Journal of Machine Learning Research 15 (2014) 2009-2053 Submitted 7/13; Revised 4/14; Published 6/14

Causal Discovery with Continuous Additive Noise Models

Jonas Peters[*] PETERS@STAT.MATH.ETHZ.CH
Seminar for Statistics, ETH Zürich
Rämistrasse 101, 8092 Zürich
Switzerland

Joris M. Mooij[*] J.M.MOOIJ@UVA.NL
Institute for Informatics, University of Amsterdam
Postbox 94323, 1090 GH Amsterdam
The Netherlands

Institute for Computing and Information Sciences, Radboud University Nijmegen
Postbox 9010, 6500 GL Nijmegen
The Netherlands

Dominik Janzing JANZING@TUEBINGEN.MPG.DE
Bernhard Schölkopf BS@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems
Spemannstraße 38, 72076 Tübingen
Germany

Editor: Aapo Hyvärinen

Abstract

We consider the problem of learning causal directed acyclic graphs from an observational joint distribution. One can use these graphs to predict the outcome of interventional experiments, from which data are often not available. We show that if the observational distribution follows a structural equation model with an additive noise structure, the directed acyclic graph becomes identifiable from the distribution under mild conditions. This constitutes an interesting alternative to traditional methods that assume faithfulness and identify only the Markov equivalence class of the graph, thus leaving some edges undirected. We provide practical algorithms for finitely many samples, RESIT (regression with subsequent independence test) and two methods based on an independence score. We prove that RESIT is correct in the population setting and provide an empirical evaluation.

Keywords: causal inference, structural equation models, additive noise, identifiability, causal minimality, Bayesian networks

1. Introduction

Many scientific questions deal with the causal structure of a data-generating process. E.g., if we know the reasons why an individual is more susceptible to a disease than others, we can hope to develop new drugs in order to cure this disease or prevent its outbreak. Recent results indicate that knowing the causal structure is also useful for classical machine learning tasks. In the two variable case, for example, knowing which is cause and which is effect has

[*] Part of this work was done while JP and JMM were with the MPI Tübingen.

implications for semi-supervised learning and covariate shift adaptation (Schölkopf et al., 2012).

We consider a p-dimensional random vector $\mathbf{X} = (X_1, \dots, X_p)$ with a joint distribution $\mathcal{L}(\mathbf{X})$ and assume that there is a true acyclic causal graph $\mathcal{G}$ that describes the data generating process (see Section 1.3). In this work we address the following problem of causal inference: given the distribution $\mathcal{L}(\mathbf{X})$ we try to infer the graph $\mathcal{G}$. A priori, the causal graph contains information about the physical process that cannot be found in properties of the joint distribution. One therefore requires assumptions connecting these two worlds. While traditional methods like PC, FCI (Spirtes et al., 2000) or score-based approaches (e.g. Chickering, 2002), that are explained in more detail in Section 2, make assumptions that enable us to recover the graph up to the Markov equivalence class, we investigate a different set of assumptions. If the data have been generated by an additive noise model (see Section 3), we will generically be able to recover the correct graph from the joint distribution.

In the remainder of this section we set up the required notation and definitions for graphs (Section 1.1), briefly introduce Judea Pearl's do-notation (Section 1.2) and use it to define our object of interest, a true causal graph (Section 1.3). We introduce structural equation models (SEMs) in Section 1.4. After discussing existing methods in Section 2, we provide the main results of this work in Section 3. We prove that for restricted additive noise models, a special class of SEMs, one can identify the graph from the joint distribution. This is possible not only for additive noise models (ANMs) but for all classes of SEMs that are able to identify graphs from a bivariate distribution, meaning they can distinguish between cause and effect. Section 4 proposes algorithms that can be used in practice, when instead of the joint distribution, we are only given i.i.d. samples. These algorithms are tested in Section 5.

This paper builds on the conference papers of Hoyer et al. (2009), Peters et al. (2011b) and Mooij et al. (2009)[1] but extends the material in several aspects. All deliberations in Section 1.3 about the true causal graph and Example 10 are novel. The presentation of the theoretical results in Section 3 is improved. In particular, we added the motivating Example 26 and Propositions 4 and 29. Example 25 provides a non-identifiable case different from the linear Gaussian example. Proposition 23 is based on Zhang and Hyvärinen (2009) and contains important necessary conditions for the failure of identifiability. In Corollary 31 we present a novel identifiability result for a class of nonlinear functions and Gaussian noise variables. Proposition 17 proves that causal minimality is satisfied if the structural equations do not contain constant functions. Section 3.3 contains results that guarantee to find the set of correct topological orderings when the assumption of causal minimality is dropped. Theorem 34 proves a conjecture from Mooij et al. (2009) by showing that given a regression and independence oracle the algorithm provided in Mooij et al. (2009) is correct. We propose a new score function for estimating the true directed acyclic graph in Section 4.2 and present two corresponding score-based methods. We provide an extended section on simulation experiments and discuss experiments on real data.

[1] Parts of Sections 1 and 2 have been taken and modified from the PhD thesis of Peters (2012).

1.1 Directed Acyclic Graphs

We start with some basic notation for graphs. Consider a finite family of random variables $\mathbf{X} = (X_1, \dots, X_p)$ with index set $\mathbf{V} := \{1, \dots, p\}$ (we use capital letters for random variables and bold letters for sets and vectors). We denote their joint distribution by $\mathcal{L}(\mathbf{X})$. We write $p_{X_1}(x)$ or simply $p(x)$ for the Radon-Nikodym derivative of $\mathcal{L}(X_1)$ either with respect to the Lebesgue or the counting measure and (sometimes implicitly) assume its existence. A graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ consists of nodes $\mathbf{V}$ and edges $\mathcal{E} \subseteq \mathbf{V}^2$ with $(v, v) \notin \mathcal{E}$ for any $v \in \mathbf{V}$. In a slight abuse of notation we identify the nodes (or vertices) $j \in \mathbf{V}$ with the variables X_j, the context should clarify the meaning. We also consider sets of variables $\mathbf{S} \subseteq \mathbf{X}$ as a single multivariate variable. We now introduce graph terminology that we require later. Most of the definitions can be found in Spirtes et al. (2000); Koller and Friedman (2009); Lauritzen (1996), for example.

Let $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ be a graph with $\mathbf{V} := \{1, \dots, p\}$ and corresponding random variables $\mathbf{X} = (X_1, \dots, X_p)$. A graph $\mathcal{G}_1 = (\mathbf{V}_1, \mathcal{E}_1)$ is called a **subgraph** of $\mathcal{G}$ if $\mathbf{V}_1 = \mathbf{V}$ and $\mathcal{E}_1 \subseteq \mathcal{E}$; we then write $\mathcal{G}_1 \leq \mathcal{G}$. If additionally, $\mathcal{E}_1 \neq \mathcal{E}$, we call $\mathcal{G}_1$ a **proper subgraph** of $\mathcal{G}$.

A node i is called a **parent** of j if $(i, j) \in \mathcal{E}$ and a **child** if $(j, i) \in \mathcal{E}$. The set of parents of j is denoted by $\mathbf{PA}_j^{\mathcal{G}}$, the set of its children by $\mathbf{CH}_j^{\mathcal{G}}$. Two nodes i and j are **adjacent** if either $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$. We call $\mathcal{G}$ **fully connected** if all pairs of nodes are adjacent. We say that there is an undirected edge between two adjacent nodes i and j if $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$. An edge between two adjacent nodes is directed if it is not undirected. We then write $i \to j$ for $(i, j) \in \mathcal{E}$. Three nodes are called an **immorality** or a **v-structure** if one node is a child of the two others that themselves are not adjacent. The **skeleton** of $\mathcal{G}$ is the set of all edges without taking the direction into account, that is all (i, j), such that $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$.

A **path** in $\mathcal{G}$ is a sequence of (at least two) distinct vertices $i_1, \dots, i_n$, such that there is an edge between i_k and i_{k+1} for all $k = 1, \dots, n-1$. If $i_k \to i_{k+1}$ for all k we speak of a **directed path** from i_1 to i_n and call i_n a **descendant** of i_1. We denote all descendants of i by $\mathbf{DE}_i^{\mathcal{G}}$ and all non-descendants of i, excluding i, by $\mathbf{ND}_i^{\mathcal{G}}$. In this work, i is neither a descendant nor a non-descendant of itself. If $i_{k-1} \to i_k$ and $i_{k+1} \to i_k$, i_k is called a **collider** on this path. $\mathcal{G}$ is called a **partially directed acyclic graph (PDAG)** if there is no directed cycle, i.e., if there is no pair (j, k) such that there are directed paths from j to k and from k to j. $\mathcal{G}$ is called a **directed acyclic graph (DAG)** if it is a PDAG and all edges are directed.

In a DAG, a path between i_1 and i_n is **blocked by a set S** (with neither i_1 nor i_n in this set) whenever there is a node i_k, such that one of the following two possibilities hold: 1. $i_k \in \mathbf{S}$ and $i_{k-1} \to i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \to i_{k+1}$ Or 2., $i_{k-1} \to i_k \leftarrow i_{k+1}$ and neither i_k nor any of its descendants is in $\mathbf{S}$. We say that two disjoint subsets of vertices $\mathbf{A}$ and $\mathbf{B}$ are ***d*-separated** by a third (also disjoint) subset $\mathbf{S}$ if every path between nodes in $\mathbf{A}$ and $\mathbf{B}$ is blocked by $\mathbf{S}$. Throughout this work, $\perp\!\!\!\perp$ denotes (conditional) independence. The joint distribution $\mathcal{L}(\mathbf{X})$ is said to be **Markov with respect to the DAG $\mathcal{G}$** if

$$\mathbf{A}, \mathbf{B} \; d\text{-sep. by } \mathbf{C} \;\Rightarrow\; \mathbf{A} \perp\!\!\!\perp \mathbf{B} \,|\, \mathbf{C}$$

for all disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$. $\mathcal{L}(\mathbf{X})$ is said to be **faithful to the DAG $\mathcal{G}$** if

$$\mathbf{A}, \mathbf{B} \; d\text{-sep. by } \mathbf{C} \;\Leftarrow\; \mathbf{A} \perp\!\!\!\perp \mathbf{B} \,|\, \mathbf{C}$$

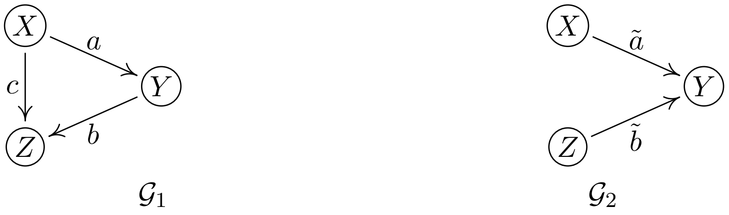


Figure 1: After fine-tuning the parameters for the two graphs, both models generate the same joint distribution.

for all disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$. A distribution satisfies **causal minimality** with respect to $\mathcal{G}$ if it is Markov with respect to $\mathcal{G}$, but not to any proper subgraph of $\mathcal{G}$. We denote by $\mathcal{M}(\mathcal{G})$ the set of distributions that are Markov with respect to $\mathcal{G}$: $\mathcal{M}(\mathcal{G}) := \{\mathcal{L}(\mathbf{X}) : \mathcal{L}(\mathbf{X}) \text{ is Markov w.r.t. } \mathcal{G}\}$. Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are **Markov equivalent** if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$. This is the case if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy the same set of d-separations, that means the Markov condition entails the same set of (conditional) independence conditions. The set of all DAGs that are Markov equivalent to some DAG (a so-called Markov equivalence class) can be represented by a **completed PDAG**. This graph satisfies $(i, j) \in \mathcal{E}$ if and only if one member of the Markov equivalence class does. Verma and Pearl (1991) showed that:

Lemma 1 *Two DAGs are Markov equivalent if and only if they have the same skeleton and the same immoralities.*

Faithfulness is not very intuitive at first glance. We now give an example of a distribution that is Markov but not faithful with respect to some DAG $\mathcal{G}_1$. This is achieved by making two paths cancel each other and creating an independence that is not implied by the graph structure.

Example 2 *Consider the two graphs in Figure 1. Corresponding to the left graph we generate a joint distribution by the following equations.* $X = N_X, Y = aX + N_Y, Z = bY + cX + N_Z$, *with normally distributed noise variables* $N_X \sim \mathcal{N}(0, \sigma_X^2)$, $N_Y \sim \mathcal{N}(0, \sigma_Y^2)$ *and* $N_Z \sim \mathcal{N}(0, \sigma_Z^2)$ *that are jointly independent. This is an example of a linear Gaussian structural equation model with graph* $\mathcal{G}_1$ *that we formally define in Section 1.4. Now, if* $a \cdot b + c = 0$, *the distribution is not faithful with respect to* $\mathcal{G}_1$ *since we obtain* $X \perp\!\!\!\perp Z$; *more precisely, it is not triangle-faithful (Zhang and Spirtes, 2008).*

Correspondingly, we generate a distribution related to graph $\mathcal{G}_2$: $X = \tilde{N}_X, Y = \tilde{a}X + \tilde{b}Z + \tilde{N}_Y, Z = \tilde{N}_Z$, *with all* $\tilde{N}_{\cdot} \sim \mathcal{N}(0, \tau_{\cdot}^2)$ *jointly independent. If we choose* $\tau_X^2 = \sigma_X^2$, $\tilde{a} = a$, $\tau_Z^2 = b^2\sigma_Y^2 + \sigma_Z^2$, $\tilde{b} = (b\sigma_Y^2)/(b^2\sigma_Y^2 + \sigma_Z^2)$ *and* $\tau_Y^2 = \sigma_Y^2 - (b^2\sigma_Y^4)/(b^2\sigma_Y^2 + \sigma_Z^2)$, *both models lead to the covariance matrix*

$$\Sigma = \begin{pmatrix} \sigma_X^2 & a\sigma_X^2 & 0 \\ a\sigma_X^2 & a^2\sigma_X^2 + \sigma_Y^2 & b\sigma_Y^2 \\ 0 & b\sigma_Y^2 & b^2\sigma_Y^2 + \sigma_Z^2 \end{pmatrix}$$

and thus to the same distribution. It can be checked that the distribution is faithful with respect to $\mathcal{G}_2$ *if* $\tilde{a}, \tilde{b} \neq 0$ *and all* $\tilde{\tau}_{\cdot} > 0$.

The distribution from Example 2 is faithful with respect to $\mathcal{G}_2$, but not with respect to $\mathcal{G}_1$. Nevertheless, for both models, causal minimality is satisfied if none of the parameters vanishes: the distribution is not Markov to any proper subgraph of $\mathcal{G}_1$ or $\mathcal{G}_2$ since removing an arrow would correspond to a new (conditional) independence that does not hold in the distribution. Note that $\mathcal{G}_2$ is not a proper subgraph of $\mathcal{G}_1$. In general, causal minimality is weaker than faithfulness:

Remark 3 *If $\mathcal{L}(\mathbf{X})$ is faithful with respect to $\mathcal{G}$, then causal minimality is satisfied.*

This is due to the fact that any two nodes that are not directly connected by an edge can be *d*-separated. Another, equivalent formulation of causal minimality reads as follows:

Proposition 4 *Consider the random vector $\mathbf{X} = (X_1, \ldots, X_p)$ and assume that the joint distribution has a density with respect to a product measure. Suppose that $\mathcal{L}(\mathbf{X})$ is Markov with respect to $\mathcal{G}$. Then $\mathcal{L}(\mathbf{X})$ satisfies causal minimality with respect to $\mathcal{G}$ if and only if $\forall X_j \; \forall Y \in \mathbf{PA}_j^{\mathcal{G}}$ we have that $X_j \not\perp\!\!\!\perp Y \,|\, \mathbf{PA}_j^{\mathcal{G}} \setminus \{Y\}$.*

Proof See Appendix A.1. ■

1.2 Intervention Distributions[2]

Given a directed acyclic graph (DAG) $\mathcal{G}$, Pearl (2009) introduces the *do*-notation as a mathematical description of interventional experiments. More precisely, $do(X_j = \tilde{p}(x_j))$ stands for setting the variable X_j randomly according to the distribution $\tilde{p}(x_j)$, irrespective of its parents, while not interfering with any other variable. Formally:

Definition 5 *Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a collection of variables with joint distribution $\mathcal{L}(\mathbf{X})$ that we assume to be absolutely continuous with respect to the Lebesgue measure or the counting measure (i.e., there exists a probability density function or a probability mass function). Given a DAG $\mathcal{G}$ over $\mathbf{X}$, we define the intervention distribution $do(X_j = \tilde{p}(x_j))$ of $X_1, \ldots, X_p$ by*

$$p\big(x_1, \ldots, x_p \,|\, do(X_j = \tilde{p}(x_j))\big) := \prod_{i \neq j}^{p} p(x_i | x_{\mathbf{PA}_i}) \cdot \tilde{p}(x_j)$$

if $p(x_1, \ldots, x_p) > 0$ and zero otherwise. Here $\tilde{p}(x_j)$ is either a probability density function or a probability mass function. Similarly, we can intervene at different nodes at the same time by defining the intervention distribution $do(X_j = \tilde{p}(x_j), j \in \mathbf{J})$ for $\mathbf{J} \subseteq \mathbf{V}$ as

$$p\big(x_1, \ldots, x_p \,|\, do(X_j = \tilde{p}(x_j), j \in \mathbf{J})\big) := \prod_{i \notin \mathbf{J}} p(x_i | x_{\mathbf{PA}_i}) \cdot \prod_{j \in \mathbf{J}} \tilde{p}(x_j)$$

if $p(x_1, \ldots, x_p) > 0$ and zero otherwise.

[2] Sections 1.2 and 1.3 are not essential for understanding the rest of the paper and can be skipped on first reading.

Here, $x_{\mathbf{PA}_i}$ denotes the tuple of all x_j for X_j being a parent of X_i in $\mathcal{G}$. Pearl (2009) introduces Definition 5 with the special case of $\tilde{p}(x_j) = \delta_{x_j,\tilde{x}_j}$, where $\delta_{x_j,\tilde{x}_j} = 1$ if $x_j = \tilde{x}_j$ and $\delta_{x_j,\tilde{x}_j} = 0$ otherwise; this corresponds to a point mass at $\tilde{x}_j$. For more details on *soft* interventions, see Eberhardt and Scheines (2007). Note that in general:

$$p(x_1, \ldots, x_p \,|\, do(X_j = \tilde{x}_j)) \neq p(x_1, \ldots, x_p \,|\, X_j = \tilde{x}_j)\,.$$

The expression $p\big(x_1, \ldots, x_p \,|\, do(X_j = \tilde{x}_j, j \in \mathbf{J})\big)$ yields a distribution over $X_1, \ldots, X_p$. If we are only interested in computing the marginal $p\big(x_i \,|\, do(X_j = \tilde{x}_j)\big)$, where X_i is not a parent of X_j, we can use the parent adjustment formula (Pearl, 2009, Theorem 3.2.2)

$$p(x_i \,|\, do(X_j = \tilde{x}_j)) = \sum_{x_{\mathbf{PA}_j}} p(x_i \,|\, \tilde{x}_j, x_{\mathbf{PA}_j})\, p(x_{\mathbf{PA}_j})\,. \tag{1}$$

1.3 True Causal Graphs[2]

In this section we clarify what we mean by a true causal graph $\mathcal{G}_c$. In short, we use this term if the results of randomized studies are determined by $\mathcal{G}_c$ and the observational joint distribution. This means that the graph and the observational joint distribution lead to causal effects that one observes in practice. Two important restrictive assumptions that we make throughout this work are *acyclicity* (the absence of directed cycles, in other words, no causal feedback loops are allowed) and *causal sufficiency* (the absence of hidden variables that are a common cause of at least two observed variables).

Definition 6 *Assume we are given a distribution $\mathcal{L}(\mathbf{X})$ over $X_1, \ldots, X_p$ and distributions $\mathcal{L}_{do(X_j=\tilde{p}(x_j), j\in\mathbf{J})}(\mathbf{X})$ for all $\mathbf{J} \subseteq \mathbf{V} = \{1, \ldots, p\}$ (think of the variables X_j having been randomized). We then call the graph $\mathcal{G}_c$ a* true causal graph *for these distributions if*

- *$\mathcal{G}_c$ is a directed acyclic graph;*
- *the distribution $\mathcal{L}(\mathbf{X})$ is Markov with respect to $\mathcal{G}_c$;*
- *for all $\mathbf{J} \subseteq \mathbf{V}$ and $\tilde{p}(x_j)$ with $j \in \mathbf{J}$ the distribution $\mathcal{L}_{do(X_j=\tilde{p}(x_j), j\in\mathbf{J})}(\mathbf{X})$ coincides with $p\big(x_1, \ldots, x_p \,|\, do(X_j = \tilde{p}(x_j), j \in \mathbf{J})\big)$, computed from $\mathcal{G}_c$ as in Definition 5.*

Definition 6 is purely mathematical if one considers $\mathcal{L}_{do(X_j=\tilde{p}(x_j), j\in\mathbf{J})}(\mathbf{X})$ as an abstract family of given distributions. But it is a small step to make the relation to the "real world". We call $\mathcal{G}_c$ the true causal graph *of a data generating process* if it is the true causal graph for the distributions $\mathcal{L}(\mathbf{X})$ and $\mathcal{L}_{do(X_j=\tilde{p}(x_j), j\in\mathbf{J})}(\mathbf{X})$, where the latter are obtained by randomizing X_j according to $\tilde{p}(x_j)$. In some situations, the precise design of a randomized experiment may not be obvious. While most people would agree on how to randomize over medical treatment procedures, there is probably less agreement how to randomize over the tolerance of a person (does this include other changes of his personality, too?). Only sometimes, this problem can be resolved by including more variables and taking a less coarse-grained point of view. We do not go into further detail since we believe that this would require philosophical deliberations, which lie beyond the scope of this work. Instead, we may explicitly add the requirement that "most people agree on what a randomized experiment should look like in this context".

In general, there can be more than one true causal DAG. If one requires causal minimality, the true causal DAG is unique.

Proposition 7 *Assume $\mathcal{L}(X_1, \ldots, X_p)$ has a density and consider all true causal DAGs $\mathbb{G} := \{\mathcal{G}_{c,1}, \ldots, \mathcal{G}_{c,m}\}$ of $X_1, \ldots, X_p$. Then there is a partial order on $\mathbb{G}$ using the subgraph property $\leq$ as an ordering. This ordering has a least element $\mathcal{G}_c$, i.e., $\mathcal{G}_c \leq \mathcal{G}_{c,i}$ for all i. This element $\mathcal{G}_c$ is the unique true causal DAG such that $\mathcal{L}(\mathbf{X})$ satisfies causal minimality with respect to $\mathcal{G}_c$.*

Proof See Appendix A.2 ■

We now briefly comment on a true causal graph's behavior when some of the variables from the joint distribution are marginalized out.

Example 8 *(i) If $X \leftarrow Z \rightarrow Y$ is the only true causal graph for X, Y and Z, there is no true causal graph for the variables X and Y (the do-statements do not coincide).*

(ii) Assume that the graph $X \rightarrow Y \rightarrow Z$ with additional $X \rightarrow Z$ is the only true causal graph for X, Y and Z and assume that $\mathcal{L}(X, Y, Z)$ is faithful with respect to this graph. Then, the only true causal graph for the variables X and Z is $X \rightarrow Z$.

(iii) If the situation is the same as in (ii) with the difference that $X \perp\!\!\!\perp Z$ (i.e., $\mathcal{L}(X, Y, Z)$ is not faithful with respect to the true causal graph), the empty graph and $Z \leftarrow X$ are also true causal graphs for X and Z.

Latent projections (Verma and Pearl, 1991) provide a formal way to obtain a true causal graph for marginalization. Cases (ii) and (iii) show that there are no purely graphical criteria that provide the *minimal* true causal graph described in Proposition 7.

The results presented in the remainder of this paper can be understood without causal interpretation. Using these techniques to infer a true causal graph, however, requires the assumption that such a true causal DAG $\mathcal{G}_c$ for the observed distribution of $X_1, \ldots, X_p$ exists. This includes the assumption that all "relevant" variables have been observed, sometimes called causal sufficiency, and that there are no feedback loops.

Richardson and Spirtes (2002) introduce a representation of graphs (so-called Maximal Ancestral Graphs, or MAGs) with hidden variables that is closed under marginalization and conditioning. The FCI algorithm (Spirtes et al., 2000) exploits the conditional independences in the data to partially reconstruct the graph. Other work concentrates on hidden variables in structural equation models (e.g., Hoyer et al., 2008; Janzing et al., 2009; Silva and Ghahramani, 2009).

1.4 Structural Equation Models

A structural equation model (SEM) (also called a functional model) is defined as a tuple $(\mathcal{S}, \mathcal{L}(\mathbf{N}))$, where $\mathcal{S} = (S_1, \ldots, S_p)$ is a collection of p equations

$$S_j: \quad X_j = f_j(\mathbf{PA}_j, N_j), \qquad j = 1, \ldots, p \tag{2}$$

and $\mathcal{L}(\mathbf{N}) = \mathcal{L}(N_1, \ldots, N_p)$ is the joint distribution of the noise variables, which we require to be jointly independent, i.e., $\mathcal{L}(\mathbf{N})$ is a product distribution. We consider SEMs only for real-valued random variables $X_1, \ldots, X_p$. The graph of a structural equation model is obtained simply by drawing direct edges from each parent to its direct effects, i.e., from each variable X_k occurring on the right-hand side of equation (2) to X_j. We henceforth assume this graph to be acyclic. According to the notation defined in Section 1.1, $\mathbf{PA}_j$ are the parents of X_j.

The $\mathbf{PA}_j$ can be considered as the direct causes of X_j. An SEM specifies how the $\mathbf{PA}_j$ affect X_j. Note that in physics (chemistry, biology, ...), we would usually expect that such causal relationships occur in time, and are governed by sets of coupled differential equations. Under certain assumptions such as stable equilibria, one can derive an SEM that describes how the equilibrium states of such a dynamical system will react to physical interventions on the observables involved (Mooij et al., 2013). We do not deal with these issues in the present paper but take the SEM as our starting point instead. We formulate the identifiability results without the notion of causality.

Pearl (2009) shows in Theorem 1.4.1 that the law $\mathcal{L}(\mathbf{X})$ generated by an SEM is Markov with respect to its graph. Reversely, there always exists an SEM that models a given distribution.[3]

Proposition 9 *Consider $X_1, \dots, X_p$ and let $\mathcal{L}(\mathbf{X})$ have a strictly positive density with respect to the Lebesgue measure and assume it is Markov with respect to $\mathcal{G}$. Then there exists an SEM $(\mathcal{S}, \mathcal{L}(\mathbf{N}))$ with graph $\mathcal{G}$ that generates the distribution $\mathcal{L}(\mathbf{X})$.*

Proof See Appendix A.3. ■

Structural equation models contain strictly more information than their corresponding graph and law and hence also more information than the family of all intervention distributions together with the observational distribution. This information sometimes helps to answer counterfactual questions, as shown in the following example.

Example 10 *Let $N_1, N_2 \sim Ber(0.5)$ and $N_3 \sim U(\{0,1,2\})$, such that the three variables are jointly independent. That is, N_1, N_2 have a Bernoulli distribution with parameter 0.5 and N_3 is uniformly distributed on $\{0,1,2\}$. We define two different SEMs, first consider:*

$$\mathcal{S}_A = \begin{cases} X_1 = N_1 \\ X_2 = N_2 \\ X_3 = (1_{N_3>0} \cdot X_1 + 1_{N_3=0} \cdot X_2) \cdot 1_{X_1 \neq X_2} + N_3 \cdot 1_{X_1 = X_2} \, . \end{cases}$$

If X_1 and X_2 have different values, depending on N_3 we either choose $X_3 = X_1$ or $X_3 = X_2$. Otherwise $X_3 = N_3$. Now, $\mathcal{S}_B$ differs from $\mathcal{S}_A$ only in the latter case:

$$\mathcal{S}_B = \begin{cases} X_1 = N_1 \\ X_2 = N_2 \\ X_3 = (1_{N_3>0} \cdot X_1 + 1_{N_3=0} \cdot X_2) \cdot 1_{X_1 \neq X_2} + (2 - N_3) \cdot 1_{X_1 = X_2} \, . \end{cases}$$

It can be checked that both SEMs generate the same observational distribution, which satisfies causal minimality with respect to the graph $X_1 \to X_3 \leftarrow X_2$. They also generate the same intervention distributions, for any possible intervention. But the two models differ in a counterfactual statement.[4] Suppose, we have seen a sample $(X_1, X_2, X_3) = (1, 0, 0)$ and

[3] A similar but weaker statement than Proposition 9 can be found in Druzdzel and van Leijen (2001); Janzing and Schölkopf (2010).

[4] Here, we make use of Judea Pearl's definition of counterfactuals (Pearl, 2009).

we are interested in the counterfactual question, what X_3 would have been if X_1 had been 0. From both $\mathcal{S}_A$ and $\mathcal{S}_B$ it follows that $N_3 = 0$, and thus the two SEMs "predict" different values for X_3 under a counterfactual change of X_1.

If we want to use an estimated SEM to predict counterfactual questions, this example shows that we require assumptions that let us distinguish between $\mathcal{S}_A$ or $\mathcal{S}_B$. In this work we exploit the additive noise assumption to infer the *structure* of an SEM. We do not claim that we can predict counterfactual statements.

Structural equation models have been used for a long time in fields like agriculture or social sciences (e.g., Wright, 1921; Bollen, 1989). Model selection, for example, was done by fitting different structures that were considered as reasonable given the prior knowledge about the system. These candidate structures were then compared using goodness of fit tests. In this work we instead consider the question of identifiability, which has not been addressed until more recently.

Problem 11 (population case) *We are given a distribution $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \ldots, X_p)$ that has been generated by an (unknown) structural equation model with graph $\mathcal{G}_0$; in particular, $\mathcal{L}(\mathbf{X})$ is Markov with respect to $\mathcal{G}_0$. Can the (observational) distribution $\mathcal{L}(\mathbf{X})$ be generated by a structural equation model with a different graph $\mathcal{G} \neq \mathcal{G}_0$? If not, we call $\mathcal{G}_0$ identifiable from $\mathcal{L}(\mathbf{X})$.*

In general, $\mathcal{G}_0$ is not identifiable from $\mathcal{L}(\mathbf{X})$: the joint distribution $\mathcal{L}(\mathbf{X})$ is certainly Markov with respect to a lot of different graphs, e.g., to all fully connected acyclic graphs. Proposition 9 states the existence of corresponding SEMs. What can be done to overcome this indeterminacy? The hope is that by using additional assumptions one obtains restricted models, in which we can identify the graph from the joint distribution. Considering graphical models, we see in Section 2.1 how the assumption that $\mathcal{L}(\mathbf{X})$ is Markov and faithful with respect to $\mathcal{G}_0$ leads to identifiability of the Markov equivalence class of $\mathcal{G}_0$. Considering SEMs, we see in Section 3 that additive noise models as a special case of restricted SEMs even lead to identifiability of the correct DAG. Also Section 2.3 contains such a restriction based on SEMs.

2. Alternative Methods

We briefly describe some existing methods and provide references for more details.

2.1 Estimating the Markov Equivalence Class: Independence-Based Methods

Conditional independence-based methods like the PC algorithm and the FCI algorithm (Spirtes et al., 2000) assume that $\mathcal{L}(\mathbf{X})$ is Markov and faithful with respect to the correct graph $\mathcal{G}_0$ (that means *all* conditional independences in the joint distribution are entailed by the Markov condition, cf. Section 1.1). Since both assumptions put restrictions only on the conditional independences in the joint distribution, these methods are not able to distinguish between two graphs that entail exactly the same set of (conditional) independences, i.e., between Markov equivalent graphs. Since many Markov equivalence classes contain more than one graph, conditional independence-based methods thus usually leave some arrows undirected and cannot uniquely identify the correct graph.

The first step of the PC algorithm determines the variables that are adjacent. One therefore has to test whether two variables are dependent given *any* other subset of variables. The PC algorithm exploits a very clever procedure to reduce the size of the condition set. In the worst case, however, one has to perform conditional independence tests with conditioning sets of up to $p-2$ variables (where p is the number of variables in the graph). Although there is recent work on kernel-based conditional independence tests (Fukumizu et al., 2008; Zhang et al., 2011), such tests are difficult to perform in practice if one does not restrict the variables to follow a Gaussian distribution, for example (e.g., Bergsma, 2004).

To prove consistency of the PC algorithm one does not only require faithfulness, but strong faithfulness (Zhang and Spirtes, 2003; Kalisch and Bühlmann, 2007). Uhler et al. (2013) argue that this is a restrictive condition. Since parts of faithfulness can be tested given the data (Zhang and Spirtes, 2008), the condition may be weakened.

From our perspective independence-based methods face the following challenges: (1) We can identify the correct DAG only up to Markov equivalence classes. (2) Conditional independence testing, especially with a large conditioning set, is difficult in practice. (3) Simulation experiments suggest, that in many cases, the distribution is close to unfaithfulness. In these cases there is no guarantee that the inferred graph(s) will be close to the original one.

2.2 Estimating the Markov Equivalence Class: Score-Based Methods

Although the roots for score-based methods for causal inference may date back even further, we mainly refer to Geiger and Heckerman (1994), Heckerman (1997) and Chickering (2002) and references therein. Given the data $\mathcal{D}$ from a vector $\mathbf{X}$ of variables, i.e., n i.i.d. samples, the idea is to assign a score $S(\mathcal{D},\mathcal{G})$ to each graph $\mathcal{G}$ and search over the space of DAGs for the best scoring graph:

$$\hat{\mathcal{G}} := \underset{\mathcal{G} \text{ DAG over } \mathbf{X}}{\operatorname{argmax}} S(\mathcal{D},\mathcal{G})\,. \tag{3}$$

There are several possibilities to define such a scoring function. Often a parametric model is assumed (e.g., linear Gaussian equations or multinomial distributions), which introduces a set of parameters $\theta \in \mathbf{\Theta}$.

From a Bayesian point of view, we may define priors $p_{pr}(\mathcal{G})$ and $p_{pr}(\theta)$ over DAGs and parameters and consider the log posterior as a score function, or equivalently (note that $p(\mathcal{D})$ is constant over all DAGs):

$$S(\mathcal{D},\mathcal{G}) := \log p_{pr}(\mathcal{G}) + \log p(\mathcal{D}|\mathcal{G})\,,$$

where $p(\mathcal{D}|\mathcal{G})$ is the marginal likelihood

$$p(\mathcal{D}|\mathcal{G}) = \int_{\mathbf{\Theta}} p(\mathcal{D}|\mathcal{G},\theta)\cdot p_{pr}(\theta)\,d\theta.$$

In this case, $\hat{\mathcal{G}}$ defined in (3) is the mode of the posterior distribution, which is usually called the *maximum a posteriori* (or MAP) estimator. Instead of a MAP estimator, one may be interested in the full posterior distribution over DAGs. This distribution can subsequently be averaged over all graphs to get a posterior of the hypothesis about the existence of a specific edge, for example.

In the case of parametric models, we call two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ *distribution equivalent* if for each parameter $\theta_1 \in \mathbf{\Theta}_1$ there is a corresponding parameter $\theta_2 \in \mathbf{\Theta_2}$, such that the distribution obtained from $\mathcal{G}_1$ in combination with θ_1 is the same as the distribution obtained from graph $\mathcal{G}_2$ with θ_2, and vice versa. It is known that in the linear Gaussian case (or for unconstrained multinomial distributions) two graphs are distribution-equivalent if and only if they are Markov equivalent. One may therefore argue that $p(\mathcal{D}|\mathcal{G}_1)$ and $p(\mathcal{D}|\mathcal{G}_2)$ should be the same for Markov equivalent graphs $\mathcal{G}_1$ and $\mathcal{G}_2$. Heckerman and Geiger (1995) discuss how to choose the prior over parameters accordingly.

Instead, we may consider the maximum likelihood estimator $\hat{\theta}$ in each graph and define a score function by using a penalty, e.g., the Bayesian Information Criterion (BIC):

$$S(\mathcal{D},\mathcal{G}) = \log p(\mathcal{D}|\hat{\theta},\mathcal{G}) - \frac{d}{2}\log n\,,$$

where n is the sample size and d the dimensionality of the parameter θ.

Since the search space of all DAGs is growing super-exponentially in the number of variables (e.g., Chickering, 2002), greedy search algorithms are applied to solve equation (3): at each step there is a candidate graph and a set of neighboring graphs. For all these neighbors one computes the score and considers the best-scoring graph as the new candidate. If none of the neighbors obtains a better score, the search procedure terminates (not knowing whether one obtained only a local optimum). Clearly, one therefore has to define a neighborhood relation. Starting from a graph $\mathcal{G}$, we may define all graphs as neighbors from $\mathcal{G}$ that can be obtained by removing, adding or reversing one edge. In the linear Gaussian case, for example, one cannot distinguish between Markov equivalent graphs. It turns out that in those cases it is beneficial to change the search space to Markov equivalence classes instead of DAGs. The greedy equivalence search (GES) (Meek, 1997; Chickering, 2002) starts with the empty graph and consists of two-phases. In the first phase, edges are added until a local maximum is reached; in the second phase, edges are removed until a local maximum is reached, which is then given as an output of the algorithm. Chickering (2002) proves consistency of this method by using consistency of the BIC (Haughton, 1988).

2.3 Estimating the DAG: LiNGAM

Kano and Shimizu (2003) and Shimizu et al. (2006) propose an inspiring method exploiting non-Gaussianity of the data.[5] Although their work covers the general case, the idea is maybe best understood in the case of two variables:

Example 12 *Suppose*

$$Y = \phi X + N, \quad N \perp\!\!\!\perp X\,,$$

where X and N are normally distributed. It is easy to check that

$$X = \tilde{\phi} Y + \tilde{N}, \quad \tilde{N} \perp\!\!\!\perp Y\,.$$

with $\tilde{\phi} = \frac{\phi \mathbf{var}(X)}{\phi^2 \mathbf{var}(X) + \sigma^2} \neq \frac{1}{\phi}$ and $\tilde{N} = X - \tilde{\phi} Y$.

[5] A more detailed tutorial can be found on `http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/papers/Shimizu13BHMK.pdf`.

If we consider non-Gaussian noise, however, the structural equation model becomes identifiable.

Proposition 13 *Let X and Y be two random variables, for which*

$$Y = \phi X + N, \quad N \perp\!\!\!\perp X,\ \phi \neq 0$$

holds. Then we can reverse the process, i.e., there exists $\psi \in \mathbb{R}$ and a noise $\tilde{N}$, such that

$$X = \psi Y + \tilde{N}, \quad \tilde{N} \perp\!\!\!\perp Y\,,$$

if and only if X and N are Gaussian distributed.

Shimizu et al. (2006) were the first to report this result. They prove it even for more than two variables using Independent Component Analysis (ICA) (Comon, 1994, Theorem 11), which itself is proved using the Darmois-Skitovič theorem (Skitovič, 1954, 1962; Darmois, 1953). Alternatively, Proposition 13 can be proved directly using the Darmois-Skitovič theorem (e.g., Peters, 2008, Theorem 2.10).

Theorem 14 (Shimizu et al., 2006) *Assume a linear SEM with graph $\mathcal{G}_0$*

$$X_j = \sum_{k \in \mathbf{PA}_j^{\mathcal{G}_0}} \beta_{jk} X_k + N_j\,, \qquad j = 1, \ldots, p\,, \tag{4}$$

where all N_j are jointly independent and non-Gaussian distributed. Additionally, for each $j \in \{1, \ldots, p\}$ we require $\beta_{jk} \neq 0$ for all $k \in \mathbf{PA}_j^{\mathcal{G}_0}$. Then, the graph $\mathcal{G}_0$ is identifiable from the joint distribution.

The authors call this model a linear non-Gaussian acyclic model (LiNGAM) and provide a practical method based on ICA that can be applied to a finite amount of data. Later, improved versions of this method have been proposed in Shimizu et al. (2011); Hyvärinen and Smith (2013).

2.4 Estimating the DAG: Gaussian SEMs with Equal Error Variances

There is another deviation from linear Gaussian SEMs that makes the graph identifiable. Peters and Bühlmann (2014) show that restricting the noise variables to have the same variance is sufficient to recover the graph structure.

Theorem 15 (Peters and Bühlmann, 2014) *Assume an SEM with graph $\mathcal{G}_0$*

$$X_j = \sum_{k \in \mathbf{PA}_j^{\mathcal{G}_0}} \beta_{jk} X_k + N_j\,, \qquad j = 1, \ldots, p\,, \tag{5}$$

where all N_j are i.i.d. and follow a Gaussian distribution. Additionally, for each $j \in \{1, \ldots, p\}$ we require $\beta_{jk} \neq 0$ for all $k \in \mathbf{PA}_j^{\mathcal{G}_0}$. Then, the graph $\mathcal{G}_0$ is identifiable from the joint distribution.

For estimating the coefficients β_{jk} and the error variance σ^2, Peters and Bühlmann (2014) propose to use a penalized maximum likelihood method (BIC). For optimization they propose a greedy search algorithm in the space of DAGs. Rescaling the variables changes the variance of the error terms. Therefore, in many applications model (5) cannot be sensibly applied. The BIC criterion, however, always allows us to compare the method's score with the score of a linear Gaussian SEM that uses more parameters and does not make the assumption of equal error variances.

3. Identifiability of Continuous Additive Noise Models

Recall that equation (2) defines the general form of an SEM: $X_j = f_j(\mathbf{PA}_j, N_j)\,, j = 1, \ldots, p$ with jointly independent variables N_i. We have seen that these models are too general to identify the graph (Proposition 9). It turns out, however, that constraining the function class leads to identifiability. As a first step we restrict the form of the function to be additive with respect to the noise variable. (Throughout this section we assume that all random variables are absolutely continuous with respect to the Lebesgue measure. Peters et al. (2011a) provide an extension for variables that are absolutely continuous with respect to the counting measure.)

Definition 16 *We define a continuous additive noise model (ANM) as a tuple $(\mathcal{S}, \mathcal{L}(\mathbf{N}))$, where $\mathcal{S} = (S_1, \ldots, S_p)$ is a collection of p equations*

$$S_j: \quad X_j = f_j(\mathbf{PA}_j) + N_j\,, \qquad j = 1, \ldots, p\,, \tag{6}$$

where the $\mathbf{PA}_j$ correspond to the direct parents of X_j, and the noise variables N_j have a strictly positive density (with respect to the Lebesgue measure) and are jointly independent. Furthermore, we assume that the corresponding graph is acyclic.

For these models causal minimality reduces to the condition that each function f_j is not constant in any of its arguments:

Proposition 17 *Consider a distribution generated by a model (6) and assume that the functions f_j are not constant in any of its arguments, i.e., for all j and $i \in \mathbf{PA}_j$ there are some $x_{\mathbf{PA}_j \setminus \{i\}}$ and some $x_i \neq x_i'$ such that*

$$f_j(x_{\mathbf{PA}_j \setminus \{i\}}, x_i) \neq f_j(x_{\mathbf{PA}_j \setminus \{i\}}, x_i')\,.$$

Then the joint distribution satisfies causal minimality with respect to the corresponding graph. Conversely, if there is a j and i such that $f_j(x_{\mathbf{PA}_j \setminus \{i\}}, \cdot)$ is constant, causal minimality is violated.

Proof See Appendix A.4 ■

Linear functions and Gaussian variables identify only the correct Markov equivalence class and not necessarily the correct graph. In the remainder of this section we establish results showing that this is an exceptional case. We develop conditions that guarantee the identifiability of the DAG. Proposition 21 indicates that this condition is rather weak.

3.1 Bivariate Additive Noise Models

We now add another assumption about the form of the structural equations.

Definition 18 *Consider an additive noise model* (6) *with two variables, i.e., the two equations $X_i = N_i$ and $X_j = f_j(X_i) + N_j$ with $\{i,j\} = \{1,2\}$. We call this SEM an* identifiable bivariate additive noise model *if the triple $(f_j, \mathcal{L}(X_i), \mathcal{L}(N_j))$ satisfies Condition 19.*

Condition 19 *The triple $(f_j, \mathcal{L}(X_i), \mathcal{L}(N_j))$ does not solve the following differential equation for all x_i, x_j with $\nu''(x_j - f(x_i))f'(x_i) \neq 0$:*

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'}\,, \tag{7}$$

Here, $f := f_j$, and $\xi := \log p_{X_i}$ and $\nu := \log p_{N_j}$ are the logarithms of the strictly positive densities. To improve readability, we have skipped the arguments $x_j - f(x_i)$, x_i, and x_i for ν, ξ, and f and their derivatives, respectively.

Zhang and Hyvärinen (2009) even allow for a bijective transformation of the data, i.e., $X_j = g_j(f_j(X_i) + N_j)$ and obtain a similar differential equation as (7). As the name in Definition 18 already suggests, we have identifiability for this class of SEMs.

Theorem 20 *Let $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, X_2)$ be generated by an identifiable bivariate additive noise model with graph $\mathcal{G}_0$ and assume causal minimality, i.e., a non-constant function f_j (Proposition 17). Then, $\mathcal{G}_0$ is identifiable from the joint distribution.*

Proof The proof of Hoyer et al. (2009) is reproduced in Appendix A.5. ■

Intuitively speaking, we expect a "generic" triple $(f_j, \mathcal{L}(X_i), \mathcal{L}(N_j))$ to satisfy Condition 19. The following proposition presents one possible formalization. After fixing $(f_j, \mathcal{L}(N_j))$ we consider the space of all distributions $\mathcal{L}(X_i)$ such that Condition 19 is violated. This space is contained in a three dimensional space. Since the space of continuous distributions is infinite dimensional, we can therefore say that Condition 19 is satisfied for "most distributions" $\mathcal{L}(X_i)$.

Proposition 21 *If for a fixed pair $(f_j, \mathcal{L}(N_j))$ there exists $x_j \in \mathbb{R}$ such that $\nu''(x_j - f(x_i))f'(x_i) \neq 0$ for all but a countable set of points $x_i \in \mathbb{R}$, the set of all $\mathcal{L}(X_i)$ for which $(f_j, \mathcal{L}(X_i), \mathcal{L}(N_j))$ does not satisfy Condition 19 is contained in a* 3*-dimensional space.*

Proof See Appendix A.6. ■

The condition $\nu''(x_j - f(x_i))f'(x_i) \neq 0$ holds for all x_i if there is no interval where f is constant and the logarithm of the noise density is not linear, for example. In the case of Gaussian variables, the differential equation (7) simplifies. We thus have the following result.

Corollary 22 *If X_i and N_j follow a Gaussian distribution and $(f_j, \mathcal{L}(X_i), \mathcal{L}(N_j))$ does not satisfy Condition 19, then f_j is linear.*

Proof See Appendix A.7. ■

Although non-identifiable cases are rare, the question remains when identifiability is violated. Zhang and Hyvärinen (2009) prove that non-identifiable additive noise models necessarily fall into one out of five classes.

Proposition 23 (Zhang and Hyvärinen, 2009) *Consider $X_j = f_j(X_i) + N_j$ with fully supported noise variable N_j that is independent of X_i and three times differentiable function f_j. Let further $\frac{d}{dx_i} f_j(x_i) \frac{d^2}{dx_j^2} \log p_{N_j}(x_j) = 0$ only at finitely many points (x_i, x_j). If there is a backward model, i.e., we can write $X_i = g_i(X_j) + M_i$ with M_i independent of X_j, then one of the following must hold.*

I. X_i is Gaussian, N_j is Gaussian and f_j is linear.

II. X_i is log-mix-lin-exp, N_j is log-mix-lin-exp and f_j is linear.

III. X_i is log-mix-lin-exp, N_j is one-sided asymptotically exponential and f_j is strictly monotonic with $f_j'(x_i) \to 0$ as $x_i \to \infty$ or as $x_i \to -\infty$.

IV. X_i is log-mix-lin-exp, N_j is generalized mixture of two exponentials and f_j is strictly monotonic with $f_j'(x_i) \to 0$ as $x_i \to \infty$ or as $x_i \to -\infty$.

V. X_i is generalized mixture of two exponentials, N_j is two-sided asymptotically exponential and f_j is strictly monotonic with $f_j'(x_i) \to 0$ as $x_i \to \infty$ or as $x_i \to -\infty$.

Precise definitions can be found in Appendix A.8. In particular, we obtain identifiability whenever the function f_j is not injective. Proposition 23 states that belonging to one of these classes is a necessary condition for non-identifiability. We now show sufficiency for two classes. The linear Gaussian case is well-known and easy to prove.

Example 24 *Let $X_2 = aX_1 + N_2$ with independent $N_2 \sim \mathcal{N}(0, \sigma^2)$ and $X_1 \sim \mathcal{N}(0, \tau^2)$. We can then consider all variables in $\mathcal{L}_2$ and project X_1 onto X_2. This leads to an orthogonal decomposition $X_1 = \tilde{a} X_2 + \tilde{N}_1$. Since for jointly Gaussian variables uncorrelatedness implies independence, we obtain a backward additive noise model. Figure 2 (left) shows the joint density and the functions for the forward and backward model.*

We also give an example of a nonidentifiable additive noise model with non-Gaussian distributions; the forward model is described by case II, and the backward model by case IV:

Example 25 *Let $X_2 = aX_1 + b + N_2$ with independent log-mix-lin-exp N_2 and X_1, i.e., we have the log-densities*

$$\xi(x) = \log p_{X_1}(x) = c_1 \exp(c_2 x) + c_3 x + c_4$$

and

$$\nu(n) = \log p_{N_2}(n) = \gamma_1 \exp(\gamma_2 n) + \gamma_3 n + \gamma_4 \, .$$

Then X_2 is a generalized mixture of exponential distributions. If and only if $c_2 = -a\gamma_2$ and $c_3 \neq a\gamma_3$ we obtain a valid backward model $X_1 = \tilde{f}_1(X_2) + \tilde{N}_1$ with log-mix-lin-exp $\tilde{N}_1$. Again, Figure 2 (right) shows the joint distribution over X_1 and X_2 and forward and backward functions.

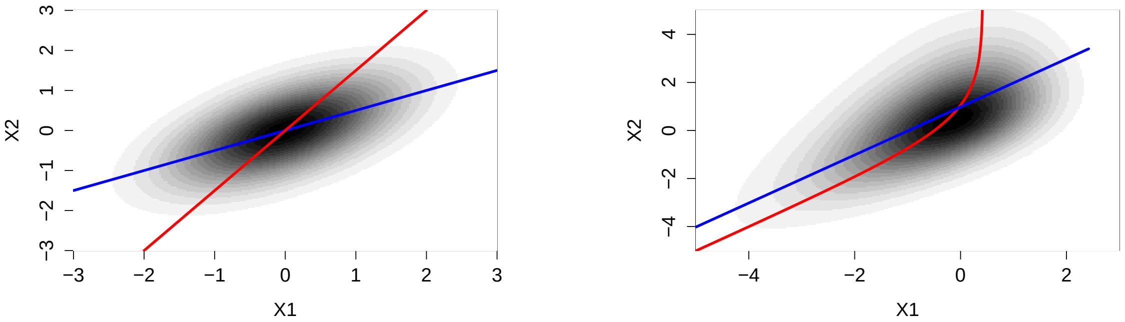


Figure 2: Joint density over X_1 and X_2 for two non-identifiable examples. The left panel shows Example 24 (linear Gaussian case) and the right panel shows Example 25 (the latter plot is based on kernel density estimation). The blue function corresponds to the forward model $X_2 = f_2(X_1) + N_2$, the red function to the backward model $X_1 = \tilde{f}_1(X_2) + \tilde{N}_1$.

Proof See Appendix A.9. ■

Example 25 shows how parameters of function, input and noise distribution have to be "fine-tuned" to yield non-identifiability (Janzing and Steudel, 2010).

It can be shown that bivariate identifiability even holds generically when feedback is allowed (i.e., if both $X \to Y$ *and* $Y \to X$), at least when assuming noise and input distributions to be Gaussian (Mooij et al., 2011).

3.2 From Bivariate to Multivariate Models

It turns out that Condition 19 also suffices to prove identifiability in the multivariate case. Assume we are given p structural equations $X_j = f_j(\mathbf{PA}_j) + N_j$ as in (6). If we fix all arguments of the functions f_j except for one parent and the noise variable, we obtain a bivariate model. One may expect that it suffices to put restrictions like Condition 19 on this triple of function, input and noise distribution. This is not the case.

Example 26 *Consider the following SEM*

$$X_1 = N_1, \quad X_2 = f_2(X_1) + N_2, \quad X_3 = f_3(X_1) + a \cdot X_2 + N_3$$

with $N_1 \sim t_{\nu=3}$, $N_2 \sim \mathcal{N}(0, \sigma_2^2)$ and $N_3 \sim \mathcal{N}(0, \sigma_3^2)$, i.e., N_1 is t-distributed with 3 degrees of freedom and N_2 and N_3 are normally distributed. X_2 and X_3 are non-Gaussian but

$$X_3 \,|\, _{X_1 = x_1} = c + a \cdot X_2 \,|\, _{X_1 = x_1} + N_3$$

is a linear Gaussian equation for all x_1. We can revert this equation and obtain the same joint distribution by an SEM of the form

$$X_1 = M_1, \quad X_2 = g_2(X_1) + b \cdot X_3 + M_2, \quad X_3 = g_3(X_1) + M_3\,,$$

for some g_1, g_2 and $M_1 \sim t_{\nu=3}$, $M_2 \sim \mathcal{N}(0, \tilde{\sigma}_2^2)$ and $M_3 \sim \mathcal{N}(0, \tilde{\sigma}_3^2)$. Thus, the DAG is not identifiable from the joint distribution.

Instead, we need to put restrictions on conditional distributions.

Definition 27 *Consider an additive noise model (6) with p variables. We call this SEM a* restricted additive noise model *if for all $j \in \mathbf{V}$, $i \in \mathbf{PA}_j$ and all sets $\mathbf{S} \subseteq \mathbf{V}$ with $\mathbf{PA}_j \setminus \{i\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_j \setminus \{i, j\}$, there is an $x_\mathbf{S}$ with $p_\mathbf{S}(x_\mathbf{S}) > 0$, s.t.*

$$\Big(f_j(x_{\mathbf{PA}_j \setminus \{i\}}, \underbrace{\cdot}_{X_i}), \mathcal{L}(X_i \,|\, X_\mathbf{S} = x_\mathbf{S}), \mathcal{L}(N_j) \Big)$$

satisfies Condition 19. Here, the underbrace indicates the input component of f_j for variable X_i. In particular, we require the noise variables to have non-vanishing densities and the functions f_j to be continuous and three times continuously differentiable.

Assuming causal minimality, we can identify the structure of the SEM from the distribution.

Theorem 28 *Let $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \ldots, X_p)$ be generated by a restricted additive noise model with graph $\mathcal{G}_0$ and let $\mathcal{L}(\mathbf{X})$ satisfy causal minimality with respect to $\mathcal{G}_0$, i.e., the functions f_j are not constant (Proposition 17). Then, $\mathcal{G}_0$ is identifiable from the joint distribution.*

Proof See Appendix A.11. ■

Our proof of Theorem 28 contains a graphical statement that turns out to be a main argument for proving identifiability for Gaussian models with equal error variances (Peters and Bühlmann, 2014). We thus state it explicitly as a proposition.

Proposition 29 *Let $\mathcal{G}$ and $\mathcal{G}'$ be two different DAGs over variables $\mathbf{X}$.*

(i) Assume that $\mathcal{L}(\mathbf{X})$ has a strictly positive density and satisfies the Markov condition and causal minimality with respect to $\mathcal{G}$ and $\mathcal{G}'$. Then there are variables $L, Y \in \mathbf{X}$ such that for the sets $\mathbf{Q} := \mathbf{PA}_L^\mathcal{G} \setminus \{Y\}$, $\mathbf{R} := \mathbf{PA}_Y^{\mathcal{G}'} \setminus \{L\}$ and $\mathbf{S} := \mathbf{Q} \cup \mathbf{R}$ we have

- *$Y \to L$ in $\mathcal{G}$ and $L \to Y$ in $\mathcal{G}'$*
- *$\mathbf{S} \subseteq \mathbf{ND}_L^\mathcal{G} \setminus \{Y\}$ and $\mathbf{S} \subseteq \mathbf{ND}_Y^{\mathcal{G}'} \setminus \{L\}$*

(ii) In particular, if $\mathcal{L}(\mathbf{X})$ is Markov and faithful with respect to $\mathcal{G}$ and $\mathcal{G}'$ (i.e., both graphs belong to the same Markov equivalence class), there are variables L, Y such that

- *$Y \to L$ in $\mathcal{G}$ and $L \to Y$ in $\mathcal{G}'$*
- *$\mathbf{PA}_L^\mathcal{G} \setminus \{Y\} = \mathbf{PA}_Y^{\mathcal{G}'} \setminus \{L\}$*

Proof See Appendix A.12. ■

If the distribution is Markov and faithful with respect to the underlying graph it is known that we can recover the correct Markov equivalence class. Chickering (1995) proves that two graphs within this Markov equivalence class can be transformed into each other by a sequence of so-called covered edge reversals. This result implies part (ii) of the proposition. Part (i) establishes a similar statement when replacing faithfulness by causal minimality.

Although Theorem 28 is stated for additive noise models, it can be seen as an example of a more general principle.

Remark 30 *Theorem 28 is not limited to restricted additive noise models. Whenever we have a restriction like Condition 19 that ensures identifiability in the bivariate case (Theorem 20), the multivariate version (Theorem 28) remains valid. The proof we provide in the appendix stays exactly the same. The algorithms in Section 4, however, use standard regression methods and therefore rely on the additive noise assumption.*

The result can therefore be used to prove identifiability of SEMs that are restricted to discrete additive noise models (Peters et al., 2011a) or post-nonlinear additive noise models (Zhang and Hyvärinen, 2009). In the latter model class we allow a bijective nonlinear distortion: $X_j = g_j\big(f_j(\mathbf{PA}_j) + N_j\big)$. These models allow for more complicated functional relationships but are harder to fit from empirical data than the additive noise models considered in this work.

We now state a specific identifiability result for Gaussian noise that we believe to constitute an important model class for applications. Tamada et al. (2011b) have already used this result for structure learning without giving an identifiability result (see also Tamada et al., 2011a). More recently, Bühlmann et al. (2013) investigate model (8) in a high-dimensional context. A bivariate version of the following corollary can be found as Lemma 6 in Zhang and Hyvärinen (2009).

Corollary 31 *(i) Let $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \dots, X_p)$ be generated by an SEM with*

$$X_j = f_j(\mathbf{PA}_j) + N_j\,,$$

with normally distributed noise variables $N_j \sim \mathcal{N}(0, \sigma_j^2)$ and three times differentiable functions f_j that are not linear in any component: denote the parents $\mathbf{PA}_j$ of X_j by $X_{k_1}, \dots, X_{k_\ell}$, then the function $f_j(x_{k_1}, \dots, x_{k_{a-1}}, \cdot, x_{k_{a+1}}, \dots, x_{k_\ell})$ is assumed to be nonlinear for all a and some $x_{k_1}, \dots, x_{k_{a-1}}, x_{k_{a+1}}, \dots, x_{k_\ell} \in \mathbb{R}^{\ell-1}$.

(ii) As a special case, let $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \dots, X_p)$ be generated by an SEM with

$$X_j = \sum_{k \in \mathbf{PA}_j} f_{j,k}(X_k) + N_j\,, \tag{8}$$

with normally distributed noise variables $N_j \sim \mathcal{N}(0, \sigma_j^2)$ and three times differentiable, nonlinear functions $f_{j,k}$.

In both cases (i) and (ii), we can identify the corresponding graph $\mathcal{G}_0$ from the distribution $\mathcal{L}(\mathbf{X})$. The statements remain true if the noise distributions for source nodes, i.e., nodes with no parents, are allowed to have a non-Gaussian density with full support on the real line $\mathbb{R}$ (the proof remains identical).

Proof See Appendix A.13. ■

Additive noise models as in (8), for which the structural equations are additive in the parents (but the noise does not need to be Gaussian) are called causal additive models (CAMs), see Bühlmann et al. (2013).

Theorem 28 requires the positivity of densities in order to make use of the intersection property of conditional independence. Peters (2014) shows that the intersection property still holds under weaker assumptions and discusses fundamental limits of causal inference when positivity is violated.

3.3 Estimating the Topological Order

We now investigate the case when we drop the assumption of causal minimality. Assume therefore that we are given a distribution $\mathcal{L}(\mathbf{X})$ from an additive noise model with graph $\mathcal{G}_0$. We cannot recover the correct graph $\mathcal{G}_0$ because we can always add edges $i \rightarrow j$ or remove edges that "do not have any effect" without changing the distribution. This is formalized by the following lemma. (This lemma may be useful in more general contexts, other than additive noise models, too.)

Lemma 32 *Let $\mathcal{L}(\mathbf{X})$ be generated by an additive noise model with graph $\mathcal{G}_0$.*

(a) For each supergraph $\mathcal{G} \geq \mathcal{G}_0$ there is an additive noise model that leads to the distribution $\mathcal{L}(\mathbf{X})$.

(b) For each subgraph $\mathcal{G} \leq \mathcal{G}_0$ such that $\mathcal{L}(\mathbf{X})$ is Markov with respect to $\mathcal{G}$ there is an additive noise model that leads to the distribution $\mathcal{L}(\mathbf{X})$. Furthermore, there is an additive noise model with unique graph $\mathcal{G}_0^{min} \leq \mathcal{G}_0$ that leads to $\mathcal{L}(\mathbf{X})$ and satisfies causal minimality.

Proof See Appendix A.14. ■

Despite this indeterminacy we can still recover the correct order of the variables. Given a permutation $\pi \in S_p$ on $\{1, \ldots, p\}$ we therefore define the fully connected DAG $\mathcal{G}_\pi^{\text{full}}$ by the DAG that contains all edges $\pi(i) \rightarrow \pi(j)$ for $i < j$.

As a direct consequence of Theorem 28 and Lemma 32 we have the following result.

Corollary 33 *Let $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \ldots, X_p)$ be generated by an additive noise model with graph $\mathcal{G}_0$. Assume that the SEM corresponding to the minimal graph $\mathcal{G}_0^{min}$ defined as in Lemma 32 (b) is a restricted additive noise model. Consider an ordering π and a restricted ANM with corresponding graph $\mathcal{G}_\pi^{full,\, min}$ (Lemma 32 (b)) that generates the distribution $\mathcal{L}(\mathbf{X})$. Theorem 28 implies that $\mathcal{G}_\pi^{full,\, min} = \mathcal{G}_0^{min}$. In this sense the set of true orderings*

$$\Pi^0 := \{\pi \in S_p \mid \mathcal{G}_\pi^{full} \geq \mathcal{G}_0^{min}\}$$

is identifiable from $\mathcal{L}(\mathbf{X})$.

This result is useful, for example, if the search over structures is performed in the space of permutations rather than in the space of DAGs (e.g. Friedman and Koller, 2003; Teyssier and Koller, 2005; Bühlmann et al., 2013).

4. Algorithms

The theoretical results do not imply an algorithm for finitely many data that is either computationally or statistically efficient. In this section we propose an algorithm called RESIT that is based on independence-tests and two simple algorithms that make use of an independence score. We prove correctness of RESIT in the population case.

4.1 Regression with Subsequent Independence Test (RESIT)

In practice, we are given i.i.d. data from the joint distribution and try to estimate the corresponding DAG. The following method is based on the fact that for each node X_i the corresponding noise variable N_i is independent of all non-descendants of X_i. In particular, for each sink node X_i we have that N_i is independent of $\mathbf{X} \setminus \{X_i\}$. We therefore propose an iterative procedure: in each step we identify and disregard a sink node. This is done by regressing each of the remaining variables on all other remaining variables and measuring the independence between the residuals and those other variables. The variable leading to the least dependent residuals is considered the sink node (Algorithm 1, lines $4-13$). This first phase of the procedure yields a topological ordering or a fully connected DAG (see Section 3.3). In the second phase we visit every node and eliminate incoming edges until the residuals are not independent anymore, see Algorithm 1, lines $15-22$.

The procedure can make use of any regression method and dependence measure, in this work we choose the p-value of the HSIC independence test (Gretton et al., 2008) as a dependence measure. Under independence, Gretton et al. (2008) provide an asymptotically correct null distribution for the test statistic times sample size. (We use moment matching to approximate this distribution by a gamma distribution.) Since under dependence the test statistic is guaranteed to converge to a value different from zero, we know that the p-value converges to zero only for dependence. As a regression method we choose linear regression, gam regression (R package `mgcv`) or Gaussian process regression (R package `gptk`).

Algorithm 1 is a slightly modified version of the one proposed in Mooij et al. (2009). In this work, we always want to obtain a graph estimate; we thus consider the node with the least dependent residuals as being the sink node, instead of stopping the search when no independence hypothesis is accepted as in Mooij et al. (2009).

Given that we have infinite data, a consistent non-parametric regression method and a perfect independence test ("independence oracle"), RESIT is correct.

Theorem 34 *Assume $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \ldots, X_p)$ is generated by a restricted additive noise model with graph $\mathcal{G}_0$ and assume that $\mathcal{L}(\mathbf{X})$ satisfies causal minimality with respect to $\mathcal{G}_0$. Then, RESIT used with a consistent non-parametric regression method and an independence oracle is guaranteed to find the correct graph $\mathcal{G}_0$ from the joint distribution $\mathcal{L}(\mathbf{X})$.*

Proof See Appendix A.15 ■

RESIT performs $\mathcal{O}(p^2)$ independence tests, which is polynomial in the number of nodes. In phase 2 of the algorithm, the removal of superfluous edges costs $\mathcal{O}(p)$. Both the independence test and the variable selection method may scale with the sample size, of course. RESIT's polynomial behavior in p may come as a surprise since problems in Bayesian network learning are often NP-hard (e.g. Chickering, 1996).

Despite this theoretical guarantee, RESIT does not scale well to a high number of nodes. Since we cannot make use of an independence oracle in practice, we have to detect dependence between a random variable and a random vector from finitely many data. The order in which the independence tests are performed (phase 2, line 16 in Algorithm 1) may lead to different results. Also, type one errors in phase 2 lead to superfluous edges in the output of the method. If the significance level of the independence test is independent of the number of variables (in the experiments we choose 5%), this effect may lead to a

Algorithm 1 Regression with subsequent independence test (RESIT)

```
**Input:** I.i.d. samples of a $p$-dimensional distribution on $(X_1, \dots, X_p)$
$S := \{1, \dots, p\}, \pi := [\ ]$
PHASE 1: Determine topological order.
**repeat**
    **for** $k \in S$ **do**
        Regress $X_k$ on $\{X_i\}_{i \in S \setminus \{k\}}$.
        Measure dependence between residuals and $\{X_i\}_{i \in S \setminus \{k\}}$.
    **end for**
    Let $k^*$ be the $k$ with the weakest dependence.
    $S := S \setminus \{k^*\}$
    $\mathrm{pa}(k^*) := S$
    $\pi := [k^*, \pi]$        ($\pi$ will be the topological order, its last component being a sink)
**until** $\#S = 0$
PHASE 2: Remove superfluous edges.
**for** $k \in \{2, \dots, p\}$ **do**
    **for** $\ell \in \mathrm{pa}(\pi(k))$ **do**
        Regress $X_{\pi(k)}$ on $\{X_i\}_{i \in \mathrm{pa}(\pi(k)) \setminus \{\ell\}}$.
        **if** residuals are independent of $\{X_i\}_{i \in \{\pi(1), \dots, \pi(k-1)\}}$ **then**
            $\mathrm{pa}(\pi(k)) := \mathrm{pa}(\pi(k)) \setminus \{\ell\}$
        **end if**
    **end for**
**end for**
**Output:** $(\mathrm{pa}(1), \dots, \mathrm{pa}(p))$
```

high structural Hamming distance between true and estimated graph for a large number of variables. Furthermore, we have to perform nonparametric regression with possibly many covariates. For high dimensions, these are both statistically hard problems that require huge sample sizes.

4.2 Independence-Based Score

Searching for sink nodes makes the method described in Section 4.1 inherently asymmetric. Mistakes made in the first iterations propagate through the whole procedure. We therefore investigate the performance of independence-based score methods. Theorem 28 ensures that if the data come from a restricted additive noise model we can fit only one (minimal) structure to the data. In order to estimate the graph structure we can test all possible DAGs and determine which DAG yields the most independent residuals. But even in the limit of infinitely many data we may find more than one DAG satisfying this constraint, some of which may not satisfy causal minimality. We therefore propose to take a penalized independence score

$$\hat{\mathcal{G}} = \operatorname*{argmin}_{\mathcal{G}} \sum_{i=1}^{p} \mathrm{DM}(\mathrm{res}_i^{\mathcal{G},\mathrm{RM}}, \mathrm{res}_{-i}^{\mathcal{G},\mathrm{RM}}) + \lambda \, \#(\mathrm{edges})\,. \tag{9}$$

Here, res_i are the residuals of node X_i, when regressing it on its parents; they depend on the graph $\mathcal{G}$ and on the regression method RM. We denote the residuals of all variables except for X_i by res_{-i} and DM denotes a measure of dependence. Note that variables $\mathbf{N} = (N_1, \ldots, N_p)$ are jointly independent if and only if each N_i is independent of $\mathbf{N} \setminus \{N_i\}$, $i = 1, \ldots, p$. We do not prove (or claim) that the minimizer of (9) is a consistent estimator for the correct DAG; we expect this to depend on the choice of DM and RM and λ.

As dependence measure we use minus the logarithm of the p-values of an independence test based on the Hilbert Schmidt Independence Criterion HSIC (Gretton et al., 2008). As regression methods we use linear regression, generalized additive models (gam) or Gaussian process regression. For the regularization parameter λ we propose to use $\log(0.05) - \log(0.01)$. This is a heuristic choice based on the following idea: we only allow for an additional edge if it allows the p-value to increase from 0.01 to 0.05 or, equivalently, by a factor of five. In practice, p-values estimated by bootstrap techniques or p-values that are smaller than computer precision can become zero and the logarithm becomes minus infinity. We therefore always consider the maximum of the computed p-value and 10^{-350}. Although our choices seem to work well in practice, we do not claim that they are optimal.

4.2.1 Brute-Force

For small graphs, we can solve equation (9) by computing the score for all possible DAGs and choose the DAG with the lowest score. Since the number of DAGs grows hyper-exponentially in the number of nodes, this method becomes quickly computationally intractable; e.g., for $p = 7$, there are $1,138,779,265$ DAGs (OEIS Foundation Inc., 2011). Nevertheless, we use this algorithm up to $p = 4$ for comparison.

4.2.2 Greedy DAG Search (GDS)

A strategy to circumvent the computational complexity of equation (9) is to use greedy search algorithms (e.g., Chickering, 2002). At each step we are given a current DAG and score neighboring DAGs that are arranged in some order (see below). Here, all DAGs are called neighbors that can be reached by an edge reversal, addition or removal. Whenever a DAG has a better score than the current DAG, we stop scoring other neighbors and exchange the latter by the former. To obtain "better" steps, in each step we consider at least p neighbors. In order to reduce the running time of the algorithm, we do not score neighboring DAGs in a completely random order but start by adding or removing edges into nodes whose residuals are highly dependent on the other residuals instead. More precisely, we are randomly sorting the nodes, choosing each node one by one with a probability proportional to the reciprocal dependence measure of its residuals. If all neighboring DAGs have a worse score than the current graph G, we nevertheless consider the best neighbor H. If H has a neighbor with a better score than G, we continue with this graph. Otherwise we stop and output G as the optimal graph. This is a simple version of tabu search (e.g. Koller and Friedman, 2009) that is used to avoid local optima. This method is not guaranteed to find the best scoring graph.

Code for the proposed methods is provided on the first and second author's homepage.

5. Experiments

The following subsections report some empirical performance of the described methods.

5.1 Experiments on Synthetic Data

For varying sample size n and number of variables p we compare the described methods. Given a value of p, we randomly choose an ordering of the variables with respect to the uniform distribution and include each of the $p(p-1)/2$ possible edges with a probability of $2/(p-1)$. This results in an expected number of p edges and can be considered as a (modestly) sparse setting. For a linear and a nonlinear setting we report the average structural Hamming distance (Acid and de Campos, 2003; Tsamardinos et al., 2006) to the true directed acyclic graph and to the true completed partially directed acyclic graph over 100 simulations. The structural Hamming distance (SHD) between two partially directed acyclic graphs counts how many edge types do not coincide. Estimating a non-edge or a directed edge instead of an undirected edge, for example, contributes an error of one to the overall distance. We also report analogous results for the structural intervention distance (SID), which has recently been proposed (Peters and Bühlmann, 2013). Given the estimated graph we can infer the intervention distribution $p(X_j \mid do(X_i = x_i))$ by parent adjustment (1). We call a pair of nodes (X_i, X_j) *good* if the intervention distribution $p(X_j \mid do(X_i = x_i))$ inferred from the estimated DAG using (1) coincides with the intervention distribution inferred from the correct DAG for all observational distributions $\mathcal{L}(\mathbf{X})$. The SID counts the number of pairs that are not good. Some methods output a Markov equivalence class instead of a single DAG. Different DAGs within such a class lead to different intervention distribution and thus different SIDs. In that case, we therefore provide the smallest and largest SID attained by members within the Markov equivalence class. As the SHD, the SID is a purely structural measure that is independent of any distribution. The rationale behind the new measure is that a reversed edge in the estimated DAG leads to more false causal effects than an additional edge does. The SHD, however, weights both errors equally.

We compare the greedy DAG search (GDS), brute-force (BF), regression with subsequent independence test (RESIT), linear non-Gaussian additive models (LINGAM), the PC algorithm (PC) with partial correlation and significance level 0.01 and greedy equivalence search (GES), see Sections 4.2.2, 4.2.1, 4.1, 2.3, 2.1 and 2.2, respectively. We also compare them with the conservative PC algorithm (CPC), suggested by Ramsey et al. (2006), and random guessing (RAND). The latter chooses a random DAG with edge inclusion probability uniformly chosen between zero and one. Its estimate does not depend on the data.

5.1.1 Linear Structural Equation Models

We first consider a linear setting as in equation (4), where the coefficients β_{jk} are uniformly chosen from $[-2, -0.1] \cup [0.1, 2]$ and the noise variables N_j are independent and distributed according to $K_j \cdot \mathrm{sign}(M_j) \cdot |M_j|^{\alpha_j}$ with $M_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $K_j \overset{\text{iid}}{\sim} \mathcal{U}([0.1, 0.5])$ and $\alpha_j \overset{\text{iid}}{\sim} \mathcal{U}([2, 4])$. The top box plot in Figure 3 compares the SHD of the estimated structure to the correct DAG for $p = 4$ and $n = 100$. All methods make use of the linear structure of the data, e.g., by performing linear regression. The brute-force method performs best, which indicates that the score function in equation (9) is a sensible choice for small graphs. Greedy DAG search performs almost equally well, it does not encounter many local optima in this setting. The constraint-based methods and greedy equivalent search perform worse. Comparing SID leads to the same conclusion (Figure 3, bottom).

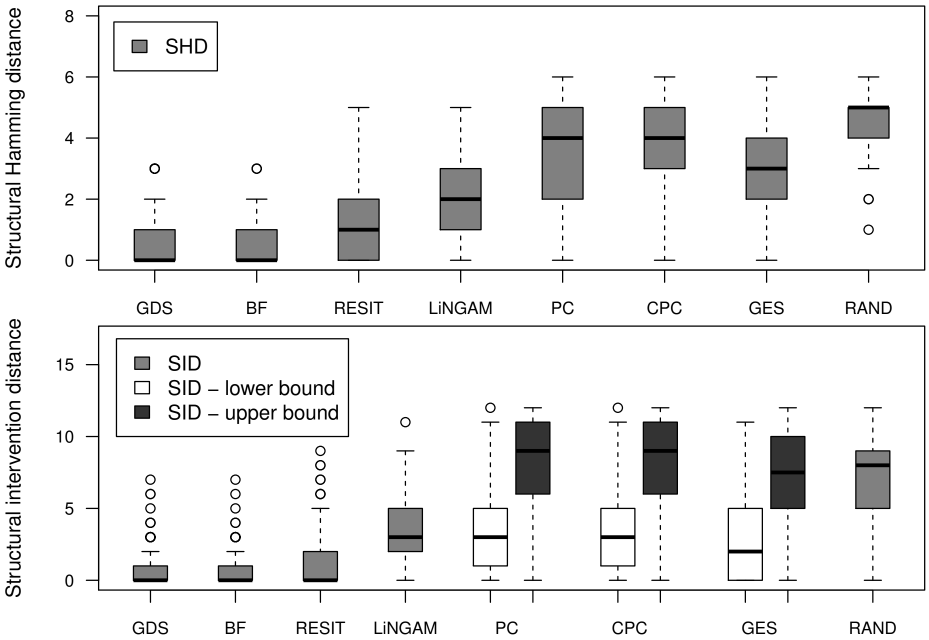


Figure 3: Box plots of the SHD between the estimated structure (either DAG or CPDAG) and the correct DAG for $p = 4$ and $n = 100$ for linear non-Gaussian SEMs (top). The SID is computed between the correct DAG and the estimated DAG (bottom). Some methods estimate only the Markov equivalence class. We then compute the SID to the "best" and to the "worst" DAG within the equivalence class; therefore a lower and an upper bound are shown.

Tables 1 and 2 provide summaries for $p \in \{4, 15\}$ and $n \in \{100, 500\}$. We additionally show distances of the estimated CPDAGs to the true CPDAGs. Therefore, if methods output a DAG instead of a CPDAG, this DAG is transformed into the CPDAG of the corresponding Markov equivalence class. For $p = 4$ and $n = 500$, GDS and brute force find almost always the correct graph (86 and 90 out of 100). RESIT and LiNGAM still perform much better than the PC methods and GES. For $p = 15$, the performance of RESIT (and GES) in relation to the other methods seems to be better when evaluating SID compared to evaluating the SHD. This indicates that the pruning (and penalization of the number of edges) does not work perfectly. Especially for RESIT with high sample size and fixed significance level, making mistakes in phase 1 leads to many edges that cannot be removed in phase 2 (and thus a large SHD). The brute-force method is not applicable to $p = 15$.

Table 1: Linear SEMs: SHD between the estimated structure and the correct DAG and SHD between the estimated CPDAG to the correct CPDAG; for both the average and the standard deviation over 100 experiments are shown (best averages are highlighted).

	GDS	BF	RESIT	LiNGAM	PC	CPC	GES	RAND
	$p = 4, n = 100$							
DAG	0.7 ± 0.9	0.6 ± 0.8	1.2 ± 1.3	1.9 ± 1.2	3.5 ± 1.5	3.6 ± 1.4	3.1 ± 1.7	4.4 ± 1.0
CPDAG	1.1 ± 1.5	0.9 ± 1.4	1.5 ± 1.7	2.4 ± 1.5	2.4 ± 1.7	2.3 ± 1.6	2.0 ± 2.0	4.3 ± 1.4
	$p = 4, n = 500$							
DAG	0.2 ± 0.6	0.1 ± 0.3	0.6 ± 0.8	0.5 ± 0.8	3.1 ± 1.4	3.2 ± 1.4	2.9 ± 1.6	4.1 ± 1.2
CPDAG	0.3 ± 0.9	0.2 ± 0.5	0.9 ± 1.3	0.8 ± 1.2	1.9 ± 1.8	1.6 ± 1.7	1.6 ± 1.9	3.9 ± 1.4
	$p = 15, n = 100$							
DAG	12.2 ± 5.3	–	25.2 ± 8.3	11.1 ± 3.7	13.0 ± 3.6	13.7 ± 3.7	12.7 ± 4.2	57.4 ± 26.4
CPDAG	13.2 ± 5.4	–	27.0 ± 8.5	12.4 ± 3.9	10.7 ± 3.5	10.8 ± 3.8	12.4 ± 4.9	58.5 ± 27.1
	$p = 15, n = 500$							
DAG	6.1 ± 6.4	–	51.2 ± 17.8	3.4 ± 2.8	10.2 ± 3.8	10.8 ± 4.2	8.7 ± 4.6	57.6 ± 24.2
CPDAG	6.8 ± 6.9	–	54.5 ± 18.5	4.5 ± 3.8	8.2 ± 4.6	7.5 ± 4.4	7.1 ± 5.6	58.9 ± 25.0

Table 2: Linear SEMs: SID to the correct DAG; the table shows average and standard deviation over 100 experiments.

GDS	BF	RESIT	LiNGAM	PC	CPC	GES	RAND
$p = 4, n = 100$							
1.0 ± 1.5	0.8 ± 1.4	1.5 ± 2.2	3.3 ± 2.1	3.4 ± 2.9 8.0 ± 3.2	3.2 ± 2.7 8.5 ± 3.2	2.9 ± 3.3 7.2 ± 3.5	7.0 ± 2.8
$p = 4, n = 500$							
0.2 ± 0.7	0.1 ± 0.4	0.3 ± 1.0	0.9 ± 1.4	2.8 ± 3.1 7.4 ± 3.4	2.3 ± 2.7 7.6 ± 3.3	2.1 ± 2.9 6.9 ± 3.6	6.3 ± 2.8
$p = 15, n = 100$							
32.3 ± 24.1	–	35.3 ± 21.2	45.1 ± 24.1	36.5 ± 21.3 63.7 ± 30.3	32.5 ± 20.2 66.4 ± 31.5	26.5 ± 18.3 37.6 ± 20.6	55.6 ± 27.1
$p = 15, n = 500$							
12.6 ± 16.3	–	18.1 ± 13.8	14.2 ± 14.6	33.6 ± 29.5 55.0 ± 32.9	23.2 ± 19.8 55.6 ± 32.4	18.1 ± 21.4 31.6 ± 22.2	57.5 ± 34.1

5.1.2 Nonlinear Structural Equation Models

We also sample data from nonlinear SEMs. We choose an additive structure as in equation (8) and sample the functions from a Gaussian process with bandwidth one. The noise variables N_j are independent and normally distributed with a uniformly chosen variance. Tables 3 and 4 show summaries for $p \in \{4, 15\}$ and $n \in \{100, 500\}$. We cannot run the brute-force method on data sets with $p = 15$. For $p = 4$, we have a similar situation as in Figure 3 with GDS and the BF method outperforming all others (RESIT performing a bit worse). Remarkably, for $p = 15$ and $n = 100$, a lot of the methods do not perform much better than random guessing when comparing the SID. The estimated CPDAG of the constraint-based methods can have very different lower and upper bounds for SID. This

Table 3: Nonlinear SEMs: SHD between the estimated structure and the correct DAG and SHD between the estimated CPDAG to the correct CPDAG; for both the average and the standard deviation over 100 experiments are shown.

	GDS	BF	RESIT	LiNGAM	PC	CPC	GES	RAND
	$p = 4, n = 100$							
DAG	1.5 ± 1.4	1.0 ± 1.0	1.7 ± 1.3	3.5 ± 1.2	3.5 ± 1.5	3.8 ± 1.4	3.5 ± 1.3	4.0 ± 1.3
CPDAG	1.7 ± 1.7	1.2 ± 1.4	2.0 ± 1.6	3.0 ± 1.4	2.9 ± 1.5	2.7 ± 1.4	3.4 ± 1.7	3.9 ± 1.4
	$p = 4, n = 500$							
DAG	0.5 ± 0.9	0.3 ± 0.5	0.8 ± 0.9	3.7 ± 1.2	3.5 ± 1.5	3.8 ± 1.5	3.3 ± 1.5	4.1 ± 1.2
CPDAG	0.6 ± 1.1	0.6 ± 1.0	1.0 ± 1.3	3.0 ± 1.7	3.1 ± 1.9	2.8 ± 1.8	3.4 ± 1.9	3.8 ± 1.6
	$p = 15, n = 100$							
DAG	14.3 ± 4.9	–	15.4 ± 5.7	15.4 ± 3.6	14.2 ± 3.5	15.5 ± 3.6	24.8 ± 6.3	56.8 ± 24.1
CPDAG	15.1 ± 5.4	–	16.5 ± 5.9	15.3 ± 4.0	13.3 ± 3.6	13.3 ± 4.0	26.4 ± 6.5	58.0 ± 24.7
	$p = 15, n = 500$							
DAG	13.0 ± 8.4	–	10.1 ± 5.7	21.4 ± 6.9	13.9 ± 4.5	15.1 ± 4.8	26.8 ± 8.5	56.1 ± 26.8
CPDAG	14.2 ± 9.2	–	11.3 ± 6.3	21.1 ± 7.3	13.7 ± 4.9	13.4 ± 5.1	28.6 ± 8.8	57.0 ± 27.3

Table 4: Nonlinear SEMs: SID to the correct DAG; the table shows average and standard deviation over 100 experiments.

GDS	BF	RESIT	LiNGAM	PC	CPC	GES	RAND
$p = 4, n = 100$							
2.0 ± 2.5	1.4 ± 1.7	2.0 ± 1.9	8.2 ± 2.8	4.7 ± 3.2 7.8 ± 3.4	4.3 ± 2.7 8.5 ± 3.2	4.7 ± 3.2 7.2 ± 3.2	6.3 ± 3.1
$p = 4, n = 500$							
0.6 ± 1.8	0.2 ± 0.8	0.9 ± 1.3	8.0 ± 2.8	4.3 ± 3.7 7.3 ± 3.2	3.7 ± 3.3 8.1 ± 3.2	3.6 ± 3.0 6.5 ± 3.3	6.6 ± 3.4
$p = 15, n = 100$							
50.6 ± 25.3	–	44.4 ± 23.9	65.0 ± 28.3	49.7 ± 24.6 68.6 ± 31.5	40.4 ± 21.6 76.7 ± 32.8	49.0 ± 27.3 53.6 ± 28.9	60.0 ± 29.9
$p = 15, n = 500$							
35.9 ± 26.8	–	24.6 ± 18.6	67.3 ± 28.1	49.9 ± 29.0 60.3 ± 31.0	36.4 ± 22.1 70.3 ± 34.6	40.2 ± 23.3 44.6 ± 24.0	58.9 ± 27.8

means that some DAGs within the equivalence class perform much better than others. (The methods do not propose any particular DAG, they treat all DAGs within the class equally.)

Figure 4 shows box plots of SHD and SID for the special case $p = 15$ and $n = 500$. This time, RESIT perform slightly better than all other methods. It makes use of the nonlinearity of the structural equations. Again, the high SHD for GES indicates that the estimate probably contains too many edges (since its SID is better than the one for the PC methods).

In conclusion, for $p = 4$, the brute force method works best for both linear and nonlinear data. Roughly speaking, for $p = 15$, LiNGAM and GDS work best in the linear non-Gaussian setting and RESIT works best for nonlinear data. If one does not know whether

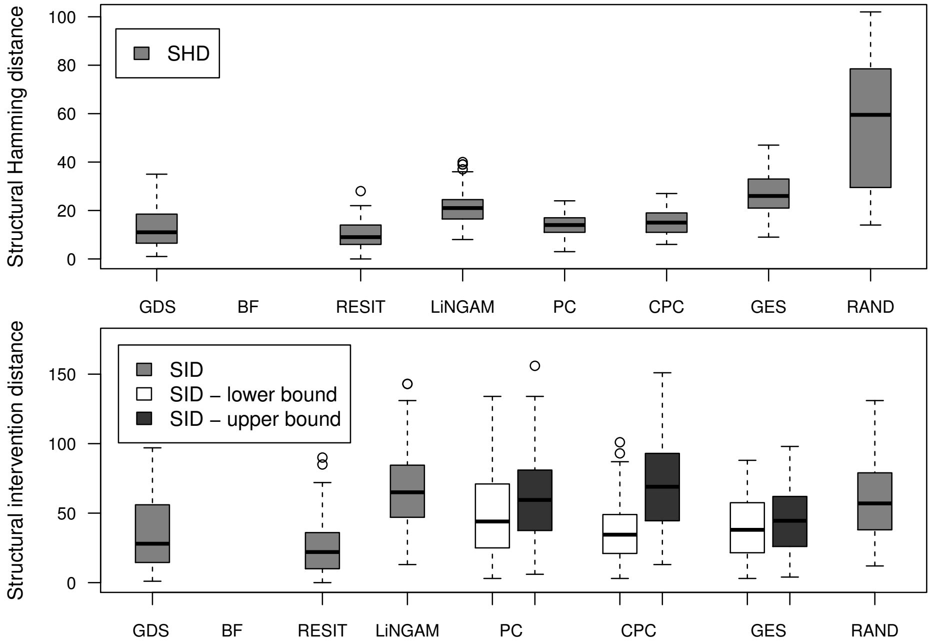


Figure 4: Similar to Figure 3: box plots of the SHD between estimated structure and correct DAG (top) and box plots of the SID to the correct DAG (bottom) for $p = 15$, $n = 500$ and nonlinear Gaussian SEMs.

the data are linear or nonlinear, GDS provides an alternative that works reasonably well in both settings.

5.2 Altitude, Temperature and Duration of Sunshine

We consider recordings of average temperature T, average duration of sunshine DS and the altitude A at 349 German weather stations (Deutscher Wetterdienst, 2008). Figure 5 shows scatter plots of all pairs. LiNGAM estimates $T \to A$, PC and CPC estimate $T \to A \leftarrow DS$,

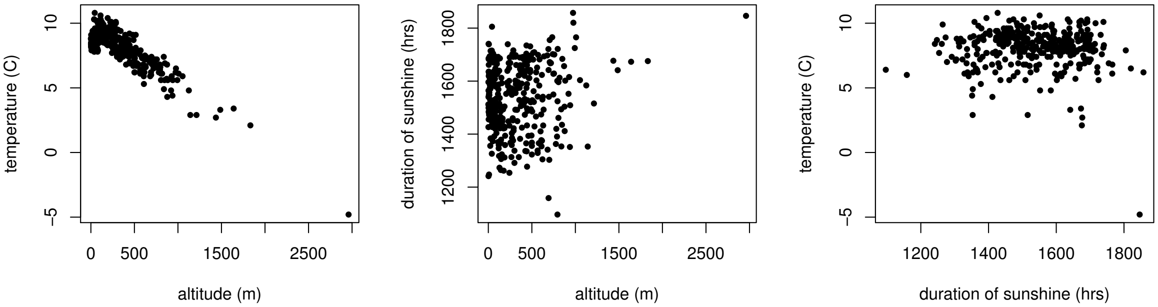


Figure 5: Scatter plots of the three pairs, altitude, temperature and duration of sunshine.

GES estimates a fully connected DAG. The brute-force estimate with linear regression obtains a score of 103.6. Since we are taking the logarithm to base 10 in equation (9), we see that the model does not fit the data well. More sensible seems the gam regression, for which both GDS and brute-force output the DAG $T \leftarrow A \rightarrow DS$ and $T \rightarrow DS$, which receives a score of 5.9. Also RESIT outputs this DAG. Although there might be a feedback between duration of sunshine and temperature through the generation of clouds, we believe that the link from sunshine to temperature should be stronger. In fact, the corresponding DAG $T \leftarrow A \rightarrow DS$ with $T \leftarrow DS$ receives the second best score. Furthermore, these data may be confounded by geographical location. Together with the possible feedback loop and a possible deviation from additive noise models this might be the reason why we do not obtain clear independence of the residuals: the HSIC between the residuals of temperature and the two others leads to a p-value of 0.012 (the other two p-values are both about 0.12). In practice, we often expect some violations of the model assumptions. This example, however, indicates that it may still possible to obtain reasonable estimates of the underlying causal structure if the violations are not too strong.

5.3 Cause-Effect Pairs

We have tested the performance of additive noise models on a collection of various cause-effect pairs, an extended version of the "Cause-effect pairs" data set described in Mooij and Janzing (2010). We used version 0.8 of this data set, which consists of observations of 86 different pairs of variables from various domains. The task is to infer which variable is the cause and which variable the effect, for each of the pairs. For example, one of the pairs consists of 349 measurements of altitude and temperature taken at different weather stations in Germany (Deutscher Wetterdienst, 2008), the same data as considered in the previous subsection. It should be obvious that here the altitude is the cause, and the temperature is the effect. The complete data set and a more detailed description of each pair can be obtained from `http://webdav.tuebingen.mpg.de/cause-effect`.

For each pair of variables (X_i, Y_i), with $i = 1, \dots, 86$, we test the two possible additive noise models that correspond with the two different possible causal directions, $X_i \rightarrow Y_i$ and $Y_i \rightarrow X_i$. For both directions, we estimate the functional relationship by performing Gaussian Process regression using the GPML toolbox (Rasmussen and Nickisch, 2010). We use the expected value of the Gaussian Process given the observations as an estimate of the functional dependence between the cause and the effect. The goodness-of-fit is then evaluated by testing independence of the residuals and the inputs. Here, we use the HSIC as an independence test and approximate the null distribution with a gamma distribution in order to obtain p-values (Gretton et al., 2005). We thus obtain two p-values for each pair, one for each possible causal direction (where a high p-value corresponds to not rejecting independence, i.e., not rejecting the causal model). We then rank the pairs according to the highest of the two p-values of the pair. Using this ranking, we can make decisions for only a subset of the pairs, starting with the pair for which the highest of the two p-values is the largest among all pairs (we say these pairs have a high rank). In this way we trade off accuracy, i.e., percentage of correct decisions, versus the amount of decisions taken.

Five of the pairs have multivariate X_i or Y_i, and we did not include those in the analysis for convenience. Furthermore, not all the pairs are independent; for example, life expectancy

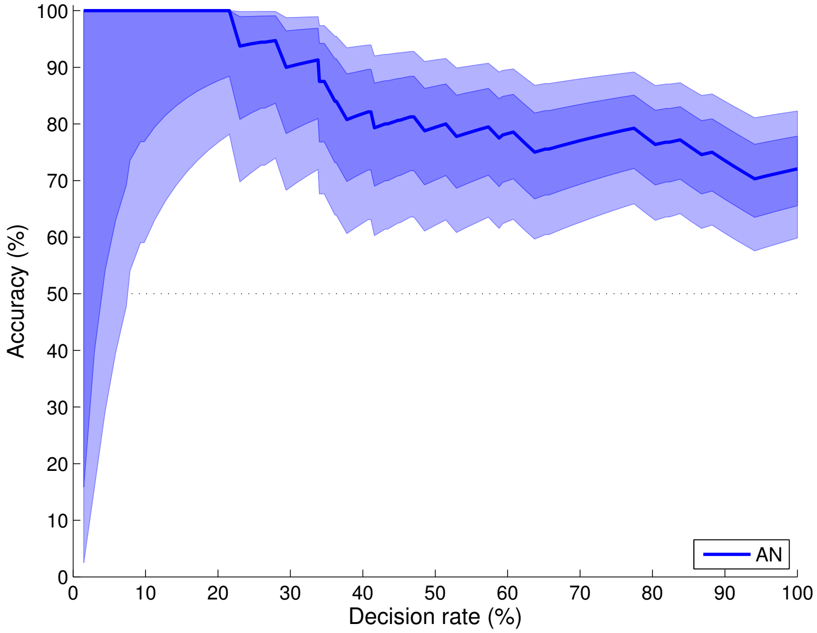


Figure 6: Results of the additive noise method on version 0.8 of the cause-effect pairs data set. After weighting, effectively 68 out of 86 pairs remained. The plot shows estimated accuracy, 68% and 95% confidence intervals for each decision rate.

versus latitude occurs more than once, but measurements were done in different years and for different gender. We therefore assigned weights to the cause-effect pairs to compensate for this when calculating the accuracy and decision rate. For example, the pair life expectancy versus latitude appears eight times (for different combinations of gender and year), hence each of these pairs is weighted down with the factor 1/8; on the other hand, the pair altitude vs. temperature at weather stations occurs only once, and therefore gets weight 1. Denoting the weight of each pair with w_i, the "effective" number of pairs becomes $\sum_{i=1}^{86} w_i = 68$. The five pairs with multivariate X_i or Y_i were given zero weight. If the set of highest-ranked pairs is denoted $\mathcal{I}$, and the set of correct decisions is denoted $\mathcal{C}$, then the *accuracy* (fraction of correct decisions) and and the *decision rate* (fraction of decisions taken) are defined as

$$\text{accuracy} = \frac{\sum_{i\in\mathcal{I}\cap\mathcal{C}} w_i}{\sum_{i\in\mathcal{I}} w_i}, \qquad \text{decision rate} = \frac{\sum_{i\in\mathcal{I}} w_i}{\sum_{i=1}^{86} w_i}.$$

The results are plotted in Figure 6. It shows the accuracy (dark blue line) as a function of the decision rate, together with confidence intervals (light blue regions). The amount of cause-effect pairs from which the accuracy can be estimated decreases proportionally to the decision rate; the accuracies reported for low decision rates therefore have higher uncertainty than the accuracies reported for high decision rates. For each decision rate, we have plotted the 68% and 95% confidence intervals for the estimated success probability assuming a binomial distribution using the Clopper-Pearson method. If for a given decision rate, the 95% confidence region lies above the line at 50%, the method performs significantly better than random guessing (for that decision rate). For example, if we take a decision for all pairs, $72 \pm 6\%$ of the decisions are correct, significantly more than random guessing. If we only take the 20% most confident decisions, all of them are correct, again significantly more than random guessing.

6. Discussion and Future Work

Apart from a few exceptions we can identify the directed acyclic graph from a bivariate distribution that has been generated by a structural equation model with continuous additive noise. Such an identifiability in the bivariate case generalizes under mild assumptions to identifiability in the multivariate case (i.e., graphs with more than two variables). This can be beneficial for the field of causal inference: if the true data generating process can be represented by a restricted structural equation model like additive noise models, the causal graph can be inferred from the joint distribution. We believe that formulating the problem using structural equation models rather than graphical models made it easier to state and exploit the assumption of additive noise. While the language of graphical models allow us to define some notion connecting a graph to the distribution (e.g., faithfulness), SEMs allow us to impose specific restrictions on the possible functional relationships between nodes and its children. This is closer in spirit to a machine learning approach where properties of function classes play a crucial role in the estimation. Both artificial and real data sets indicate that methods based on restricted structural equation models can outperform traditional constraint-based methods.

We have proposed two methods for estimating the graph from finitely many data. RESIT iteratively identifies sink nodes. Another method optimizes a score that reflects the independence of residuals. Although the score seems to be suitable to detect the correct graph structure, it remains unclear how to find the best scoring DAG when an exhaustive search is infeasible. We investigated the possibility to search this space by greedily choosing best-scoring neighbors (GDS). Multiple random initializations may decrease the chance that the greedy DAG search gets stuck in local optima by the additional cost of computational complexity. We further believe that the proposed score may benefit from an extended version of HSIC that is able to estimate mutual independence instead of pairwise independence. Recently, Nowzohour and Bühlmann (2013) have suggested a penalized likelihood based score for bivariate models. They estimate the noise distribution and use the BIC for penalization. In principle this idea can again be combined with a brute-force search as in Section 4.2.1 or a greedy DAG search as in Section 4.2.2.

Making the methods applicable to larger graphs ($p > 20$) remains a major challenge. Also, studying the statistical properties of the methods (for example, establishing consistency) is an important task for future research.

Acknowledgements

We thank Peter Bühlmann and Markus Kalisch for helpful discussions and Patrik Hoyer for the collaboration initiating the idea of nonlinear additive noise models (Hoyer et al., 2009). The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-

2013) under REA grant agreement no 326496. JM was supported by NWO, the Netherlands Organization for Scientific Research (VENI grant 639.031.036). We thank the anonymous reviewers for their insightful comments.

Appendix A. Proofs

We now provide all proofs that have been omitted in the main text.

A.1 Proof of Proposition 4

Proof "if": Assume that causal minimality is not satisfied. Then, there is an X_j and a $Y \in \mathbf{PA}_j^{\mathcal{G}}$, such that $\mathcal{L}(\mathbf{X})$ is also Markov with respect to the graph obtained when removing the edge $Y \to X_j$ from $\mathcal{G}$.
"only if": If $\mathcal{L}(\mathbf{X})$ has a density, the Markov condition is equivalent to the Markov factorization (Lauritzen, 1996, Theorem 3.27). Assume that $Y \in \mathbf{PA}_j^{\mathcal{G}}$ and $X_j \perp\!\!\!\perp Y \,|\, \mathbf{PA}_j^{\mathcal{G}} \setminus \{Y\}$. Then $P(\mathbf{X}) = P(X_j|\mathbf{PA}_j^{\mathcal{G}} \setminus \{Y\}) \prod_{k\neq j} P(X_k|\mathbf{PA}_k^{\mathcal{G}})$, which implies that $\mathcal{L}(\mathbf{X})$ is Markov w.r.t. $\mathcal{G}$ without $Y \to X_j$. ■

A.2 Proof of Proposition 7

Proof We will prove that for all $\mathcal{G}_1$ and $\mathcal{G}_2$ in $\mathbb{G}$ there is DAG $\mathcal{G} \in \mathbb{G}$ such that $\mathcal{G} \leq \mathcal{G}_1$ and $\mathcal{G} \leq \mathcal{G}_2$. This implies the existence of a least element since the set $\mathbb{G}$ is finite. Consider any node X_i and denote the $\mathcal{G}_1$-parents by $X_{j_1}, \ldots, X_{j_r}, X_{k_{r+1}}, \ldots, X_{k_{r+s}}$ and the $\mathcal{G}_2$-parents by $X_{j_1}, \ldots, X_{j_r}, X_{\ell_{r+1}}, \ldots, X_{\ell_{r+t}}$, such that $\{k_{r+1}, \ldots, k_{r+s}\}$ and $\{\ell_{r+1}, \ldots, \ell_{r+t}\}$ are disjoint sets. Here, $X_{j_1}, \ldots, X_{j_r}$ are the joint parents in $\mathcal{G}_1$ and $\mathcal{G}_2$. We have for all $x_{j_1}, \ldots, x_{j_r}$, $x_{k_{r+1}}, \ldots, x_{k_{r+s}}$ and $x_{\ell_{r+1}}, \ldots, x_{\ell_{r+t}}$ (at which the density p is strictly positive) that

$$\begin{aligned}
&p(X_i \,|\, X_{j_1} = x_{j_1}, \ldots, X_{j_r} = x_{j_r}, X_{k_{r+1}} = x_{k_{r+1}}, \ldots, X_{k_{r+s}} = x_{k_{r+s}}) \\
&= p\big(X_i \,|\, do(X_{j_1} = x_{j_1}, \ldots, X_{j_r} = x_{j_r}, X_{k_{r+1}} = x_{k_{r+1}}, \ldots, X_{k_{r+s}} = x_{k_{r+s}}, \\
&\qquad\qquad\qquad\qquad\qquad\qquad X_{\ell_{r+1}} = x_{\ell_{r+1}}, \ldots, X_{\ell_{r+t}} = x_{\ell_{r+t}})\big) \\
&= p(X_i \,|\, X_{j_1} = x_{j_1}, \ldots, X_{j_r} = x_{j_r}, X_{\ell_{r+1}} = x_{\ell_{r+1}}, \ldots, X_{\ell_{r+t}} = x_{\ell_{r+t}}) =: (*)\,.
\end{aligned}$$

This implies

$$(*) = p(X_i \,|\, X_{j_1} = x_{j_1}, \ldots, X_{j_r} = x_{j_r})\,.$$

Set the variables $X_{j_1}, \ldots, X_{j_r}$ to be the $\mathcal{G}$-parents of node X_i and repeat for all nodes X_i. The distribution $\mathcal{L}(\mathbf{X})$ is Markov w.r.t. graph $\mathcal{G}$ by its construction. Note that all proper subgraphs of a true causal DAG with respect to which $\mathcal{L}(\mathbf{X})$ is Markov are again true causal DAGs. This proves the statement about causal minimality. ■

A.3 Proof of Proposition 9

Proof Let $N_1, \cdots, N_p$ be independent and uniformly distributed between 0 and 1. We then define $X_j = f_j(\mathbf{PA}_j, N_j)$ with

$$f_j(x_{\mathbf{PA}_j}, n_j) = F^{-1}_{X_j|\mathbf{PA}_j = x_{\mathbf{PA}_j}}(n_j)\,,$$

where $F^{-1}_{X_j|\mathbf{PA}_j = x_{\mathbf{PA}_j}}$ is the inverse cdf of X_j given $\mathbf{PA}_j = x_{\mathbf{PA}_j}$. ■

A.4 Proof of Proposition 17

Proof Assume causal minimality is not satisfied. We can then find a j and $i \in \mathbf{PA}_j$ with $X_j = f_j(X_{\mathbf{PA}_j \setminus \{i\}}, X_i) + N_j$ that does not depend on X_i if we condition on all other parents $\mathbf{PA}_j \setminus \{i\}$ (Proposition 4). Let us denote $\mathbf{PA}_j \setminus \{X_i\}$ by X_A. For the function f_j it follows that $f_j(x_A, x_i) = c_{x_A}$ for $\mathcal{L}(X_A, X_i)$-almost all (x_A, x_i). Indeed, assume without loss of generality that $\mathbf{E}N_j = 0$, take the mean of $X_j \,|\, \mathbf{PA}_j^{\mathcal{G}_0} = (x_A, x_i)$ and use e.g. (2b) from Dawid (1979). The continuity of f_j implies that f_j is constant in its last argument.

The converse statement follows from Proposition 4, too. ■

A.5 Proof of Theorem 20

Proof To simplify notation we write $X := X_i$ and $Y := X_j$ (see Definition 18). If $\mathcal{G}_0$ is the empty graph, $X \perp\!\!\!\perp Y$. On the other hand, if the graph is not empty, $X \perp\!\!\!\perp Y$ would be a violation of causal minimality. We can therefore now assume that the graph is not empty and $X \not\perp\!\!\!\perp Y$. Let us assume that the graph is not identifiable and we have

$$p_n(y - f(x))p_x(x) = p(x, y) = p_{\tilde{n}}(x - g(y))p_y(y)\,. \tag{10}$$

Set

$$\pi(x, y) := \log p(x, y) = \nu(y - f(x)) + \xi(x)\,, \tag{11}$$

and $\tilde{\nu} := \log p_{\tilde{n}}$, $\eta := \log p_y$. From the r.h.s. of Equation (10) we find $\pi(x, y) = \tilde{\nu}(x - g(y)) + \eta(y)$, implying

$$\frac{\partial^2 \pi}{\partial x \partial y} = -\tilde{\nu}''(x - g(y))g'(y) \quad \text{and} \quad \frac{\partial^2 \pi}{\partial x^2} = \tilde{\nu}''(x - g(y))\,.$$

We conclude

$$\frac{\partial}{\partial x}\left(\frac{\partial^2 \pi / \partial x^2}{\partial^2 \pi / (\partial x \partial y)}\right) = 0\,. \tag{12}$$

Using Equation (11) we obtain

$$\frac{\partial^2 \pi}{\partial x \partial y} = -\nu''(y - f(x))f'(x)\,, \tag{13}$$

and

$$\frac{\partial^2 \pi}{\partial x^2} = \frac{\partial}{\partial x}\left(-\nu'(y - f(x))f'(x) + \xi'(x)\right) = \nu''(f')^2 - \nu' f'' + \xi''\,, \tag{14}$$

where we have dropped the arguments for convenience. Combining Equations (13) and (14) yields

$$\frac{\partial}{\partial x}\left(\frac{\frac{\partial^2\pi}{\partial x^2}}{\frac{\partial^2\pi}{\partial x\partial y}}\right) = -\,2f'' + \frac{\nu' f'''}{\nu'' f'} - \xi'''\frac{1}{\nu'' f'} + \frac{\nu'\nu''' f''}{(\nu'')^2} - \frac{\nu'(f'')^2}{\nu''(f')^2} - \xi''\frac{\nu'''}{(\nu'')^2} + \xi''\frac{f''}{\nu''(f')^2}\,.$$

Due to equation (12) this expression must vanish and we obtain the differential equation (7)

$$\xi''' = \xi''\left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'}\right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu'\nu''' f'' f'}{\nu''} - \frac{\nu'(f'')^2}{f'}$$

by term reordering. This contradicts the assumption that the distribution is generated by an identifiable bivariate additive noise model, see Condition 19. ■

A.6 Proof of Proposition 21

Proof Let the notation be as in Theorem 20 and let y be fixed such that $\nu''(y-f(x))f'(x) \neq 0$ holds for all but countably many x. Given f, ν, we obtain a linear inhomogeneous differential equation (DE) for ξ:

$$\xi'''(x) = \xi''(x)G(x,y) + H(x,y)\,, \tag{15}$$

where G and H are defined by

$$G := -\frac{\nu''' f'}{\nu''} + \frac{f''}{f'}$$

and

$$H := -2\nu'' f'' f' + \nu' f''' + \frac{\nu'\nu''' f'' f'}{\nu''} - \frac{\nu'(f'')^2}{f'}\,,$$

see proof of Theorem 20. Setting $z := \xi''$ we have $z'(x) = z(x)G(x,y) + H(x,y)$. Given that such a function z exists, it is given by

$$z(x) = z(x_0)e^{\int_{x_0}^{x} G(\tilde{x},y)d\tilde{x}} + \int_{x_0}^{x} e^{\int_{\hat{x}}^{x} G(\tilde{x},y)d\tilde{x}} H(\hat{x},y)d\hat{x}\,. \tag{16}$$

Then z is determined by $z(x_0)$ since we can extend Equation (16) to the remaining points. The set of all functions ξ satisfying the linear inhomogenous DE (15) is a 3-dimensional affine space: Once we have fixed $\xi(x_0), \xi'(x_0), \xi''(x_0)$ for some arbitrary point x_0, ξ is completely determined. Given fixed f and ν, the set of all ξ admitting a backward model is contained in this subspace. ■

A.7 Proof of Corollary 22

Proof Similarly to how (12) was derived, under the assumption of the existence of a reverse model one can derive

$$\frac{\partial^2\pi}{\partial x\partial y}\cdot\frac{\partial}{\partial x}\left(\frac{\partial^2\pi}{\partial x^2}\right) = \frac{\partial^2\pi}{\partial x^2}\cdot\frac{\partial}{\partial x}\left(\frac{\partial^2\pi}{\partial x\partial y}\right)\,.$$

Now using (13) and (14), we obtain

$$(-\nu''f')\cdot\frac{\partial}{\partial x}\left(\nu''(f')^2-\nu'f''+\xi''\right)=\left(\nu''(f')^2-\nu'f''+\xi''\right)\cdot\frac{\partial}{\partial x}\left(-\nu''f'\right),$$

which reduces to

$$-2(\nu''f')^2f''+\nu''f'\nu'f'''-\nu''f'\xi'''=-\nu'f''\nu'''(f')^2+\xi''\nu'''(f')^2+\nu''\nu'(f'')^2-\nu''f''\xi''.$$

Substituting the assumptions $\xi'''=0$ and $\nu'''=0$ (and hence $\nu''=C$ everywhere with $C\neq 0$ since otherwise ν cannot be a proper log-density) yields

$$\nu'\big(y-f(x)\big)\cdot\left(f'f'''-(f'')^2\right)=2C(f')^2f''-f''\xi''.$$

Since $C\neq 0$ there exists an α such that $\nu'(\alpha)=0$. Then, restricting ourselves to the submanifold $\{(x,y)\in\mathbb{R}^2 : y-f(x)=\alpha\}$ on which $\nu'=0$, we have

$$0=f''(2C(f')^2-\xi'').$$

Therefore, for all x in the open set $[f''\neq 0]$, we have $(f'(x))^2=\xi''/(2C)$, which is a constant, so $f''=0$ on $[f''\neq 0]$: a contradiction. Therefore, $f''=0$ everywhere. ■

A.8 Definitions of Proposition 23

Definition 35 *(Zhang and Hyvärinen, 2009) A one-dimensional distribution that is absolutely continuous with respect to the Lebesgue measure and has positive density p is called:*

- log-mix-lin-exp *if there are c_1, c_2, c_3, c_4 with $c_1<0$ and $c_2c_3>0$ such that*

$$\log p(x)=c_1\exp(c_2x)+c_3x+c_4,$$

- one-sided asymptotically exponential *if there is $c\neq 0$ such that*

$$\frac{d}{dx}\log p(x)\to c$$

  as $x\to-\infty$ or $x\to\infty$,

- two-sided asymptotically exponential *if there are $c_1\neq 0$ and $c_2\neq 0$ such that*

$$\frac{d}{dx}\log p(x)\to c_1$$

  as $x\to-\infty$ and

$$\frac{d}{dx}\log p(x)\to c_2$$

  as $x\to\infty$

- *and a* generalized mixture of two exponentials *if there are $d_1, d_2, d_3, d_4, d_5, d_6$ with $d_4>0$, $d_3>0$, $d_1d_5>0$ and $d_2<-\frac{d_1}{d_5}$ such that*

$$\log p(x)=d_1x+d_2\log(d_3+d_4\exp(d_5x))+d_6.$$

A.9 Proof of Example 25

Proof Our starting point is the assumption of nonidentifiability. In other words, we can describe the joint distribution of x and y both as an additive noise model where X causes Y, and as an additive noise model where Y causes X. Using the same notation as in Theorem 20, this means that:

$$\xi(x) + \nu\big(y - f(x)\big) = \eta(y) + \tilde{\nu}\big(x - g(y)\big) \qquad \forall x, y \in \mathbb{R}\,. \tag{17}$$

Case II in Proposition 23 (reproduced from Table 1 in Zhang and Hyvärinen, 2009) states that if both ξ and ν are log-mix-lin-exp and f is affine, then there could be an unidentifiable model. Let us verify whether that is indeed the case. We take

$$\begin{aligned}\xi(x) &= c_1 \exp(c_2 x) + c_3 x + c_4 \\ \nu(n) &= \gamma_1 \exp(\gamma_2 n) + \gamma_3 n + \gamma_4 \\ f(x) &= ax + b\end{aligned}$$

with $a \neq 0$ ($a = 0$ is the degenerate case with X and Y independent).

We can rewrite (17) as follows, by substituting x with $x + g(y)$:

$$c_1 e^{c_2(x+g(y))} + c_3(x+g(y)) + c_4 + \gamma_1 e^{\gamma_2(y-ax-ag(y)-b)} + \gamma_3(y-ax-ag(y)-b) + \gamma_4 = \eta(y) + \tilde{\nu}(x)\,. \tag{18}$$

Differentiating with respect to x yields

$$c_1 c_2 e^{c_2(x+g(y))} + c_3 - a\gamma_1\gamma_2 e^{\gamma_2(y-ax-ag(y)-b)} - \gamma_3 a = \tilde{\nu}'(x)\,. \tag{19}$$

Differentiating with respect to y yields

$$c_1 c_2^2 e^{c_2(x+g(y))} g'(y) - a\gamma_1\gamma_2^2 e^{\gamma_2(y-ax-ag(y)-b)}(1 - ag'(y)) = 0\,.$$

This can only be satisfied for all x if $c_2 = -a\gamma_2$. In that case:

$$-ac_1 g'(y) + \gamma_1 e^{\gamma_2(y-b)}(1 - ag'(y)) = 0\,.$$

Rewriting:

$$ag'(y) = \frac{\gamma_1 e^{\gamma_2(y-b)}}{c_1 + \gamma_1 e^{\gamma_2(y-b)}}\,.$$

Integrating:

$$g(y) = -\frac{1}{c_2}\ln(-c_1 - \gamma_1 e^{\gamma_2(y-b)}) + \frac{C}{c_2}\,.$$

Note that

$$e^{c_2 g(y)} = -\frac{1}{c_1 + \gamma_1 e^{\gamma_2(y-b)}} e^{-C}\,.$$

Substituting into (19) yields

$$-c_2 e^{-C} e^{c_2 x} + c_3 - \gamma_3 a = \tilde{\nu}'(x)\,.$$

Integrating yields

$$-e^{-C} e^{c_2 x} + (c_3 - \gamma_3 a)x + \delta_4 = \tilde{\nu}(x)\,,$$

which is also log-mix-lin-exp with parameters $\delta_1 = -e^{-C}$, $\delta_2 = c_2$, $\delta_3 = c_3 - \gamma_3 a$, δ_4. Substituting into (18):

$$g(y)(c_3 - \gamma_3 a) + \gamma_3 y + c_4 - \gamma_3 b + \gamma_4 - \delta_4 = \eta(y)\,,$$

i.e.,

$$\eta(y) = \left(-\frac{1}{c_2}\ln(-c_1 - \gamma_1 e^{\gamma_2(y-b)}) + \frac{C}{c_2}\right)(c_3 - \gamma_3 a) + \gamma_3 y + c_4 - \gamma_3 b + \gamma_4 - \delta_4\,.$$

This gives an inequality constraint: $c_3 \neq a\gamma_3$. $\eta(y)$ is a generalized mixture of exponentials distribution with parameters $d_1 = \gamma_3$, $d_2 = -\frac{c_3 - a\gamma_3}{c_2}$, $d_3 = -c_1$, $d_4 = -\gamma_1 e^{-\gamma_2 b}$, $d_5 = \gamma_2$, $d_6 = C\frac{c_3 - a\gamma_3}{c_2} + c_4 - \gamma_3 b + \gamma_4 - \delta_4$. One can check that all constraints on the parameters of the generalized mixture of exponentials are satisfied. Choosing C appropriately allows for normalizing the log-density. One can also easily verify that with these choices of $\tilde{\nu}(x)$ and $\eta(y)$, equation (17) holds, and therefore this gives an example of a nonidentifiable additive noise model. ∎

A.10 Some Lemmata

The following four statements are all plausible and their proof is mostly about technicalities. The reader may skip to the next section and use the lemmata whenever needed. For random variables A and B we use $A\,|_{\,B=b}$ to denote the random variable A after conditioning on $B = b$ (assuming densities exist and B has positive density at b).

Lemma 36 *Let $Y \in \mathcal{Y}, N \in \mathcal{N}, \mathbf{Q} \in \mathcal{Q}, \mathbf{R} \in \mathcal{R}$ be random variables whose joint distribution is absolutely continuous with respect to some product measure ($\mathbf{Q}$ and $\mathbf{R}$ can be multivariate) and with density $p_{Y,\mathbf{Q},\mathbf{R},N}(y,\mathbf{q},\mathbf{r},n)$. Let $f : \mathcal{Y} \times \mathcal{Q} \times \mathcal{N} \to \mathbb{R}$ be a measurable function. If $N \perp\!\!\!\perp (Y, \mathbf{Q}, \mathbf{R})$ then for all $\mathbf{q} \in \mathcal{Q}, \mathbf{r} \in \mathcal{R}$ with $p_{\mathbf{Q},\mathbf{R}}(\mathbf{q},\mathbf{r}) > 0$:*

$$f(Y,\mathbf{Q},N)\,|_{\,\mathbf{Q}=\mathbf{q},\mathbf{R}=\mathbf{r}} \overset{\mathcal{L}}{=} f(Y\,|_{\,\mathbf{Q}=\mathbf{q},\mathbf{R}=\mathbf{r}}, \mathbf{q}, N)\,.$$

A formal proof of this statement can be found in Peters et al. (2011b, Lemma 2).

Lemma 37 *Let $\mathcal{L}(\mathbf{X})$ be generated according to a SEM as in (2) with corresponding DAG $\mathcal{G}$ and consider a variable $X \in \mathbf{X}$. If $\mathbf{S} \subseteq \mathbf{ND}_X^{\mathcal{G}}$ then $N_X \perp\!\!\!\perp \mathbf{S}$.*

Proof Write $\mathbf{S} = \{S_1, \ldots, S_k\}$. Then

$$\mathbf{S} = \left(f_{S_1}(\mathbf{PA}_{S_1}^{\mathcal{G}}, N_{S_1}), \ldots, f_{S_k}(\mathbf{PA}_{S_k}^{\mathcal{G}}, N_{S_k})\right).$$

Again, one can substitute the parents of S_i by the corresponding functional equations and proceed recursively. After finitely many steps one obtains $\mathbf{S} = f(N_{T_1}, \ldots, N_{T_l})$, where $\{T_1, \ldots, T_l\}$ is the set of *all* ancestors of nodes in $\mathbf{S}$, which does not contain X. Since all noise variables are jointly independent we have $N_X \perp\!\!\!\perp \mathbf{S}$. ∎

With the intersection property of conditional independence (e.g., 1.1.5 in Pearl, 2009), Proposition 4 has the following corollary that we formalize as a lemma.

Lemma 38 *Consider the random vector* $\mathbf{X}$ *and assume that the joint distribution has a (strictly) positive density. Then* $\mathcal{L}(\mathbf{X})$ *satisfies causal minimality with respect to* $\mathcal{G}$ *if and only if* $\forall B \in \mathbf{X}\ \forall A \in \mathbf{PA}_B^{\mathcal{G}}$ *and* $\forall \mathbf{S} \subset \mathbf{X}$ *with* $\mathbf{PA}_B^{\mathcal{G}} \setminus \{A\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_B^{\mathcal{G}} \setminus \{A\}$ *we have that*

$$B \not\perp\!\!\!\perp A \mid \mathbf{S}\,.$$

Proof The "if" part is immediate. For the "only if" let us denote $\mathbf{P} := \mathbf{PA}_B^{\mathcal{G}} \setminus \{A\}$ and $\mathbf{Q} := \mathbf{S} \setminus (\mathbf{PA}_B^{\mathcal{G}} \setminus \{A\})$, such that $\mathbf{S} = \mathbf{P} \cup \mathbf{Q}$. Observe that $B \not\perp\!\!\!\perp A \,|\, \mathbf{P}$ (see Proposition 4) implies $B \not\perp\!\!\!\perp (\{A\} \cup \mathbf{Q}) \,|\, \mathbf{P}$. From the Markov condition we have $B \perp\!\!\!\perp \mathbf{Q} \,|\, (\mathbf{P} \cup \{A\})$. The intersection property of conditional independence yields $B \not\perp\!\!\!\perp A \,|\, (\mathbf{P} \cup \mathbf{Q})$. ■

A.11 Proof of Theorem 28

Proof We assume that there are two restricted additive noise models (see Definition 27) that both induce $\mathcal{L}(\mathbf{X})$, one with graph $\mathcal{G}$, the other with graph $\mathcal{G}'$. We will show that $\mathcal{G} = \mathcal{G}'$. Consider the variables L, Y from Proposition 29 (i) and define the sets $\mathbf{Q} := \mathbf{PA}_L^{\mathcal{G}} \setminus \{Y\}$, $\mathbf{R} := \mathbf{PA}_Y^{\mathcal{G}'} \setminus \{L\}$ and $\mathbf{S} := \mathbf{Q} \cup \mathbf{R}$. At first, we consider any $\mathbf{s} = (\mathbf{q}, \mathbf{r})$ and write $L^* := L \,|\, _{\mathbf{S}=\mathbf{s}}$ and $Y^* := Y \,|\, _{\mathbf{S}=\mathbf{s}}$. Lemma 37 gives us $N_L \perp\!\!\!\perp (Y, \mathbf{S})$ and $N_Y \perp\!\!\!\perp (L, \mathbf{S})$ and we can thus apply Lemma 36. From $\mathcal{G}$ we find

$$L^* = f_L(\mathbf{q}, Y^*) + N_L, \qquad N_L \perp\!\!\!\perp Y^*$$

and from $\mathcal{G}'$ we have

$$Y^* = g_Y(\mathbf{r}, L^*) + N_Y, \qquad N_Y \perp\!\!\!\perp L^*\,.$$

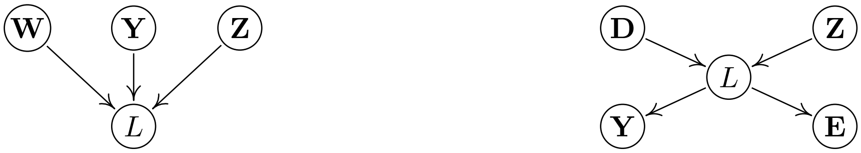
This contradicts Theorem 20 since according to Definition 27 we can choose $\mathbf{s} = (\mathbf{q}, \mathbf{r})$ such that $(f_L(\mathbf{q}, \cdot), \mathcal{L}(Y^*), \mathcal{L}(N_L))$ and $(g_Y(\mathbf{r}, \cdot), \mathcal{L}(L^*), \mathcal{L}(N_Y))$ satisfy Condition 19. ■

A.12 Proof of Proposition 29

Proof Since DAGs do not contain any cycles, we always find nodes that have no descendants (start a directed path at some node: after at most $\#\mathbf{X} - 1$ steps we reach a node without a child). Eliminating such a node from the graph leads to a DAG, again; we can discard further nodes without children in the new graph. We repeat this process for all nodes that have no children in both $\mathcal{G}$ and $\mathcal{G}'$ and have the same parents in both graphs. If we end up with no nodes left, the two graphs are identical which violates the assumption of the proposition. Otherwise, we end up with a smaller set of variables that we again call $\mathbf{X}$, two smaller graphs that we again call $\mathcal{G}$ and $\mathcal{G}'$ and a node L that has no children in $\mathcal{G}$ and either $\mathbf{PA}_L^{\mathcal{G}} \neq \mathbf{PA}_L^{\mathcal{G}'}$ or $\mathbf{CH}_L^{\mathcal{G}'} \neq \emptyset$. We will show that this leads to a contradiction. Importantly, because of the Markov property of the distribution with respect to $\mathcal{G}$, all other nodes are independent of L given $\mathbf{PA}_L^{\mathcal{G}}$:

$$L \perp\!\!\!\perp \mathbf{X} \setminus (\mathbf{PA}_L^{\mathcal{G}} \cup \{L\}) \mid \mathbf{PA}_L^{\mathcal{G}}\,. \tag{20}$$

To make the arguments easier to understand, we introduce the following notation (see also Fig. 7): we partition $\mathcal{G}$-parents of L into $\mathbf{Y}, \mathbf{Z}$ and $\mathbf{W}$. Here, $\mathbf{Z}$ are also $\mathcal{G}'$-parents of L, $\mathbf{Y}$ are $\mathcal{G}'$-children of L and $\mathbf{W}$ are not adjacent to L in $\mathcal{G}'$. We denote with $\mathbf{D}$ the $\mathcal{G}'$-parents of L that are not adjacent to L in $\mathcal{G}$ and by $\mathbf{E}$ the $\mathcal{G}'$-children of L that are not adjacent to L in $\mathcal{G}$. Thus: $\mathbf{PA}_L^{\mathcal{G}} = \mathbf{Y} \cup \mathbf{Z} \cup \mathbf{W}$, $\mathbf{CH}_L^{\mathcal{G}} = \emptyset$, $\mathbf{PA}_L^{\mathcal{G}'} = \mathbf{Z} \cup \mathbf{D}$, $\mathbf{CH}_L^{\mathcal{G}'} = \mathbf{Y} \cup \mathbf{E}$.



part of $\mathcal{G}$

part of $\mathcal{G}'$

Figure 7: Nodes adjacent to L in $\mathcal{G}$ and $\mathcal{G}'$

Consider $\mathbf{T} := \mathbf{W} \cup \mathbf{Y}$. We distinguish two cases:

Case (i): $\mathbf{T} = \emptyset$.
Then there must be a node $D \in \mathbf{D}$ or a node $E \in \mathbf{E}$, otherwise L would have been discarded.

1. If there is a $D \in \mathbf{D}$ then (20) implies $L \perp\!\!\!\perp D \,|\, \mathbf{S}$ for $\mathbf{S} := \mathbf{Z} \cup \mathbf{D} \setminus \{D\}$, which contradicts Lemma 38 (applied to $\mathcal{G}'$).

2. If $\mathbf{D} = \emptyset$ and there is $E \in \mathbf{E}$ then $E \perp\!\!\!\perp L \,|\, \mathbf{S}$ holds for $\mathbf{S} := \mathbf{Z} \cup \mathbf{PA}_E^{\mathcal{G}'} \setminus \{L\}$ (see graph $\mathcal{G}$), which also contradicts Lemma 38 (note that $\mathbf{Z} \subseteq \mathbf{ND}_E^{\mathcal{G}'}$ to avoid cycles).

Case (ii): $\mathbf{T} \neq \emptyset$.
Then $\mathbf{T}$ contains a "$\mathcal{G}'$-youngest" node with the property that there is no directed $\mathcal{G}'$-path from this node to any other node in $\mathbf{T}$. This node may not be unique.

1. Suppose that some $W \in \mathbf{W}$ is such a youngest node. Consider the DAG $\tilde{\mathcal{G}}'$ that equals $\mathcal{G}'$ with additional edges $Y \to W$ and $W' \to W$ for all $Y \in \mathbf{Y}$ and $W' \in \mathbf{W} \setminus \{W\}$. In $\tilde{\mathcal{G}}'$, L and W are not adjacent. Thus we find a set $\tilde{\mathbf{S}}$ such that $\tilde{\mathbf{S}}$ d-separates L and W in $\tilde{\mathcal{G}}'$; indeed, one can take $\tilde{\mathbf{S}} = \mathbf{PA}_L^{\tilde{\mathcal{G}}'}$ if $W \notin \mathbf{DE}_L^{\tilde{\mathcal{G}}'}$ and $\tilde{\mathbf{S}} := \mathbf{PA}_W^{\tilde{\mathcal{G}}'}$ if $L \notin \mathbf{DE}_W^{\tilde{\mathcal{G}}'}$. Then also $\mathbf{S} = \tilde{\mathbf{S}} \cup \{\mathbf{Y}, \mathbf{Z}, \mathbf{W} \setminus \{W\}\}$ d-separates L and W in $\tilde{\mathcal{G}}'$.

   Indeed, all $Y \in \mathbf{Y}$ are already in $\tilde{\mathbf{S}}$ in order to block $L \to Y \to W$. Suppose there is a $\tilde{\mathcal{G}}'$-path that is blocked by $\tilde{\mathbf{S}}$ and unblocked if we add Z and W' nodes to $\tilde{\mathbf{S}}$. How can we unblock a path by including more nodes? The path ($L \cdots V_1 \cdots U_1 \cdots W$ in Fig. 8) must contain a collider V_1 that is an ancestor of a Z with $V_1, \ldots, V_m, Z \notin \tilde{\mathbf{S}}$ and corresponding nodes U_i for a W' node. Choose V_1 and U_1 on the given path so close to each other such that there is no such collider in between. If there is no V_1, choose U_1 closest to L, if there is no U_1, choose V_1 closest to W. Now the path $L \leftarrow Z \cdots V_1 \cdots U_1 \cdots W' \to W$ is unblocked given $\tilde{\mathbf{S}}$, which is a contradiction to the assumption that $\tilde{\mathbf{S}}$ d-separates L and W.

   But then $\mathbf{S}$ d-separates L and W in $\mathcal{G}'$, too (there are less paths), and we have $L \perp\!\!\!\perp W \mid \mathbf{S}$, which contradicts Lemma 38 (applied to $\mathcal{G}$).

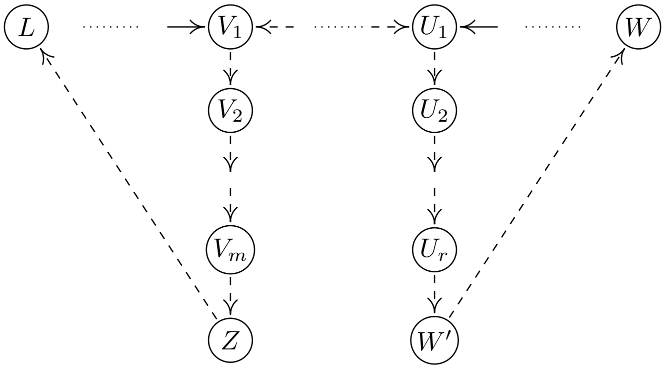


Figure 8: Assume the path $L \cdots V_1 \cdots U_1 \cdots W$ is blocked by $\tilde{\mathbf{S}}$, but unblocked if we include Z and W'. Then the dashed path is unblocked given $\tilde{\mathbf{S}}$.

2. Therefore, the $\mathcal{G}'$-youngest node in $\mathbf{T}$ must be some $Y \in \mathbf{Y}$.
   Define $\mathbf{Q} := \mathbf{PA}_L^{\mathcal{G}} \setminus \{Y\}$, $\mathbf{R} := \mathbf{PA}_Y^{\mathcal{G}'} \setminus \{L\}$ and $\mathbf{S} := \mathbf{Q} \cup \mathbf{R}$. Clearly, $\mathbf{S} \subseteq \mathbf{ND}_L^{\mathcal{G}} \setminus \{Y\}$ since L does not have any descendants in $\mathcal{G}$. Further, $\mathbf{S} \subseteq \mathbf{ND}_Y^{\mathcal{G}'} \setminus \{L\}$ because Y is the $\mathcal{G}'$-youngest under all $\mathbf{W}$ and $\mathbf{Y} \setminus \{Y\}$ by construction and any directed path from Y to $Z \in \mathbf{Z}$ would introduce a cycle in $\mathcal{G}'$. Ergo, $\{Y\} \cup \mathbf{S} \subseteq \mathbf{ND}_L^{\mathcal{G}}$ and $\{L\} \cup \mathbf{S} \subseteq \mathbf{ND}_Y^{\mathcal{G}'}$.

The variables L and Y and the sets $\mathbf{Q}, \mathbf{R}$ and $\mathbf{S}$ satisfy the conditions required in statement (i) of Proposition 29.

Statement (ii) follows as a special case since for Markov equivalent graphs, $\mathbf{W}, \mathbf{D}$ and $\mathbf{E}$ are all empty. Consider the $\mathcal{G}'$-youngest node Y. In order to avoid v-structures appearing in $\mathcal{G}$ and not in $\mathcal{G}'$ all nodes $Z \in \mathbf{Z}$ are directly connected to the $\mathcal{G}'$-youngest Y. And to avoid cycles, those nodes $Z \in \mathbf{Z}$ are $\mathcal{G}'$-parents of Y. The node Y cannot have other parents except for the ones in $\mathbf{Y}$ and $\mathbf{Z}$ since this would introduce v-structures in $\mathcal{G}'$ (with collider Y) that do not appear in $\mathcal{G}$. ■

A.13 Proof of Corollary 31

Proof We only prove (i) since (ii) is a special case. Causal minimality is satisfied because of Proposition 17. We can then assume that the statement is false and apply the same argument as in Theorem 28. This yields the two equations

$$\begin{aligned} L^* &= f_L(\mathbf{q}, Y^*) + N_L, \qquad && N_L \perp\!\!\!\perp Y^* \qquad \text{and} \\ Y^* &= g_Y(\mathbf{r}, L^*) + N_Y, \qquad && N_Y \perp\!\!\!\perp L^* \,. \end{aligned}$$

Let us define $f := f_L(\mathbf{q}, \cdot)$ and $g := g_Y(\mathbf{r}, \cdot)$. Because of independence of N_Y and L^* we have

$$0 = \frac{\partial^2 \log p(\ell^*, n_y)}{\partial n_y \, \partial \ell^*} = \frac{\partial^2 \log p(y^*, n_\ell)}{\partial n_y \, \partial \ell^*} = \frac{\partial^2 \log p(y^*) + \partial^2 \log p(n_\ell)}{\partial n_y \, \partial \ell^*}$$

with $y^* = g(\ell^*) + n_y$ and $n_\ell = \ell^* - f(g(\ell^*) + n_y)$. Ergo, for all ℓ^* and y^* we have

$$0 = \frac{\partial^2 \log p_{Y^*}(y^*)}{(\partial y^*)^2} g'(\ell^*) - \frac{1}{\sigma_{N_L}^2} f'(y^*)^2 g'(\ell^*) + \frac{1}{\sigma_{N_L}^2} f'(y^*) + \frac{\ell^* - f(y^*)}{\sigma_{N_L}^2} f''(y^*) g'(\ell^*) \,. \quad (21)$$

If there is a ℓ^* with $g'(\ell^*) = 0$, (21) implies that f' is constantly zero which is not the case. Exchanging the role of L^* and Y^* yields $f'(y^*) \neq 0$. Since N_L is Gaussian, Proposition 23 implies that f is linear. This contradicts the assumption of nonlinearity. For completeness, however, we give a direct proof that is similar to Lemma 6 in Zhang and Hyvärinen (2009). Dividing (21) by $g'(\ell^*) f'(y^*)$ yields

$$0 = \frac{\partial^2 \log p_{Y^*}(y^*)}{(\partial y^*)^2} \frac{1}{f'(y^*)} - \frac{1}{\sigma_{N_L}^2} f'(y^*) + \frac{1}{\sigma_{N_L}^2} \frac{1}{g'(\ell^*)} + \frac{\ell^* - f(y^*)}{\sigma_{N_L}^2} \frac{f''(y^*)}{f'(y^*)}$$

and therefore (take the derivative with respect to ℓ^*) $\frac{f''(y^*)}{f'(y^*)} \equiv a_1$ which means $f'(y^*) = a_2 \exp(a_1 y^*)$ with $a_1, a_2 \neq 0$ because f is nonlinear. But then $\frac{1}{g'(\ell^*)} + a_1 \ell^* \equiv a_3$ is constant and using $a_1 f(y^*) = f'(y^*) + a_4$ for some a_4 we have

$$\frac{\partial^2 \log p_{Y^*}(y^*)}{(\partial y^*)^2} - \frac{2}{\sigma_{N_L}^2} f'(y^*)^2 + \frac{a_3 - a_4}{\sigma_{N_L}^2} f'(y^*) = 0 \,,$$

which implies $\log p_{Y^*}(y^*) \to \infty$ for either $y^* \to \infty$ or $y^* \to -\infty$. Obviously, this cannot be the case. This proves the corollary. ■

A.14 Proof of Lemma 32

Proof For (a) suppose that $\mathcal{G}$ has an additional edge from X_i to X_j compared to $\mathcal{G}_0$. We can then change the corresponding structural equation $X_j = f_j(\mathbf{PA}_j^{\mathcal{G}_0}) + N_j$ into $X_j = \tilde{f}_j(\mathbf{PA}_j^{\mathcal{G}_0}, X_i) + N_j$ where $\tilde{f}_j$ equals f_j in the first $\#\mathbf{PA}_j^{\mathcal{G}_0}$ components and $\tilde{f}_j$ is constant in the last component.

We now prove statement (b). Let $\mathcal{G} \leq \mathcal{G}_0$ such that $\mathcal{L}(\mathbf{X})$ is Markov with respect to $\mathcal{G}$. Suppose $i \in \mathcal{G}$ with $\mathbf{PA}_i^{\mathcal{G}} \subsetneq \mathbf{PA}_i^{\mathcal{G}_0}$. Denote $X_B = \mathbf{PA}_i^{\mathcal{G}_0} \setminus \mathbf{PA}_i^{\mathcal{G}}$. Since $\mathbf{PA}_i^{\mathcal{G}_0} \subseteq \mathbf{ND}_i^{\mathcal{G}_0} \subseteq \mathbf{ND}_i^{\mathcal{G}}$, we have from the Markov property that $X_i \perp\!\!\!\perp X_B \,|\, \mathbf{PA}_i^{\mathcal{G}}$. Analogously to the proof of Proposition 17 this implies that the (continuous) function f_i in the corresponding structural equation $X_i = f_i(\mathbf{PA}_i^{\mathcal{G}_0}) + N_i$ must be constant in X_B. We can therefore define the corresponding structural equation in $\mathcal{G}$ to be $X_i = f_i(\mathbf{PA}_i^{\mathcal{G}}, x_B) + N_i$ for some arbitrary x_B. Structural equations for variables with identical parent sets do not need to be changed. Now suppose $\mathcal{G}_1 \leq \mathcal{G}_0$ and $\mathcal{G}_2 \leq \mathcal{G}_0$. Then there is an additive noise model with graph $\mathcal{G}_{12} \leq \mathcal{G}_0$ that leads to $\mathcal{L}(\mathbf{X})$, where $\mathcal{G}_{12}$ has precisely the edges that appear in both $\mathcal{G}_1$ and $\mathcal{G}_2$. This follows by noting that the intersection property implies that $X_i \perp\!\!\!\perp (\mathbf{PA}_i^{\mathcal{G}_0} \setminus \mathbf{PA}_i^{\mathcal{G}_{12}}) \,|\, \mathbf{PA}_i^{\mathcal{G}_{12}}$, and hence f_i is constant in $(\mathbf{PA}_i^{\mathcal{G}_0} \setminus \mathbf{PA}_i^{\mathcal{G}_{12}})$. (This step is not necessarily true for densities that are not strictly positive.) The partial ordering $\leq$ defined by the subgraph property therefore has a unique least element $\mathcal{G}_0^{min}$, which satisfies causal minimality by Proposition 17. ■

A.15 Proof of Theorem 34

Proof For the correct graph, we know that N_i is independent of all ancestor variables X_j since the latter can be expressed in terms of noise variables without N_i. The correct sink nodes therefore lead to independence in step 7 of Algorithm 1. We will now show that "wrong sinks", that is nodes who are not sinks in the correct graph $\mathcal{G}_0$ do not lead to independent residuals in the first iteration of Phase 1. It follows by induction that this is true for any later iteration, too. Suppose that node Y is not a sink in $\mathcal{G}_0$ but leads to independent residuals (step 7). Since Y is not a sink in $\mathcal{G}_0$, Y has children in $\mathcal{G}_0$. Call Z the $\mathcal{G}_0$-youngest child, that is there is no directed path from Z to any other child of Y. Disregard all descendants of Z and denote the remaining set of variables $\mathbf{S} := \mathbf{X} \setminus \{Y, Z, \mathbf{DE}_Z^{\mathcal{G}_0}\}$. It therefore follows that

$$\mathbf{DE}_Z^{\mathcal{G}_0} \perp\!\!\!\perp Y \,|\, \mathbf{S} \cup \{Z\} \,. \tag{22}$$

Because Y leads to independent residuals we can think of a graph $\mathcal{G}$ in which all variables are parents of Y. From Equation (22) it follows that $Y = g_Y(\mathbf{S}, Z) + \tilde{N}_Y$ with $\tilde{N}_Y \perp\!\!\!\perp (\mathbf{S}, Z)$. We then proceed similarly as in the proof of Theorem 28 and find from $\mathcal{G}_0$ that

$$Z \,|\, _{\mathbf{S}=\mathbf{s}} = f_Z(s_{\mathbf{PA}_Z^{\mathcal{G}_0}}, Y \,|\, _{\mathbf{S}=\mathbf{s}}) + N_Z \,.$$

From $\mathcal{G}$ we conclude that

$$Y \,|\, _{\mathbf{S}=\mathbf{s}} = g_Y(\mathbf{s}, Z \,|\, _{\mathbf{S}=\mathbf{s}}) + \tilde{N}_Y \,.$$

Again, this contradicts Theorem 20. The correctness of Phase 2 follows from causal minimality and Lemma 38. ■

References

S. Acid and L. M. de Campos. Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 18:445–490, 2003.

W. P. Bergsma. *Testing Conditional Independence for Continuous Random Variables*, 2004. EURANDOM-report 2004-049.

K. A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.

P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *ArXiv e-prints (1207.5136)*, 2013.

D. M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 1995.

D. M. Chickering. Learning Bayesian networks is NP-complete. In *Learning from Data: Artificial Intelligence and Statistics V*. Springer-Verlag, 1996.

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

P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36: 287–314, 1994.

G. Darmois. Analyse générale des liaisons stochastiques. *Revue de l'Institut International de Statistique*, 21:2–8, 1953.

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B*, 41(1):1–31, 1979.

Deutscher Wetterdienst. Climate data. `http://www.dwd.de/`, 2008.

M. J. Druzdzel and H. van Leijen. Causal reversibility in Bayesian networks. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(1):45–62, 2001.

F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

N. Friedman and D. Koller. Being Bayesian about Bayesian network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125, 2003.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 2008.

D. Geiger and D. Heckerman. Learning Gaussian networks. In *Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 1994.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 2008.

D. M. A. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16:342–355, 1988.

D. Heckerman. A Bayesian approach to causal discovery. Technical report, Microsoft Research (MSR-TR-97-05), 1997.

D. Heckerman and D. Geiger. Likelihoods and parameter priors for Bayesian networks. Technical report, Microsoft Research (MSR-TR-95-54), 1995.

P. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008.

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 21 (NIPS)*, 2009.

A. Hyvärinen and S. M. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, 14:111–152, 2013.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56:5168–5194, 2010.

D. Janzing and B. Steudel. Justifying additive-noise-model based causal discovery via algorithmic information theory. *Open Systems and Information Dynamics*, 17:189–212, 2010.

D. Janzing, J. Peters, J. M. Mooij, and B. Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, Tokyo, Japan, 2003.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, 2009.

S. Lauritzen. *Graphical Models.* Oxford University Press, New York, 1996.

C. Meek. *Graphical Models: Selecting Causal and Statistical Models.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1997.

J. M. Mooij and D. Janzing. Distinguishing between cause and effect. *Journal of Machine Learning Research W&CP*, 6:147–156, 2010.

J. M. Mooij, D. Janzing, J. Peters, and B. Schölkopf. Regression by dependence minimization and its application to causal inference. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009.

J. M. Mooij, D. Janzing, T. Heskes, and B. Schölkopf. On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems 24 (NIPS)*, 2011.

J. M. Mooij, D. Janzing, and B. Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. In *Proceedings of the 29th Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

C. Nowzohour and P. Bühlmann. Score-based causal learning in additive noise models. *ArXiv e-prints (1311.6359)*, 2013.

OEIS Foundation Inc. The on-line encyclopedia of integer sequences. `http://oeis.org/A003024`, 2011.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

J. Peters. Asymmetries of time series under inverting their direction. Diploma Thesis, University of Heidelberg, 2008. `http://stat.ethz.ch/people/jopeters`.

J. Peters. *Restricted Structural Equation Models for Causal Inference*. PhD thesis, ETH Zurich and MPI for Intelligent Systems, 2012. `http://dx.doi.org/10.3929/ethz-a-007597940`.

J. Peters. On the intersection property of conditional independence and its application to causal discovery. *ArXiv e-prints (1403.0408)*, 2014.

J. Peters and P. Bühlmann. Structural intervention distance (SID) for evaluating causal graphs. *ArXiv e-prints (1306.1043)*, 2013.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101:219–228, 2014.

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: 2436–2450, 2011a.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011b.

J. Ramsey, J. Zhang, and P. Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

C. E. Rasmussen and H. Nickisch. Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research*, 11:3011–3015, 2010.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4): 962–1030, 2002.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

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

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. Hoyer, and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

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

V.P. Skitovič. Linear forms in independent random variables and the normal distribution law (in Russian). *Izvestiia AN SSSR, Ser. Matem.*, 18:185–200, 1954.

V.P. Skitovič. Linear combinations of independent random variables and the normal distribution law. *Select. Transl. Math. Stat. Probab.*, 2:211–228, 1962.

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

Y. Tamada, S. Imoto, H. Araki, M. Nagasaki, C. G. Print, S. D. Charnock-Jones, and S. Miyano. Estimating genome-wide gene networks using nonparametric Bayesian network models on massively parallel computers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):683–697, 2011a.

Y. Tamada, S. Imoto, and S. Miyano. Parallel algorithm for learning optimal Bayesian network structure. *Journal of Machine Learning Research*, 12:2437–2459, 2011b.

M. Teyssier and D. Koller. Ordering-based search: a simple and effective algorithm for learning Bayesian networks. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *Annals of Statistics*, 41(2):436–463, 2013.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 1991.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2003.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271, 2008.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.