \mathcal{N}\big(\mu(\boldsymbol{X}_{pa_i}),\ \sigma^2(\boldsymbol{X}_{pa_i})\big).$$

Example 2 *In certain applications, it may be reasonable to assume*

$$X_i \mid \boldsymbol{X}_{pa_i} \sim Poisson\big(\theta(\boldsymbol{X}_{pa_i})\big),$$

where θ is a function describing the rate of certain phenomena. Such a model is common in applications when X_i represents a number of events occurring in a certain time period.

We introduce a causal model (we call it the conditionally parametric causal model or CPCM) where the structural equation has the following form:

$$\begin{aligned} X_i = f_i(\mathbf{X}_{pa_i}, \varepsilon_i) = F^{-1}\big(\varepsilon_i; \theta(\mathbf{X}_{pa_i})\big), \quad \varepsilon_i \sim U(0,1), \\ \text{or equivalently } X_i \mid \mathbf{X}_{pa_i} \sim F\big(\theta(\mathbf{X}_{pa_i})\big), \end{aligned} \tag{1}$$

where F is a known distribution function with a vector of parameters $\theta(\mathbf{X}_{pa_i})$.

1.1 Setup and notation

We adapt the usual notation of graphical models (e.g., Spirtes et al., 2001). We consider a DAG (directed acyclic graph) $\mathcal{G} = (V, E)$ with a finite set of vertices (nodes) $V = \{1, \dots, d\}$ and a set of directed edges E, and write $pa_i(\mathcal{G})$, $ch_i(\mathcal{G})$ and $an_i(\mathcal{G})$ for parents, children and ancestors of the node i, respectively. In addition, we say that the node $i \in V$ is a source node if $pa_i(\mathcal{G}) = \emptyset$, notation $i \in Source(\mathcal{G})$. Given a random vector $\mathbf{X} = (X_i)_{i \in V}$ over some probability space with distribution $F_{\mathbf{X}}$, we identify the vertices $j \in V$ with the variables X_j. We omit the argument $\mathcal{G}$ if evident from the context.

We frequently use the concept of an exponential family, which is a class of probability distributions whose probability density function can be expressed as:

$$p(x; \theta) = h_1(x) h_2(\theta) e^{\sum_{i=1}^{q} \theta_i T_i(x)}, \tag{2}$$

where h_1, h_2, T_i are measurable functions. We call T_i a *sufficient* statistic, h_1 a base measure, and h_2 a normalizing function. Note that T_i are only unique up to a linear transformation. Many well-known distribution families belong to the exponential family, including the Gaussian, Poisson, Binomial, and Gamma distributions. We assume that q is minimal in the sense that we cannot write $p(x; \theta)$ using only $q - 1$ parameters; see Appendix A.1 that provides more information and detailed description.

We use capital F for distributions and small p for densities. A random variable Z that is uniformly distributed on $(0, 1)$ is denoted as $Z \sim U(0, 1)$. Support of a random variable Z is denoted as $supp(Z)$. We denote a random vector $\mathbf{X}_S = \{X_s \colon s \in S\}$ for $S \subseteq V$.

1.2 Related work

Many papers address the problem of the identifiability of the causal structure (for a review, see Glymour et al. (2019)). Shimizu et al. (2006) show identifiability for the linear non-Gaussian additive models (LiNGaM), where $X_i = \beta \mathbf{X}_{pa_i} + \varepsilon_i$ for non-Gaussian noise variables ε_i. Bühlmann et al. (2014) explore causal additive models (CAM) of the form $X_i = \sum_{j \in pa_i} g_j(X_j) + \varepsilon_i$ for smooth functions g_j. Hoyer et al. (2009) and Peters et al. (2014) develop a framework for additive noise models (ANM), where $X_i = g(\mathbf{X}_{pa_i}) + \varepsilon_i$. Under certain (not too restrictive) conditions on g, the authors show the identifiability of such models (Peters et al., 2014, Corollary 31) and propose an algorithm estimating $\mathcal{G}$ (for a review on ANM, see Mooij et al. (2016)). All these frameworks assume that the variance of $X_i \mid \mathbf{X}_{pa_i}$ does not depend on $\mathbf{X}_{pa_i}$. This is a crucial aspect of the identifiability results.

Zhang and Hyvärinen (2009) introduce a generalization known as the post-nonlinear model, defined by $X_i = g_1\big(g_2(\mathbf{X}_{pa_i}) + \varepsilon_i\big)$, with an invertible link function g_1. Park and Raskutti (2015, 2017) reveal identifiability in discrete models in which $var[X_i \mid \mathbf{X}_{pa_i}]$ is a quadratic function of $\mathbb{E}[X_i \mid \mathbf{X}_{pa_i}]$. If $X_i \mid \mathbf{X}_{pa_i}$ has a Poisson or binomial distribution, such a condition is satisfied. They also provide an algorithm based on comparing dispersions for estimating a DAG in polynomial time. Other algorithms have also been proposed, with comparable speed and different assumptions on the conditional densities (Gao et al., 2020). Galanti et al. (2020); Poinsot et al. (2024) consider the neural SCM with representation $X_i = g_1\big(g_2(\mathbf{X}_{pa_i}), \varepsilon_i\big)$, where g_1 and g_2 are assumed to be neural networks.

Recently, location-scale models of the form $X_i = g_1(\mathbf{X}_{pa_i}) + g_2(\mathbf{X}_{pa_i})\varepsilon_i$ have garnered attention. Immer et al. (2023) demonstrated that bivariate non-identifiable location-scale

models must satisfy a specific differential equation. Strobl and Lasko (2022) explored the problem of estimating patient-specific root causes in location-scale models. Additionally, Khemakhem et al. (2021) provided more detailed identifiability results under Gaussian noise ε_i in the bivariate case using autoregressive flows. Xu et al. (2022) investigated a more restricted location-scale model, dividing the range of the predictor variable into a finite set of bins and fitting an additive model in each bin. Klippert and Marx (2025) considered causal discovery in skewed location-scale models.

Further, several different algorithms for estimating causal graphs have been proposed, working with different assumptions (Janzing and Schölkopf, 2010; Nowzohour and Bühlmann, 2016; Marx and Vreeken, 2019; Tagasovska et al., 2020; Krali, 2025). They are often based on Kolmogorov complexity or independence between certain functions in a deterministic scenario.

Constraint-based methods, like the PC and FCI algorithms (Spirtes et al., 2001, 2013), are considered a gold standard for causal discovery. They utilize sequential independence testing for causal discovery, consistently estimating the Markov equivalence class. While these methods are powerful, they rely heavily on the accuracy of the conditional independence tests, making them sensitive to statistical errors and often resulting in many edges remaining unoriented.

A few authors assume that causal Markov kernels lie in a parametric family of distributions. Janzing et al. (2009) consider the case in which the density of $X_i \mid \mathbf{X}_{pa_i}$ lies in a second-order exponential family and the variables are a mixture of discrete and continuous random variables. Park and Park (2019) concentrate on a specific subclass of model (1), where F lies in a discrete family of generalized hypergeometric distributions; that is, the family of random variables in which the mean and variance have a polynomial relationship. To the best of our knowledge, there does not exist any study in the literature, that provides identifiability results in the case in which F lies in a general class of the exponential family. This is the focus of this paper.

The structure of the paper is as follows. Section 2 introduces the main definitions and motivation in a bivariate case. Section 3 presents identifiability results for the causal structure in the bivariate case, and Section 4 discusses the multivariate extension. In Section 5, we propose an algorithm for estimating the causal graph under assumption (1). Section 6 contains an extensive simulation study. We provide three appendices: Appendix A includes formal definition of Exponential family and some omitted technical content; Appendix B details the experiments and Appendix C contains all proofs.

2. Bivariate Conditionally Parametric Causal Models

We focus on the bivariate SCM in this section, with multivariate extensions in Section 4. The following definition describes the restriction on the SCM, assuming $X_2 \mid X_1$ has the conditional distribution F with parameters $\theta(X_1) \in \mathbb{R}^q$ for some $q \in \mathbb{N}$.

Definition 1 *We define the bivariate* ***conditionally parametric causal model*** *(bivariate CPCM(F)) with graph $X_1 \to X_2$ by two assignments:*

$$X_1 = \varepsilon_1, \qquad X_2 = F^{-1}\big(\varepsilon_2; \theta(X_1)\big), \tag{3}$$

where $\varepsilon_1 \perp\!\!\!\perp \varepsilon_2$ *are noise variables,* ε_2 *is uniformly distributed, and* F^{-1} *is the quantile function of a distribution function* F *with* q *parameters* $\theta(X_1) = \big(\theta_1(X_1), \dots, \theta_q(X_1)\big)^\top$.

We assume that θ_i *represent measurable functions, where at least one of the functions* $\theta_1, \dots, \theta_q$ *is non-constant on the support of* X_1.

We impose no restrictions on the marginal distribution of the cause. Note that we implicitly assume causal minimality (Zhang and Spirtes, 2010), as we assume that θ is non-constant.

The Gaussian model introduced in Example 1 is equivalent to the $CPCM(F)$ model with F being the Gaussian distribution and $\theta(X_1) = (\mu(X_1), \sigma(X_1))^\top$.

2.1 $CPCM(F_1, F_2, \dots, F_k)$ Models

2.1.1 Motivation

Occam's razor posits that $F_{\text{effect|cause}}$ should be "simpler" than $F_{\text{cause|effect}}$. In model-based approaches for causal discovery, we define a "simple distribution" as one that belongs to a pre-defined class of distributions $\mathcal{F}$. In ANM (Peters et al., 2014), $\mathcal{F}$ consists of all distributions that can be expressed as the sum of a function of the cause and a noise term. In CPCM, $\mathcal{F}$ is a given parametric family of distributions. If $\mathcal{F}$ is sufficiently small, we achieve identifiability of the causal graph $\mathcal{G}$ because both $F_{\text{effect|cause}}$ and $F_{\text{cause|effect}}$ cannot lie in $\mathcal{F}$.

However, the choice of $\mathcal{F}$ is crucial, especially when dealing with mixtures of discrete and continuous distributions. Suppose we observe data as shown in Figure 1. To handle such cases, $\mathcal{F}$ needs to include both continuous and discrete distributions. If we define $\mathcal{F}$ as a class of $CPCM(F)$ with continuous F (e.g., Gaussian), then $F_{X_2|X_1}$ can never lie in $\mathcal{F}$. Conversely, choosing discrete F leads to $F_{X_1|X_2} \notin \mathcal{F}$.

To accommodate a wide range of applications with various conditional distributions, we define $\mathcal{F}$ as the union of $CPCM(F_1), \dots, CPCM(F_k)$ models. By selecting $F_1, \dots, F_k$ as a collection of "standard simple well-known distributions" (such as Gaussian, Gamma, Poisson, etc., see Section 5.3), $\mathcal{F}$ is composed of "standard simple (conditional) distributions" with a wide range of possible supports, forms, and properties. We refer to this as the $CPCM(F_1, F_2, \dots, F_k)$ model.

2.1.2 Definition

Definition 2 *Let* $F_1, \dots, F_k$ *be a collection of distribution functions, each parameterized by a* q_i*-dimensional parameter, where* $q_i \in \mathbb{N}$ *and* $i = 1, \dots, k$. *A pair of dependent random variables* (X_1, X_2) *follows the* $CPCM(F_1, F_2, \dots, F_k)$ *model if there exists an* $i \in \{1, \dots, k\}$ *such that* (X_1, X_2) *follows a* $CPCM(F_i)$ *model. Specifically, either:*

$$X_1 = \varepsilon_1, \quad X_2 = F_i^{-1}\big(\varepsilon_2; \theta_2(X_1)\big), \quad \varepsilon_2 \sim U(0,1), \quad \varepsilon_1 \perp\!\!\!\perp \varepsilon_2, \quad \textit{or}$$
$$X_2 = \varepsilon_2, \quad X_1 = F_i^{-1}\big(\varepsilon_1; \theta_1(X_2)\big), \quad \varepsilon_1 \sim U(0,1), \quad \varepsilon_1 \perp\!\!\!\perp \varepsilon_2,$$

for some $i \in \{1, \dots, k\}$, *where* $\theta_1(\cdot)$ *and* $\theta_2(\cdot)$ *are suitable parameter functions of dimension* q_i, *that are measurable and non-constant.*

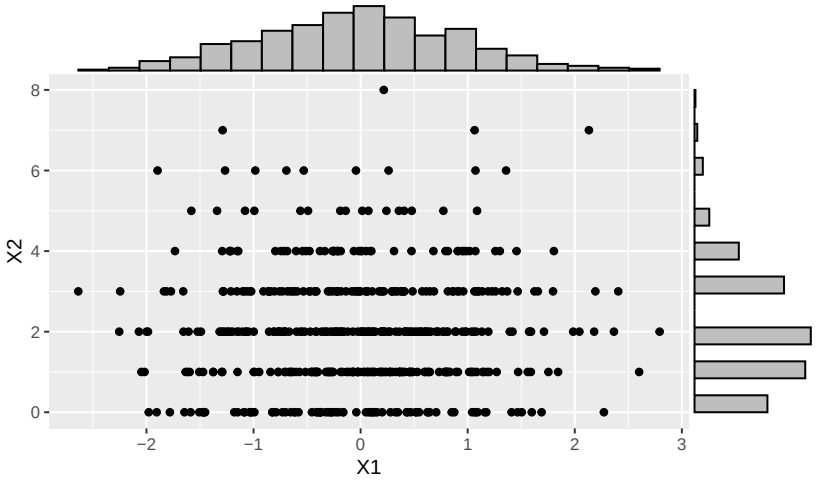

Figure 1: A dataset generated as follows: $X_1 \sim N(0,1)$, $X_2 \sim \text{Poisson}(|X_1|)$; in other words, X_1, X_2 follow $CPCM(F)$ with a Poisson F and DAG $X_1 \to X_2$.

We can potentially combine other well-known models, such as ANM and discrete QVF models; however, we will focus solely on $CPCM$ classes for the remainder of this paper. Note the distinction between ANM and CPCM models: the former assumes that $\theta(X)$ corresponds to the mean, while the latter allows $\theta(X)$ to represent any distributional characteristic; however, CPCM imposes additional assumptions on the noise.

Extending the definition of $CPCM(F)$ to allow multiple data-generating mechanisms induces the risk of unidentifiability. If the class $CPCM(F_1, \ldots, F_k)$ is too large, both $F_{\text{effect|cause}}$ and $F_{\text{cause|effect}}$ may lie within it. In the following section, we show that this is typically not the case as long as $F_1, \ldots, F_k$ belong to the exponential family of distributions.

3. Identifiability results

Identifiability is a prerequisite for causal discovery. We now examine the identifiability of the causal graph: can the true causal structure be inferred from the joint distribution under our CPCM model?

Definition 3 (Identifiability) *Let $F_{(X_1,X_2)}$ be a distribution that has been generated according to the $CPCM(F_1, \ldots, F_k)$ model with graph $X_1 \to X_2$. We say that the causal graph is identifiable from the joint distribution (equivalently, that the model is identifiable) if there does not exist $\tilde{\theta}$ and a pair of random variables $\tilde{\varepsilon}_2 \perp\!\!\!\perp \tilde{\varepsilon}_1$, where $\tilde{\varepsilon}_1$ is uniformly distributed, such that the model $X_2 = \tilde{\varepsilon}_2, X_1 = F_i^{-1}\big(\tilde{\varepsilon}_1; \tilde{\theta}(X_2)\big)$ for some $i \in \{1, \ldots, k\}$ generates the same distribution $F_{(X_1,X_2)}$.*

3.1 Identifiability in $CPCM(F)$

First, we discuss the Gaussian case. Recall that in the additive Gaussian model, where $X_2 = f(X_1) + \varepsilon_2$, $\varepsilon_2 \sim N(0, \sigma^2)$, the identifiability holds if and only if f is non-linear (Hoyer et al., 2009). We provide a different result with both mean *and* variance as functions of the cause. A similar result is found in (Khemakhem et al., 2021, Theorem 1) in the context

of autoregressive flows and where only a sufficient condition for identifiability is provided. Another similar problem is studied in Immer et al. (2023) and Strobl and Lasko (2022), both of which discuss identifiability in general location-scale models.

Theorem 4 (Gaussian case) *Let (X_1, X_2) admit the $CPCM(F)$ model with graph $X_1 \to X_2$, where F is the Gaussian distribution function with parameters $\theta(X_1) = \big(\mu(X_1), \sigma(X_1)\big)^\top$ as in Example 1.*

Let p_{ε_1} be the density of ε_1, which is absolutely continuous with full support $\mathbb{R}$. Let $\mu(x), \sigma(x)$ be two times differentiable. Then, the causal graph is identifiable from the joint distribution if and only if there do not exist $a, c, d, e, \alpha, \beta \in \mathbb{R}$, $a \geq 0, c > 0, \beta > 0$, such that

$$\frac{1}{\sigma^2(x)} = ax^2 + c, \qquad \frac{\mu(x)}{\sigma^2(x)} = d + ex, \tag{4}$$

for all $x \in \mathbb{R}$ and

$$p_{\varepsilon_1}(x) \propto \sigma(x) e^{-\frac{1}{2}\left[\frac{(x-\alpha)^2}{\beta^2} - \frac{\mu^2(x)}{\sigma^2(x)}\right]}, \tag{5}$$

where $\propto$ represents an equality up to a constant (here, p_{ε_1} is a valid density function if and only if $\frac{1}{\beta^2} > \frac{e^2}{c}\mathbb{1}[a = 0]$). Specifically, if $\sigma(x)$ is constant (case $a = 0$), then the causal graph is identifiable, unless $\mu(x)$ is linear and p_{ε_1} is the Gaussian density.

The proof is provided in Appendix C.1. Moreover, a visual example of an unidentifiable Gaussian model with $a = c = d = e = \alpha = \beta = 1$ can be found in Appendix C.1, Figure 9.

Theorem 4 indicates that the non-identifiability holds only in the "special case," when $\frac{\mu(x)}{\sigma^2(x)}, \frac{-1}{2\sigma^2(x)}$ are linear and quadratic, respectively. Note that natural parameters of a Gaussian distribution are $\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}$, and sufficient statistics of the Gaussian distribution have a linear and quadratic form (for the definition of the exponential family, natural parameter and sufficient statistic, see Appendix A.1). We show that such connections between non-identifiability and sufficient statistics hold in the more general context of the exponential family.

Proposition 5 (General case, one parameter) *Let $q = 1$. Let (X_1, X_2) admit the $CPCM(F)$ model with graph $X_1 \to X_2$, where F lies in the exponential family of distributions with a sufficient statistic T. The causal graph is identifiable if at least one of the following conditions is met:*

1. *Supports of X_1 and X_2 differ.*
2. *There do not exist $a, b \in \mathbb{R}$, such that*

$$\theta(x) = a\,T(x) + b, \qquad \forall x \in supp(X_1). \tag{6}$$

3. *There does not exist $c \in \mathbb{R}$, such that*

$$p_{X_1}(x) \propto \frac{h_1(x)}{h_2[\theta(x)]} e^{cT(x)}, \qquad \forall x \in supp(X_1), \tag{7}$$

where h_1 is a base measure of F and h_2 is the normalizing function of F defined in Appendix A.1.

Idea of the proof While this proposition follows as a special case of Theorem 7, we outline the key ideas behind the proof.

If the graph is *not* identifiable, there exists a function $\tilde{\theta}$, such that causal models $X_1 = \varepsilon_1, X_2 = F^{-1}\big(\varepsilon_2; \theta(X_1)\big)$, and $X_2 = \varepsilon_2, X_1 = F^{-1}\big(\varepsilon_1; \tilde{\theta}(X_2)\big)$ generate the same joint distribution.

1) If the supports of X_1 and X_2 differ, then X_1 trivially can not be written as $X_1 = F^{-1}\big(\varepsilon_1; \tilde{\theta}(X_2)\big)$ since the support of any distribution in the exponential family is fixed. This immediately rules out non-identifiability in such cases. At the population level, even a difference on a measure-zero set is sufficient to ensure identifiability.

2) Assuming the causal graph is not identifiable, we decompose the joint density:

$$p_{(X_1,X_2)}(x,y) = p_{X_1}(x)p_{X_2|X_1}(y \mid x) = p_{X_2}(y)p_{X_1|X_2}(x \mid y), \qquad x, y \in supp(X_1). \tag{8}$$

Since F belongs to the exponential family, we rewrite the conditional densities using notation from (2):

$$p_{X_2|X_1}(y \mid x) = h_1(y)h_2[\theta(x)]\exp[\theta(x)T(y)], \quad p_{X_1|X_2}(x \mid y) = h_1(x)h_2[\tilde{\theta}(y)]\exp[\tilde{\theta}(y)T(x)].$$

Substituting this into (8) gives:

$$\begin{aligned} &p_{X_1}(x)h_1(y)h_2[\theta(x)]\exp[\theta(x)T(y)] = p_{X_2}(y)h_1(x)h_2[\tilde{\theta}(y)]\exp[\tilde{\theta}(y)T(x)], \\ &\underbrace{\log\left\{\frac{p_{X_1}(x)h_2[\theta(x)]}{h_1(x)}\right\}}_{f(x)} + \underbrace{\log\left\{\frac{h_1(y)}{p_{X_2}(y)h_2[\tilde{\theta}(y)]}\right\}}_{g(y)} = \tilde{\theta}(y)T(x) - \theta(x)T(y), \end{aligned} \tag{9}$$

where the second equation follows from taking the logarithm of both sides after dividing by h_1 and h_2. Differentiating both sides with respect to x and y, and fixing y such that $T'(y) \neq 0$, we obtain $\theta'(x) = \frac{\tilde{\theta}'(y)}{T'(y)}T'(x)$. Integrating this equation with respect to x leads to (6). Note that we do not need to assume differentiability of θ, and we can get around it by applying Lemma 18.

3) equality (9) implies $f(x)+g(y) = a_1T(x)+a_2T(y)+a_3$ for some constants $a_1, a_2, a_3 \in \mathbb{R}$. Therefore, fixing y yields $f(x) = \log\left\{\frac{p_{X_1}(x)h_2[\theta(x)]}{h_1(x)}\right\} = a_1T(x) + const.$ Rewriting this directly yields (7). ■

Intuitively, condition (6) rules out joint distributions that are symmetric with respect to the transformed axis $aT(x) + b$, similar to how symmetry around $y = a + bx$ causes non-identifiability in the Gaussian ANM case (Zhang and Hyvärinen, 2009). In the $CPCM(F)$ setting, the parameter $\theta(x)$ affects the distribution through the sufficient statistic $T(x)$. If $\theta(x)$ takes the form $T(x)$ (or $aT(x) + b$, since a sufficient statistic is defined only up to an affine transformation), the joint law has this symmetry, making both causal directions consistent with the data. By violating (6), the symmetry is broken and the true causal direction becomes identifiable.

As a consequence of Proposition 5, we extend the results demonstrated by Park and Raskutti (2015) and Park and Park (2019) for a Poisson DAG model. These authors established the identifiability of a Poisson DAG model, where all variables (including source

variables) given their parents follow a Poisson distribution. We present an analogous result, relaxing the restriction on the source variables.

Consequence 6 • *Let (X_1, X_2) admit the $CPCM(F)$ model with graph $X_1 \to X_2$, where F is the Poisson distribution function with rate λ. Then, the causal graph is not identifiable if and only if*

$$\lambda(x) = e^{ax+b}, \quad P(X_1 = x) \propto \frac{e^{\lambda(x)+cx}}{x!}, \quad \forall x \in \{0, 1, 2, \dots\},$$

for some $a < 0, b, c \in \mathbb{R}$.

- *Let (X_1, X_2) admit the $CPCM(F)$ model with graph $X_1 \to X_2$, where F is the Pareto distribution function. Then, the causal graph is not identifiable if and only if*

$$\theta(x) = a\log(x) + b, \quad p_{X_1}(x) \propto \frac{1}{[a\log(x)+b]x^{c+1}}, \quad \forall x \geq 1,$$

for some $a, b, c > 0$.

- *Let (X_1, X_2) admit the $CPCM(F)$ model with graph $X_1 \to X_2$, where F is Bernoulli distribution function. Then, the causal graph is identifiable if and only if $supp(X_1) \neq supp(X_2)$.*

The proof is provided in Appendix C.2, together with definitions of the distribution functions. Observe that if $a = 0$, then $X_1 \perp\!\!\!\perp X_2$ and the (empty) graph is trivially identifiable.

Note that in the first two bullet points of Consequence 6, we have three free parameters: a, b, and c. The non-identifiability of the graph in a Bernoulli model arises from the fact that the joint distribution of (X_1, X_2) can be fully characterized by only three parameters.

3.2 Identifiability in $CPCM(F_1, \dots, F_k)$ models

We generalize Proposition 5 to the general case. The following theorem establishes that $CPCM(F_1, \dots, F_k)$ models are "typically" identifiable, except for a finite-dimensional set within the space of all possible distributions.

Theorem 7 *Let (X_1, X_2) follow the $CPCM(F_1, \dots, F_k)$ model with graph $X_1 \to X_2$, where $F_1, \dots, F_k$ belong to the exponential family of distributions with corresponding sufficient statistics $T_m = (T_{m,1}, \dots, T_{m,q_m})^\top$, $m = 1, \dots, k$. Following Definition 2, let $\tilde{m} \in \{1, \dots, k\}$ be the index such that $X_2 = F_{\tilde{m}}^{-1}\big(\varepsilon_2; \theta_2(X_1)\big)$.*

The causal graph is identifiable if for all $m \in \{1, \dots, k\}$, at least one of the following holds:

- *$supp(F_m) \neq supp(X_1)$.*
- *The function θ_2 is not a linear combination of the sufficient statistics $T_{m,1}, \dots, T_{m,q_m}$, i.e., there do not exist coefficients $a_{i,j}, b_i \in \mathbb{R}$ for $i = 1, \dots, q_{\tilde{m}}$ and $j = 1, \dots, q_m$ such that*

$$\theta_{2,i}(x) = \sum_{j=1}^{q_m} a_{i,j} T_{m,j}(x) + b_i, \quad \forall x \in \mathrm{supp}(X_1), \quad \forall i \in \{1, \dots, q_{\tilde{m}}\}. \tag{10}$$

- *There do not exist constants* $c_1, \dots, c_{q_m} \in \mathbb{R}$ *such that the density of* X_1 *satisfies*

$$p_{X_1}(x) \propto \frac{h_{m,1}(x)}{h_{\tilde{m},2}[\theta_2(x)]} e^{\sum_{i=1}^{q_m} c_i T_{m,i}(x)}, \quad \forall x \in \text{supp}(X_1), \tag{11}$$

  where $h_{m,1}$ *is a base measure associated with* F_m *and* $h_{\tilde{m},2}$ *is the normalizing function of* $F_{\tilde{m}}$*, both defined in Appendix A.1.*

Consequentially, the space of non-identifiable distributions is contained in a $\tilde{d}$*-dimensional space, where*

$$\tilde{d} \leq \sum_{m \in \{1,\dots,k\}: supp(F_m) = supp(X_1)} (q_m + 1)(q_{\tilde{m}} + 1) - 1. \tag{12}$$

The proof is provided in Appendix C.3. It is insightful to examine the dimension of the space of unidentifiable distributions for different choices of F. In one-parameter cases, such as in Consequence 6, the set of all unidentifiable distributions lies within a three-dimensional space, similar to the case for ANM (Peters et al., 2014, Proposition 21). For the Gaussian $CPCM(F)$ model, Theorem 4 shows that this dimension is 6, despite Theorem 7 initially suggesting $(2+1)(2+1)-1 = 8$. In the proof of Theorem 4, we showed that two of these coefficients must be zero, confirming that (12) provides only an upper bound and the actual dimension can be smaller.

Since the space of all distributions is infinite-dimensional (assuming infinite support), one can argue that identifiability holds for "most distributions," regardless of the choice of F. However, when F has many parameters, the model often lies "close" to an unidentifiable case, making the finite sample inference significantly more challenging.

Consequence 8 *Suppose that* $supp(X_1) = \mathbb{R}$*,* $supp(X_2) = \{0, 1, \dots\}$ *such as on Figure 1, and let* (X_1, X_2) *admit the* $CPCM(F_1, F_2)$ *model with graph* $X_1 \to X_2$*, where* F_1 *is a Gaussian distribution and* F_2 *is a Poisson distribution with rate parameter* λ*. The causal graph is identifiable if and only if there do not exist constants* $a_1, a_2, b, c_1, c_2 \in \mathbb{R}$*,* $a_1, c_1 < 0$*, such that:*

$$\lambda(x) = e^{a_1 x^2 + a_2 x + b}, \quad p_{X_1}(x) \propto e^{c_1 x^2 + c_2 x}, \qquad \forall x \in \mathbb{R}.$$

Details and more examples are provided in Appendix C.4.

4. Multivariate case $d \geq 2$

We extend the theory to the case with possibly more than two variables, $\mathbf{X} = (X_1, \dots, X_d)^\top$.

Definition 9 *Let* $\{F_1, \dots, F_k\}$ *be a collection of distribution functions with* $q_1, \dots, q_k$ *parameters, respectively, where* $q_1, \dots, q_k \in \mathbb{N}$*. We define a conditionally parametric causal model* $CPCM(F_1, \dots, F_k)$ *with an underlying DAG* $\mathcal{G}$ *as a collection of equations:*

$$X_j = \begin{cases} \varepsilon_j, & \text{if } j \in Source(\mathcal{G}), \\ F^{-1}_{\pi(j)}\big(\varepsilon_j; \theta_j(\mathbf{X}_{pa_j})\big), \text{where } \pi(j) \in \{1, \dots, k\}, & \text{if } j \notin Source(\mathcal{G}), \end{cases}$$

where $(\varepsilon_1, \dots, \varepsilon_d)^\top$ *is a collection of jointly independent random variables with* $\varepsilon_j \sim U(0,1)$ *for all* $j \notin Source(\mathcal{G})$*, and* θ_j *are non-constant functions in any of their arguments.*

Simply said, we assume that $X_j \mid \mathbf{X}_{pa_j}$ is distributed according to distribution $F_{\pi(j)}$ with parameters $\theta_j(\mathbf{X}_{pa_j})$, where $F_{\pi(j)}$ is either $F_1, F_2, \ldots,$ or F_k. Although we implicitly assume causal minimality (Zhang and Spirtes, 2010), we do not require the stronger assumption of faithfulness (Uhler et al., 2013).

The question of the identifiability of $\mathcal{G}$ in the multivariate case is in order. Here, it is not satisfactory to consider the identifiability of each pair of $X_i \to X_j$ separately. Each pair X_i, X_j needs to have an identifiable causal relation *conditioned* on other variables $\mathbf{X}_S$. Such an observation was first made by Peters et al. (2014) in the context of additive noise models. We now provide a more precise statement in the context of $CPCM(F_1, \ldots, F_k)$.

Definition 10 *We say that the $CPCM(F_1, \ldots, F_k)$ is pairwise identifiable, if for all $i, j \in V$, $S \subseteq V$, such that $i \in pa_j$ and $pa_j \setminus \{i\} \subseteq S \subseteq nd_j \setminus \{i, j\}$, there exists $\boldsymbol{x}_S$: $p_S(\boldsymbol{x}_S) > 0$, which satisfies that a bivariate model defined as $X = \tilde{\varepsilon}_X, Y = F_j^{-1}\big(\tilde{\varepsilon}_Y, \tilde{\theta}(X)\big)$ is identifiable (in the sense of Definition 3), where $F_{\tilde{\varepsilon}_X} = F_{X_i \mid \boldsymbol{X}_S = \boldsymbol{x}_S}$ and $\tilde{\theta}(x) = \theta_j(\boldsymbol{x}_{pa_j \setminus \{i\}}, x)$, $x \in supp(X)$.*

Lemma 11 *Let $F_{\boldsymbol{X}}$ be generated by the pairwise identifiable $CPCM(F_1, \ldots, F_k)$ with DAG $\mathcal{G}$. Then, $\mathcal{G}$ is identifiable from the joint distribution.*

The proof follows as a consequence of Theorem 28 in Peters et al. (2014) and is provided in Appendix C.5.

Consequence 12 (Multivariate Gaussian case) *Suppose that $\boldsymbol{X} = (X_1, \ldots, X_d)$ follow $CPCM(F)$ with a Gaussian distribution function F. This corresponds to $X_j \mid \boldsymbol{X}_{pa_j} \sim N\big(\mu_j(\boldsymbol{X}_{pa_j}), \sigma_j^2(\boldsymbol{X}_{pa_j})\big)$ for all $j = 1, \ldots, d$ and for some functions μ_j, σ_j. In other words, we assume that the data-generation process has the following form:*

$$X_j = \mu_j(\boldsymbol{X}_{pa_j}) + \sigma_j(\boldsymbol{X}_{pa_j})\, \varepsilon_j, \qquad \text{where } \varepsilon_j \text{ is Gaussian.}$$

Potentially, source nodes can have arbitrary distributions. Combining Theorem 4 and Lemma 11, the causal graph $\mathcal{G}$ is identifiable if the functions $\theta_j(\boldsymbol{x}) := \big(\mu_j(\boldsymbol{x}), \sigma_j(\boldsymbol{x})\big)^\top, \boldsymbol{x} \in \mathbb{R}^{|pa_j(\mathcal{G})|}$, $j = 1, \ldots, d$, are not in the form (4) in any of their arguments.

5. Inference

5.1 CPCM algorithm - bivariate case

Our CPCM methodology is based on selecting an appropriate causal model (in our case, the choice of collection $\{F_1, \ldots, F_k\}$) and a measure of a model fit. In the following subsections, we measure the model fit by exploiting the principle of independence between the cause and the mechanism.

We say that a DAG $\mathcal{G}$ is **plausible** under $CPCM(F)$ model if the joint distribution *can* be generated via $CPCM(F)$ model with graph $\mathcal{G}$. The Algorithm 1 describes the main steps to test the plausibility and estimation of $\mathcal{G}$ in the bivariate case.

An estimation of $\theta(X_1)$ in Step 1a) is discussed in detail in Appendix A.2. It can be performed using any suitable machine learning method, such as GAM, GAMLSS, random forests, or neural networks (Green and Silverman, 1994; Stasinopoulos and Rigby, 2007). For

Algorithm 1: CPCM(F) - bivariate case

Data: Random sample $(x_{1,1}, x_{2,1})^\top, \ldots, (x_{1,n}, x_{2,n})^\top$
Result: Estimate $\hat{\mathcal{G}}$ and plausibility of graphs $X_1 \to X_2$ and $X_2 \to X_1$
Step 0) Test independence between X_1 and X_2.
Step 1) Determine plausibility of $X_1 \to X_2$ using the following:
1a) Estimate $\theta(X_1)$ in $X_2 = F^{-1}(\varepsilon_2; \theta(X_1))$; compute $\hat{\varepsilon}_2 := F(X_2; \hat{\theta}(X_1))$.
1b) Test independence between $\hat{\varepsilon}_2$ and X_1. If the p-value is larger than $\alpha = 0.05$, mark $X_1 \to X_2$ as *plausible*.
Step 2: Repeat Step 1 for $X_2 \to X_1$.
Forced estimate (Choose the direction with the higher residual independence):
Return $X_1 \to X_2$ if p-value$(\hat{\varepsilon}_2, X_1) >$ p-value$(\hat{\varepsilon}_1, X_2)$, else return $X_2 \to X_1$.
Conservative Estimate:
- If Step 0 fails to reject independence, return $\hat{\mathcal{G}} = \emptyset$.
- If exactly one of the two graph directions is plausible, return it as $\hat{\mathcal{G}}$.
- If both directions are plausible, return `"Unidentifiable case"`.
- If neither is plausible, return `"Assumptions not fulfilled"`.

the independence test in Step 1b), one may use the HSIC test (kernel-based Hilbert–Schmidt independence criterion; Pfister et al. (2018)) or a copula-based test (Genest et al., 2019).

The conservative estimate $\hat{\mathcal{G}}$ helps guard against unfulfilled assumptions or unidentifiable cases. If it returns "Assumptions not fulfilled," it means that we were unable to fit the $CPCM(F)$ model in either direction, suggesting that the variables do not follow the $CPCM(F)$ model. In this case, one should consider increasing the complexity (i.e., the number of parameters) of F. This is discussed further in Section 5.3.

If it returns "Unidentifiable case," this means that we were able to fit the $CPCM(F)$ model in both directions. This could indicate that the sample size is too small or that we are in an unfortunate unidentifiable case, such as the one described in Consequence 8. In this case, one should consider decreasing the complexity (i.e., the number of parameters) of F.

While the warnings from the conservative estimate $\hat{\mathcal{G}}$ are useful, we often require a single estimate of $\hat{\mathcal{G}}$ for comparison with other benchmark methods.

5.1.1 Extension to $CPCM(F_1, \ldots, F_k)$

The following adjustment to Step 1 in Algorithm 1 can be applied to accommodate the $CPCM(F_1, \ldots, F_k)$ model, given a collection $\{F_1, \ldots, F_k\}$.

***CPCM*$(F_1, \ldots, F_k)$**

Step 1) Determine plausibility of $X_1 \to X_2$ using the following:

Step 1a) Estimate the set $S := \{j \in \{1, \ldots, k\} : supp(F_j) = supp(X_2)\}$. If empty, return "STOP: Inappropriate choice of F".

Step 1b) For all $j \in \hat{S}$, estimate $\theta(X_1)$ in a model $X_2 = F_j\big(\varepsilon_2; \theta(X_1)\big)$, and compute probability transform $\hat{\varepsilon}_2^j := F_j\big(X_2; \hat{\theta}(X_1)\big)$.

Step 1c) Compute the p-value of an independence test between $\hat{\varepsilon}_2^j$ and X_1 for all $j \in \hat{S}$. Choose the largest p-value. Direction $X_1 \to X_2$ is marked as *plausible* if this p-value is larger than $\alpha = 0.05$.

By estimating the set S in Step 1a), we can preliminarily filter out the F_j choices that are evidently unsuitable, such as fitting Gaussian model when the data are discrete. We can apply a simple heuristic under the assumption that $\text{supp}(X_2)$ and $\text{supp}(F_j)$ for all j are one of the following: 1) $\mathbb{R}$ (Gaussian), 2) $\mathbb{R}^+$ (Gamma), 3) $[0, 1]$ (Beta), or 4) $\mathbb{N}$(Poisson). The heuristic is as follows: if the number of unique values in $(x_{2,1}, \ldots, x_{2,n})$ is fewer than $n/10$ and the values lie in $\mathbb{N}$, we set $\text{supp}(X_2) = \mathbb{N}$. Otherwise, if all values lie within $[0, 1]$, we set $\text{supp}(X_2) = [0, 1]$. To distinguish between $\mathbb{R}$ and $\mathbb{R}^+$, we use the skewness of the distribution: if the skewness is close to 0, the distribution resembles a Gaussian distribution, so we set $\text{supp}(X_2) = \mathbb{R}$. Otherwise, the distribution resembles a Gamma distribution, so we set $\text{supp}(X_2) = \mathbb{R}^+$.

5.2 Multivariate case: $CPCM(F)$ as a score-based DAG optimization

The forced estimate of $\mathcal{G}$ in Algorithm 1 can be seen as a score-based algorithm; defining the score of a graph as a p-value of the corresponding independence test, we simply compare the scores of graphs $X_1 \to X_2$ and $X_2 \to X_1$. In the following, we formalize this idea, leading to generalizing the CPCM inference to multivariate case.

Following the ideas of Rajendran et al. (2021); Nowzohour and Bühlmann (2016); Peters et al. (2014), we use the following penalized independence score:

$$\hat{\mathcal{G}} = \underset{\mathcal{G} \in DAG(d)}{\arg\min}\ s(\mathcal{G}) = \underset{\mathcal{G} \in DAG(d)}{\arg\min}\ \rho(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_d) + \lambda(\text{Number of edges in } \mathcal{G}), \tag{13}$$

where ρ represents some dependence measure, λ is a hyperparameter, $DAG(d)$ is the set of all DAGs over $V = \{1, \ldots, d\}$ and $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_d$ are noise estimators obtained by estimating $\theta_i(\mathbf{X}_{pa_i(\mathcal{G})})$ and putting $\hat{\varepsilon}_i := F\big(X_i; \hat{\theta}_i(\mathbf{X}_{pa_i(\mathcal{G})})\big)$ analogously to Algorithm 1.

With regard to choice of ρ, we use minus the logarithm of the p-value of the independence test (Genest et al., 2019) and $\lambda = 2$. These choices appear to work well in practice, but we do not provide any theoretical justification of their optimality.

Analogously to the bivariate case, we can define that a DAG $\mathcal{G}$ is **plausible** if the p-value of the corresponding independence test between $(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_d)$ is larger than level $\alpha \in (0, 1)$. If every $\mathcal{G} \in DAG(d)$ is not plausible, it suggests that the variables do not follow the $CPCM(F)$ model. In this case, one should consider increasing the complexity (i.e., the

number of parameters) of F; this is discussed in more detail in Section 5.3.4. On the contrary to the bivariate case, many DAGs can be plausible in an identifiable $CPCM(F)$ model. In particular, if a true causal graph $\mathcal{G}$ is plausible, than any $\tilde{\mathcal{G}} \supseteq \mathcal{G}$ should be plausible.

It is important to note that in a bivariate case, score based estimate (13) is equal to the forced output of Algorithm 1, given $\lambda = -\infty$.

5.2.1 $CPCM(F_1, \dots, F_k)$ extension

Similarly to the extension of Algorithm 1, the following adjustment can be applied to accommodate the $CPCM(F_1, \dots, F_k)$ model, given a collection $\{F_1, \dots, F_k\}$:

$$s(\mathcal{G}) = \min_{j_1 \in \hat{S}_1, \dots, j_d \in \hat{S}_d} \rho(\hat{\varepsilon}_1^{j_i}, \dots, \hat{\varepsilon}_d^{j_d}) + \lambda(\text{Number of edges in } \mathcal{G}), \tag{14}$$

where $\hat{S}_i$ are estimates of $S_i := \{j \in \{1, \dots, k\} : supp(X_i) = supp(F_j)\}$, and where $\hat{\varepsilon}_i^{j_i} := F_{j_i}\big(X_i; \hat{\theta}_i(\mathbf{X}_{pa_i(\mathcal{G})})\big)$. If $supp(F_i) \neq supp(F_j)$ for all $i \neq j$, we end up with a single evaluation of the score function for each possible $\mathcal{G}$. Finally, analogously to $CPCM(F)$, we say that DAG $\mathcal{G}$ is plausible if there exist $j_1 \in \hat{S}_1, \dots, j_d \in \hat{S}_d$ such that the p-value of the independence test between $(\hat{\varepsilon}_1^{j_i}, \dots, \hat{\varepsilon}_d^{j_d})$ is larger than level $\alpha \in (0, 1)$.

5.2.2 Consistency

Mooij et al. (2016) demonstrates consistency of a causal discovery algorithm in ANMs. We establish an analogous result for $CPCM(F_1, \dots, F_k)$.

Proposition 13 *Let (X_1, X_2) follow an identifiable $CPCM(F_1, \dots, F_k)$ with DAG $\mathcal{G}$. Then, our score based algorithm presented in Section 5.2 is consistent, meaning that*

$$\hat{\mathcal{G}} \xrightarrow{P} \mathcal{G} \ \text{ as } n \to \infty,$$

given that we employ a "suitable" estimation procedure for the estimation $\hat{\varepsilon}_i$, we use HSIC score as our choice of ρ and consistent estimators $\hat{S}_i$ (for definitions, proof and more details, see Appendix A.2).

Additionally, in the *unidentifiable* case, Algorithm 1 detects non-identifiability with high probability, as stated in the following lemma.

Lemma 14 (Algorithm 1 under unidentifiability) *Let (X_1, X_2) follow an unidentifiable model $CPCM(F_1, \dots, F_k)$. Assume that Algorithm 1 employs the HSIC independence test in step 1b) and that the regression function is estimated perfectly, i.e., using oracle estimator $\hat{\theta} = \theta$. Then, for sufficiently large n, Algorithm 1 outputs `"Unidentifiable case"` with probability at least $1 - 2\alpha$.*

Extending Lemma 14 to the non-oracle setting or Proposition 13 beyond the bivariate case remains technically challenging and is left for future work. Although Pfister et al. (2018) proved the asymptotic validity of bootstrap-based p-values for the multivariate HSIC test, estimation error in $\hat{\theta}$ propagates to the residuals and alters the variability of the HSIC statistic, complicating its asymptotic analysis. A rigorous generalization would require deriving asymptotic distributions for higher-order kernel statistics under estimated residuals,

which is beyond the current theoretical scope of our work. Empirically, however, as shown in Section 6.2, the resulting p-values remain well-calibrated even in the presence of estimation error.

5.2.3 Scalability and greedy algorithms for larger dimensions

The main disadvantage of the proposed method is that we have to go through all graphs $\mathcal{G} \in DAG(d)$, which is possible only for very small d. However, even though graph learning is typically NP-hard (Chickering et al., 2004), numerous algorithms have been proposed to speed up the process (Chickering, 2002; Silander and Myllymäki, 2006; Ramsey et al., 2016; Nandy et al., 2018; Rajendran et al., 2021).

The **naive-edge-greedy** algorithm (Chickering, 2002) is one such algorithm that iteratively adds or removes edges to minimize the CPCM score (13). Another is the **RESIT** algorithm (Peters et al., 2014), which first estimates a topological ordering and then prunes redundant edges via independence tests. RESIT can be naturally adapted to the CPCM framework (Algorithm 4 in Appendix B.1) and retains large-sample consistency guarantees (Lemma 17 in Appendix B.1). However, its performance tends to deteriorate in higher dimensions where early ordering errors and type I error introduce spurious edges.

To combine the strengths of both approaches, we introduce a hybrid **RESIT-greedy** algorithm that merges Phase 1 of RESIT with the edge-pruning strategy of greedy search; see Algorithm 2. In Phase 1, the topological ordering is estimated by iteratively removing the node whose residuals $\hat{\varepsilon}_i$ exhibit the weakest dependence on the remaining variables; this node is then appended to the ordering. Alternative procedures for estimating the topological order also exist (Gnecco et al., 2020). In Phase 2, instead of performing independence tests, we apply a greedy edge-removal procedure guided by the CPCM score: starting from the fully connected graph consistent with the estimated ordering, we iteratively remove the edge whose deletion yields the largest decrease in the CPCM score, stopping when no further improvement is possible.

In Appendix B.1, we compare these algorithms in terms of accuracy and computational time. The results (unsurprisingly) show that the exact method achieves the highest accuracy but is computationally feasible only for small graphs ($d < 5$). RESIT and naive-greedy methods are the most scalable but exhibit worst accuracy. The RESIT-greedy algorithm strikes a balance: it consistently outperforms both RESIT and naive-greedy in terms of accuracy while maintaining reasonable computational efficiency for moderate dimensions ($d < 10$), making it a practical choice in such cases.

For much larger dimensions, a hybrid approach is recommended: using e.g. PC algorithm to estimate the skeleton and applying the CPCM algorithm only to smaller subgroups of unoriented edges. Similar strategy has been discussed, for example, in (Goudet et al., 2017), though a detailed exploration is beyond the scope of this paper.

5.3 Choice of the collection $\{F_1, \dots, F_k\}$

5.3.1 Why $\{F_1, \dots, F_k\}$ cannot be chosen in a data-driven way

Selecting the collection $\{F_1, \dots, F_k\}$ is a crucial step in our approach. Unfortunately, there is no principled, general data-driven way to select this collection without access to alternative data. As discussed in Section 2.1.1, alternative methods such as ANM, QVF, bQCD, or

Algorithm 2: RESIT-greedy algorithm

Input: Random sample of $(X_1, \ldots, X_d)$
Phase 1 (topological order): Obtain topological order π via RESIT (Algorithm 4).
Phase 2 (edge removal): Initialize $\mathcal{G} \leftarrow \{(j \to i) : j \text{ precedes } i \text{ in } \pi\}$.
repeat
  $S \leftarrow s(\mathcal{G})$, where $s(\cdot)$ is the score defined in Equation (13) or (14);
  $e^\star \leftarrow \arg\min_{e \in \mathcal{G}} s(\mathcal{G} \setminus \{e\})$;
  if $s(\mathcal{G} \setminus \{e^\star\}) < S$ **then** $\mathcal{G} \leftarrow \mathcal{G} \setminus \{e^\star\}$;
until ***no improvement***;
Output: $\mathcal{G}$

IGCI all predefine the notion of a "simple" distribution. This predefinition is necessary; otherwise, Occam's razor loses its meaning. The following lemma formalizes this idea.

Lemma 15 *Suppose that the joint distribution $F_{(X_1,X_2)}$ is generated according to the model $CPCM(F_2)$ with graph $X_1 \to X_2$, where F_2 is a distribution function belonging to the exponential family.*

Then, there exists F_1 such that the model $CPCM(F_1)$ with graph $X_2 \to X_1$ also generates $F_{(X_1,X_2)}$. In other words, there exists F_1 such that the causal graph in $CPCM(F_1, F_2)$ is not identifiable from the joint distribution.

The proof (provided in Appendix C.6) is based on the specific choice of F_1, such that its sufficient statistic is equal to the parameter θ_2 from the original model $CPCM(F_2)$.

It is important to note that such an F_1 often results in a rather non-standard distribution. While the notion of a "standard" distribution is somewhat philosophical, some distributions are objectively considered standard due to physical motivations; for example, the Gaussian (by the central limit theorem) or the Poisson (which arises when counting independent events). This motivates the definition of a "standard set of well-known distributions," which we discuss in Section 5.3.3.

5.3.2 Unfair game issue

Even if the theory suggests that it is not possible to fit a $CPCM(F_1, \ldots, F_k)$ in both causal directions, this is only an asymptotic result and may not hold in practice with finite samples. Overparameterized models may overfit and capture spurious patterns, while underparameterized ones may violate key assumptions. To ensure a fair comparison, we recommend selecting all F_i with the same number of parameters ($q_i = q_j$). For example, choosing F_1 with one parameter and F_2 with three ($q_1 = 1$, $q_2 = 3$) introduces bias, since the more flexible model is more likely to fit the data regardless of its correctness (unless we are in the asymptotic regime). We refer to this issue as an "unfair game."

5.3.3 Practical choices of the set of "standard well-known distributions"

We define nested sets of "standard set of well-known distributions"; set $\mathscr{S}_1$ containing distributions with one parameter, and a more complex set $\mathscr{S}_2$ containing distributions with two

parameters. We do this to avoid the "unfair game" issue. Both sets should be rich enough to contain a wide range of distributions with different supports and characteristics, but should not contain many distributions with the same support in order to avoid unidentifiable setups.

For practical purposes, we restrict our attention to distributions that are implemented in `mgcv` package in `family.mgcv` (Wood, 2017). These include:

1. $\mathscr{S}_1$ consists of the following one parameter distributions: Gaussian with fixed variance, Poisson, Pareto and Exponential distribution.

2. $\mathscr{S}_2$ consists of the following two parameter distributions: Gaussian, Negative binomial, Generalized Pareto and Gamma distribution.

These choices are tactically made such that every distribution in the family $\mathscr{S}_1$ is a special case of some distribution in $\mathscr{S}_2$; for example, the Poisson distribution arises as a special (limiting) case of the Negative Binomial distribution (see Casella and Berger, 2024, p. 96).

We emphasize that many other choices are possible, and our collections are neither exhaustive nor immutable. They may be adapted for specific applications where alternative distributions could be regarded as "standard."

5.3.4 $\mathscr{S}_1$ or $\mathscr{S}_2$? Sequential approach

We propose a sequential approach for the selection between $\mathscr{S}_1$ and $\mathscr{S}_2$. We first choose the (least complex) set $\mathscr{S}_1$; if this choice is "not appropriate", we then extend the set of distributions to $\mathscr{S}_1 \cup \mathscr{S}_2$. If $\mathscr{S}_1 \cup \mathscr{S}_2$ is also "not appropriate", we may further extend the model class by introducing $\mathscr{S}_3$: the class of standard three-parameter distributions. This hierarchical selection process incrementally increases the complexity of the conditional distributions until at least one graph is plausible.

We define that $\mathscr{S}_1$ is "not appropriate" if there does not exist any *plausible* graph $\mathcal{G} \in DAG(d)$ under $CPCM(\mathscr{S}_1)$. This is formally defined in Algorithm 3.

Algorithm 3: Sequential approach for the choice between $\mathscr{S}_1$ and $\mathscr{S}_2$

Exact version: For each $\mathcal{G} \in \mathrm{DAG}(d)$, infer whether it is plausible under $CPCM(\mathscr{S}_1)$ (as discussed in Section 5.2). If no graph is plausible, return $\mathscr{S}_1 \cup \mathscr{S}_2$; otherwise, return $\mathscr{S}_1$.

Greedy version: When employing a (RESIT-)greedy algorithm, scores are evaluated for multiple candidate DAGs. For each candidate, test plausibility under $CPCM(\mathscr{S}_1)$; if none are plausible, return $\mathscr{S}_1 \cup \mathscr{S}_2$; otherwise, return $\mathscr{S}_1$.
Note: The plausibility check is effectively free, as it relies on the same dependence measure $\rho(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_d)$ already computed in the score.

Fast version: Estimate $\hat{\mathcal{G}}$. Test its plausibility under $CPCM(\mathscr{S}_1)$; if $\hat{\mathcal{G}}$ is not plausible return $\mathscr{S}_1 \cup \mathscr{S}_2$; otherwise, return $\mathscr{S}_1$.

By sequentially expanding $\mathscr{S}_1$ by $\mathscr{S}_2$, and potentially by $\mathscr{S}_3$, we ensure that the simplest adequate model is chosen while avoiding unnecessary complexity and unfair game issue. As an example when our data are continuous uni-modal with full support, consider that a Gaussian model with fixed variance is fitted in both directions $X \to Y$ and $Y \to X$. If both

graphs are not plausible, we expand the model to location-scale Gaussian model. Potentially, if both directions are again not plausible in this model, we add a shape parameter and consider Generalized Gaussian model (Nadarajah, 2005).

Consequence 16 *Let (X_1, X_2) follow an $CPCM(F)$ model with DAG $\mathcal{G}$, for some $F \in \mathscr{S}_1 \cup \mathscr{S}_2$. Assume the conditions of Proposition 13 hold: namely, identifiability of $CPCM(\mathscr{S}_1 \cup \mathscr{S}_2)$, employing a "suitable" estimation procedure and the HSIC score. Then, our score-based algorithm, with the collection $\{F_1, \dots, F_k\}$ chosen via the Sequential approach (Exact version), is consistent: $\hat{\mathcal{G}} \xrightarrow{P} \mathcal{G}$, as $n \to \infty$.*

6. Experiments

The `R` code for the presented algorithms, simulations, and application is available at https://github.com/jurobodik/Causal_CPCM.git.

- In Section 6.1, we investigate the robustness of our approach against misspecifications of the choice of F (the data are generated with F being an exponential distribution but we use different choices for F such as Gaussian or Pareto).
- In Section 6.2, we empirically validate Consequence 6 and Proposition 13, showing that the distribution of the p-values (scores) produced by Algorithm 1 across different causal graphs is approximately uniform.
- In Section 6.3 and Section 6.4, we compare our methodology with state-of-the-art methods using bivariate and multivariate benchmark datasets, respectively.
- In Section 6.5 we consider a toy example on real-world data.
- In Appendix B.1, we evaluate the performance of different greedy algorithms. This simulation study suggests that RESIT-greedy achieves better accuracy than standard greedy or RESIT methods, but at the cost of slightly higher computational complexity, making it a practical choice for moderately sized graphs ($d < 10$).
- In Appendix B.2, we empirically evaluate the Sequential approach, compared to oracle choice of F.

Implementation details. In all experiments, we estimate $\theta(X_i)$ using Generalized Additive Models (GAM, Wood et al. (2016)). We use the HSIC as an independence test and approximate the null distribution with a gamma distribution in order to obtain p-values (Pfister et al., 2018). In Section 6.2, we use the conservative estimate in Algorithm 1; elsewhere, for comparability with other methods, we use the Forced/Score-based estimate (recall that the forced estimate is a special case of the score-based estimate when $d = 2$). For the sequential approach, the exact version is applied with the exact CPCM algorithm, the fast version with RESIT, and the greedy version with RESIT-greedy or edge-greedy algorithms.

6.1 Robustness against a misspecification of F

We consider the structural model $X_1 \to X_2$, where $X_1 \sim N(2,1)^+$ (Gaussian truncated to $x > 0$) and

$$X_2 \mid X_1 \sim \text{Exponential}\big(\theta(X_1)\big), \tag{15}$$

for some non-negative function θ. This corresponds to model (3) with F equal to the exponential distribution, a special case of the Gamma family with a fixed shape parameter.

The simulation evaluates how misspecifying F affects causal graph estimation. We generate $n = 1000$ samples from (15), apply our framework for several candidate families, and record the proportion of correctly identified directions (using Forced Algorithm 1), averaged over 100 repetitions.

Results are shown in Table 1. The method is robust when F resembles the exponential distribution in support and tail behavior, but accuracy drops sharply for e.g. $F =$ Gaussian. Surprisingly, $F =$ Gamma, the correctly specified family, performs relatively poorly due to over-parameterization: the reverse model $X_2 \to X_1$ can often fit well with two parameters. Using three-parameter families typically reduces accuracy to about 50%, no better than a coin toss, showing the importance of controlling model complexity in causal discovery.

Family F	#param	$\theta(x) =$ random	$\theta(x) = x$	$\theta(x) = x^2 + 1$	$\theta(x) = e^x/2$
Exponential (oracle)	1	0.99	0.98	0.99	1.00
Gamma, fixed scale	1	0.96	0.96	1.00	0.99
Pareto, shifted support	1	0.99	0.99	1.00	1.00
Gumbel, fixed scale	1	0.36	0.00	0.01	0.00
Gaussian, fixed σ	1	0.01	0.00	0.01	0.00
Gamma	2	0.96	0.73	0.64	0.79
Gumbel	2	0.68	0.87	0.91	0.97
Gaussian	2	0.69	0.07	0.32	0.29

Table 1: Accuracy of CPCM estimations for different distribution families F, with the Exponential distribution as the ground truth. "Random" $\theta(x)$ is generated via Gaussian processes as detailed in Appendix B.3.

6.2 Empirical validation of Consequence 6, Proposition 13 and p-value distribution in Pareto model

As in Consequence 6, we consider the CPCM Pareto model $X_1 \to X_2$ with

$$p_{X_1}(x) \propto \frac{1}{[\log(x)+1]x^2}, \quad X_2 \mid X_1 \sim \text{Pareto}\big(\theta(X_1)\big), \quad \theta(x) = x^\gamma \log(x) + 1, \tag{16}$$

where $\gamma \in \mathbb{R}$ measures deviation from the unidentifiable case ($\gamma = 0$). For $\gamma > 0$, the causal graph should be identifiable; for $\gamma < 0$, θ is nearly constant and X_1, X_2 are almost independent.

Using $\gamma \in \{-2, -1, 0, 1, 2\}$, we apply the Conservative Algorithm 1 with Pareto F and sample size $n = 500$. Note that Algorithm 1 has five possible outcomes: 1) $X_1 \perp\!\!\!\perp X_2$, 2) $X_1 \to X_2$, 3) $X_2 \to X_1$, 4) "unidentifiable setup" (both directions appear to be plausible) and 5) "Assumptions not fulfilled" (neither direction appears to be plausible).

- Table 4 in Appendix B.3 shows the results averaged over 100 repetitions. The results align with the theory: if $\gamma = 0$, we typically estimate both directions to be plausible. If $\gamma > 0$, we tend to estimate the correct direction $X_1 \to X_2$; if $\gamma < 0$, we tend to estimate an empty graph since X_1 and X_2 are (nearly) independent.

- Figure 6 in Appendix B.3 shows the distributions of the p-values from the independence test in Step 1b) of Algorithm 1, using data simulated with $\gamma = 0$ and $\gamma = 2$. The distribution of the p-values appears roughly uniform on $(0, 1)$ in the $\gamma = 0$ case and in the correct direction $X \to Y$, while the p-values in the direction $Y \to X$ are typically very close to 0 in the identifiable setup $\gamma = 2$.

- Figure 7 in Appendix B.3 shows an empirical validation of the consistency result from Proposition 13. For a range of sample sizes n, we generate the dataset using $\gamma = 1$ and $\gamma = 2$ as the hyperparameters, and we compute the percentage (out of 100 repetitions) of correctly estimated causal direction. The algorithm appears to achieve near-perfect performance for large sample sizes.

6.3 Comparison with baseline methods: bivariate case

We compare our method with LOCI (Immer et al., 2023), HECI (Xu et al., 2022), RESIT (Peters et al., 2014), bQCD (Tagasovska et al., 2020), IGCI with Gaussian and uniform reference measures (Janzing and Schölkopf, 2010) and Slope (Marx and Vreeken, 2019). As in Mooij et al. (2016), we use the accuracy for forced decisions as our evaluation metric. Details can be found in Appendix B.3.

We consider seven benchmark datasets, described below. The first five datasets are taken directly from Tagasovska et al. (2020) and described in Appendix B.3. They consist of additive and location-scale Gaussian pairs of the form $X_2 = \mu(X_1) + \sigma(X_1)\varepsilon_2$, where $\varepsilon_2 \sim \mathcal{N}(0, 1)$. In each case, we generate $X_1 \sim N(0, \sqrt{2})$.

1. **LSg (Location-Scale Gaussian)**: Here, μ and σ are nonlinear functions simulated using Gaussian processes.

2. **LSs (Location-Scale Sigmoid)**: In this setup, μ and σ are sigmoid functions.

3. **ANMg (Additive Noise Model)**: Nonlinear additive noise models generated similarly to LSg, but with constant $\sigma(X_1) = \sigma \sim U(1/5, \sqrt{2/5})$.

4. **ANMs (Additive Noise Model)**: Nonlinear additive noise models generated similarly to LSs, but with constant $\sigma(X_1) = \sigma \sim U(1/5, \sqrt{2/5})$.

5. **MNs (Multiplicative Noise)**: Nonlinear multiplicative noise models generated similarly to LSs, but with $\mu(X_1) = 0$.

In addition to the models from Tagasovska et al. (2020), we consider two more setups:

6. **POISg (Poisson Model)**: $X_2 \sim \text{Pois}\big(\lambda(X_1)\big)$, where λ is generated using Gaussian processes similar to σ in LSg. Observe that X_2 is discrete, creating error in some methods.

7. **PARg (Pareto Model)**: $X_2 \sim \text{Pareto}\big(\theta(X_1)\big)$, where θ is generated using Gaussian processes similar to σ in LSg.

For each of the seven setups, we simulate 100 pairs with $n = 1000$ data points each. One realization of each model can be found in Figure 8 in Appendix B.

The results are presented in Table 2. We conclude that our estimator performs well across all considered datasets, effectively handling the mix of discrete, continuous and heavy-tailed variables. Under the Gaussian location-scale setups, it provides comparable results to LOCI and bQCD, which are specifically developed for such cases.

Method	**ANMg**	**ANMs**	**MNs**	**LSg**	**LSs**	**POISg**	**PARg**
CPCM (Seq. choice) Forced Algorithm 1	**100**	**99**	**88**	**98**	**92**	**94**	**90**
LOCI (NN H)	**100**	**100**	**99**	**91**	**85**	79	0
HECI	**98**	**43**	**29**	**96**	**54**	—	100
ANM-RESIT	**100**	**100**	39	51	11	0	12
bQCD (m=3)	**100**	**79**	**99**	**100**	**98**	97	34
IGCI (Gauss)	100	99	99	97	100	0	0
IGCI (Unif)	31	35	12	36	28	0	100
Slope	22	25	9	12	15	0	100

Table 2: Accuracy (%) of different estimators on the simulated datasets. Similar results (excluding the first row and the last two columns) can also be found in Immer et al. (2023). Bold entries indicate cases where high accuracy is expected because the data-generating mechanism aligns with the method's modeling assumptions.

6.4 Comparison with baseline methods: multivariate case

We compare our proposed CPCM method with several widely used causal discovery algorithms implemented in `R`. These include LINGAM (Shimizu et al., 2006), ANM (RESIT) (Peters et al., 2014; Hoyer et al., 2009), PC (constraint-based method that tests for conditional independences, Kalisch et al. (2012)), and GES (score-based algorithm that greedily searches over equivalence classes of DAGs using the BIC score, Ramsey et al. (2016)). We also include a random baseline that selects a DAG uniformly at random from the space of all DAGs on d nodes. Many other methods, such as NOTEARS (Zheng et al., 2018) or DAG-GNN (Yu et al., 2019), would also be appropriate comparisons. However, we restrict our evaluation to methods directly available in `R` to ensure consistency and reproducibility within our implementation framework. More details can be found in Section B.3.

We generate data from 4 different scenarios:

- **Linear (non-gaussian)**: $X_i = \sum_{k \in pa(i)} X_k + \varepsilon_k$, where $\varepsilon_k \overset{i.i.d}{\sim} Exp(1)$,
- **Nonlinear ANM**: $X_i \sim N(\mu(\mathbf{X}_{pa_i}), 1)$ where $\mu(\mathbf{X}_{pa_i}) = \sum_{k \in pa(i)} sin(X_k) + \frac{1}{2} X_k + \varepsilon_k$, where $\varepsilon_k \overset{i.i.d}{\sim} N(0, 1)$
- **CPCM(Exp)**: $X_i \sim Exp(\lambda(\mathbf{X}_{pa_i}))$, where $\lambda(\mathbf{X}_{pa_i}) = \sum_{k \in pa(i)} |X_k| \vee 0.1$,
- **CPCM(Exp, Gauss)**: each node is, with probability $\frac{1}{2}$, generated either according to the nonlinear ANM case or the $CPCM(Exp)$ case.

The underlying DAG $\mathcal{G}$ is generated uniformly at random by sampling a random ordering of the d variables and including each of the $\frac{d(d-1)}{2}$ admissible edges with probability $\frac{2}{d-1}$, resulting in an expected total of d edges (following Peters et al. (2014)). We report results for $d = 5$; results for $d = 10$ are similar and omitted for brevity. For each of the four scenarios, we simulate 50 graphs with $n = 1000$ data points each. We compare the methods using the Structural Intervention Distance (SID, Peters and Bühlmann (2013)). For fairness, undirected edges in the estimated graph are counted as correct if the true edge exists in either direction.

In the linear case, it performs nearly as well as LINGAM, which is tailored to linear non-Gaussian models. In the nonlinear ANM scenario, it is only slightly less accurate than ANM (Peters et al., 2014). The comparison between RESIT and its greedy variant highlights that in lower dimensions, RESIT-greedy tends to yield better accuracy. Most importantly, in the non-Gaussian $CPCM(F)$ settings, our method clearly outperforms all baselines, underscoring its robustness and suitability for non-additive and heterogeneous functional forms, where standard approaches often fail.

Method / Scenario (Case $d = 5$)	**Linear model**	**Nonlin. ANM**	$\boldsymbol{CPCM(F_1)}$ $F_1 = \text{Exp}$	$\boldsymbol{CPCM(F_1, F_2)}$ $F_1 = \text{Exp}$, $F_2 = \text{Gauss}$
CPCM (Seq. app., RESIT-greedy)	0.22	0.72	**2.16**	**0.92**
LINGAM	**0.12**	4.72	4.52	5.51
ANM (RESIT)	1.10	1.34	12.21	6.21
ANM (RESIT-greedy)	**0.12**	**0.66**	11.81	6.84
PC (gaussCItest)	1.90	2.20	4.28	3.15
PC (HSIC)	**0.12**	1.50	2.32	1.92
GES (Gaussian score)	1.96	2.34	4.24	3.15
Random	7.20	7.58	7.50	7.57

Table 3: Average SID over 100 repetitions between estimated and true graphs under different methods and scenarios. Bold values are the best (lowest) across methods.

6.5 Illustration using a motor insurance dataset

Understanding the causal relationships between driver and vehicle characteristics is a key step in insurance analytics, as it informs both risk assessment and premium setting. We

demonstrate the advantages of CPCM using a subset of the French MTPL motor insurance dataset (Sarpal, 2025), restricted to four variables: "ClaimNb" records the number of claims during the policy period (integer between 1 and 16) and lending itself naturally to a Poisson model. "VehPower" and "VehAge" are vehicles engine power and age, and "Exposure" is the duration that the policy was active. Exponential/Gamma distribution is a natural model for variables "VehPower", "VehAge", and "Exposure" due to their exponential-type shapes. Since the dataset contains almost a million observations, we focus on a subset of the first $n = 1000$ records; results based on different random subsamples are provided in Appendix B.3.

Fitting an LINGAM/ANM to this dataset is problematic for two reasons. First, the discrete nature of "ClaimNb" violates the continuous additive noise assumption used by most ANM methods (Peters et al., 2011). Second, the other variables are skewed, non-Gaussian and heteroskedastic, making CPCM much more natural choice.

Using the sequential family-selection approach within CPCM and the RESIT-greedy algorithm, we obtain the graph shown in Figure 2. In this case, family $\mathscr{S}_1$ was selected, as the resulting graph was marked as plausible. In contrast, applying LiNGAM or ANM-RESIT yields markedly different structures: LiNGAM suggests Exposure $\to$ VehAge $\to$ VehPower, while ANM-RESIT recovers only VehAge $\to$ VehPower.

Although the ground truth for this dataset is not perfectly clear, the results provide evidence that CPCM can recover more plausible causal structures in mixed-type insurance data than existing ANM-based or linear approaches. By accommodating discrete outcomes, non-Gaussian noise, heteroskedasticity and heavy-tails, CPCM produces graphs that align better with domain knowledge and avoid the misspecifications that can arise in more restrictive frameworks. On the other hand, CPCM can be computationally demanding for larger datasets, especially when sequential family selection is combined with many candidate parent sets. Furthermore, as with most observational causal discovery methods, CPCM relies on the assumption of causal sufficiency, which can be restrictive in practical applications.

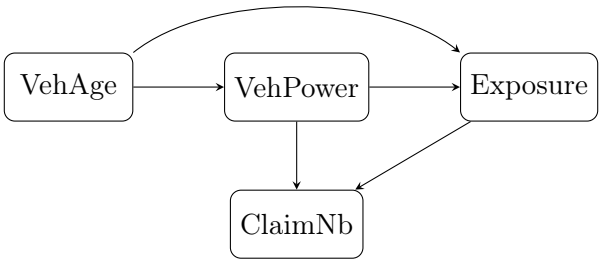

Figure 2: Causal graph estimated by CPCM on the French motor insurance dataset subset.

7. Conclusion and future research

We introduced a new family of models for causal inference called Conditionally Parametric Causal Models (CPCM), designed to flexibly accommodate a broad range of variable types and distributional forms. Our primary theoretical contributions lie in establishing the identifiability conditions for the causal structure within this framework. Specifically, we have demonstrated that the bivariate $CPCM(F)$ models are identifiable, with exceptions

arising only when the parameters of F take the form of a linear combination of its sufficient statistics. Furthermore, we have provided detailed characterizations of identifiability across various cases such as Gaussian, Poisson, and Pareto, significantly broadening the scope of identifiable models beyond existing literature. We also explained the multivariate extensions of these results.

We complement these results with two consistent estimation algorithms for CPCM-based causal graph recovery. Experiments show competitive performance in Gaussian location–scale models, while retaining the ability to operate in much broader distributional settings, including heavy-tailed, continuous, discrete or even a mixture of these.

CPCM also connects naturally to invariant causal prediction (Peters et al. (2016); Kook et al. (2024)), offering promising directions for distribution-aware causal feature selection, as the framework of target-variable causal modeling provides a natural environment for embedding the CPCM ideology (Bodik and Chavez-Demoulin, 2025). Extensions to time series settings (Bodik et al., 2024; Assaad et al., 2022) or uncertainty quantification under distribution shift (Liu et al., 2021; Bodik et al., 2025) would further broaden its applicability. Integrating these ideas and validating CPCM in diverse applied domains (Gamella et al., 2025) are key avenues for future work.

Conflict of interest and data availability

The open-source implementation of the methods discussed in this manuscript together with the data used can be found in the supplementary package or at `https://github.com/jurobodik/Causal_CPCM.git`.

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.

Acknowledgments

This study was supported by the Swiss National Science Foundation under grant number 201126.

Appendix

This appendix is organized as follows:

- Appendix A provides a detailed definition and assumptions concerning the Exponential family, which were omitted from the main text for clarity. It also introduces the notion of F-suitability of an estimator and includes an in-depth discussion and proof of Proposition 13.

- Appendix B presents additional experiments and elaborates on certain implementation details.

- Appendix C contains all theoretical results along with their corresponding proofs.

A. Exponential family and Proposition 13

A.1 Exponential family

The exponential family is a set of probability distributions whose probability density function can be expressed in the following form:

$$f(x;\theta) = h_1(x)h_2(\theta)\exp\big[\sum_{i=1}^{q}\theta_i T_i(x)\big], \tag{2}$$

where h_1, T_i are real functions and $h_2 : \mathbb{R}^q \to \mathbb{R}^+$ is a vector-valued function. We call T_i a *sufficient* statistic, h_1 a base measure, and h_2 a normalizing (or partition) function.

Often, form (2) is called a canonical form and

$$f(x;\theta) = h_1(x)h_2(\theta)\exp\big[\sum_{i=1}^{q}h_{3,i}(\theta)T_i(x)\big],$$

where $h_{3,i} : \mathbb{R}^q \to \mathbb{R}, i = 1,\dots,q$, is called its *reparametrization* (natural parameters are a specific form of the reparametrization). We always work only with a canonical form (attention for Gaussian distribution, where the standard form is not in the canonical form).

Numerous important distributions lie in the exponential family of distributions, such as Gaussian, log-normal, Poisson, Pareto (with fixed support), Weibull, chi-squared, multinomial, Binomial, Gamma, and Beta distributions, to name a few.

It is important to note that functions in (2) are *not* uniquely defined. For example, T_i is unique up to a linear transformation.

The support of f is fixed and does not depend on θ. Potentially, T_i and h_1 do not have to be defined outside of this support; however, we typically overlook this fact (or possibly define $h_1(x) = T_i(x) = 0$ for x where these functions are not defined). We additionally assume that the support is nontrivial in the sense that it contains at least two distinct values.

Without loss of generality, we assume that q is minimal in the sense that $f(x;\theta)$ cannot be expressed using only $q-1$ parameters. The sufficient statistics $T_1,\dots,T_q$ are then linearly independent in the following sense: there exist points $x_1,\dots,x_q \in \operatorname{supp}(f)$ such that the matrix

$$\begin{pmatrix} T_1(x_1) & \cdots & T_q(x_1) \\ \dots & \ddots & \dots \\ T_1(x_q) & \cdots & T_q(x_q) \end{pmatrix} \tag{17}$$

has full rank. Moreover, $T_1,\dots,T_q$ are affinly independent in the following sense: there exist $y_0, y_1, \dots, y_q \in supp(f)$, such that a matrix

$$\begin{pmatrix} T_1(y_1) - T_1(y_0) & \cdots & T_q(y_1) - T_q(y_0) \\ \dots & \ddots & \dots \\ T_1(y_q) - T_1(y_0) & \cdots & T_q(y_q) - T_q(y_0) \end{pmatrix} \tag{18}$$

has full rank. In this paper (Lemma 18) we assume affine independence of $T_1,\dots,T_q$, i.e., that condition (18) holds.

Since the notions of linear and affine independence used here are nonstandard, we illustrate them with a simple example. Let $T_1(x) = x$ and $T_2(x) = x^2$, corresponding to the sufficient statistics of a Gaussian distribution. Then matrices (17) and (18) become:

$$M_1 = \begin{pmatrix} x_1 & x_1^2 \\ x_2 & x_2^2 \end{pmatrix}, \quad M_2 = \begin{pmatrix} y_1 - y_0 & y_1^2 - y_0^2 \\ y_2 - y_0 & y_2^2 - y_0^2 \end{pmatrix},$$

both of which are full-rank for the choices $(x_1, x_2) = (1, 2)$ and $(y_0, y_1, y_2) = (0, 1, 2)$, for instance. The same analogously holds for all distributions considered in this paper.

A.2 F-suitability: Proposition 13, Consequence 16 and Lemma 14

A.2.1 Suitability of an estimator

In the following, we define an F-suitable estimator for a given distribution function F with parameters θ. This is a modification of the concept of a suitable estimator for the conditional expectation discussed in (Mooij et al., 2016, Appendix A.2). In case when F is location-distribution (such as Gaussian distribution with fixed variance), our results fully align with (Mooij et al., 2016).

Let $(X_i, Y_i)_{i=1}^n$ be a random sample from (X, Y). We say that the estimator $\hat{\theta}$ is **F-suitable for** $X \to Y$ if the following conditions are satisfied:

- **(Existence of a point-wise limit)** There exists θ such that $\hat{\theta}(x) \xrightarrow{P} \theta(x)$ as $n \to \infty$ for all $x \in supp(X)$. Moreover, if it is possible to write $Y = F^{-1}(\varepsilon_2; \tilde{\theta}(X))$, with $X \perp\!\!\!\perp \varepsilon_2 \sim Unif(0, 1)$, then this limit is equal to $\theta = \tilde{\theta}$.

- **(Weak residual consistency)** It holds that

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \hat{\varepsilon}_i)^2\right) = 0, \tag{19}$$

  where $\varepsilon_i := F(Y_i; \theta(X_i))$ and $\hat{\varepsilon}_i := F(Y_i; \hat{\theta}(X_i)), i = 1, \ldots, n$ and where the expectation is taken with respect to the distribution of the random sample.

We say that the estimator $\hat{\theta}$ is **F-suitable** if it is F-suitable for both $X \to Y$ and $Y \to X$. We simply write that $\hat{\theta}$ is "suitable" if F is evident from the context.

A.2.2 Literature review - which estimators are suitable?

Location family: when F is location-family distribution, such as a Gaussian distribution with fixed variance, property (19) reduces to classical notion of a weak universal consistency: $\lim_{n\to\infty} \mathbb{E}[||\theta(X) - \hat{\theta}(X)||^2] = 0$. Such consistency have been already discussed also in relation to causal discovery (Zhang et al., 2015; Uemura et al., 2022; Keropyan et al., 2023). The weak universal consistency has been established for various estimators under appropriate smoothness assumptions on θ. Examples of such estimators include:

- Kernel estimators (see Theorem 5.1 in Györfi et al. (2002))

- Smoothing spline GAM estimators (see Chapter 14.2 in Györfi et al. (2002), Claeskens et al. (2009) or Wood et al. (2016))
- Neural networks (see Theorem 16.1 in Györfi et al. (2002) or Drews and Kohler (2024); Heiss (2024)).

Importantly, these consistency results apply regardless of the causal direction. In the anti-causal direction, we have $X = \theta(Y) + \varepsilon$, where $\theta(Y) = \mathbb{E}[X \mid Y]$ and $\varepsilon \not\perp\!\!\!\perp Y$, $\mathbb{E}[\varepsilon \mid Y] = 0$. Note that $\varepsilon \not\perp\!\!\!\perp Y$ holds if and only if the model is identifiable. For such case, the same form of weak consistency remains valid for the estimators listed above (again, under appropriate smoothness assumptions).

Location-scale family: when F is a location-scale distribution (e.g. Gaussian), several consistency results has also been established for various estimators under appropriate smoothness and regularity assumptions, see e.g. Fan and Yao (1998), or Immer et al. (2023); Siegfried et al. (2023); Le et al. (2005).

More general families: Consistency results for non-Gaussian families are less explored in nonparametric settings, with most existing literature focusing on empirical evidence. GAMLSS (Stasinopoulos and Rigby, 2007) offers a broad class of estimators, with Rigby and Stasinopoulos (2025) presenting extensive empirical evidence on simulations and hundreds of real-data examples demonstrating consistency. A few theoretical results are avaliable for GAM estimators; e.g. Theorem 1 in Chen and Samworth (2015) establishes its almost sure universal consistency for a general one-dimensional exponential family of distributions for $Y \mid X$, assuming certain smoothness and convexity/monotonicity conditions on θ. Wood et al. (2016) discusses general framework for smoothing parameter estimation for models with regular likelihoods constructed in terms of unknown smooth functions of covariates. Mammen and van de Geer (1997) discusses the consistency under partial linearity assumption.

A.2.3 HSIC score

For the definition and discussion of the HSIC score, see (Mooij et al., 2016, Appendix A.1). We use the same notation and implicitly use bounded non-negative Lipschitz-continuous kernels such that their product is characteristic, as in the "Data recycling" scenario in (Mooij et al., 2016, Corollary 21).

A.2.4 Proofs of Proposition 13, Consequence 16 and Lemma 14

Proposition 13 *Let (X_1, X_2) follow an identifiable $CPCM(F_1, \ldots, F_k)$ with DAG $\mathcal{G}$. Then, our score-based algorithm presented in Section 5.2 is consistent, meaning that*

$$\hat{\mathcal{G}} \xrightarrow{P} \mathcal{G} \text{ as } n \to \infty,$$

given that we employ a "suitable" estimation procedure for the estimator $\hat{\varepsilon}_i$, we use HSIC score as our choice of ρ and consistent estimates of S_i.

Proof The proof mostly aligns with the proof of Corollary 21 in Mooij et al. (2016). We use the notation $X = X_1$ and $Y = X_2$.

If $X \perp\!\!\!\perp Y$, then $\hat{\mathcal{G}}$ converges to an empty graph, since any other graph has a score of at least λ and $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_d$ are independent by definition. From now on, without loss of generality, let $Y = F_j^{-1}(\varepsilon_2; \theta(X)), \varepsilon_2 \perp\!\!\!\perp X$ for some $j \in \{1, \ldots, k\}$. Denote by $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ the observed data. In the following, we compare the asymptotic scores of graphs $X \to Y$ and $Y \to X$.

Graph $X \to Y$: Define the "population residual" $E_Y := F_j(Y; \theta(X))$ and the "estimated residual" $\hat{\varepsilon}_Y^i = (F_i(y_l; \hat{\theta}(x_l)))_{l=1}^n$, where $i = 1, \ldots, k$. For the true $i = j$, we omit the superscript and write $\hat{\varepsilon}_Y = (F_j(y_l; \hat{\theta}(x_l)))_{l=1}^n$. By construction, $X \perp\!\!\!\perp E_Y$ which implies $\mathrm{HSIC}(X, E_Y) = 0$ (due to Lemma 12 in Mooij et al. (2016)).

Now, since the estimator $\hat{\theta}$ satisfies (19), we can use the argument presented in (Mooij et al., 2016, Theorem 20), and obtain $\widehat{HSIC}(\mathbf{x}, \hat{\varepsilon}_y) \xrightarrow{P} \mathrm{HSIC}(X, E_Y)$.

Since $\hat{S}_2$ is consistent, we can find n_0 such that for all $n \geq n_0$ holds $P(j \in \hat{S}_2) \geq 1 - \delta$ for given $\delta > 0$.

Putting everything together, with probability larger than $1 - \delta$ holds

$$\begin{aligned} s(X \to Y) &= \min_{j_2 \in \hat{S}_2} \rho(\mathbf{x}, \hat{\varepsilon}_y^{j_2}) + \lambda \leq \rho(\mathbf{x}, \hat{\varepsilon}_y) + \lambda \\ &= \widehat{HSIC}(\mathbf{x}, \hat{\varepsilon}_y) + \lambda \xrightarrow{P} \mathrm{HSIC}(X, E_Y) + \lambda = \lambda. \end{aligned}$$

By sending $\delta \to 0$, we obtain $s(X \to Y) \xrightarrow{P} \lambda$.

Graph $Y \to X$: Define the "population residual" $E_X^j := F_j(X; \theta_j(Y))$ and the "estimated residual" $\hat{\varepsilon}_X^j = (F_j(x_i; \hat{\theta}_j(y_i)))_{i=1}^n$, where θ_j is the limit of $\hat{\theta}_j$ defined in the definition of F_j-suitability.

Due to the assumption of identifiability, $Y \not\perp\!\!\!\perp E_X^j$ for all $j \in \{1, \ldots, k\}$. Therefore, using Lemma 12 in Mooij et al. (2016), we have $\mathrm{HSIC}(Y, E_X^j) > 0$. Again, since the estimator $\hat{\theta}$ satisfies (19), we can use the argument presented in (Mooij et al., 2016, Theorem 20), and obtain $\widehat{HSIC}(\mathbf{y}, \hat{\varepsilon}_X^j) \xrightarrow{P} \mathrm{HSIC}(Y, E_X^j)$.

Putting everything together

$$\begin{aligned} s(Y \to X) &= \min_{j \in \hat{S}_1} \rho(\mathbf{y}, \hat{\varepsilon}_x^j) + \lambda \geq \min_{j \in \{1, \ldots, k\}} \widehat{HSIC}(\mathbf{y}, \hat{\varepsilon}_X^j) + \lambda \\ &\xrightarrow{P} \min_{j \in \{1, \ldots, k\}} \mathrm{HSIC}(Y, E_X^j) + \lambda > \lambda, \qquad \text{as } n \to \infty. \end{aligned}$$

Therefore, the score for the correct direction is asymptotically smaller than that for the wrong causal direction as $n \to \infty$, hence the procedure is consistent. ■

Consequence 16 *Let (X_1, X_2) follow an $CPCM(F)$ model with DAG $\mathcal{G}$, for some $F \in \mathscr{S}_1 \cup \mathscr{S}_2$. Assume the conditions of Proposition 13 hold: namely, identifiability of $CPCM(\mathscr{S}_1 \cup \mathscr{S}_2)$, a suitable estimation procedure, and the use of the HSIC score. Then, our score-based algorithm, with the collection $\{F_1, \ldots, F_k\}$ chosen via the Sequential approach (Exact or Fast version), is consistent: $\hat{\mathcal{G}} \xrightarrow{P} \mathcal{G}$, as $n \to \infty$.*

Proof If $F \in \mathscr{S}_1$, the result follows directly from Proposition 13. Similarly, if $F \in \mathscr{S}_2$ and the Sequential approach returns $\mathscr{S}_1 \cup \mathscr{S}_2$, the proposition again directly applies.

It remains to consider the case where $F \in \mathscr{S}_2$ but the Sequential approach returns only $\mathscr{S}_1$. We will show that, as $n \to \infty$, this event occurs with probability tending to zero. In other words, when $F \in \mathscr{S}_2$, the Sequential approach will return $\mathscr{S}_1 \cup \mathscr{S}_2$ with probability tending to one.

If $X \perp\!\!\!\perp Y$, then $\hat{\mathcal{G}}$ converges to an empty graph regardless of the choices of F. Hence, without loss of generality, assume non-empty causal graph. Consider graph $\mathcal{G} = X_1 \to X_2$, and take some $F_1 \in \mathscr{S}_1$ (that is, wrong distribution $F_1 \neq F$). In this case, following the notation of the proof of Proposition 13, we have

$$\widehat{\mathrm{HSIC}}(\mathbf{x}_1, \hat{\varepsilon}^1_{X_2}) \xrightarrow{P} \mathrm{HSIC}(X_1, E^1_{X_2}) > 0, \tag{20}$$

where $\hat{\varepsilon}^j_{X_2} := (F_j(X_2; \hat{\theta}_j(X_1)))_{l=1}^n$ is the vector of residuals obtained by applying a "suitable" estimation procedure, and the "population residual" $E^j_{X_2} := F_j(X_2; \theta(X_1))$. This convergence follows from the same reasoning as in the '$Y \to X$' case in the proof of Proposition 13.

In particular, (20) implies that the variance of the empirical HSIC statistic vanishes:

$$\mathrm{Var}\left(\widehat{\mathrm{HSIC}}(\mathbf{x}_1, \hat{\varepsilon}^1_{X_2})\right) \to 0 \quad \text{as } n \to \infty.$$

Therefore, according to the construction of confidence intervals for the HSIC test in (Pfister et al., 2018, Section 3.2.2), the $(1-\alpha)$ confidence interval will eventually lie entirely within an interval of the form $[\mathrm{HSIC}(X_1, E^1_{X_2}) - \delta, \mathrm{HSIC}(X_1, E^1_{X_2}) + \delta]$, for arbitrarily small $\delta > 0$. Since the limiting HSIC value is strictly positive, this interval will exclude 0 for large enough n, yielding p-values smaller than α. As a result, the DAG $\mathcal{G}$ will be rejected as plausible.

The same argument applies to the case $\mathcal{G} = X_2 \to X_1$ and for any $F_j \in \mathscr{S}_1$, $j = 1, \ldots, k$. Therefore, for sufficiently large n, no DAG will be deemed plausible, and the Sequential approach will return the full set $\mathscr{S}_1 \cup \mathscr{S}_2$ with high probability. ■

Lemma 14 *Let (X_1, X_2) follow an unidentifiable CPCM$(F_1, \ldots, F_k)$ with DAG $\mathcal{G}$. Then, for sufficiently large n, Algorithm 1 outputs "Unidentifiable case" with probability at least $1 - 2\alpha$, given that we employ a HSIC independence test in step 1b) of Algorithm 1, and assume access to an oracle regression estimator $\hat{\theta} = \theta$.*

Proof If the true model is unidentifiable, independence holds in both directions: $X_1 \perp\!\!\!\perp \varepsilon_2$ and $X_2 \perp\!\!\!\perp \varepsilon_1$. Given an oracle regression estimator $\hat{\theta} = \theta$, we have $\hat{\varepsilon}_i = \varepsilon_i$ for $i = 1, 2$. Therefore, Algorithm 1 reduces to performing two HSIC independence tests and returning "Unidentifiable case" if both tests fail to reject the null.

The HSIC test has asymptotically correct level under the null hypothesis of independence (Pfister et al. (2018), Theorem 3.8). Thus, for sufficiently large sample size n, the probability that the HSIC test fails to reject the null in each direction is at least $1 - \alpha$. By the union bound, the probability that both tests fail to reject is at least $(1-\alpha)^2 > 1 - 2\alpha$. Hence, Algorithm 1 returns "Unidentifiable case" with probability larger than $1 - 2\alpha$. ■

B. Experiments details and additional plots

B.1 Exact, Naive-greedy, RESIT, and RESIT-greedy algorithms: definitions and comparison

We consider the following algorithms for estimating the underlying causal graph from observational data using the CPCM score function (13):

- **Exact search:** This algorithm evaluates the CPCM score for all DAGs on d nodes and selects the one with the lowest score. Since the number of DAGs grows super-exponentially with d (e.g., 29,281 DAGs for $d = 5$), exact search is computationally feasible only for graphs with $d \leq 4$.

- **Naive-edge-greedy:** Starting from an empty DAG, this algorithm iteratively explores neighboring DAGs by adding or removing a single edge. At each step, it selects the neighboring graph with the lowest CPCM score and replaces the current graph if the score improves. The procedure stops when no further improvement is possible. While simple and scalable, this greedy approach lacks theoretical guarantees and may get stuck in local minima, unless we assume some advanced notions of convexity over the space of all DAGs.

- **RESIT (Regression with Subsequent Independence Test):** RESIT first estimates a topological ordering by iteratively selecting the variable whose residual is least dependent on the remaining variables. In the second phase, it removes superfluous edges using conditional independence tests. See Algorithm 4 for details. The procedure is computationally efficient and comes with statistical guarantees (see Lemma 17). However, empirical performance in smaller sample sizes tends to be worse than that of greedy algorithms, particularly due to the accumulation of errors in the ordering phase and false positives (type I errors) in the Phase 2.

- **RESIT-greedy:** This hybrid algorithm combines the topological ordering phase of RESIT with the edge-pruning phase of naive-greedy search. After estimating the ordering, it starts with a fully connected DAG consistent with the order and iteratively removes edges that lead to the largest improvement in the CPCM score, until no further improvement is possible. See Algorithm 2.

Peters et al. (2014) showed that, in the population case, the RESIT algorithm is consistent under identifiable additive noise models, assuming a consistent nonparametric regression method and a perfect independence oracle. The same reasoning applies directly to CPCM models.

Lemma 17 (Consistency of RESIT under CPCM) *Let* $\mathbf{X}$ *be generated by an identifiable* $CPCM(F_1, \ldots, F_k)$ *model with underlying DAG* $\mathcal{G}_0$*. Then, the RESIT algorithm, when applied with consistent estimators* $\hat{\theta}_k$ *and an independence oracle, is guaranteed to recover the true graph* $\mathcal{G}_0$ *from the distribution of* $\mathbf{X}$*.*

Proof A direct consequence of Theorem 34 in Peters et al. (2014). Note that we implicitly assume the causal minimality condition for $CPCM(F_1, \ldots, F_k)$ as stated in Definition 9. ■

Algorithm 4: Regression with Subsequent Independence Test (RESIT; Peters et al. (2014)), modified for $CPCM(F_1, \dots, F_k)$

Input: Random sample of $(X_1, \dots, X_d)$; candidate families $F_1, \dots, F_k$

Pre-step (support gating).

For each node i, estimate $\hat{S}_i := \{\, j \in \{1, \dots, k\} : \ \mathrm{supp}(X_i) = \mathrm{supp}(F_j) \,\}$; `// See Section 5.1.1. This filters discrete vs. continuous etc.`

If $\hat{S}_i = \emptyset$ return "Assumptions not fulfilled".

Phase 1: Determine a topological order π

$M \leftarrow \{1, \dots, d\}$; $\pi \leftarrow [\,]$;

while $M \neq \emptyset$ **do**

  foreach $v \in M$ **do**

    foreach $m \in \hat{S}_v$ **do**

      Fit parameters $\hat{\theta}_v^{(m)}$ in $X_v \mid X_{M\setminus\{v\}} \sim F_m(\,\theta_v(\cdot)\,)$;

      Compute residuals $\hat{\varepsilon}_v^{(m)} := F_m\Big(X_v;\ \hat{\theta}_v^{(m)}\big(X_{M\setminus\{v\}}\big)\Big)$;

      Compute dependence score $s_v^{(m)} := \rho\Big(\hat{\varepsilon}_v^{(m)}\ ;\ X_{M\setminus\{v\}}\Big)$; `// ρ small ⇔ near independence. Implemention uses HSIC score`

    $s_v \leftarrow \min_{m \in \hat{S}_v} s_v^{(m)}$; `// Best fitting distribution` F_m

  $v^\star \leftarrow \arg\min_{v \in M} s_v$; `// most source-like variable`

  $M \leftarrow M \setminus \{v^\star\}$; $\mathrm{pa}(v^\star) \leftarrow M$;

  prepend $v^\star$ to π

Phase 2: Prune superfluous edges

for $t = 2$ **to** d **do**

  $v \leftarrow \pi_t$; $C \leftarrow \mathrm{pa}(v)$;

  foreach $\ell \in C$ **do**

    foreach $m \in \hat{S}_v$ **do**

      Fit $\hat{\theta}_{v,-\ell}^{(m)}$ in $X_v \mid X_{C\setminus\{\ell\}} \sim F_m(\,\theta_v(\cdot)\,)$;

      Compute residuals $\hat{\varepsilon}_v^{(m)} := F_m\Big(X_v;\ \hat{\theta}_{v,-\ell}^{(m)}\big(X_{C\setminus\{\ell\}}\big)\Big)$;

      Compute $p_\ell^{(m)} := \mathsf{IndepP}\Big(\hat{\varepsilon}_v^{(m)}\ ;\ \{X_{\pi_1}, \dots, X_{\pi_{t-1}}\}\Big)$; `// p-value of the test of` $H_0 : \varepsilon_v^{(m)} \perp\!\!\!\perp \{X_{\pi_1}, \dots, X_{\pi_{t-1}}\}$

    $p_\ell := \max_{m \in \hat{S}_v} p_\ell^{(m)}$;

    if $p_\ell \geq \alpha$ *using* $\alpha = 0.05$ **then**

      $C \leftarrow C \setminus \{\ell\}$; `// remove` ℓ `if residuals are independent`

  $\mathrm{pa}(v) \leftarrow C$

Output: $(\mathrm{pa}(1), \dots, \mathrm{pa}(d))$

B.1.1 Experiments: comparison of different greedy algorithms

Experimental setup: We generate random DAGs uniformly over d nodes with p edges, where $p \sim \text{Exp}(1/d)$ and capped at $\frac{d(d-1)}{2}$. On average, each DAG contains approximately d edges. For a given DAG $\mathcal{G}$, we simulate data from the $CPCM(F)$ model, $F = Exponential$, defined as follows:

$$X_i \sim \text{Exp}\left(\lambda(\mathbf{X}_{pa_i(\mathcal{G})})\right), \quad \text{with} \quad \lambda(\mathbf{X}_{pa_i(\mathcal{G})}) = \frac{1}{\sum_{j \in pa_i(\mathcal{G})} |X_j|} = \frac{1}{\mathbb{E}[X_i \mid \mathbf{X}_{pa_i}]}.$$

If $pa_i(\mathcal{G}) = \emptyset$, then $X_i \sim N(0,1)$. Using $n = 1000$ samples, we estimate $\mathcal{G}$ using the score-based estimator with score defined in Equation (14), with a fixed function class $\mathscr{S}_1$ (rather than the sequential approach) to allow for a fair comparison in both accuracy and computational cost. We compare the Exact, Naive-greedy, RESIT, and RESIT-greedy methods. We evaluate their performance using the Structural Intervention Distance (SID, Peters and Bühlmann (2013)), computing $\text{SID}(\mathcal{G}, \hat{\mathcal{G}})$ for each estimated graph.

Results: Figure 3 presents the average normalized $\frac{SID}{d}$ over 50 repetitions.

- The exact method achieves, unsurprisingly, the lowest SID, with the highest computational cost.
- Both RESIT and naive-greedy exhibit similar performance. The greedy approach achieves slightly lower SID on average but requires slightly more computation time. Note that for $d \geq 8$, both methods perform as badly as Trivial algorithm (empty graph).
- RESIT-greedy serves as a middle ground: it significantly improves over RESIT/naive-greedy methods in terms of SID while incurring higher computational cost.

These experiments highlight the trade-offs between statistical performance and computational efficiency among the evaluated methods. While the exact method yields the most accurate graph recovery, its scalability is limited. In contrast, RESIT and naive-greedy offer faster but less accurate alternatives, with performance deteriorating as graph complexity increases. RESIT-greedy provides a promising compromise, achieving lower SID than standard greedy methods at a moderate computational cost. **Overall, RESIT-greedy seems to be a practical choice in graphs with size** $4 < d < 10$.

B.1.2 Statistical scalability: sample size needs to grow with increasing dimension

Conducting independence tests and performing nonparametric regression becomes increasingly difficult as the number of covariates grows. In high-dimensional settings, these tasks require substantially larger sample sizes for reliable estimation. Similar limitations are observed in other methods such as ANM-RESIT, LOCI, and bQCD. The experiment below illustrates how the sample size required for consistent estimation increases systematically with the dimension d.

Experimental setup. For each sample size n and dimension d, we generated a uniformly random DAG with d nodes and $d-1$ edges using the `bnlearn` package in R. Data

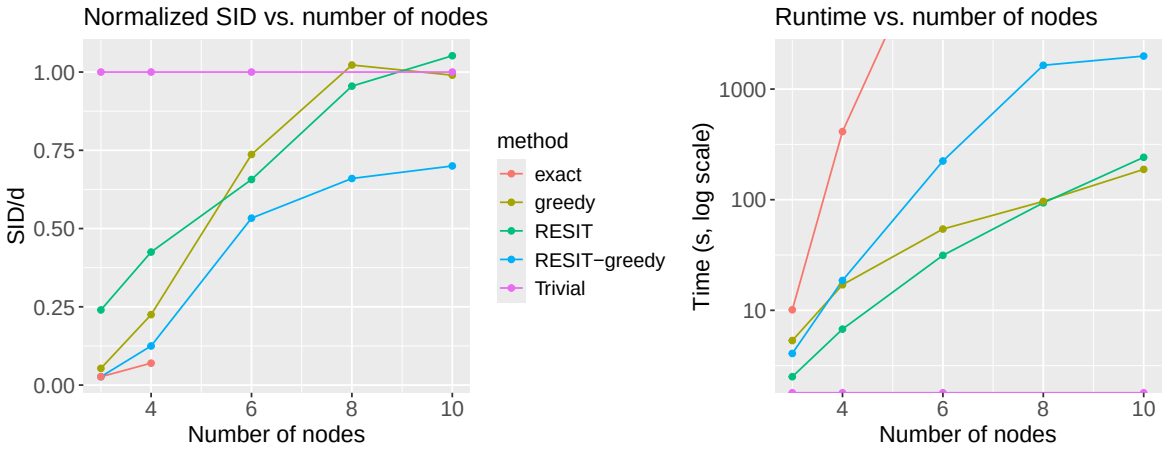

Figure 3: Performance of different greedy algorithms from Section B.1.1, measured by the normalized Structural Interventional Distance (SID). Here, the `Trivial` algorithm always returns an empty graph. Runtime was measured on a machine with an Intel Core i5-6300U 2.5 GHz processor and 16 GB of RAM.

were drawn from the structural equation model $X_i = f_i(\mathbf{X}_{pa_i}) + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$ and $f_i(x) = c_i x^\top x$ with $c_i \sim \text{Unif}(0.5, 1.5)$. The DAG was estimated using the $CPCM(F)$ algorithm (13) with F set to a Gaussian distribution of fixed variance, and structure learning was performed via the naive-greedy algorithm. Estimation accuracy was evaluated using the Structural Intervention Distance (SID). We also generate a baseline guess by generating random DAGs with $d-1$ edges uniformly over the space of all DAGs. Each configuration was repeated 100 times.

Results. Figure 4 reports the ratio between the average SID obtained by the $CPCM(F)$ algorithm and that of a random-DAG baseline. As the dimension d increases, considerably larger samples are required to achieve ratios below 0.5, which correspond to meaningful structural recovery. Specifically, we observe ratios below 0.5 for approximately $n = 100$ when $d = 2$–3, $n = 500$ for $d = 4$, $n = 1000$ for $d = 5$, and $n = 2000$ for $d = 6$. These results indicate that **the sample size required for reliable estimation grows rapidly** (potentially exponentially) **with the dimension of X**.

B.2 Sequential approach vs oracle: empirical performance

We evaluate the empirical performance of the sequential approach for selecting the distribution family, comparing it to $CPCM(F)$ with access to the true (oracle) distribution F.
Experimental setup: The data is generated as follows:

$$X_1 \sim \mathcal{N}(0,1), \quad X_2 = F^{-1}(\varepsilon, \theta(X)), \quad \varepsilon \perp\!\!\!\perp X,\ \varepsilon \sim \mathcal{U}(0,1),$$

where $F \in \mathscr{S}_s$ belongs to either the one-parameter family $\mathscr{S}_1$ or the two-parameter family $\mathscr{S}_2$.

When $s = 1$, F is, with equal probability $(\frac{1}{4})$, either a Gaussian distribution with fixed variance, a Poisson, Pareto, or Exponential distribution. When $s = 2$, F is, again with equal

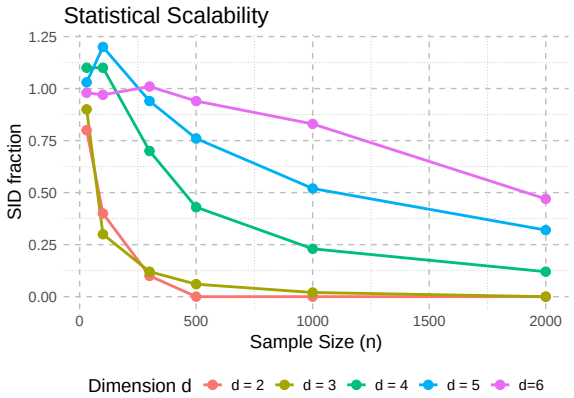

Figure 4: Scalability experiment from Section B.1.2. The plot shows the ratio between the average SID obtained by the $CPCM(F)$ algorithm (edge-greedy version) and that of a random-DAG baseline, across 100 repetitions for different values of n and d. Lower ratios indicate better structural recovery, while values near one correspond to random performance. The results suggest that the sample size required for accurate estimation grows rapidly with the dimension of $\mathbf{X}$.

probability, a Gaussian, Negative Binomial, Generalized Pareto, or Gamma distribution. All parameters are generated as a transformation of a random smooth polynomial using

$$\theta(x) = 0.5 + 4.5 \cdot \frac{\tanh\left(b(x)^\top u\right) + 1}{2}, \qquad u_j \sim \text{Unif}(-2, 2), \quad j = 1, 2, 3,$$

where $b(x)$ denotes the natural spline basis with three degrees of freedom. This construction yields a smooth parameter function $\theta(x) \in [0.5, 5]$, ensuring that the conditional distribution $X_2 \mid X_1$ varies continuously with x and that the variance does not explode.

We estimate the graph $\mathcal{G} = 1 \to 2$ using both $CPCM(\text{Seq.app})$ and $CPCM(F)$ with oracle knowledge of F, for various sample sizes n, repeating each experiment 100 times for both $s = 1$ and $s = 2$. Figure 5 summarizes the results across sample sizes.

Results. For $s = 1$, we observe that $CPCM(\text{Seq.app})$ typically performs equivalently to oracle $CPCM(F)$ for $n > 100$. For $s = 2$, the sequential approach tends to select the simpler class $\mathscr{S}_1$ instead of the true two-parameter family $\mathscr{S}_2$ at smaller sample sizes. Nevertheless, the performance gap between the sequential approach and the oracle method remains almost negligible. These results indicate that **the sequential approach is performing nearly as well as the oracle** $CPCM(F)$, provided that $F \in \mathscr{S}_1 \cup \mathscr{S}_2$.

B.3 Details about sections 6.1, 6.2, 6.3, 6.4 and 6.5

In **Section 6.1**, we generated random functions $\theta(x)$ from zero-mean Gaussian processes with squared-exponential covariance kernel $k(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$, where $\sigma = 1$ and $\ell = 0.2$; same choices as in Section 6.3 taken from Tagasovska et al. (2020). Each realization was then shifted to ensure positivity, $\theta(x) \leftarrow \theta(x) - \min_x \theta(x) + 0.5$.

For **Section 6.2**, the additional plots are presented in Table 4 and Figures 6 and 7.

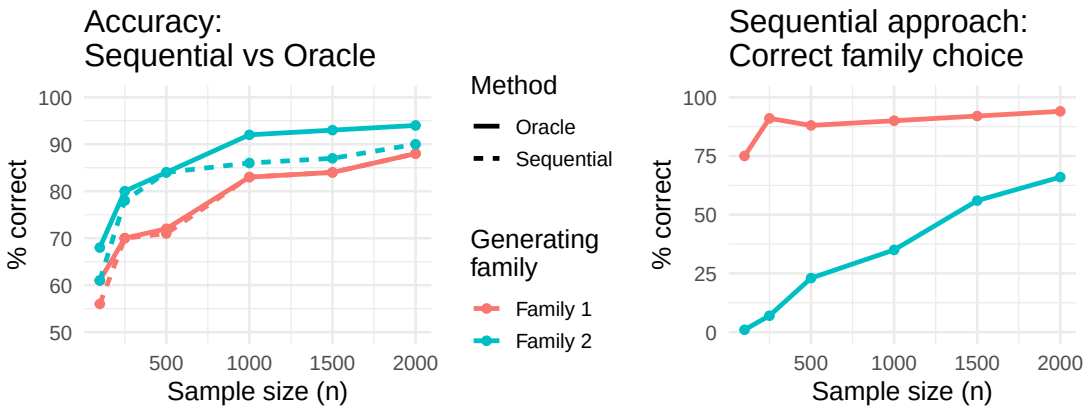

Figure 5: Performance of the sequential approach. Left: percentage of simulations where $\hat{\mathcal{G}} = 1 \to 2$. Right: percentage of simulations in which $CPCM$(Seq.app) and $CPCM(F)$ with oracle F are equivalent.

γ	$X_1 \to X_2$	$X_2 \to X_1$	Empty graph	Both directions appear plausible	Neither direction appears plausible
-2	0	0	**96**	2	2
-1	3	2	0	**95**	0
0	7	1	0	**92**	0
1	7	5	0	**86**	2
2	**93**	0	0	14	3

Table 4: Simulation results for the CPCM model using the Pareto distribution function F. The table displays the percentage of cases for each type of graph structure estimated by Conservative Algorithm 1 with the model specified in (16), across various values of the hyperparameter $\gamma \in \mathbb{R}$. The columns indicate the frequency of each graph structure being estimated, with the highest frequency in each row highlighted in bold.

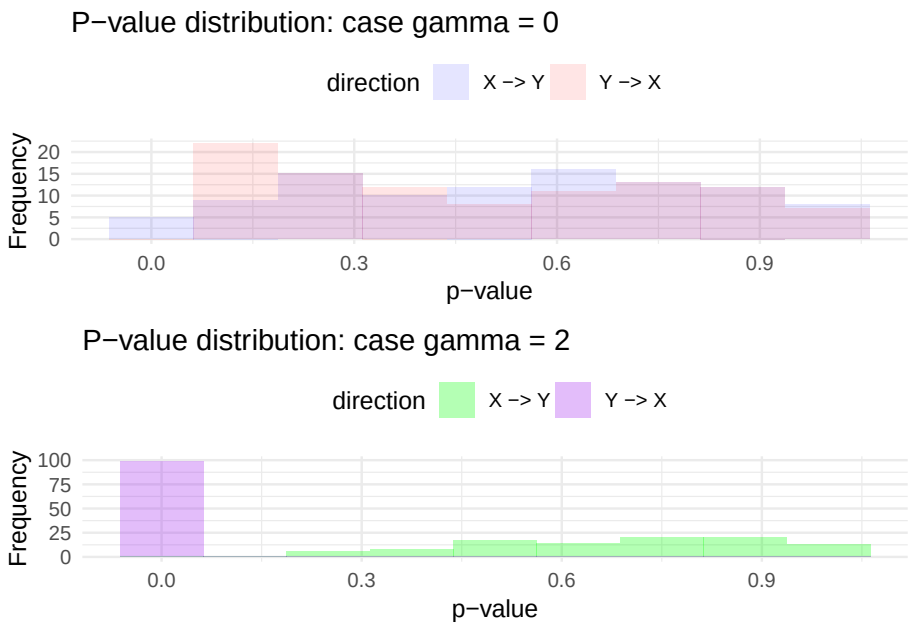

Figure 6: (Simulations 6.2). Distributions of the p-values from the independence test in Step 1b) of Algorithm 1, for model (16) with $\gamma = 0$ and $\gamma = 2$.

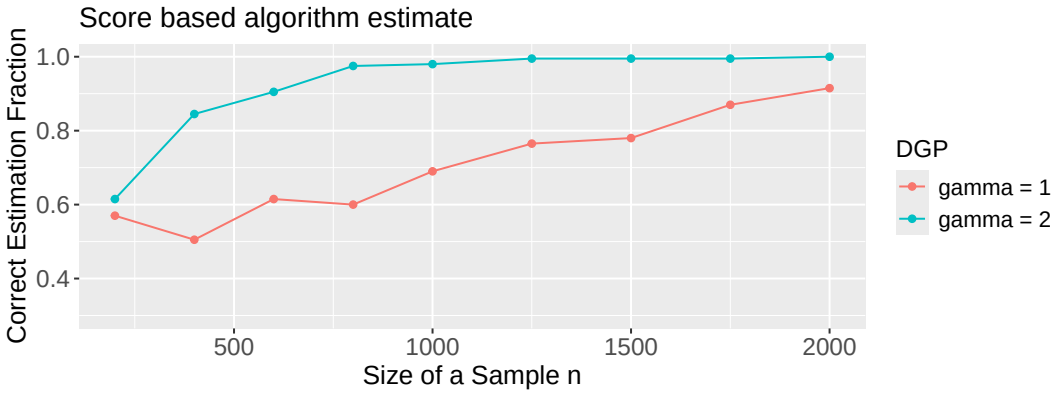

Figure 7: (Simulations 6.2). The plot displays the percentage of correctly estimated causal directions across a range of sample sizes n, using model (16) with hyperparameters $\gamma = 1$ and $\gamma = 2$. As n increases, the algorithm demonstrates near-perfect performance, affirming the theoretical consistency of the proposed method.

In **Section 6.3**, the experiments were inspired by Tagasovska et al. (2020) and implementations of other baseline methods are also taken from Tagasovska et al. (2020) and Immer et al. (2023).

- For LOCI, we use the default format with neural network estimations and subsequent independence testing (also denoted as $NN - LOCI_H$) (Immer et al., 2023).
- For IGCI, we use the original implementation from Janzing and Schölkopf (2010) with slope-based estimation with Gaussian and uniform reference measures.
- For RESIT, we use the GP regression and the HSIC independence test with a threshold value of $\alpha = 0.05$.
- For the slope algorithm, we use the implementation of Marx and Vreeken (2019), with the local regression included in the fitting process.
- For comparisons with other methods such as PNL, GPI-MML, ANM, Sloppy, GR-AN, EMD, GRCI, see Section 3.2 in Tagasovska et al. (2020) and Section 5 in Immer et al. (2023).

The random functions were generated in the same way as in Tagasovska et al. (2020). Specifically, Models 1, 3, 6, and 7 are realizations of zero-mean Gaussian processes with a squared-exponential covariance kernel $k(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$, where $\sigma = 1$ and $\ell = 0.2$. Each realization was then shifted to ensure positivity, $\theta(x) \leftarrow \theta(x) - \min_x \theta(x) + 0.5$. Models 2, 4, and 5 use an injective (monotone) nonlinear transformation $m(x) = C \frac{B(x+A)}{1+|B(x+A)|}$, where $A \sim \text{Unif}[-2, 2]$, B follows a two–point mixture with $B \sim \text{Unif}[0.5, 2]$ with probability 0.5 and $B \sim \text{Unif}[-2, -0.5]$ otherwise, and $C \sim 1 + \text{Exp}(4)$. This parametrization produces smooth saturating nonlinear relationships that can be either increasing or decreasing, thereby capturing a wide range of monotone functional dependencies between the parent and child variables. Figure 8 shows an example of datasets generated via different models from Section 6.3.

In **Section 6.4**, all baseline methods are implemented using the `pcalg` package (Kalisch et al., 2012):

- For the PC algorithm, we consider three variants: (1) the default Gaussian conditional independence test `gaussCItest`, (2) the HSIC-based test (Pfister et al., 2018), and (3) the copula-based test (Genest et al., 2019), which is omitted from the table as it yielded consistently inferior results. We always used significance level $\alpha = 0.05$.
- The GES algorithm is implemented with the Gaussian observational BIC score and the default penalty parameter.
- LiNGAM with default parameters.
- For the ANM baseline, we ensure fairness by adopting the same components as in our CPCM implementation; namely, the GAM estimator and the HSIC dependence measure.

Regarding **Section 6.5**, Table 5 shows the relative frequencies with which each edge was recovered across repeated subsamples of the motor insurance dataset.

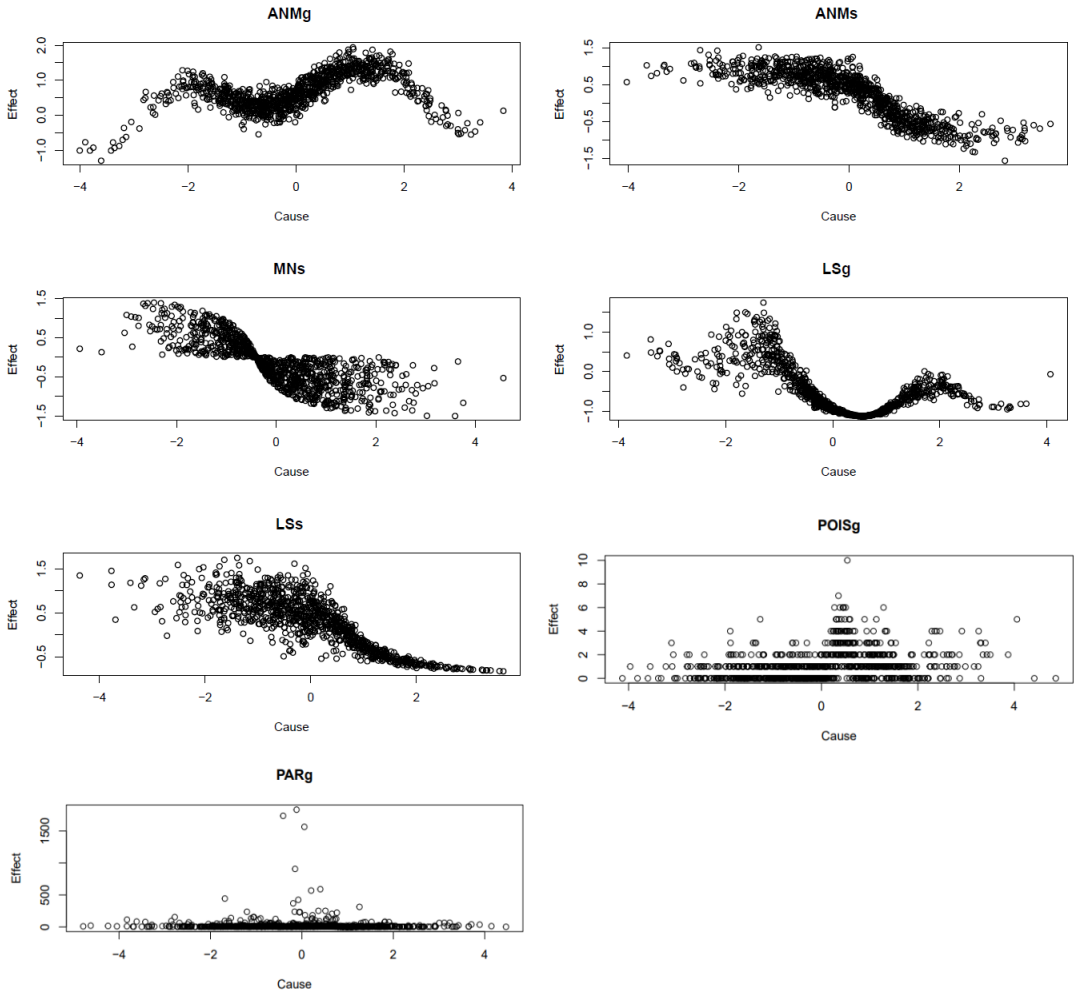

Figure 8: Simulations 6.3. An example of datasets generated via different models.

Edge	Share (%)
VehAge → ClaimNb	60
Exposure → ClaimNb	48
VehPower → ClaimNb	42
VehPower → Exposure	44

Edge	Share (%)
VehAge → Exposure	36
Exposure → VehAge	64
VehPower → VehAge	54
VehAge → VehPower	30

Table 5: Relative frequency (in %) with which each directed edge was recovered by CPCM across 50 random subsamples of the French MTPL motor insurance dataset (shown only those with more than 25% share).

C. Proofs

C.1 Proof of Theorem 4

Theorem 4 *Let (X_1, X_2) admit the $CPCM(F)$ model with graph $X_1 \to X_2$, where F is the Gaussian distribution function with parameters $\theta(X_1) = \big(\mu(X_1), \sigma(X_1)\big)^\top$. Let p_{ε_1} be the density of ε_1 that is absolutely continuous with full support $\mathbb{R}$. Let $\mu(x), \sigma(x)$ be two times differentiable.*

Then, the causal graph is identifiable from the joint distribution if and only if there do not exist $a, c, d, e, \alpha, \beta \in \mathbb{R}$, $a \geq 0, c > 0, \beta > 0$, such that

$$\frac{1}{\sigma^2(x)} = ax^2 + c, \qquad \frac{\mu(x)}{\sigma^2(x)} = d + ex, \tag{4}$$

for all $x \in \mathbb{R}$ and

$$p_{\varepsilon_1}(x) \propto \sigma(x) e^{-\frac{1}{2}\left[\frac{(x-\alpha)^2}{\beta^2} - \frac{\mu^2(x)}{\sigma^2(x)}\right]}, \tag{5}$$

where $\propto$ represents an equality up to a constant (here, p_{ε_1} is a valid density function if and only if $\frac{1}{\beta^2} \neq \frac{e^2}{c}\mathbb{1}[a = 0]$). Specifically, if $\sigma(x)$ is constant (case $a = 0$), then the causal graph is identifiable unless $\mu(x)$ is linear and p_{ε_1} is the Gaussian density.

Proof We opt for proving this theorem from scratch, without using Theorem 7. An interested reader can try to use Theorem 7 instead. For clarity regarding the indexes, we use the notation $X = X_1, Y = X_2$.

First, we show that if the causal graph is not identifiable, then $\mu(x)$ and $\sigma(x)$ must satisfy (4). Let $p_{(X,Y)}$ be the density function of (X, Y). Since the causal graph is not identifiable, there exist two CPCM models that generate $p_{(X,Y)}$: the CPCM model with $X \to Y$ and the function $\theta(x) = \big(\mu(x), \sigma^2(x)\big)^\top$ and the CPCM model with $Y \to X$ and the function $\tilde{\theta}(y) = \big(\tilde{\mu}(y), \tilde{\sigma}^2(y)\big)^\top$.

We decompose (corresponding to the direction $X \to Y$)

$$p_{(X,Y)}(x, y) = p_X(x) p_{Y|X}(y \mid x) = p_X(x) \phi\big(y; \theta(x)\big),$$

where $\phi\big(y; \theta(x)\big)$ is the Gaussian density function with parameters $\theta(x) = \big(\mu(x), \sigma^2(x)\big)^\top$. We rewrite this in the other direction:

$$p_{(X,Y)}(x, y) = p_Y(y) p_{X|Y}(x \mid y) = p_Y(y) \phi\big(x; \tilde{\theta}(y)\big).$$

We take the logarithm of both equations and rewrite them in the following manner:

$$\log[p_X(x)] + \log\left\{\frac{1}{\sqrt{2\pi\sigma^2(x)}}e^{\frac{-[y-\mu(x)]^2}{2\sigma^2(x)}}\right\} = \log[p_Y(y)] + \log\left\{\frac{1}{\sqrt{2\pi\tilde{\sigma}^2(y)}}e^{\frac{-(x-\tilde{\mu}(y))^2}{2\tilde{\sigma}^2(y)}}\right\} \text{ and}$$

$$\log[p_X(x)] - \log\sigma(x) - \frac{1}{2}\frac{[y-\mu(x)]^2}{\sigma^2(x)} = \log[p_Y(y)] - \log\tilde{\sigma}(y) - \frac{1}{2}\frac{[x-\tilde{\mu}(y)]^2}{\tilde{\sigma}^2(y)}. \tag{21}$$

Calculating on both sides $\frac{\partial^4}{\partial^2 x \partial^2 y}$, we obtain

$$\frac{\sigma''(x)\sigma(x) - 3\sigma'(x)'\sigma(x)}{\sigma^4(x)} = \frac{\tilde{\sigma}''(y)\tilde{\sigma}(y) - 3\tilde{\sigma}'(y)\tilde{\sigma}'(y)}{\tilde{\sigma}^4(y)}.$$

Since this has to hold for all x, y, both sides need to be constant (let us denote this constant by $a \in \mathbb{R}$).

Differential equation $\sigma''(x)\sigma(x) - 3\sigma'(x)\sigma'(x) = a\,\sigma^4(x)$ has solution $\sigma(x) = \frac{1}{\sqrt{a(x+b)^2+c}}$ for x , such that $a(x+b)^2 + c > 0$.

Plugging this result into (21) and calculating on both sides $\frac{\partial^3}{\partial^2 x \partial y}$, we obtain

$$\mu''(x)(a(x+b)^2 + c) + \mu'(x)(4ax + 4ab) + \mu(x)2a = 2ab. \tag{22}$$

Equation (22) is another differential equation with a solution $\mu(x) = \frac{d+ex}{a(x+b)^2+c} + b$, for some $d, e \in \mathbb{R}$ for all x: $\sigma(x) > 0$.

Next, we show that it is necessary that $b = 0$. If we show $b = 0$, then $\mu(x)$ and $\sigma^2(x)$ are exactly in the form (4). We plug the representations $\mu(x) = \frac{d+ex}{a(x+b)^2+c} + b, \sigma(x) = \frac{1}{\sqrt{a(x+b)^2+c}}$ and $\tilde{\mu}(x) = \frac{\tilde{d}+\tilde{e}x}{\tilde{a}(x+\tilde{b})^2+\tilde{c}} + \tilde{b}, \tilde{\sigma}(x) = \frac{1}{\sqrt{\tilde{a}(x+\tilde{b})^2+\tilde{c}}}$ into (21). Thus, we obtain

$$\begin{aligned}
&\log[p_X(x)] + \frac{1}{2}\log[a(x+b)^2 + c] \\
&- \frac{1}{2}\left[y^2(ax^2 + 2abx + ab^2 + c) + y(2d + 2ex) + \frac{1}{a(x+b)^2 + c}\right] \\
&= \log[p_Y(y)] + \frac{1}{2}\log[\tilde{a}(y+\tilde{b})^2 + \tilde{c}] \\
&- \frac{1}{2}\left[x^2(\tilde{a}y^2 + 2\tilde{a}\tilde{b}y + \tilde{a}\tilde{b}^2 + \tilde{c}) + x(2\tilde{d} + 2\tilde{e}y) + \frac{1}{\tilde{a}(y+\tilde{b})^2 + \tilde{c}}\right].
\end{aligned}$$

We can re-write the last expression as

$$\begin{aligned}
h_X(x) + h_Y(y) &= \frac{1}{2}[y^2(ax^2 + 2abx) + 2yex] - \frac{1}{2}[x^2(\tilde{a}y^2 + 2\tilde{a}\tilde{b}y) + 2x\tilde{e}y)] \\
&= \frac{1}{2}[x^2y^2(a - \tilde{a}) + xy(2aby - 2\tilde{a}\tilde{b}x + e - \tilde{e})],
\end{aligned} \tag{23}$$

where

$$h_X(x) = \log[p_X(x)] + \frac{1}{2}\log[a(x+b)^2 + c] + \frac{1}{2}\frac{1}{a(x+b)^2 + c} - 2x\tilde{d} - x^2(\tilde{a}\tilde{b}^2 + \tilde{c}),$$

$$h_Y(y) = -\log[p_Y(y)] - \frac{1}{2}\log[\tilde{a}(y+\tilde{b})^2 + \tilde{c}] + \frac{1}{2}[\frac{1}{\tilde{a}(y+\tilde{b})^2+\tilde{c}} - 2yd - y^2(ab^2+c)].$$

Since the left-hand side of (23) is in additive form, the right side also needs to have an additive representation. However, that is only possible if $a - \tilde{a} = 0$ and $2aby - 2\tilde{a}\tilde{b}x + e - \tilde{e} = 0$. Therefore, we necessarily have $a = \tilde{a}$ and either $a = 0$ or $b = \tilde{b} = 0$. The case $a = 0$ corresponds to a constant σ and, hence, also $b = \tilde{b} = 0$. We have shown that $\mu(x)$ and $\sigma^2(x)$ have to satisfy (4).

Next, we show that if the causal graph is not identifiable, then the density of $p_X(x)$ has form (5). Plugging the form of $\mu(x)$ and $\sigma^2(x)$ into (21), we obtain

$$\begin{aligned}\log[p_X(x)] - \log\sigma(x) - \frac{1}{2}\left(y - \frac{d+ex}{ax^2+c}\right)^2(ax^2+c) \\ = \log[p_Y(y)] - \log\tilde{\sigma}(y) - \frac{1}{2}\left(x - \frac{\tilde{d}+\tilde{e}y}{ay^2+\tilde{c}}\right)^2(ay^2+\tilde{c}).\end{aligned}$$

We rewrite

$$\begin{aligned}\log[p_X(x)] - \log\sigma(x) + \frac{1}{2}\left[\tilde{c}x^2 - 2x\tilde{d} + \frac{(d+ex)^2}{ax^2+c}\right] \\ = \log[p_Y(y)] - \log\tilde{\sigma}(y) + \frac{1}{2}\left[cy^2 - 2yd + \frac{(\tilde{d}+\tilde{e}y)^2}{ay^2+\tilde{c}}\right].\end{aligned} \tag{24}$$

Since this has to hold for all $x, y \in \mathbb{R}$, both sides of (24) need to be constant and we obtain $\log[p_X(x))] \propto \log\sigma(x) - \frac{1}{2}\left[\tilde{c}x^2 - 2x\tilde{d} + \frac{(d+ex)^2}{ax^2+c}\right]$. Hence,

$$p_X(x) \propto \sigma(x)e^{-\frac{1}{2}\left[\tilde{c}x^2 - 2x\tilde{d} + \frac{(d+ex)^2}{ax^2+c}\right]} = \sigma(x)e^{-\frac{1}{2}\left[\frac{(x-\alpha)^2}{\beta^2} - \frac{\mu^2(x)}{\sigma^2(x)}\right]},$$

where $\beta = 1/\sqrt{\tilde{c}}$ and $\alpha = \frac{\tilde{d}}{\tilde{c}}$. The condition $\frac{1}{\beta^2} > \frac{e^2}{c}\mathbb{1}[a=0]$ arises from the fact that if $a = 0$ and $\frac{1}{\beta^2} \leq \frac{e^2}{c}$, then $p_X(x)$ is not a density function. This is because

$$\sigma(x)e^{-\frac{1}{2}\left[\frac{(x-\alpha)^2}{\beta^2} - \frac{\mu^2(x)}{\sigma^2(x)}\right]} \propto e^{-\frac{1}{2}\left[x^2\left(\frac{1}{\beta^2} - \frac{e^2}{c}\right) + x\left(-2\frac{\alpha}{\beta^2} - \frac{2de}{c}\right)\right]}$$

for all $x \in \mathbb{R}$. This expression is integrable is and only if the coefficient at x^2 is positive.

Finally, we deal with the other direction: we show that if μ and σ satisfy (4) and p_{ε_X} has form (5), then the causal graph is not identifiable. Assume that a, c, d, e are given. Define $\tilde{a} = a, \tilde{e} = e$ and select $\tilde{c}, \tilde{d} \in \mathbb{R}$, such that $\tilde{c} > 0, \tilde{c} \neq \frac{e^2}{c}\mathbb{1}[a=0]$. Define $\frac{1}{\tilde{\sigma}^2(y)} = \tilde{a}y^2 + \tilde{c}$, $\frac{\tilde{\mu}(y)}{\tilde{\sigma}^2(y)} = \tilde{d} + \tilde{e}y$. Moreover, define

$$\begin{aligned}p_X(x) &\propto \sigma(x)e^{-\frac{1}{2}\left[\tilde{c}\left(x - \frac{\tilde{d}}{\tilde{c}}\right)^2 - \frac{\mu^2(x)}{\sigma^2(x)}\right]} \quad \text{and} \\ p_Y(y) &\propto \tilde{\sigma}(y)e^{-\frac{1}{2}\left[c\left(x - \frac{d}{c}\right)^2 - \frac{\tilde{\mu}^2(y)}{\tilde{\sigma}^2(y)}\right]}.\end{aligned}$$

Note that regardless of the coefficients, these are valid density functions (with one exception when $\tilde{c} = \frac{e^2}{c}$ and $a = 0$, which is why we selected $\tilde{c} \neq \frac{e^2}{c}\mathbb{1}[a=0]$). In case of $a = 0$, this is the classical Gaussian distribution density function.

Using these values, we obtain the equality

$$p_X(x)p_{Y|X}(y \mid x) = p_Y(y)p_{X|Y}(x \mid y), \forall x, y \in \mathbb{R},$$

or more precisely,

$$\sigma(x)e^{-\frac{1}{2}\left[\tilde{c}\left(x-\frac{\tilde{d}}{\tilde{c}}\right)^2-\frac{\mu^2(x)}{\sigma^2(x)}\right]}\frac{1}{\sqrt{2\pi}\sigma(x)}e^{-\frac{1}{2}\frac{[y-\mu(x)]^2}{\sigma^2(x)}} \propto \tilde{\sigma}(y)e^{-\frac{1}{2}\left[c\left(x-\frac{d}{c}\right)^2-\frac{\tilde{\mu}^2(y)}{\tilde{\sigma}^2(y)}\right]}\frac{1}{\sqrt{2\pi}\tilde{\sigma}(y)}e^{-\frac{1}{2}\frac{(x-\tilde{\mu}(y))^2}{\tilde{\sigma}^2(y)}}.$$

Since this holds for all $x, y \in \mathbb{R}$, we found a valid backward model. The density in (5) uses the notation $\alpha = \frac{\tilde{d}}{\tilde{c}}$ and $\beta = 1/\sqrt{\tilde{c}}$. ■

An example of the joint distribution of X_1, X_2 with $a = c = d = e = \alpha = \beta = 1$ is depicted in Figure 9.

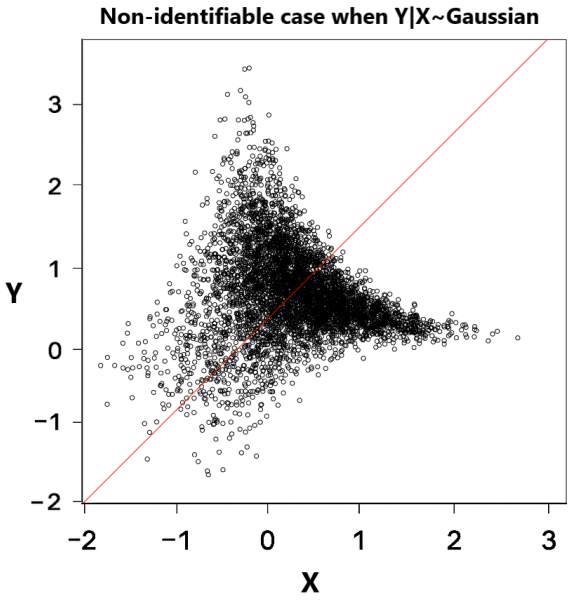

Figure 9: Random sample from a joint distribution of (X_1, X_2), where X_1 has the marginal density (5) and $X_2 \mid X_1 \sim N\big(\mu(X_1), \sigma^2(X_1)\big)$ with μ, σ defined in (4) with constants $a = c = d = e = \alpha = \beta = 1$. The distribution function is symmetric according to the $x = y$ axis (red line).

C.2 Proof of Consequence 6

Consequence 6 • *Let (X_1, X_2) admit the CPCM(F) model with graph $X_1 \to X_2$, where F is the (discrete) Poisson distribution function. Then, the causal graph is not identifiable if and only if*

$$\lambda(x) = e^{ax+b}, \quad P(X_1 = x) \propto \frac{e^{e^{ax+b}+cx}}{x!}, \quad \forall x \in \mathbb{N}_0, \tag{25}$$

for some $a < 0, b, c \in \mathbb{R}$.

- *Let (X_1, X_2) admit the $CPCM(F)$ model with graph $X_1 \to X_2$, where F is the Pareto distribution function. Then, the causal graph is not identifiable if and only if*

$$\theta(x) = a\log(x) + b, \quad p_{X_1}(x) \propto \frac{1}{[a\log(x)+b]x^{c+1}}, \quad \forall x \geq 1, \tag{26}$$

for some $a, b, c > 0$.

- *Let (X_1, X_2) admit the $CPCM(F)$ model with graph $X_1 \to X_2$, where F is Bernoulli distribution function. Then, the causal graph is identifiable if and only if $supp(X_1) \neq \{0, 1\}$.*

Proof : **First bullet-point:** Poisson distribution has one parameter and it can be written as $h_1(x) = 1/x!, h_2(x) = e^{-e^x}, T(x) = x$. Note that we do not use classical form of density function but its reparametrisation where $\theta(x) = \log(\lambda(x))$ where λ is the classical rate parameter.

Plugging this into Proposition 5, we directly obtain (25) with possible $a \in \mathbb{R}$. It is not hard to see that $\frac{e^{e^{ax+b}+cx}}{x!}$ is integrable if and only if $a < 0$. In such a case, the backward model exist and has a form $\tilde{\theta}(y) = ay + c$ and $P(X_2 = y) \propto \frac{e^{e^{ay+c}+by}}{y!}$.

Second bullet-point: Pareto distribution has one parameter and it can be written as $h_1(x) = 1, h_2(\theta) = 1/\theta, T(x) = log(x)$.

Plugging this into Proposition 5, we directly obtain (26) with possible $a, b, c \in \mathbb{R}$. It is not hard to see that the density is integrable if and only if $a, c, b > 0$, in which case, the backward model exist and has a form $\tilde{\theta}(y) = \tilde{a}\log(y) + \tilde{b}$, $p_{X_2}(x) \propto \frac{1}{[\tilde{a}\log(x)+\tilde{b}]x^{\tilde{c}+1}}$ for $\tilde{a} = a, \tilde{b} = c, \tilde{c} = b$.

Third bullet-point: If $supp(X_1) \neq \{0, 1\}$, then the causal graph is identifiable due to the first bullet-point in Proposition 5. Next, consider $supp(X_1) = \{0, 1\}$. In this case, we can always write a backward model for the Bernoulli distribution. Let $P(X_1 = X_2 = 0) = p_0$, $P(X_1 = 0, X_2 = 1) = p_{0,1}$, $P(X_1 = 1, X_2 = 0) = p_{1,0}$ and $P(X_1 = X_2 = 1) = p_1$ for $p_0, p_{0,1}, p_{1,0}, p_1 > 0$ and $p_0 + p_{0,1} + p_{1,0} + p_1 = 1$. We can define $X_2 \mid X_1 \sim Bernoulli(\theta(X_1))$ as $\theta(0) = p_{0,1}$ and $\theta(1) = p_1$. On the other hand, we can define $X_1 \mid X_2 \sim Bernoulli(\tilde{\theta}(X_2))$ as $\tilde{\theta}(0) = p_{1,0}$ and $\tilde{\theta}(1) = p_1$. Since both models produce the same joint distribution, the causal model is not identifiable for any values of $p_0, p_{0,1}, p_{1,0}, p_1$. ■

C.3 Proof of Theorem 7

Before we prove Theorem 7, we show the following auxiliary lemma.

Lemma 18 *Let $n \in \mathbb{N}$ and $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$. Let $f_1, \ldots, f_n, g_1, \ldots, g_n$ be non-constant functions on $\mathcal{X}, \mathcal{Y}$, respectively, such that $f_1(x)g_1(y) + \cdots + f_n(x)g_n(y)$ is additive in x, y—that is, there exist functions f and g, such that*

$$f_1(x)g_1(y) + \cdots + f_n(x)g_n(y) = f(x) + g(y), \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

Then, there exist (not all zero) constants $a_1, \ldots, a_n, c \in \mathbb{R}$, such that $\sum_{i=1}^n a_i f_i(x) = c$ for all $x \in \mathcal{X}$. Specifically for $n = 2$, it holds that $f_1(x) = af_2(x) + c$ for some $a, c \in \mathbb{R}$.

Moreover, assume that for some $q < n$, functions $g_1, \dots, g_q$ are affinly independent—that is, there exist $y_0, y_1, \dots, y_q \in \mathcal{Y}$, such that a matrix

$$M = \begin{pmatrix} g_1(y_1) - g_1(y_0) & \cdots & g_q(y_1) - g_q(y_0) \\ \dots & \ddots & \dots \\ g_1(y_q) - g_1(y_0) & \cdots & g_q(y_q) - g_q(y_0) \end{pmatrix} \tag{27}$$

has full rank. Then, for all $i = 1, \dots, q$ there exist constants $a_{q+1}, \dots, a_n, c \in \mathbb{R}$, such that $f_i(x) = \sum_{j=q+1}^{n} a_j f_j(x) + c$ for all $x \in \mathcal{X}$.

Proof Fix $y_1, y_2 \in \mathcal{Y}$, such that $g_1(y_1) \neq g_1(y_2)$. Then, we have for all $x \in \mathcal{X}$

$$\begin{aligned} f_1(x)g_1(y_1) + \cdots + f_n(x)g_n(y_1) &= f(x) + g(y_1), \\ f_1(x)g_1(y_2) + \cdots + f_n(x)g_n(y_2) &= f(x) + g(y_2), \end{aligned}$$

and subtraction of these equalities yields

$$f_1(x)[g_1(y_1) - g_1(y_2)] + \cdots + f_n(x)[g_n(y_1) - g_n(y_2)] = g(y_1) - g(y_2).$$

Defining $a_i = g_i(y_1) - g_i(y_2)$ and $c = g(y_1) - g(y_2)$ yields the first result (with $a_1 \neq 0$).

Now, we prove the "Moreover" part. Consider equalities

$$\begin{aligned} f_1(x)g_1(y_0) + \cdots + f_n(x)g_n(y_0) &= f(x) + g(y_0), \\ f_1(x)g_1(y_1) + \cdots + f_n(x)g_n(y_1) &= f(x) + g(y_1), \\ &\dots \\ f_1(x)g_1(y_q) + \cdots + f_n(x)g_n(y_q) &= f(x) + g(y_q), \end{aligned}$$

where $y_0, \dots, y_q$ are defined, such that matrix (27) has full rank. Subtracting from each equality, the first equality yields

$$\begin{aligned} f_1(x)[g_1(y_1) - g_1(y_0)] + \cdots + f_n(x)[g_n(y_1) - g_n(y_0)] &= g(y_1) - g(y_0) \\ &\dots \\ f_1(x)[g_1(y_q) - g_1(y_0)] + \cdots + f_n(x)[g_n(y_q) - g_n(y_0)] &= g(y_q) - g(y_0). \end{aligned}$$

Using matrix formulation, this can be rewritten as

$$M \begin{pmatrix} f_1(x) \\ \cdots \\ f_q(x) \end{pmatrix} = \begin{pmatrix} g(y_1) - g(y_0) - \sum_{j=q+1}^{n} f_j(x)[g_j(y_1) - g_j(y_0)] \\ \cdots \\ g(y_q) - g(y_0) - \sum_{j=q+1}^{n} f_j(x)[g_j(y_q) - g_j(y_0)] \end{pmatrix}. \tag{28}$$

Multiplying both sides by M^{-1} indicates that $f_i(x), i = 1, \dots, q$ are nothing else than a linear combination of $f_{q+1}(x), \dots, f_n(x)$, which is what we wanted to show. ■

Theorem 7 *Let (X_1, X_2) follow the $CPCM(F_1, \dots, F_k)$ model with graph $X_1 \to X_2$, where $F_1, \dots, F_k$ belong to the exponential family of distributions with corresponding sufficient statistics $T_m = (T_{m,1}, \dots, T_{m,q_m})^\top$, $m = 1, \dots, k$. Following Definition 2, let $\tilde{m} \in \{1, \dots, k\}$ be the index such that $X_2 = F_{\tilde{m}}^{-1}\big(\varepsilon_2; \theta_2(X_1)\big)$.*

The causal graph is identifiable if for all $m \in \{1, \dots, k\}$, at least one of the following holds:

- $supp(F_m) \neq supp(X_1)$.
- *The function θ_2 is not a linear combination of the sufficient statistics $T_{m,1}, \ldots, T_{m,q_m}$, i.e., there do not exist coefficients $a_{i,j}, b_i \in \mathbb{R}$ for $i = 1, \ldots, q_{\tilde{m}}$ and $j = 1, \ldots, q_m$ such that*

$$\theta_{2,i}(x) = \sum_{j=1}^{q_m} a_{i,j} T_{m,j}(x) + b_i, \quad \forall x \in \text{supp}(X_1), \quad \forall i \in \{1, \ldots, q_{\tilde{m}}\}. \tag{10}$$

- *There do not exist constants $c_1, \ldots, c_{q_m} \in \mathbb{R}$ such that the density of X_1 satisfies*

$$p_{X_1}(x) \propto \frac{h_{m,1}(x)}{h_{\tilde{m},2}[\theta_2(x)]} e^{\sum_{i=1}^{q_m} c_i T_{m,i}(x)}, \quad \forall x \in \text{supp}(X_1), \tag{11}$$

where $h_{m,1}$ is a base measure associated with F_m and $h_{\tilde{m},2}$ is the normalizing function of $F_{\tilde{m}}$, both defined in Appendix A.1.

Consequentially, the space of non-identifiable distributions is contained in a $\tilde{d}$-dimensional space, where

$$\tilde{d} = \sum_{m \in \{1,\ldots,k\}: supp(F_m) = supp(X_1)} (q_m + 1)(q_{\tilde{m}} + 1) - 1. \tag{12}$$

Proof

If the $CPCM(F_1, \ldots, F_k)$ is *not* identifiable, then there exists $m \in \{1, \ldots, k\}$ and functions θ_1 and θ_2, such that models

$$X_1 = \varepsilon_1, X_2 = F_{\tilde{m}}^{-1}(\varepsilon_2, \theta_2(X_1)), \text{ and } X_2 = \varepsilon_2, X_1 = F_m^{-1}(\varepsilon_1, \theta_1(X_2)) \tag{29}$$

generate the same joint density function. For simplifying the notation, let $m = 1$ and $\tilde{m} = 2$.

1) Trivially, X_1 can not be generated as $X_1 = F_1^{-1}(\varepsilon_1, \theta_1(X_2))$ if $supp(F_1) \neq supp(X_1)$.

2) For a contradiction, we show that θ_2 is a linear combination of $T_{1,1}, \ldots, T_{1,q_m}$. Decompose the joint density as

$$p_{(X_1,X_2)}(x, y) = p_{X_1}(x) p_{X_2|X_1}(y \mid x) = p_{X_2}(y) p_{X_1|X_2}(x \mid y), \quad x \in supp(X_1), y \in supp(X_2). \tag{30}$$

Since F_1 and F_2 lie in the exponential family of distributions, we use the notation from Appendix A.1 and rewrite it as

$$\begin{aligned} p_{X_2|X_1}(y \mid x) &= h_{1,1}(y) h_{1,2}[\theta_2(x)] e^{\sum_{i=1}^{q_2} \theta_{2,i}(x) T_{2,i}(y)}, \\ p_{X_1|X_2}(x \mid y) &= h_{2,1}(x) h_{2,2}[\theta_1(y)] e^{\sum_{i=1}^{q_1} \theta_{1,i}(y) T_{1,i}(x)}. \end{aligned}$$

After a logarithmic transformation of both sides of (30), we obtain

$$\begin{aligned} \log[p_{(X_1,X_2)}(x, y)] &= \log[p_{X_1}(x)] + \log[h_{1,1}(y)] + \log\{h_{1,2}[\theta_2(x)]\} + \sum_{i=1}^{q_2} \theta_{2,i}(x) T_{2,i}(y) \\ &= \log[p_{X_2}(y)] + \log[h_{2,1}(x)] + \log\{h_{2,2}[\theta_1(y)]\} + \sum_{i=1}^{q_1} \theta_{1,i}(y) T_{1,i}(x). \end{aligned} \tag{31}$$

Define $f(x) = \log[p_{X_1}(x)] + \log\{h_{1,2}[\theta_2(x)]\} - \log[h_{2,1}(x)]$ and $g(y) = \log[h_{1,1}(y)] - \log[p_{X_2}(y)] + \log\{h_{2,2}[\theta_1(y)]\}$. Then, equality (31) reads as

$$f(x) + g(y) = \sum_{i=1}^{q_1} \theta_{1,i}(y) T_{1,i}(x) - \sum_{i=1}^{q_2} T_{2,i}(y) \theta_{2,i}(x). \tag{32}$$

Finally, we use Lemma 18. We know that functions $T_{2,i}$ are affinly independent in the sense presented in Lemma 18 (see (18) in Appendix A.1). Therefore, Lemma 18 gives us that $\theta_{2,i}, i = 1, \dots, q_2$ are only a linear combination of $T_{1,j}, j = 1, \dots, q_1$, which is what we wanted to show.

3) For a contradiction, we show that p_{X_1} must have a form (11). Let us rewrite equation (32) into

$$\begin{aligned} \log[p_{X_1}(x)] = &- \log\{h_{1,2}[\theta_2(x)]\} + \log[h_{2,1}(x)] - g(y) \\ &+ \sum_{i=1}^{q_1} \theta_{1,i}(y) T_{1,i}(x) - \sum_{i=1}^{q_2} T_{2,i}(y) \theta_{2,i}(x). \end{aligned} \tag{33}$$

Fix $y \in supp(F_2)$. Using the form of $\theta_{2,i}$ from the previous bullet-point, we can write

$$\begin{aligned} &\sum_{i=1}^{q_1} \theta_{1,i}(y) T_{1,i}(x) - \sum_{i=1}^{q_2} T_{2,i}(y) \theta_{2,i}(x) = \sum_{i=1}^{q_1} \theta_{1,i}(y) T_{1,i}(x) - \sum_{i=1}^{q_2} T_{2,i}(y) \Big[\sum_{j=1}^{q_1} a_{i,j} T_{1,j}(x) + b_i \Big] \\ &= \sum_{i=1}^{q_1} c_i T_{1,i}(x) + d, \end{aligned}$$

where $c_i = \theta_{1,i}(y) - \sum_{j=1}^{q_2} \sum_{k=1}^{q_1} T_{2,i}(y) a_{i,j}$ and $d = \sum_{j=1}^{q_2} b_j T_{2,j}(y)$. Therefore, equation 33 can be written as

$$\log[p_{X_1}(x)] = - \log\{h_{1,2}[\theta_2(x)]\} + \log[h_{2,1}(x)] + \sum_{i=1}^{q_1} c_i T_{1,i}(x) + [d - g(y)].$$

Applying exponential on both sides, we obtain (11).

Part "Consequentially": We have shown that if (29) holds, then $\text{supp}(F_m) = \text{supp}(X_1)$, and the joint density $p_{(X_1,X_2)}$ is uniquely determined by the coefficients $a_{i,j}, b_i, c_j \in \mathbb{R}$, where $i = 1, \dots, q_{\tilde{m}}$ and $j = 1, \dots, q_m$.

By counting the number of these coefficients, we find that there are $(q_m + 1)(q_{\tilde{m}} + 1) - 1$ of them, with the "-1" term accounting for the normalization of the density function. Consequently, (12) follows by summing over all $m \in \{1, \dots, k\}$. ■

C.4 Proof of Consequence 8

Consequence 8 • *Suppose that $supp(X_1) = \mathbb{R}$, $supp(X_2) = \{0, 1, \dots\}$ such as on Figure 1, and let (X_1, X_2) admit the $CPCM(F_1, F_2)$ model with graph $X_1 \to X_2$, where F_1 is a Gaussian distribution and F_2 is a Poisson distribution with rate parameter λ. The*

causal graph is identifiable if and only if there do not exist constants $a_1, a_2, b, c_1, c_2 \in \mathbb{R}$, $a_1, c_1 < 0$, such that for all $x \in \mathbb{R}$

$$\lambda(x) = e^{a_1 x^2 + a_2 x + b}, \quad p_{X_1}(x) \propto e^{c_1 x^2 + c_2 x}.$$

- *Let (X_1, X_2) admit the $CPCM(F)$ model with graph $X_1 \to X_2$, where F is a Gamma distribution with parameters $\theta = (\alpha, \beta)^\top$. If there do not exist constants $a, b, c, d, e, f \in \mathbb{R}$ such that*

$$\alpha(x) = a\log(x) + bx + c, \quad \beta(x) = d\log(x) + ex + f, \quad \forall x > 0,$$

then the causal graph is identifiable.

- *Let (X_1, X_2) admit the $CPCM(F_1, F_2)$ model, where F_1 is a Gamma distribution with parameters $\theta_1 = (\alpha_1, \beta_1)^\top$ and F_2 is a Beta distribution with parameters $\theta_2 = (\alpha_2, \beta_2)^\top$. If there do not exist constants $a_i, b_i, c_i, d_i, e_i, f_i \in \mathbb{R}$, $i = 1, 2$, such that for all $x \in (0, 1)$*

$$\alpha_1(x) = a_1\log(x) + b_1 x + c_1, \quad \beta_1(x) = d_1\log(x) + e_1 x + f_1,$$
$$\alpha_2(x) = a_2\log(x) + b_2\log(1-x) + c_2, \quad \beta_2(x) = d_2\log(x) + e_2\log(1-x) + f_2,$$

then the causal graph is identifiable.

Proof Poisson distribution has one parameter and it can be written as $h_1(x) = 1/x!, h_2(x) = e^{-e^x}, T(x) = x$. Note that we do not use classical form of density function but its reparametrisation where $\theta(x) = \log(\lambda(x))$ where λ is the classical rate parameter. Theorem 7 gives us that the causal graph is identifiable if there do not exist constants $a_1, a_2, b, c_1, c_2 \in \mathbb{R}$, such that for all $x \in \mathbb{R}$

$$\lambda(x) = e^{a_1 x^2 + a_2 x + b}, \quad p_{X_1}(x) \propto e^{c_1 x^2 + c_2 x},$$

then the causal graph is identifiable. It is a simple exercise to prove that the joint distribution is integrable if and only if $a_1, c_1 < 0$.

The second and the third bullet-point follow directly from Theorem 7, noting the following:

- The density function of the **Gamma distribution** with parameters $\theta = (\alpha, \beta)^\top$ is given by $p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, $x > 0$. The sufficient statistics are $[T_1(x), T_2(x)] = [\log(x), x]$.

- The density function of the **Beta distribution** with parameters $\theta = (\alpha, \beta)^\top$ is given by $p(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$. The sufficient statistics are $[T_1(x), T_2(x)] = [\log(x), \log(1-x)]$.

■

C.5 Proof of Lemma 11

Lemma 11 *Let $F_{\boldsymbol{X}}$ be generated by the $CPCM(F_1,\dots,F_k)$ with DAG $\mathcal{G}$ and with density $p_{\boldsymbol{X}}$. Assume that for all $i,j \in \mathcal{G}$, $S \subseteq V$, such that $i \in pa_j$ and $pa_j \setminus \{i\} \subseteq S \subseteq nd_j \setminus \{i,j\}$, there exist $\boldsymbol{x}_S$: $p_S(\boldsymbol{x}_S) > 0$, such that a bivariate model defined as $X = \tilde{\varepsilon}_X, Y = F_j^{-1}\big(\tilde{\varepsilon}_Y, \tilde{\theta}(X)\big)$ is identifiable (in the sense of Definition 3), where $F_{\tilde{\varepsilon}_X} = F_{X_i|\boldsymbol{X}_S=\boldsymbol{x}_S}$ and $\tilde{\theta}(x) = \theta_j(\boldsymbol{x}_{pa_j\setminus\{i\}}, x)$, $x \in supp(X)$. Then, $\mathcal{G}$ is identifiable from the joint distribution.*

Proof Let there be two $CPCM(F_1,\dots,F_k)$ models, with causal graphs $\mathcal{G} \neq \mathcal{G}'$, that both generate $F_{\mathbf{X}}$. From Proposition 29 in Peters et al. (2014) (recall that we assume causal minimality of $CPCM(F_1,\dots,F_k)$), there exist variables $L, Y \in \{X_1,\dots,X_d\}$, such that

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

For this S, select $\mathbf{x}_S$ in accordance to the condition in the theorem. Below, we use the notation $\mathbf{x}_S = (\mathbf{x}_q, \mathbf{x}_r)$ where $q \in \mathbf{Q}, r \in \mathbf{R}$. Now, we use Lemma 36 and Lemma 37 from (Peters et al., 2014). Since $Y \to L$ in $\mathcal{G}$, we define a bivariate SCM as [1]

$$Y^\star = \tilde{\varepsilon}_{Y^\star}, \qquad L^\star = F_L^{-1}\big(\varepsilon_L; \theta_L(Y^\star)\big),$$

where $\tilde{\varepsilon}_{Y^\star} \overset{D}{=} Y \mid \{\mathbf{X}_S = \mathbf{x}_S\}$ and $\varepsilon_L \perp\!\!\!\perp Y^\star, \varepsilon_L \sim U(0,1)$. This is a bivariate CPCM with $Y^\star \to L^\star$. However, the same holds for the other direction: Since $L \to Y$ in $\mathcal{G}'$, we can also define a bivariate SCM in the following manner:

$$L^\star = \tilde{\varepsilon}_{L^\star}, \qquad Y^\star = F_Y^{-1}\big(\varepsilon_Y; \theta_Y(L^\star)\big),$$

where $\tilde{\varepsilon}_{L^\star} \overset{D}{=} L \mid \{\mathbf{X}_S = \mathbf{x}_S\}$ and $\varepsilon_Y \perp\!\!\!\perp L^\star, \varepsilon_Y \sim U(0,1)$. We obtained a bivariate CPCM with $L^\star \to Y^\star$, which is a contradiction with the pairwise identifiability. Hence, $\mathcal{G} = \mathcal{G}'$. ■

C.6 Proof of Lemma 15

Lemma 15 *Suppose that the joint distribution $F_{(X_1,X_2)}$ is generated according to the model $CPCM(F_2)$ with graph $X_1 \to X_2$, where F_2 is a distribution function belonging to the exponential family.*

Then, there exists F_1 such that the model $CPCM(F_1)$ with graph $X_2 \to X_1$ also generates $F_{(X_1,X_2)}$. In other words, there exists F_1 such that the causal graph in $CPCM(F_1,F_2)$ is not identifiable from the joint distribution.

Proof The idea of the proof is the following: we select F_1, such that its sufficient statistic is equal to θ_2.

Let us denote the original model as

$$X_1 = \varepsilon_1, X_2 = F_2^{-1}\big(\varepsilon_2, \theta_2(X_1)\big), \varepsilon_2 \sim U(0,1), \varepsilon_1 \perp\!\!\!\perp \varepsilon_2,$$

1. Informally, we consider $Y^\star := Y \mid \{\mathbf{X}_S = \mathbf{x}_S\}$ and $L^\star := L \mid \{\mathbf{X}_S = \mathbf{x}_S\}$.

where (using notation from Appendix A.1) the conditional density function has a form:

$$p_{X_2|X_1}(y \mid x) = h_{2,1}(y)h_{2,2}[\theta_2(x)] \exp[\theta_2(x)T_2(y)].$$

We define F_1 from an exponential family in the following manner: consider the sufficient statistic $T_1(x) = \theta_2(x)$ for all x in support of X_1 and choose $h_{1,1}(x) = p_{X_1}(x)h_{2,2}[\theta_2(x)]$ and $h_{1,2}(y) = \frac{h_{2,1}(y)}{p_{X_2}(y)}$ for all y in support of X_2. Then, a model where

$$X_2 = \varepsilon_2, X_1 = F_1^{-1}\big(\varepsilon_1, \theta_1(X_2)\big), \varepsilon_1 \sim U(0,1), \varepsilon_1 \perp\!\!\!\perp \varepsilon_2,$$

for a specific choice $\theta_1(y) = T_2(y)$ has the following conditional density function:

$$p_{X_1|X_2}(x \mid y) = h_{1,1}(x)h_{1,2}[\theta_1(y)] \exp[\theta_1(y)T_1(x)] = \frac{p_{X_1}(x)}{p_{X_2}(y)} h_{2,1}(y)h_{2,2}[\theta_2(x)] \exp[\theta_2(x)T_2(y)].$$

Therefore, the joint distribution is equal in both models, since

$$p_{X_1}(x)h_{2,1}(y)h_{2,2}[\theta_2(x)] \exp[\theta_2(x)T_2(y)] = p_{X_2}(y)\frac{p_{X_1}(x)}{p_{X_2}(y)} h_{2,1}(y)h_{2,2}[\theta_2(x)] \exp[\theta_2(x)T_2(y)]$$
$$p_{X_1}(x)p_{X_2|X_1}(y \mid x) = p_{X_2}(y)p_{X_1|X_2}(x \mid y).$$

We found $CPCM(F_1)$ model with graph $X_2 \to X_1$ that generates the same distribution. This completes the proof. ■

References

C. Assaad, E. Devijver, and E. Gaussier. Discovery of extended summary graphs in time series. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 96–106. PMLR, 01–05 Aug 2022.

J. Bodik and V. Chavez-Demoulin. Structural restrictions in local causal discovery: identifying direct causes of a target variable. *Biometrika*, page asaf042, 06 2025. ISSN 1464-3510. URL https://doi.org/10.1093/biomet/asaf042.

J. Bodik, M. Paluš, and Z. Pawlas. Causality in extremes of time series. *Extremes*, 27 (1):67–121, 2024. doi: 10.1007/s10687-023-00479-5. URL https://doi.org/10.1007/s10687-023-00479-5.

J. Bodik, Y. Huang, and B. Yu. Cross-world assumption and refining prediction intervals for individual treatment effects, 2025. URL https://arxiv.org/abs/2507.12581. arXiv:2507.12581.

P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014. doi: 10.1214/14-aos1260.

G. Casella and R. L. Berger. *Statistical Inference*. Chapman and Hall/CRC, 2nd edition, 2024. doi: 10.1201/9781003456285.

Y. Chen and R. J. Samworth. Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(4):729–754, 2015. doi: 10.1111/rssb.12137.

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

D. M. Chickering, D. Heckerman, and Ch. Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

G. Claeskens, T. Krivobokova, and J. D. Opsomer. Asymptotic properties of penalized spline estimators for generalized additive models. *Statistica Sinica*, 19(2):621–640, 2009.

S. Drews and M. Kohler. On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent. *Annals of the Institute of Statistical Mathematics*, 76(3):361–391, 2024.

J. Fan and Q. Yao. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660, 09 1998. ISSN 0006-3444. doi: 10.1093/biomet/85.3.645.

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

J. L. Gamella, M. Blohsfeld, M. Hein, et al. Causal chambers as a real-world physical testbed for AI methods. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-024-00964-x. Case studies include causal discovery on physical systems.

M. Gao, Y. Ding, and B. Aragam. A polynomial-time algorithm for learning nonparametric causal graphs. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

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.

N. Gnecco, N. Meinshausen, J. Peters, and S. Engelke. Causal discovery in heavy-tailed models. *The Annals of Statistics*, 49, 2020. doi: 10.1214/20-AOS2021.

O. Goudet, D. Kalainathan, P. Caillou, D. Lopez-Paz, I. Guyon, M. Sebag, A. Tritas, and P. Tubaro. Learning functional causal models with generative neural networks. *arXiv preprint arXiv:1709.05321*, 2017.

P. Green and B. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall/CRC, 1994. ISBN 9780412300400. URL https://www.routledge.com/Nonparametric-Regression-and-Generalized-Linear-Models-A-roughness-penalty/Green-Silverman/p/book/9780412300400.

S. Greenland, J. Pearl, and J.M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10:37–48, 1999. URL https://pubmed.ncbi.nlm.nih.gov/9888278/.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer New York, NY, 2002. doi: 10.1007/b97848.

J. Heiss. *Inductive Bias of Neural Networks and Selected Applications*. PhD thesis, ETH Zurich, Zurich, 2024. URL https://www.research-collection.ethz.ch/handle/20.500.11850/699241.

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., 2009. 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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 14316–14332. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/immer23a.html.

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.

D. Janzing, X. Sun, and B. Schölkopf. Distinguishing cause and effect via second order exponential models. *ArXiv e-prints (0910.5561)*, 2009. doi: 10.48550/ARXIV.0910.5561.

M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012. doi: 10.18637/jss.v047.i11.

G. Keropyan, D. Strieder, and M. Drton. Rank-based causal discovery for post-nonlinear models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 7849–7870. PMLR, 2023.

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, pages 3520–3528. PMLR, 2021. URL https://proceedings.mlr.press/v130/khemakhem21a.html.

D. Klippert and A. Marx. Skewness-robust causal discovery in location-scale noise models, 2025. URL https://arxiv.org/abs/2511.14441.

L. Kook, S. Saengkyongam, A. Lundborg, T. Hothorn, and J. Peters. Model-based causal feature selection for general response types. *Journal of the American Statistical Association*, 120(550):1090–1101, 2024. doi: 10.1080/01621459.2024.2395588.

M. Krali. Causal discovery in heavy-tailed linear structural equation models via scalings, 2025. URL https://arxiv.org/abs/2502.13762.

Q. Le, A. Smola, and S. Canu. Heteroscedastic gaussian process regression. In *ICML*, pages 489–496, 01 2005. doi: 10.1145/1102351.1102413.

C. Liu, X. Sun, J. Wang, T. Li, T. Qin, W. Chen, and T. Liu. Learning causal semantic representation for out-of-distribution prediction, 2021. URL https://openreview.net/forum?id=xyGFYKIPTDJ.

E. Mammen and S. van de Geer. Penalized quasi-likelihood estimation in partial linear models. *Annals of Statistics*, 25(3):1014–1035, 1997.

A. Marx and J. Vreeken. Telling cause from effect using mdl-based local and global regression. *Knowledge and Information Systems*, 60(3):1277–1305, 2019. doi: 10.1007/s10115-018-1286-7.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213, pages 726–751. PMLR, 11–14 Apr 2023.

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.

S. Nadarajah. A generalized normal distribution. *Journal of Applied Statistics*, 32(7):685–694, 2005. doi: 10.1080/02664760500079464.

A. Nandy, A. Hauser, and M. H. Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *Annals of Statistics*, 46(6A):3151–3183, 2018.

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 H. Park. Identifiability of generalized hypergeometric distribution (ghd) directed acyclic graphical models. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 158–166. PMLR, 2019. URL https://proceedings.mlr.press/v89/park19a.html.

G. Park and G. Raskutti. Learning large-scale poisson dag models based on overdispersion scoring. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/fccb60fb512d13df5083790d64c4d5dd-Paper.pdf.

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.

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

J. Pearl and D. Mackenzie. *The Book of Why*. Penguin Books, 2019. URL http://bayes.cs.ucla.edu/WHY/.

J. Peters and P. Bühlmann. Structural intervention distance (sid) for evaluating causal graphs. *arXiv preprint arXiv:1306.1043*, 2013.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12): 2436–2450, 2011. doi: 10.1109/TPAMI.2011.71.

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.

A. Poinsot, A. Leite, N. Chesneau, M. Sébag, and M. Schoenauer. Learning structural causal models through deep generative models: methods, guarantees, and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. doi: 10.24963/ijcai.2024/907.

G. Rajendran, B. Kivva, M. Gao, and B. Aragam. Structure learning in polynomial time: Greedy algorithms, bregman information, and exponential families. *NeurIPS, Conference Paper 7430*, 2021. doi: 10.48550/arXiv.2110.04719.

J. Ramsey, M. Glymour, R. Scheines, and P. Spirtes. The fast greedy equivalence search algorithm for learning high-dimensional graphical causal models. *Computational Statistics & Data Analysis*, 103:29–39, 2016.

R. A. Rigby and M. D. Stasinopoulos. Gamlss: Generalized additive models for location scale and shape, 2025. URL https://www.gamlss.com/. Accessed: 15 March 2025.

K. Sarpal. fremtpl2 – french motor third-party liability insurance claims. Kaggle dataset, 2025. URL https://www.kaggle.com/datasets/karansarpal/fremtpl2-french-motor-tpl-insurance-claims. Accessed: 2025-08-08; Includes freMTPL2freq.csv and freMTPL2sev.csv.

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.

S. Siegfried, L. Kook, and T. Hothorn. Distribution-free location-scale regression. *The American Statistician*, 77(4):345–356, 2023. doi: 10.1080/00031305.2023.2203177.

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

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.

P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias, 02 2013.

D.M. Stasinopoulos and R.A. Rigby. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46, 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.

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, pages 9311–9323, 2020. URL https://proceedings.mlr.press/v119/tagasovska20a.html.

K. Uemura, T. Takagi, T. Takayuki, H. Yoshida, and S. Shimizu. A multivariate causal discovery based on post-nonlinear model. In *Proceedings of the Conference on Causal Learning and Reasoning*, pages 826–839. PMLR, 2022.

C. Uhler, G. Raskutti, P. Buhlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013. doi: 10.1214/12-AOS1048.

M. Wang, X. Shen, and W. Pan. Causal discovery with generalized linear models through peeling algorithms. *Journal of Machine Learning Research*, 25(310):1–49, 2024. URL http://jmlr.org/papers/v25/23-1228.html.

S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017.

S. N. Wood, N. Pya, and B. Säfken. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563, 2016. doi: 10.1080/01621459.2016.1180986.

S. Xu, O. A. Mian, A. Marx, and J. Vreeken. Inferring cause and effect in the presence of heteroscedastic noise. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 24615–24630, 2022. URL https://proceedings.mlr.press/v162/xu22f.html.

Y. Yu, X. Zheng, A. Anandkumar, and Y. Yue. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

J. Zhang and P. Spirtes. Intervention, determinism, and the causal minimality condition. *Synthese*, 175(1):39–56, 2010. doi: 10.1007/s11229-010-9751-1.

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. doi: 10.5555/1795114.1795190.

K. Zhang, Z. Wang, J. Zhang, and B. Schölkopf. On estimation of functional causal models: General results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

X. Zheng, B. Aragam, P. Ravikumar, and E. Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.