# Conditions and Assumptions for Constraint-based Causal Structure Learning

**Kayvan Sadeghi**                                                    K.SADEGHI@UCL.AC.UK
*Department of Statistical Science*
*University College London*
*London, United Kingdom*

**Terry Soo**                                                        MATH@TERRYSOO.COM
*Department of Statistical Science*
*University College London*
*London, United Kingdom*

## Abstract

We formalize constraint-based structure learning of the "true" causal graph from observed data when unobserved variables are also existent. We provide conditions for a "natural" family of constraint-based structure-learning algorithms that output graphs that are Markov equivalent to the causal graph. Under the faithfulness assumption, this natural family contains all exact structure-learning algorithms. We also provide a set of assumptions, under which any natural structure-learning algorithm outputs Markov equivalent graphs to the causal graph. These assumptions can be thought of as a relaxation of faithfulness, and most of them can be directly tested from (the underlying distribution) of the data, particularly when one focuses on structural causal models. We specialize the definitions and results for structural causal models.

**Keywords:**  ancestral graphs; causal discovery; constraint-based structure learning; faithfulness; structural causal models

## 1. Introduction

Inferring causal relationships has always been one of the main objectives in many fields of study, ranging from philosophy to science (Holland, 1986). In the past decades, *causal models* (Pearl, 2009) have been used to infer causal relationships from observed data. Extensive research has been conducted on defining, interpreting, and applying causal models, and these models have become mainstream in statistics and computer science (Pearl, 1988; Spirtes et al., 2000; Dawid, 2000; Hitchcock, 2018). Today, a very popular method for inferring causal relationships is based on the use of statistical models over graphs with nodes that are random variables representing the quantities of interest. Edges indicate probabilistic dependence among these variables conditional on some other variables in the graph, which with some additional assumptions can be interpreted as causal relationships. These models are called *graphical (Markov) models* (Lauritzen, 1996).

*Structural causal models* (SCMs), also known as *structural equation models*, are one of the most used causal models (Spirtes et al., 2000; Pearl, 1988). SCMs have been generally

defined for directed acyclic graphs (DAGs) (i.e., a Bayesian network), where an arrow implies a "direct" cause. However, for the purpose of causal inference, it may not be possible to determine or measure all common causes. For this reason, SCMs associated to DAG models with latent variables (represented by ancestral graphs (Richardson and Spirtes, 2002) without undirected edges) have been defined (Spirtes et al., 2000; Zhang, 2008).

A main goal of causal inference, based on graphical models, is causal discovery; in this framework we are concerned with finding graphs, using only observed data, that represent the causal structure, which is represented by an unknown "true" causal graph. Algorithms for constructing or selecting such a graph are known as *structural learning* algorithms. There has been a surge in defining different structural learning algorithms; they are generally categorized as constraint-based (Spirtes et al., 2000), score-based (Chickering, 2003) –which work on the factorization of density–, or hybrid (Tsamardinos et al., 2006) –which is a combination of these.

We are concerned with *constraint-based* structure-learning algorithms for the general case of ancestral graphs without undirected edges. These algorithms use conditional independence statements, which are theoretically testable from observed data to find graphs that, at best, are Markov equivalent to the true causal graph. A main sufficient condition for the algorithms to be *correct* or *work*, that is, to provide an output that is Markov equivalent to the causal graph, is faithfulness (Zhang and Spirtes, 2008; Woodward, 1998; Steel, 2006; Colombo and Maathuis, 2014); this sufficient condition relates the independencies induced by the distribution (of the SCM) and the separations induced by the causal graph. However, faithfulness to the unknown causal graph, or similar assumptions (Zhang and Spirtes, 2003), are generally not testable. In addition, the (sufficient) assumption of faithfulness is in some circumstances strong, and can be relaxed; see Section 5.

In this paper, we formalize the problem of constraint-based structure learning from observed data. Most constraint-based structure-learning algorithms contain two tasks which may be performed in parallel or separately: finding the undirected edges of the graph, the skeleton; and directing the edges of the skeleton. We show that under the assumption of faithfulness and using a formalization of faithfulness, provided in Sadeghi (2017), if the output of such algorithms works, then pair of conditions known as *ordered stability* must be satisfied with respect to the output; see Section 2.3 and Remark 12. Ordered stability is a condition that informs us how the edges of the graph should be directed and we propose the class of *natural* structure-learning algorithms, which is one that should respect stability when directing the edges; see Definition 6. We stress that in this paper, we do not provide specific algorithms that could be implemented, but concentrate on providing conditions that structure-learning algorithms should satisfy. More importantly, we provide clear assumptions, many of which are directly testable from the distribution, under which a natural structure-learning algorithm generates the correct output.

We provide the theory for the general class of models under the assumption that the distribution is Markovian to the true causal graph; see Theorem 25. The assumptions are the following: the distribution satisfies ordered stability and the converse of the pairwise Markov property with respect to the true causal graph, and also the distribution satisfies what we call path-stability.

We will make precise our assumptions later, and we note that they can be thought of as a relaxation of faithfulness, and in particular, in the case of DAGs our assumptions are

logically implied by faithfulness; see Corollary 23. We also specialize the definitions and results for SCMs; see Theorem 42, and, in particular, provide testable sufficient conditions for the converse of the pairwise Markov property to be satisfied. We hope that the framework provided here will be a useful theoretical tool in defining appropriate SCMs and concrete algorithms with strong theoretical backing.

The structure of the paper is as follows: In the next section, we provide known definitions and results needed for the theory presented in this paper. In Section 3, we formalize the settings of the problem. Section 4 deals with the theory under the faithfulness assumption. In Section 5, we introduce the structure learning theory without the faithfulness assumption. Section 6 specializes on causal structure learning for SCMs. We end the paper with a summary and discussion in Section 7.

## 2. Preliminaries

In this section, we provide the basic definitions and concepts as well as the known results needed for the paper.

### 2.1 Graph theoretical definitions and concepts

We usually refer to a graph as an ordered pair $G = (V, E)$, where $V$ is the *node* set and $E$ is the *edge* set. When nodes $i$ and $j$ are the endpoints of an edge, we call them *adjacent*, and write $i \sim j$, and otherwise $i \nsim j$. Graphs in this paper are *simple*, that is, they have at most one edge between a pair of nodes. We consider two types of edges: *arrows* $(i \longrightarrow j)$ and *bidirected edges* or *arcs* $(i \leftarrow\!\!\rightarrow j)$. We do not consider graphs that have simultaneous third type of edge: *undirected edges* or *lines* $(i \text{───} j)$. The completely undirected graph, obtained from a graph $G$, is called the *skeleton* of $G$, denoted by $\mathrm{sk}(G)$; it is obtained by removing all arrowheads from the edges of the graph $G$. We say that we *direct* the edges of a skeleton by assigning arrowheads at the edges, which results in a graph.

A *subgraph* of a graph $G_1$ is graph $G_2$ such that $V(G_2) \subseteq V(G_1)$ and $E(G_2) \subseteq E(G_1)$ and the assignment of endpoints to edges in $G_2$ is the same as in $G_1$. An *induced subgraph* by nodes $A \subseteq V$ is a subgraph that contains all and only nodes in $A$ and all edges between two nodes in $A$.

A *walk* is a list $\langle v_0, e_1, v_1, \ldots, e_k, v_k \rangle$ of nodes and edges such that for $1 \leq i \leq k$, the edge $e_i$ has endpoints $v_{i-1}$ and $v_i$. A *path* is a walk with no repeated node or edge. When we define a path, we only write the nodes (and not the edges). A *cycle* is a walk with no repeated nodes or edges except for $v_0 = v_k$.

We call the first and the last nodes *endpoints* of the path and all other nodes *inner nodes*. A path can also be seen as a certain type of connected subgraph of $G$; a *subpath* of a path $\pi$ is an induced connected subgraph of $\pi$.

For an arrow $j \longrightarrow i$, we say that the arrow is *from $j$ to $i$*. We also call $j$ a *parent* of $i$, $i$ a *child* of $j$ and we use the notation $\mathrm{pa}(i)$ for the set of all parents of $i$ in the graph. In the cases of $i \longrightarrow j$ or $i \leftarrow\!\!\rightarrow j$ we say that there is an arrowhead at $j$ or pointing to $j$. A path $\langle i = i_0, i_1, \ldots, i_n = j \rangle$ is *directed* from $i$ to $j$ if all $i_k i_{k+1}$ edges are arrows pointing from $i_k$ to $i_{k+1}$. If there is a directed path from $j$ to $i$, then node $j$ is an *ancestor* of $i$ and $i$ is a *descendant* of $j$. We denote the set of ancestors of $i$ by $\mathrm{an}(i)$; unlike some authors, we do not assume that $i \in \mathrm{an}(i)$.

A *tripath* is a path with three nodes. Note that Sadeghi (2013) used the term V-configuration for such a path. We follow Kiiveri et al. (1984) and most texts by letting a *V-configuration* be a tripath with non-adjacent endpoints. The inner node $t$ in each of the three tripaths

$$i \longrightarrow t \longleftarrow j, \; i \longleftrightarrow t \longleftarrow \; j, \; i \longleftrightarrow t \longleftrightarrow j$$

is a *collider* (or a collider node) and the inner node of any other tripath is a *non-collider* (or a non-collider node) on the tripath or, more generally, on any path of which the tripath is a subpath; i.e., a node is a collider if two arrowheads meet. Notice that in some causal inference literature, the term *V-structure* is used for a collider V-configuration. A path is called a *collider* path if all its inner nodes are colliders.

In this paper, we are mainly concerned with *directed ancestral graphs*; these are *ancestral graphs* (AGs) (Richardson and Spirtes, 2002) without lines, i.e., graphs with arrows and arcs without a directed cycle and without an arc with one endpoint being an ancestor of the other. Sometimes we simply refer to such graphs as "graphs," dropping any modifiers. Some results we use in this paper have originally been proven for the more general classes of chain mixed graphs (CMGs) and anterial graphs (AnGs) (Sadeghi, 2016) as well as acyclic directed mixed graphs (ADMGs) (Richardson, 2003), all of which contain directed ancestral graphs.

The class of directed ancestral graphs also contains *bidirected graphs* (BGs), containing only bidirected edges, and *directed acyclic graphs* (DAGs), containing only arrows and being acyclic. DAGs have been particularly useful to describe causal Markov relations; see for example Kiiveri et al. (1984); Pearl (1988); Lauritzen and Spiegelhalter (1988); Geiger et al. (1990); Spirtes et al. (2000). For an extensive discussion on the subclasses of acyclic graphs and their relationships and hierarchy, see Lauritzen and Sadeghi (2018).

## 2.2 Independence models and their properties

An *independence model* $\mathcal{J}$, over a finite set $V$, is a set of triples $\langle X, Y \mid Z \rangle$, called *independence statements*, where $X$, $Y$, and $Z$ are disjoint subsets of $V$; $Z$ may be empty, but $\langle \varnothing, Y \mid Z \rangle$ and $\langle X, \varnothing \mid Z \rangle$ are always included in $\mathcal{J}$. The independence statement $\langle X, Y \mid Z \rangle$ is read as "$X$ is independent of $Y$ given $Z$." Independence models may have a probabilistic interpretation, but this is not necessarily the case. Similarly, not all independence models can be easily represented by graphs. For further discussion on general independence models, see Studený (2005).

In order to define probabilistic independence models, consider a set $V$ and a collection of random variables $\{X_\alpha\}_{\alpha \in V}$ with state spaces $\{\mathcal{X}_\alpha\}_{\alpha \in V}$ and joint distribution $P$. We let $X_A = \{X_v\}_{v \in A}$ for each subset $A$ of $V$. For disjoint subsets $A$, $B$, and $C$ of $V$ we use the short notation $A \perp\!\!\!\perp B \mid C$ to denote that $X_A$ is *conditionally independent of $X_B$ given $X_C$* (Dawid, 1979; Lauritzen, 1996), that is, for any measurable $S \subseteq \mathcal{X}_A$ and $P$-almost all $x_B$ and $x_C$, we have

$$\mathbb{P}(X_A \in S \mid X_B = x_B, X_C = x_C) = \mathbb{P}(X_A \in S \mid X_C = x_C),$$

where we make suitable interpretations in terms of regular conditional probabilities when necessary; see for example Chang and Pollard (1997). We can now induce an independence

model $\mathcal{J}(P)$ by letting

$$\langle A, B \,|\, C\rangle \in \mathcal{J}(P) \iff A \perp\!\!\!\perp B \,|\, C \text{ w.r.t. } P.$$

Similarly we use the notation $A \not\perp\!\!\!\perp B \,|\, C$ for $\langle A, B \,|\, C\rangle \notin \mathcal{J}(P)$. In this paper, we present the results for "$P$," but in fact all the conditions and assumptions are on $\mathcal{J}(P)$. We remark that all the results and conditions could be presented using an abstract independence model $\mathcal{J}$.

In order to use graphs to represent independence models, the notion of *separation* in a graph is fundamental. A path $\pi$ is *connecting* given $C$ if every collider node of $\pi$ is in $C$ or an ancestor of a node in $C$, and every non-collider node is not in $C$. For pairwise disjoint subsets $(A, B, C)$, we write $A \perp B \,|\, C$ if there are no connecting paths between $A$ and $B$ given $C$, and say that $A$ and $B$ are *separated given $C$*. We will also use the notation $A \not\perp B \,|\, C$ for $A$ and $B$ not separated given $C$. This separation criterion is generally known as the *m-separation*, used in Richardson and Spirtes (2002), Wermuth (2011), and Wermuth and Sadeghi (2012), for ancestral graphs and summary graphs, and is a generalization of the *d-separation* of Pearl (1988).

If $A$, $B$, or $C$ has only one member $\{i\}$, $\{j\}$, or $\{k\}$, for better readability, we write $\langle i, j \,|\, k\rangle \in \mathcal{J}$ instead of $\langle \{i\}, \{j\} \,|\, \{k\}\rangle \in \mathcal{J}$; and similarly for $i \perp j \,|\, k$ and $i \perp\!\!\!\perp j \,|\, k$. We also write $A \perp B$ when $C = \varnothing$; and similarly $A \perp\!\!\!\perp B$.

A graph $G$ induces an independence model $\mathcal{J}(G)$ by separation, letting

$$\langle A, B \,|\, C\rangle \in \mathcal{J}(G) \iff A \perp B \,|\, C \text{ in } G.$$

A graph is called *maximal* if the absence of an edge between $i$ and $j$ corresponds to a conditional separation statement for $i$ and $j$, i.e., there exists for some $C$ a statement of form $i \perp j \,|\, C$. In addition, we call two graphs $G$ and $H$ *Markov equivalent* if $\mathcal{J}(G) = \mathcal{J}(H)$. Conditions for Markov equivalence for most classes of graphs are known; see Verma and Pearl (1990); Ali et al. (2009); Wermuth and Sadeghi (2012). Notice that two Markov equivalent maximal graphs have the same skeleton. We require conditions for Markov equivalence of the classes of graphs discussed here. A collider path $\pi = \langle i, B, j\rangle$ is called a *minimal collider path* if $i \nsim j$ and there is no proper subset $B' \subset B$ such that $\langle i, B', j\rangle$ is a collider path between $i$ and $j$.

**Lemma 1 (Zhao et al. (2004))** *Two maximal ancestral graphs (and consequently maximal directed ancestral graphs) are Markov equivalent if and only if they have the same skeletons and minimal collider paths.*

For DAGs, there is a simpler condition for Markov equivalence:

**Lemma 2 (Verma and Pearl (1990))** *Two DAGs are Markov equivalent if and only if they have the same skeletons and collider V-configurations.*

An independence model $\mathcal{J}$ over a set $V$ is a *semi-graphoid* if it satisfies the four following properties for disjoint subsets $A$, $B$, $C$, and $D$ of $V$:

1. $\langle A, B \,|\, C\rangle \in \mathcal{J}$ if and only if $\langle B, A \,|\, C\rangle \in \mathcal{J}$ (*symmetry*);

2. if $\langle A, B \cup D \,|\, C\rangle \in \mathcal{J}$, then $\langle A, B \,|\, C\rangle \in \mathcal{J}$ and $\langle A, D \,|\, C\rangle \in \mathcal{J}$ (*decomposition*);

3. if $\langle A, B \cup D \,|\, C \rangle \in \mathcal{J}$, then $\langle A, B \,|\, C \cup D \rangle \in \mathcal{J}$ and $\langle A, D \,|\, C \cup B \rangle \in \mathcal{J}$ (*weak union*);

4. if $\langle A, B \,|\, C \cup D \rangle \in \mathcal{J}$ and $\langle A, D \,|\, C \rangle \in \mathcal{J}$, then $\langle A, B \cup D \,|\, C \rangle \in \mathcal{J}$ (*contraction*).

Notice that the reverse implication of contraction clearly holds by decomposition and weak union. A semi-graphoid for which the reverse implication of the weak union property holds is said to be a *graphoid*; that is, it also satisfies

5. if $\langle A, B \,|\, C \cup D \rangle \in \mathcal{J}$ and $\langle A, D \,|\, C \cup B \rangle \in \mathcal{J}$, then $\langle A, B \cup D \,|\, C \rangle \in \mathcal{J}$ (*intersection*).

Furthermore, a graphoid or semi-graphoid for which the reverse implication of the decomposition property holds is said to be *compositional*, that is, it also satisfies

6. if $\langle A, B \,|\, C \rangle \in \mathcal{J}$ and $\langle A, D \,|\, C \rangle \in \mathcal{J}$, then $\langle A, B \cup D \,|\, C \rangle \in \mathcal{J}$ (*composition*).

Another important property of independence models is *singleton-transitivity* (also called *weak transitivity* in Pearl (1988)). For $i$, $j$, and $k$, single elements in $V$,

7. if $\langle i, j \,|\, C \rangle \in \mathcal{J}$ and $\langle i, j \,|\, C \cup \{k\} \rangle \in \mathcal{J}$, then $\langle i, k \,|\, C \rangle \in \mathcal{J}$ or $\langle j, k \,|\, C \rangle \in \mathcal{J}$ (singleton-transitivity).

Probabilistic independence models are always semi-graphoids (Pearl, 1988), whereas the converse is not necessarily true; see Studený (1989). If $P$ has strictly positive density, the induced independence model is always a graphoid; see, for example, Proposition 3.1 in Lauritzen (1996). See also Peters (2015) for a necessary and sufficient condition for the intersection property to hold. If the distribution $P$ is a regular multivariate Gaussian distribution, then $\mathcal{J}(P)$ is a singleton-transitive compositional graphoid; for example see Studený (2005) and Pearl (1988). For this reason, a different axiomatization of singleton-transitive compositional graphoid is called a *Gaussoid* in Lněnička and Matúš (2007). On the other hand, separation in graphs satisfies all these properties; see Propositions 1 and 2 in Sadeghi (2017). If $\mathcal{J}(P)$ satisfies any of these properties, then we may simply say that "P satisfies that property."

### 2.3 Ordered upward- and downward-stabilities

The definitions and results below are a special case of those in Section 4 of Sadeghi (2017). Consider a *partial order* $\leq$ over the set $V$. If $a \leq b$ or $b \leq a$, then $a$ and $b$ are *comparable*; otherwise they are *incomparable*. Henceforth, we speak of "orders" referring to "partial orders."

We say that a graph $G = (V, E)$ admits a *valid order* $\leq$ if for nodes $i$ and $j$ of $G$, $i \longrightarrow j$ implies that $i > j$; and $i \leftarrow\!\!\!\rightarrow j$ implies that $i$ and $j$ are incomparable. Notice that this specifies the partial order via its cover relations. There may be many different orders that are valid for a graph. However, we obtain a unique ordering for a graph as follows: Given a graph $G$, if $i \notin \mathrm{an}(j)$ and $j \notin \mathrm{an}(i)$, then set $i$ and $j$ to be incomparable. Otherwise, let $i > j$ when $i \longrightarrow j$. We call this ordering the *minimal order* for $G$ since it gives the fewest possible comparable pairs of nodes. It can also be seen that if $G$ is ancestral, then the minimal order for $G$ is a valid order.

We now exploit the ordering for independence models in order to define two other properties in addition to the seven properties defined in Section 2.2 (namely singleton-transitive compositional graphoid axioms). We say that an independence model $\mathcal{J}$ over the

set $V$ respectively satisfies *ordered upward-* and *downward-stability* w.r.t. an order $\leq$ of $V$ if the following hold:

8. if $\langle i,j \,|\, C \rangle \in \mathcal{J}$, then $\langle i,j \,|\, C \cup \{k\} \rangle \in \mathcal{J}$ for every $k \in V \setminus \{i,j\}$ such that $i < k$ or $j < k$ (ordered upward-stability);

9. if $\langle i,j \,|\, C \rangle \in \mathcal{J}$, then $\langle i,j \,|\, C \setminus \{k\} \rangle \in \mathcal{J}$ for every $k \in V \setminus \{i,j\}$ such that $i \not< k$, $j \not< k$, and $l \not< k$ for every $l \in C \setminus \{k\}$ (ordered downward-stability).

Sometimes we refer to the pair of Conditions 8 and 9, together, simply as *ordered stability*.

Ordered upward-stability is a generalization of a modification of *upward stability*, defined in Fallat et al. (2017), and *strong union*, defined in Pearl and Paz (1985), for undirected graphs. A similar concept was used in Claassen and Heskes (2012) for the class of DAGs for inferring causal relations; see Lemma 2 of the mentioned manuscript.

Sadeghi (2017) proved that the independence model induced by ancestral graphs satisfies ordered upward- and downward-stability w.r.t. their minimal order. If an independence model $\mathcal{J}$ satisfies ordered upward- and downward-stability w.r.t. the minimal order of a graph $G$, then we simply say that "$\mathcal{J}$ satisfies upward- and downward-stability w.r.t. $G$." In addition, similar to compositional graphoids, if $\mathcal{J}(P)$ satisfies ordered upward- or downward-stability, then we may simply say that "$P$ satisfies that property."

## 2.4 Markov and faithful independence models

For a graph $G = (V,E)$, a probability distribution defined over $V$ satisfies the *global Markov property* w.r.t. $G$, or is simply *Markovian* to $G$, if for disjoint subsets $A$, $B$, and $C$ of $V$ it holds that

$$A \perp B \,|\, C \implies A \perp\!\!\!\perp B \,|\, C,$$

or equivalently $\mathcal{J}(G) \subseteq \mathcal{J}(P)$. Notice that every probability distribution over $V$ is Markovian to the complete graph with the node set $V$.

We say that $P$ and a graph $G$ are *faithful* if $\mathcal{J}(P) = \mathcal{J}(G)$. If $P$ and $G$ are faithful, then we may sometimes also say that $P$ is *faithful to* $G$ or vice versa, although in principle faithfulness is a symmetric relation. Thus, if $P$ and $G$ are faithful, then in addition to $P$ being Markovian to $G$, every conditional independence statement corresponds to a separation in $G$. Notice that, originally in Spirtes et al. (2000), faithfulness was defined only to be the opposite direction to the Markov property. If there is a graph $G$ and a distribution $P$ that are faithful, then we say that $P$ is *graphical* and $G$ is *probabilistic*.

For a given probability distribution $P$, we define the *skeleton* of $P$, denoted by $\mathrm{sk}(P)$, to be the undirected graph that is obtained from $P$ as follows: we define the node set of $\mathrm{sk}(P)$ to be $V$, and for every pair of nodes $i,j$, we check whether $i \perp\!\!\!\perp j \,|\, C$ holds for some $C \subseteq V \setminus \{i,j\}$; if it does not then we draw an edge between $i$ and $j$. The resulting undirected graph is the same as the one obtained by the first step of the SGS algorithm (Glymour et al., 1987).

Suppose that there exists an order $\leq$ over $V$. We can uniquely direct the edges of $\mathrm{sk}(P)$ based on $\leq$ in order to define a $G(P, \leq)$ induced by $P$ and $\leq$. It can be seen that $G(P, \leq)$ is ancestral.

We are only interested in orderings that are minimal orders of graphs. Given $P$, we define an ordering $\leq$ to be *P-compatible* if $\leq$ is the minimal order for $G(P, \leq)$.

If $P$ is Markovian to a graph $G$, then $\mathrm{sk}(P)$ is a subgraph of $\mathrm{sk}(G)$; see Sadeghi (2017). Hence, if $P$ is Markovian to a graph $G$ such that $\mathrm{sk}(G) = \mathrm{sk}(P)$, then $G$ has the fewest number of edges among those to which $P$ is Markovian. We say that $P$ is *minimally Markovian* to a graph $G$ if $P$ is Markovian to $G$ and $\mathrm{sk}(G) = \mathrm{sk}(P)$ (Sadeghi, 2017). Notice that only minimally Markovian independence models to a graph can also be faithful to the graph.

**Remark 3 (Related versions of minimality)** *There are several similarly defined minimality assumptions in the literature. Notice that we do not define the concept of minimal Markovian as an assumption in this paper, since minimality is implied by some other assumptions used in this paper; see Corollary 15.*

*The definition given here is subtly different from the (causal) minimality assumption (Pearl, 2009; Spirtes et al., 2000; Zhang and Spirtes, 2008; Neapolitan, 2004), which holds if $P$ is Markovian to $G$, but not to any proper subgraph of $G$ — in fact, it can be seen that being minimal Markovian implies causal minimality. It is also different from the P-minimality assumption (Zhang, 2012), which states that no proper I-structure of $G$ is Markovian to $P$, where I-structure of $G$ is any subgraph of $G$ whose induced independence model is a subset of that of $G$. Again, being minimal Markovian implies P-minimality since there could be a graph with fewer edges than $G$ that is not an I-structure of $G$.*

*The sparsest Markov representation (SMR) assumption (Raskutti and Uhler, 2018) for $G$ and $P$ (originally defined for DAGs) states that $P$ is Markovian to $G$ and every $G'$ to which $P$ is Markovian and not Markov equivalent to $G$ has more edges than $G$ does. A similar concept is the frugality assumption (Forster et al., 2018). In the special case of interest here, where $G$ is Markovian to $P$, it states that $G$ satisfies frugality if it is in the set of sparsest Markov representations. If $G$ and $P$ satisfy SMR, then $G$ is not necessarily minimally Markovian to $P$ (see Figure 5 of Forster et al. (2018) as an example). It is also possible to create two non-Markov equivalent graphs with the same skeleton, and a $P$ that is Markovian to both. This implies that the minimal Markovian assumption does not imply SMR either.*

We make use of the necessary and sufficient conditions for faithfulness to directed ancestral graphs, proven in Sadeghi (2017) for the more general case of anterial graphs:

**Proposition 4 (Sadeghi (2017))** *Let $P$ be a probability distribution defined over $X_V$. Then $P$ is graphical if and only if*

*(A) $P$ is a singleton-transitive compositional graphoid; and*

*(B) there exists a P-compatible order $\lesssim$ over $V$ w.r.t. which $P$ satisfies ordered downward- and upward-stability.*

*In addition, if (A) and (B) are satisfied, then $P$ is faithful to $G(P, \lesssim)$.*

## 2.5 Structural causal models

Here we define the *structural causal models* (SCMs) (also known as the *structural equation models* (SEMs)) (Pearl, 2009; Spirtes et al., 2000) for the class of directed ancestral graphs, as introduced in Richardson and Spirtes (2002). Consider a graph $G$ with $N$ nodes, which

in the context of causal inference may be referred to as the "true causal graph." An SCM $\mathfrak{C}$ associated with $G$ is defined as a collection of $N$ equations

$$X_i = \phi_i(X_{\text{pa}(i)}, \epsilon_i), \quad i \in \{1, \ldots, N\}, \tag{1}$$

where $\text{pa}(i)$ is defined on $G$, and, for any subsets $A$ and $B$, we require that $\epsilon_A \perp\!\!\!\perp \epsilon_B$ if and only if, in $G$, there is no arc between any node in $A$ and any node in $B$. Notice that in the more widely-used case where $G$ is a DAG, all the $\epsilon_i$ are jointly independent. Sometimes the $X_i$ are called *endogenous variables* and the *noises* $\epsilon_i$ are also called *exogenous variables* (which are generally assumed to be latent). For both mathematical and causal discussions on SCMs with DAGs, see Peters et al. (2017). In the case where $\phi_i$ represent multivariate linear regressions, SCMs are also known as the system of linear equations (Cox and Wermuth, 1996).

Notice that the minimal order of $G$ induces an order for $X_i$s in $\mathfrak{C}$. In addition, by using the equations iteratively, it is easy to see that each $X_i$ can be written as a function of $\epsilon_{\text{an}(i) \cup \{i\}}$. An SCM also entails a unique joint distribution $P$ over the variables $(X_1, \ldots, X_N)$.

A more complete representation of the graph associated with an SCM is obtained by including the random variables $\epsilon_i$. These graphs, called *augmented graphs*, are denoted by $G(X, \epsilon)$. We construct the augmented graph in our setting as follows: we denote the nodes in $G$ by the original notation $\{X_1, \ldots, X_N\}$; we add the nodes $\{\epsilon_1, \ldots, \epsilon_N\}$ and add arrows from each $\epsilon_i$ to $X_i$; we also remove all arcs between $X_i$s, and instead of every $X_i X_j$-arc, add an arc between $\epsilon_i$ and $\epsilon_j$.

For example, in Figure 1, the graph on the right, is the augmented graph of the, graph on the left. An SCM associated to these graphs, in which $\phi_i$ are linear is as follows:

$$X_4 = \epsilon_4, \quad X_3 = \epsilon_3, \quad X_2 = \alpha X_4 + \epsilon_2, \quad X_1 = \beta X_3 + \epsilon_1.$$



Figure 1: A graph $G$ (left) and its augmented graph $G(X, \epsilon)$ (right).

The above setting and construction of the augmented graph is in line with the idea originally presented in Richardson and Spirtes (2002) for linear functions and Gaussian distributions, although unlike in the mentioned manuscript, noises corresponding to endpoints of variables adjacent by an arc are *required* to be dependent. This setting is different from the dominant setting in the literature (e.g. Peters et al. (2017)) for directed ancestral graphs, where for every arc $X_i X_j$ it is assumed that there exists an exogenous variable $\epsilon_{ij}$, where, in the directed acyclic augmented graph, it points to both $X_i$ and $X_j$, and that $\epsilon_{ij}$s are mutually independent.

**Remark 5** *Our setting is more general than the other alternative presented above. Notice that we can write a noise $\epsilon_i$ as a tuple $(\epsilon_{is})_{s \in S}$, where $S$ is the set of nodes connected to $i$ by an arc. However, by writing $\epsilon_{ij} = (\epsilon_i, \epsilon_j)$, we would not obtain independent noises in the other setting.*

## 3. General setting for structure learning

We start by assuming the existence of an *unknown* distribution $P$, which is Markovian to an unknown *(true) causal graph* $G_0$. Indeed, this setting requires additional models and assumptions to be considered "causal." This stems from the fact that, in the causal model, interventional probabilities and observational conditional probabilities coincide. For more discussion on this, see Pearl (2009); Spirtes et al. (2000); Peters et al. (2017). Nevertheless, the setting presented here is what is required for causal discovery using conditional independence under any causal model.

The main goal of structure learning is to use observational data to find a graph that belongs to the Markov equivalence class of $G_0$. The distribution $P$ induces an independence model $\mathcal{J}(P)$ whose elements we can test for using the observed data. Although we can, in principle, test each statement in $\mathcal{J}(P)$, testing for all independence statements is computationally intractable. In a constraint-based structural learning algorithm, one uses some of the independence statements in $\mathcal{J}(P)$ to construct a graph, which is (hopefully) in the Markov equivalence class of $G_0$. Notice that, in this paper, we concern ourselves only with the oracle setting, i.e., assuming we have an oracle that will always correctly tell us whether or not a given conditional independence relation holds in the distribution.

Indeed the exact relationship between $P$ (or more specifically $\mathcal{J}(P)$) and the constructed graph depends on the specific algorithm used. Here, we propose the following class of generic algorithms.

**Definition 6 (Natural structure-learning algorithm)** *Let $P$ be a joint probability distribution on $X_V$. We say that a natural constraint-based structure-learning algorithm is an algorithm whose output graph with node set $V$, denoted by $G(P)$, has the following properties:*

1. *It holds that $\mathrm{sk}(G(P)) = \mathrm{sk}(P)$;*

2. *The distribution $P$ satisfies ordered downward- and upward-stability w.r.t. $G(P)$, that is, $G(P) = G(\mathcal{J}(P), \leq)$, for a partial order $\leq$, w.r.t. which $P$ satisfies ordered downward- and upward-stability.*

In short, the diagram below shows the relationship between the concepts used in this setting:

$$G_0 \xleftarrow{\text{Markovian}} P \xrightarrow{\text{conditional independence}} \mathcal{J}(P) \xrightarrow{\text{Algorithm}} G(P) \overset{\text{Markov equivalent}}{\sim} G_0$$

In the same fashion as many constraint-based structural learning algorithms, the skeleton of the output of a natural structure-learning algorithm is the same as $\mathrm{sk}(P)$. Under the faithfulness assumption of $P$ and $G_0$, we will show in Section 4 that not only is $G(P)$ Markov equivalent to $G_0$, but also $P$ satisfies ordered downward- and upward-stability w.r.t. every graph Markov equivalent to $G_0$; thus the class of natural structure-learning algorithms

contains all well-known "exact" algorithms that work under faithfulness, so we use term "natural." Under alternative conditions, provided in Section 5, we show that the output $G(P)$ is Markov equivalent to $G_0$. Notice again that we will always use $G(P)$ to denote an output of a natural structure-learning algorithm.

The algorithms in the literature, generally generate a completed partially directed acyclic graph (CPDAGs) (Andersson et al., 1997; Spirtes et al., 2000) or partial ancestral graphs (PAGs) (Richardson and Spirtes, 2002; Ali et al., 2009), which are a representative of the Markov equivalence class of $G_0$ with the fewest arrowheads. Here, we are not concerned with making a CPDAG or a PAG from $G(P)$.

We assume, without loss of generality, that the causal graph $G_0$ is maximal: the reason is that if $G_0$ is not maximal, then there is always a maximal graph that is Markov equivalent to $G_0$, which could be used to find graphs Markov equivalent to $G_0$; see Richardson and Spirtes (2002); Sadeghi and Lauritzen (2014). Note that the first condition of the natural structure-learning algorithm in Definition 6 also ensures that the generated $G(P)$ is maximal.

Since Markov equivalent maximal graphs have the same skeleton, the obtained $G(P)$ must have the same skeleton with $G_0$. We are looking for conditions, under which this holds.

Let $\mathrm{an}(i,j) = (\mathrm{an}(i) \cup \mathrm{an}(j)) \setminus \{i,j\}$. We say that $P$ is *converse pairwise Markovian* to $G_0$ if

$$i \perp\!\!\!\perp j \mid \mathrm{an}(i,j) \Rightarrow i \nsim j \text{ in } G_0; \tag{2}$$

note that the converse of this condition is the well-known pairwise Markov property (Lauritzen, 1996). To ensure the equality of the mentioned skeletons, we need the following lemma:

**Lemma 7** *Suppose that the distribution $P$ is Markovian to the causal graph $G_0$, and that the following two conditions are satisfied:*

*(a) $P$ is converse pairwise Markovian to $G_0$;*

*(b) if there exists $C$, not containing $i$ and $j$, such that $i \perp\!\!\!\perp j \mid C$, then $i \perp\!\!\!\perp j \mid \mathrm{an}(i,j)$.*

*Then $\mathrm{sk}(G_0) = \mathrm{sk}(G(P))$, where $G(P)$ is an output of a natural structure-learning algorithm.*

**Proof** We have on $G_0$ the equivalence

$$i \nsim j \text{ in } G_0 \iff i \perp\!\!\!\perp j \mid \mathrm{an}(i,j),$$

where assumption (a) gives the reverse direction, and the forward direction follows from the fact that $P$ is Markovian to $G_0$, and $\mathrm{an}(i,j)$ always separates non-adjacent $i$ and $j$. Since $\mathrm{sk}(P) = \mathrm{sk}(G(P))$, we have the equivalence

$$i \nsim j \text{ in } G(P) \iff \exists C \text{ s.t. } i \perp\!\!\!\perp j \mid C.$$

Hence, assumption (b) allows us to tie these two conditions together and obtain that

$$i \nsim j \text{ in } G(P) \iff i \nsim j \text{ in } G_0,$$

from which the desired result follows. ■

**Remark 8** *The two conditions of Lemma 7 are equivalent to the adjacency-faithfulness condition (Ramsey et al., 2006; Zhang and Spirtes, 2008) (defined for DAGs), which states that for every edge between $i$ and $j$ in $G_0$, there are no independence statements $i \perp\!\!\!\perp j \mid C$ for any $C$. This breakdown of the adjacency faithfulness assumption allows us to later provide useful sufficient conditions (in general and for SCMs) that ensure that these two conditions are satisfied; see also the discussion below.*

We need some additional assumptions on the distributions to determine when the assumptions of Lemma 7 hold; i.e., when an existing edge implies dependence given the joint ancestors, and when an arbitrary conditional independence of a pair implies their conditional independence given their joint ancestors. A condition under which these assumptions hold is faithfulness, as will be discussed in the next section. However, weaker conditions will be provided in Section 5. Specific conditions for SCMs will also be provided in Section 6.

## 4. Structure learning under faithfulness

Faithfulness of $P$ and $G_0$ is a main assumption in the literature for structure learning. It is well-known that under faithfulness, $P$ is *identifiable*, i.e., an ancestral graph (or a DAG) (in reality a PAG (or a CPDAG)) Markov equivalent to $G_0$ can be identified (Spirtes et al., 2000; Zhang and Spirtes, 2008). Under faithfulness, if $P$ is faithful to $G_0$, then any structure-learning algorithm should provide an output $G(P)$ that is faithful to $P$, in the setting where an oracle can be queried regarding the validity of statements of conditional independence.

By Proposition 4, if the assumption of $P$ being graphical holds, then $P$ is singleton-transitive compositional graphoid and there is an ordering $\leq$ of the nodes of $\text{sk}(P)$ w.r.t. which $P$ satisfies ordered downward- and upward-stability. Notice that among all these conditions required for graphicality, ordered downward- and upward-stability are the only ones that are related to, and in fact determine, the direction of the edges of $\text{sk}(P)$; these considerations also motivate the definition of natural structure-learning algorithms in Definition 6.

We work towards a proof, that under faithfulness, a natural structure-learning algorithm works.

**Proposition 9** *Suppose that $P$ is faithful to the causal graph $G_0$. The output of a natural structure-learning algorithm satisfies $\text{sk}(G_0) = \text{sk}(G(P))$.*

**Proof** We show that the conditions of Lemma 7 hold.

(a) If $i \perp\!\!\!\perp j \mid \text{an}(i,j)$, then faithfulness gives, $i \perp j \mid \text{an}(i,j)$ in $G_0$, which implies that $i \nsim j$ in $G_0$.

(b) If there exists a $C$ such that $i \perp\!\!\!\perp j \mid C$, then faithfulness gives, $i \perp j \mid C$ in $G_0$, which implies that $i \nsim j$ in $G_0$. Since $P$ is Markovian to $G_0$, we have that $i \perp\!\!\!\perp j \mid \text{an}(i,j)$. ∎

**Proposition 10** *Suppose that $P$ is faithful to the causal graph $G_0$. All potential outputs of natural structure-learning algorithms are Markov equivalent, and they are Markov equivalent to $G_0$.*

**Proof** Since $P$ is graphical, by Proposition 4, it is singleton-transitive compositional graphoid. Let $G(P)$ and $H(P)$ be two potential graphs generated by a natural structure-learning algorithm. By Proposition 9, we have that $\mathrm{sk}(G_0) = \mathrm{sk}(G(P)) = \mathrm{sk}(H(P))$.

In addition, $P$ satisfies ordered upward- and downward-stability w.r.t. both $G(P)$ and $H(P)$. Hence, again by Proposition 4, $P$ is faithful to both. Therefore, these graphs are all Markov equivalent. ∎

In addition, the other direction of the above result also holds:

**Proposition 11** *Suppose that $P$ is faithful to $G_0$. If $G$ is Markov equivalent to $G_0$, then $P$ satisfies ordered upward- and downward-stability w.r.t. $G$.*

**Proof** The graph $G$ and $P$ have to be faithful. Therefore, by Proposition 4, we know that $P$ satisfies ordered upward- and downward-stability w.r.t. $G$. ∎

**Remark 12** *Therefore, ordered downward- and upward-stability w.r.t. $G(P)$ are also necessary conditions of an output of a structure-learning algorithm that generates Markov equivalent graphs to $G_0$. This justifies Definition 6, when $P$ is graphical. Hence, any "exact" (as opposed to approximate) constraint-based algorithms that work under the assumption of faithfulness (such as the SGS and PC algorithms (Spirtes et al., 2000)) must (and, in fact, do) ensure that, w.r.t. their output, $P$ satisfies ordered downward- and upward-stability.*

## 5. Conditions and assumptions for the structure-learning algorithm

Recall that the general goal of constraint-based structure learning is to consider algorithms for which the output $G(P)$ is Markov equivalent to (the unknown) causal graph $G_0$. In Section 4, we saw that under the assumption of faithfulness of $P$ and $G_0$, this goal is achieved for the class of algorithms falling under Definition 6, the class of natural structure-learning algorithms. However, in fact, faithfulness is not necessary for this goal, as will be shown below in Example 13; see also Example 18 for when one focuses solely on the class of DAGs.

**Example 13 (Ordered stability and faithfulness)** *Consider the graph in Figure 2, and assume that it is $G_0$ in the setting of this manuscript. Let $\mathcal{J}(P)$ be the set of statements implied by the global Markov property plus $i \perp\!\!\!\perp j \mid k$. First of all, notice that $P$ does not satisfy singleton-transitivity. Notice also that $P$ is minimally Markovian to $G_0$, but $P$ and $G_0$ are not faithful. In addition, $P$ satisfies ordered upward- and downward-stability w.r.t. $G_0$.*

*However, all graphs, including the true causal graph $G_0$, w.r.t. which $P$ satisfies ordered downward- and upward-stability are Markov equivalent. To see this, suppose that there is a graph $H$ with the same skeleton, w.r.t. which $P$ satisfies ordered upward- and downward-stability, and that $H$ is not Markov equivalent to $G_0$. If, in $H$, $k$ is non-collider then $k > \ell$ or $k > i$, which, by ordered upward-stability, implies $i \perp\!\!\!\perp \ell \mid k$; but this is not in $\mathcal{J}(P)$. Now, if in $H$, the node $\ell$ is collider, then $k, j > \ell$, which by ordered downward-stability and conditional independence $k \perp\!\!\!\perp j \mid \ell$, implies $k \perp\!\!\!\perp j$, which is again not in $\mathcal{J}(P)$.*
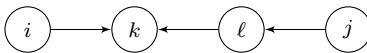
Figure 2: $G_0$

Below, we find assumptions under which a natural structure-learning algorithm works; these assumptions can be thought of as a relaxation of faithfulness.

### 5.1 Equivalence of the skeletons of $G_0$ and $G(P)$

We first need to provide assumptions so that $G_0$ and $G(P)$ have the same skeletons. One of the main assumptions we employ is that $P$ satisfies ordered downward- and upward-stability w.r.t. $G_0$.

**Theorem 14** *Consider a distribution $P$ that is Markovian to the causal graph $G_0$ satisfying the following conditions:*

*(i) $P$ is converse pairwise Markovian to $G_0$;*

*(ii) $P$ satisfies ordered upward- and downward-stability w.r.t. $G_0$.*

*Then $\mathrm{sk}(G_0) = \mathrm{sk}(G(P))$.*

**Proof** It suffices to show that assumption (b) of Lemma 7 is satisfied. Suppose $i \perp\!\!\!\perp j \,|\, C$. Consider all the nodes in $C \setminus \mathrm{an}(i,j)$. Since $P$ satisfies ordered downward-stability w.r.t. $G_0$, one can remove the node $k_0$ with a lowest order in this set to obtain $i \perp\!\!\!\perp j \,|\, C \setminus \{k_0\}$. Inductively, we will obtain $i \perp\!\!\!\perp j \,|\, C \cap \mathrm{an}(i,j)$. Now, by the use of ordered upward-stability, we add other nodes in $\mathrm{an}(i,j)$ to the conditioning set, and obtain $i \perp\!\!\!\perp j \,|\, \mathrm{an}(i,j)$. ∎

**Corollary 15** *Consider a distribution $P$ that satisfies the conditions of Theorem 14. Then $P$ is minimally Markovian to $G_0$.*

**Remark 16** *Notice that there exist cases where there is a unique Markov equivalent class of graphs including $G_0$ to which the distribution $P$ is Markovian, but the skeleton of $G_0$ is not the same as the skeleton of $P$; see Forster et al. (2018), where such examples are used as a motivation for the assumption of frugality; see Remark 3. In such cases, conditions of Theorem 14 (or equivalently adjacency-faithfulness) are not satisfied. This suggests that there could be extensions to the setting of this paper.*

### 5.2 Markov equivalence of $G(P)$ and $G_0$

In this section, we will describe under what additional testable conditions $G_0$ and $G(P)$ are Markov equivalent. We provide the conditions and results both for the case where $G_0$ is a DAG and the structure-learning algorithm only generates DAGs, and also for the general case where $G_0$ is a directed ancestral graph and the structure-learning algorithm generates such graphs.

We formalize the problem using the following concept: We say that a distribution $P$ satisfies the *uniqueness property* (up to Markov equivalence) when the following holds:

maximal graphs with skeleton $\mathrm{sk}(P)$ w.r.t. which $P$ satisfies ordered upward- and downward-stability are Markov equivalent. For the special case of DAGs, we say that a distribution $P$ satisfies the *DAG-uniqueness property* if the graphs concerned in the definition of the uniqueness property are restricted to DAGs.

The distribution $P$ satisfying the uniqueness property is equivalent to the potential outputs of the structure-learning algorithm being unique up to Markov equivalence. This is an obvious requirement of the structure-learning algorithm so that the algorithm cannot generate two graphs that are not Markov equivalent.

In order to ensure that the output of a natural structure-learning algorithm is correct, we need the following important lemma:

**Lemma 17** *Consider a distribution $P$ that is Markovian to $G_0$. Suppose that the conditions of Theorem 14 are satisfied; that is, $P$ is converse pairwise Markovian to $G_0$ and $P$ satisfies ordered upward- and downward-stability w.r.t. $G_0$. Let $G(P)$ be the output of a natural structure-learning algorithm.*

*(a) If $P$ satisfies the uniqueness property, then $G(P)$ is Markov equivalent to $G_0$.*

*(b) If $G_0$ is a DAG, then $P$ satisfying the DAG-uniqueness property implies that a DAG output $G(P)$ is Markov equivalent to $G_0$.*

**Proof**  The proof follows from Theorem 14. ∎

In what follows, we provide and study explicit conditions under which the uniqueness property is satisfied.

By Corollary 15, we know that, under the assumptions for $P$, which we have so far utilized, $P$ is minimally Markovian to $G_0$. In Example 18, we observe that, if we replace the faithfulness assumption with the assumption of being minimal Markovian, then the uniqueness property no longer holds.

**Example 18** *Let $\mathcal{J}(P) = \{1 \perp\!\!\!\perp 4 \mid \{2, 3\},\ 1 \perp\!\!\!\perp 4 \mid 3,\ 2 \perp\!\!\!\perp 3 \mid 4\}$. For example, there is an SCM that induces such an independence model: Let $\epsilon_1, \epsilon_2, \epsilon_3$, and $\epsilon_4$ be i.i.d. Define*

$$X_1 = \max(X_2, X_3, \epsilon_1); \quad X_2 = \max(X_4, \epsilon_2); \quad X_3 = \max(2X_4, \epsilon_3); \quad X_4 = \epsilon_4.$$

*It can be seen that $(X_i)_{i \in \{1, \ldots, 4\}}$ induces $\mathcal{J}(P)$.*

*First notice that $P$ does not satisfy the uniqueness property: w.r.t. both graphs $G_1$ and $G_2$ in Figure 3(b) and 3(c), $\mathcal{J}(P)$ satisfies ordered upward- and downward-stability. However, these two graphs are not Markov equivalent.*

*Since $P$ satisfies the compositional graphoid axioms (but not singleton-transitivity), this also shows that compositional graphoid plus ordered upward- and downward-stability are not in general sufficient for the uniqueness property to hold.*

*However, if we assume that the structure-learning algorithm is only DAG-generating, then there is a unique graph w.r.t. which $P$ satisfies ordered upward- and downward-stability (and consequently the DAG-uniqueness property holds).*

From Example 13, we see that singleton-transitivity is not necessary for the uniqueness property. We provide a (weaker) condition, which together with assumptions of Theorem 14, are still sufficient for uniqueness.
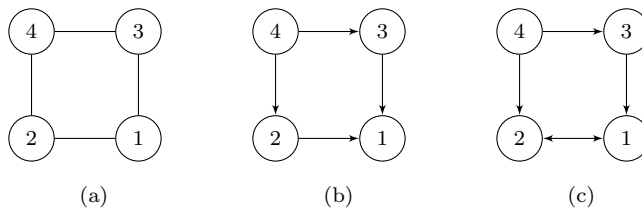
Figure 3: (a) The skeleton of $P$ with no arrows. (b) A DAG $G_1$. (c) A graph $G_2$ with arcs.

We say that an independence model $\mathcal{J}$ is *path-unstable* if there exists a path

$$\langle i = i_0, i_1, \cdots, i_r, k, j \rangle, \quad i \not\sim j \text{ and } i_r \sim j \text{ for every } r \geq 1$$

in $\mathrm{sk}(\mathcal{J})$ such that $i \perp\!\!\!\perp j \mid U \cup C$ and $i \perp\!\!\!\perp j \mid U \cup C \cup \{k\}$, for some $C$ disjoint from $\{i, j, k\}$, where $U = \{i_1, \cdots, i_r\}$ or $U = \emptyset$ when $r = 0$. An independence model that is not path-unstable is called *path-stable*.

In the case where the structure-learning algorithm is DAG-generating, the required condition is simpler. We say that an independence model $\mathcal{J}$ is *V-unstable* if there exists a V-configuration $\langle i, k, j \rangle$ in $\mathrm{sk}(\mathcal{J})$ such that $i \perp\!\!\!\perp j \mid C$ and $i \perp\!\!\!\perp j \mid C \cup \{k\}$, for some $C$ disjoint from $\{i, j, k\}$. An independence model that is not V-unstable is called *V-stable*. In the same fashion as before, if $\mathcal{J}(P)$ satisfies path- or V-stability then we simply say that "$P$ satisfies these conditions."

**Remark 19** *As will be seen in the proof of Proposition 24, the path that appears in the definition of a path-unstable $P$ can be thought of as the skeleton of (and can be directed to become) a* discriminating path *(Ali et al., 2009) from $i$ to $j$; i.e., a path $\langle i = i_0, i_1, \cdots, i_r, k, j \rangle$ that is a collider path, and $i_n \in \mathrm{pa}(j)$, for every $1 \leq n \leq r$.*

**Remark 20** *The well-known orientation-faithfulness assumption (which is defined as a part of the restricted-faithfulness assumptions) (Ramsey et al., 2006) states that for all V-configurations $\langle j, l, k \rangle$ and all subsets $S \subset V \setminus \{j, k\}$ such that $j$ is d-connected to $k$ given $S$ it holds that $j \not\perp\!\!\!\perp k \mid S$. Notice that, unlike orientation-faithfulness, V-stability (and also path stability) are purely defined on the distribution $P$ and not on the unknown causal graph $G_0$. Indeed, if the conditions of Theorem 14 are satisfied to ensure that $\mathrm{sk}(G_0) = \mathrm{sk}(G(P))$, then orientation-faithfulness implies V-stability. However, even under these extra assumptions, the other direction does not hold; see Example 21 below.*

**Example 21** *Consider the true causal graph $G_0$ in Figure 4. Let $\mathcal{J}$ contain all the conditional independence statements implied by the global Markov property on $G_0$, with the addition of $j \perp\!\!\!\perp k \mid s$. We have that $\mathrm{sk}(\mathcal{J}) = \mathrm{sk}(G_0)$. By looking at $\mathrm{sk}(\mathcal{J})$, it can be seen that V-stability is satisfied, but orientation-faithfulness is not: Consider the V-configuration $\langle j, l, k \rangle$. We have that $j$ is d-connected to $k$ given $s$, but that $j \perp\!\!\!\perp k \mid s$.*

V-stability is weaker than singleton-transitivity:

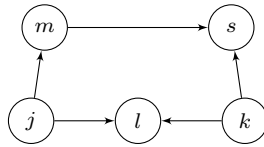**Proposition 22** *Singleton-transitivity implies V-stability.*

Figure 4: A directed acyclic graph $G_0$.

**Proof** Towards a contradiction, suppose that there is a V-configuration $\langle i, k, j \rangle$ in $\mathrm{sk}(P)$ such that $i \perp\!\!\!\perp j \mid C$ and $i \perp\!\!\!\perp j \mid C \cup \{k\}$, for some $C$ disjoint from $\{i, j, k\}$. By singleton-transitivity, $i \perp\!\!\!\perp k \mid C$ or $k \perp\!\!\!\perp j \mid C$, which is absurd since both $ik$ and $jk$ are edges in $\mathrm{sk}(P)$. ∎

In addition, the converse of the above result does not hold; see Example 13. Hence, we have the following corollary:

**Corollary 23** *If $P$ is faithful to a graph, then it satisfies V-stability.*

**Proof** The proof follows from the fact that $P$ being graphical implies it satisfies singleton-transitivity (Proposition 4). ∎

Also notice that singleton-transitivity and path-stability do not imply each other (for one direction, see Example 13 again), but there are considerably fewer independence statements to test for for path-stability than for singleton-transitivity. In a similar fashion, there could be cases where $P$ is graphical, but path-stability is not satisfied: one can take the skeleton of a discriminating path and direct the edges (with the path not being collider) so that the induced independence model (by separation) is path unstable. However, notice again that unlike faithfulness, path-stability is purely defined on $P$, and hence, in principle, testable. We now have the following relationship to the uniqueness property:

**Proposition 24** *Let $P$ be a distribution.*

*(a) If $P$ is path-stable, then it satisfies the uniqueness property.*

*(b) If $P$ is V-stable, then it satisfies the DAG-uniqueness property.*

**Proof** We start with the proof for the case of DAGs, which is more elementary than the general case.

(b) Suppose that $P$ satisfies ordered upward- and downward-stability w.r.t. two orders $\leq'$ and $\leq''$. Towards a contradiction, suppose that the two DAGs, $G(P, \leq')$ and $G(P, \leq'')$ are not Markov equivalent.

By Lemma 2, there is a collider V-configuration $\langle i, k, j \rangle$, say, in $G(P, \leq')$, which is a non-collider in $G(P, \leq'')$ (say $k >'' j$). Since $G(P, \leq')$ is maximal, $i \perp\!\!\!\perp j \mid C$, for some $C$. By ordered downward-stability w.r.t. $\leq'$, we obtain $i \perp\!\!\!\perp j \mid C'$, for some $C'$ such that $k \notin C'$. Now by ordered upward-stability w.r.t. $\leq''$, we obtain $i \perp\!\!\!\perp j \mid C' \cup \{k\}$. This shows that $P$ is V-unstable, which is a contradiction.

(a) Suppose that $P$ satisfies ordered upward- and downward-stability w.r.t. two orders $\leq'$ and $\leq''$. Towards a contradiction, suppose that the two maximal graphs $G(P, \leq')$ and $G(P, \leq'')$ are not Markov equivalent.

By Lemma 1, there is a minimal collider path $\pi$, say, in $G(P, \leq')$, which is not a minimal collider path in $G(P, \leq'')$. Consider a shortest minimal collider subpath $\pi'$ of $\pi$ in $G(P, \leq')$ that is not a minimal collider in $G(P, \leq'')$. Furthermore, we can assume that $\pi'$ is not a collider path in $G(P, \leq'')$, since if $\pi'$ is a collider path, but not minimal, we consider the minimal collider subpath $\pi''$ in $G(P, \leq'')$. Note that $\pi''$ is not a minimal collider in $G(P, \leq')$. We can then continue the proof using $\pi''$ instead of $\pi'$, and the roles of $G(P, \leq')$ and $G(P, \leq'')$ reversed.

Let $i$ and $j$ be the endpoints of $\pi'$ which is not a collider in $G(P, \leq'')$. Consider a node $k$, which is a non-collider node on $\pi'$. We show that $k$ is adjacent to $i$ or $j$ on $\pi'$. Towards a contradiction, suppose that it is not adjacent to either. Consider a subpath of $\pi'$ given by $\langle i', h, k, \ell, j' \rangle$. Recall that the endpoints of a minimal collider path are not adjacent. Hence $h \sim \ell$, otherwise, $\langle h, k, \ell \rangle$ is a shorter minimal collider subpath in $G(P, \leq')$ that is not a minimal collider in $G(P, \leq'')$. Similarly, $h \sim j'$ and $i' \sim \ell$. Since $\pi'$ is a minimal collider, the $h\ell$-edge cannot be an arc, and the $hj$- and $i\ell$-edge have to be arrows from $h$ to $j'$ and $\ell$ to $i'$, respectively. Since directed cycles are not permissible in an ancestral graph, it is easy to rule out the existence of an arrow from $h$ to $\ell$ and vice-versa; we already ruled out the possibility that the $h\ell$-edge is an arc, and we require that $h \sim \ell$, which is absurd.

Without loss of generality, assume that $k$ is adjacent to $j$ on $\pi'$. Let $\pi' = \langle i = i_0, i_1, \cdots, i_r, k, j \rangle$, where $r$ may be 0. If $r > 0$, then by minimality, $i_r \sim j$ and $i_r \longrightarrow j$. From an inductive argument along $\pi'$, if $r > 0$, we conclude that $\pi'$ is a discriminating path from $i$ to $j$.

Since $G(P, \leq')$ is maximal, $i \perp\!\!\!\perp j \mid C$, for some $C$; moreover, $k \notin \mathrm{an}(i)$ since otherwise $i$ and $j$ are not separated. Since $k$ is a collider in $G(P, <')$, we have $k \notin \mathrm{an}(j)$. Therefore, by ordered downward-stability w.r.t. $\leq'$, we obtain $i \perp\!\!\!\perp j \mid C'$, for some $C'$ such that $k \notin C'$. By ordered upward-stability w.r.t. $\leq'$, and the previously established discrimination, we obtain $i \perp\!\!\!\perp j \mid U \cup C'$, where $U = \{i_1, \cdots, i_r\}$ (and if $r = 0$, then $U = \emptyset$). Now by ordered upward-stability w.r.t. $\leq''$, we obtain $i \perp\!\!\!\perp j \mid U \cup C' \cup \{k\}$, since $k$ is not a collider node on $G(P, \leq'')$. Thus $P$ is path-unstable, which is a contradiction. ∎

This leads to the main result, which provides conditions under which a natural structure-learning algorithm works.

**Theorem 25** *Suppose that $P$ satisfies the conditions of Theorem 14, that is, it is Markovian and converse pairwise Markovian to $G_0$, and satisfies ordered upward- and downward-stability w.r.t. $G_0$. Let $G(P)$ be the output of a natural structure-learning algorithm.*

*(a) If $P$ is also path-stable, then $G(P)$ is Markov equivalent to $G_0$.*

*(b) If $G_0$ is a DAG and $P$ is V-stable, then a DAG-output $G(P)$ is Markov equivalent to $G_0$.*

**Proof** The proof follows from Proposition 24 and Lemma 17. ∎

As an example for the theorem above, we can revisit our motivating Example 13 to observe that although faithfulness is not satisfied, all the conditions of Theorem 25, and, in particular, V-stability are satisfied.

However, notice that, under the conditions presented here, ordered downward- and upward-stability w.r.t. $G_0$ are not necessary in the sense that there are cases where ordered downward- and upward-stability are not satisfied w.r.t. $G_0$, but an output of the algorithm is Markov equivalent to $G_0$. This stems from the fact that there could be two Markov equivalent graphs and a $P$ such that $P$ satisfies ordered downward- and upward-stability w.r.t. one but not the other; see example below.

**Example 26** *Consider the causal graph $G_0$ and its Markov equivalent graph $G_1$ below. Let $\mathcal{J}$ (which, in principle could be induced by a probability distribution) contain all the independencies induced by the global Markov property plus $k \perp\!\!\!\perp m \mid l$ and additional independence statements implied by applying ordered upward-stability w.r.t. $G_1$ and semi-graphoid axioms. It can be checked that $k \not\perp\!\!\!\perp m \mid \{i, l\}$. Hence, $\mathcal{J}$ does not satisfy ordered upward-stability w.r.t. $G_0$. However, by definition, $\mathcal{J}$ satisfies ordered upward-stability w.r.t. $G_1$. Notice also that, in this example, $V$-stability (and path-stability) are satisfied.*



(a)                                               (b)

Figure 5: (a) $G_0$, (b) $G_1$

The above discussion suggests that one can replace the condition "$P$ satisfying ordered upward- and downward-stability w.r.t. $G_0$" in Theorem 25 with "$P$ satisfying ordered upward- and downward-stability w.r.t. a graph in the Markov equivalence class of $G_0$" in order to make the condition necessary.

## 6. Structure learning for structural causal models

First, we give a proof for the global Markov property for general SCMs, which allow for the possibility of arcs. Then we find conditions on the functions of the SCM to ensure the converse pairwise Markov property; see (2).

### 6.1 Global Markov property for general SCMs

The main assumption used in the general setting for structure learning is that $P$ is Markovian to the causal graph $G_0$. A main reason for popularity of SCMs for causal learning is that this condition is automatically satisfied for the joint distribution of an SCM. Indeed, the joint distribution of an SCM with mutually independent errors being Markovian to the DAG $G_0$ is a well-known result (Verma and Pearl, 1988; Pearl, 2009), but we are not aware of such a result in an explicit and non-parametric form when unobserved variables are existent; see Koster (1999) for a similar result in the Gaussian case and with a slightly different type of graph. Recently, in Bongers et al. (2021), a similar result is proven for directed

SCMs, but for a less general setting when one focuses on ancestral graphs; see Remark 5. See also Forre and Mooij (2017) for various equivalent notions of the Markov property.

**Theorem 27** *The joint distribution $P$ of an SCM is Markovian to $G_0$.*

The proof of the above theorem requires some definitions and results that are not used in the subsequent sections. We provide the proof together with these definitions and results in the remainder of this subsection. Readers may skip the proof without loss of continuity.

We first need some primary definitions and results mostly from Richardson (2003). The *district* of node $i$, denoted by $\mathrm{dis}(i)$, is the set of nodes connected to $i$ by a collider path on which every edge is an arc. We also define a set $A$ to be *ancestral* if it is closed under the ancestor relation, i.e., if $A \cup \mathrm{an}(A) = A$. If $A$ is an ancestral set in a graph $G$, and $i$ is a node in $A$ that has no children in $A$, then we define the *Markov blanket* of $i$ w.r.t. the induced subgraph $G[A]$ on $A$ by

$$\mathrm{mb}(i, A) = \mathrm{pa}_{G[A]}(\mathrm{dis}_{G[A]}(i)) \ \cup \ (\mathrm{dis}_{G[A]}(i) \setminus \{i\}).$$

We use the notation $\mathrm{pst}(i)$ for the set of nodes with an order larger that $i$. We say that a probability distribution satisfies the *ordered local Markov property* w.r.t. the ordering $\leq$, if for any node $i$, and ancestral set $A$ such that $i \in A \subseteq \mathrm{pst}(i)$, we have

$$i \perp\!\!\!\perp [A \setminus (\mathrm{mb}(i, A) \cup \{i\})] \,|\, \mathrm{mb}(i, A).$$

We have the following equivalence, which does not require any extra assumptions, originally proven for ADMGs (which is a more general class than directed ancestral graphs):

**Lemma 28 (Richardson (2003))** *For a valid ordering $\leq$ of the nodes of a graph $G$, a distribution $P$ satisfying the ordered local Markov property w.r.t. $\leq$ is equivalent to $P$ being Markovian to $G$.*

In this section, we also utilize the concept of marginalization over probability distributions and graphs. Let $\alpha(P, M)$ be the marginalized distribution over $M \subset V$ so that $\alpha(P, M)$ is defined over $X_{V \setminus M}$. Notice that the induced independence model of $\alpha(P, M)$ is $\mathcal{J}(\alpha(P, M)) = \{A \perp\!\!\!\perp B \,|\, D : (A \cup B \cup D) \cap M = \varnothing\}$.

By marginalizing over the node set $M$ of graph $G$, a new graph $\alpha(G, M)$ is obtained from which the node set $M$ is removed: For every path $\pi$ in $G$ between $i, j \notin M$, whose inner nodes are all non-collider and in $M$, we generate an edge in $\alpha(G, M)$, which preserves the (lack of) arrowheads pointing to $i$ and $j$ in $\pi$; We apply this repeatedly until no new edges are being generated, and then remove all the nodes in $M$; for more details of this algorithm, refer to Richardson and Spirtes (2002); Sadeghi (2013). The following result is well-known (Richardson and Spirtes, 2002; Sadeghi, 2013):

**Lemma 29** *For a graph $G$ and disjoint node sets $A$, $B$, $C$, and $M$, it holds that*

$$A \perp B \,|\, C \text{ in } \alpha(G, M) \iff A \perp B \,|\, C \text{ in } G.$$

We shall need the following corollary of this result:

**Lemma 30** *Suppose that a distribution $P$ is Markovian to the graph $G$. Then $\alpha(P, M)$ is Markovian to $\alpha(G, M)$.*

**Proof** $P$ being Markovian to $G$ implies that $A \perp B \mid C$ in $G$ implies that $A \perp\!\!\!\perp B \mid C$. $A \perp B \mid C$ in $\alpha(G, M)$ implies that $(A \cup B \cup C) \cap M = \varnothing$, and also $A \perp B \mid C$ (by Lemma 29). $(A \cup B \cup C) \cap M = \varnothing$ implies that $A \perp\!\!\!\perp B \mid C$ exists in $\mathcal{J}(\alpha(P, M))$. Therefore, $A \perp B \mid C$ in $\alpha(G, M)$ implies $A \perp\!\!\!\perp B \mid C$ exists in $\mathcal{J}(\alpha(P, M))$. ∎

We are now ready to provide the proof of Theorem 27.

**Proof of Theorem 27** Consider the augmented graph $G_0(X, \epsilon)$ and the joint distribution $P_{X,\epsilon}$. Consider the ordering similar to the ordering in the SCM for $X_i$s, and, for $\epsilon_i$s, let $X_{i+1} > \epsilon_i > X_i$. We show that $P_{X,\epsilon}$ satisfies the ordered local Markov property w.r.t. this ordering in $G_0(X, \epsilon)$. Hence by Lemma 28, we obtain that $P_{X,\epsilon}$ is Markovian to $G_0(X, \epsilon)$. Consider a node in $G_0(X, \epsilon)$. We have two cases:

First, assume that this node is of $\epsilon$ form, i.e., say it is $\epsilon_i$. Consider also $A \subseteq \mathrm{pst}(\epsilon_i)$. We have that $\mathrm{mb}(\epsilon_i, A) = \{\epsilon_j \in A : X_j \in \mathrm{dis}(X_i) \text{ in } G_0\}$. If $\mathrm{mb}(\epsilon_i, A) = A$, then $\epsilon_i \perp\!\!\!\perp [A \setminus (\mathrm{mb}(\epsilon_i, A) \cup \{\epsilon_i\})] \mid \mathrm{mb}(\epsilon_i, A)$ is trivial; otherwise there is no edge between $\mathrm{mb}(\epsilon_i, A)$ and $A \setminus (\mathrm{mb}(\epsilon_i, A) \cup \{\epsilon_i\})$ in $G_0(X, \epsilon)$, which, by joint independence of such $\epsilon_j$, implies that $\{\epsilon_i\} \cup \mathrm{mb}(\epsilon_i, A) \perp\!\!\!\perp A \setminus (\mathrm{mb}(\epsilon_i, A) \cup \{\epsilon_i\})$. This, by the weak union property, implies

$$\epsilon_i \perp\!\!\!\perp [A \setminus (\mathrm{mb}(\epsilon_i, A) \cup \{\epsilon_i\})] \mid \mathrm{mb}(\epsilon_i, A).$$

Secondly, assume that this node is of $X$ form, i.e., say it is $X_i$. Consider also $A \subseteq \mathrm{pst}(X_i) = (\epsilon_i, X_{i+1}, \epsilon_{i+1}, \ldots, X_n, \epsilon_n)$. We have that $\mathrm{mb}(X_i, A) = \{\epsilon_i\} \cup \{X_j : X_j \in \mathrm{pa}_{G_0[A]}(X_i)\}$. However, given $\mathrm{mb}(X_i, A)$, we have that $X_i$ is deterministic; hence, it is independent of all disjoint random vectors including $A \setminus (\mathrm{mb}(X_i, A) \cup \{X_i\})$.

Finally, since the inner nodes of $\langle X_i, \epsilon_i, \epsilon_j, X_j \rangle$ are non-collider in $\alpha(G_0(X, \epsilon))$, it can be seen that $\alpha(G_0(X, \epsilon), \epsilon) = G_0$. Hence, $P_{X,\epsilon}$ being Markovian to $G_0(X, \epsilon)$, by Lemma 30, implies that $P$ is Markovian to $G_0$. ∎

## 6.2 Conditions and assumptions for structure learning of SCMs

The general setting for structure learning for SCMs is the same as in the previous sections, and, in fact, the case of SCM could be considered a special case of the theory presented above: We start by assuming the existence of an *unknown* SCM $\mathfrak{C}$ associated to an *unknown* causal graph $G_0$ as described in Section 2.5. For structure learning (from observational data), we find a graph that belongs to the Markov equivalence class of $G_0$. As discussed, the SCM $\mathfrak{C}$ entails a joint distribution $P$, which is still unknown. Therefore, the relationships can be illustrated in the following diagram:

$$G_0 \leftarrow \mathfrak{C} \rightarrow P \overset{\text{conditional independence}}{\rightarrow} \mathcal{J}(P) \overset{\text{Algorithm}}{\rightarrow} G(P) \overset{\text{Markov equivalent}}{\sim} G_0$$

We begin our discussion with a few examples. In general, the independence models of the SCMs are not faithful to $G_0$. Although there are many examples where faithfulness fails, the following simple and perhaps familiar example will help to motivate further considerations regarding the converse pairwise Markov property.

**Example 31 (Failure of faithfulness)** *Consider the following SCM where* $2 \longrightarrow 1$, *and* *2 is the only parent of 1. Suppose* $X_2 = \epsilon_2$ *and* $X_1 = X_2 \oplus \epsilon_1 \bmod 2$, *where* $\epsilon_2$ *and* $\epsilon_1$ *are independent Bernoulli random variables with parameter* $p = \frac{1}{2}$. *It is easy to verify that* $X_2$ *is also independent of* $X_1$.

*For a continuous example, we exploit of the invariance of Lebesgue measure on the circle. Take* $\epsilon_1$ *and* $X_2 = \epsilon_2$ *to be independent continuous real-valued random variables. Let* $F$ *and* $G$ *be the cumulative distribution functions for* $X_2$ *and* $\epsilon_1$, *respectively. Set* $X_1 = F(X_2) \oplus G(\epsilon_1) \bmod 1$. *It is easy to verify that* $X_2$ *is independent of* $X_1$.

Motivated by Example 31 and the elementary fact that a random rotation that is independent of a standard bivariate normal remains independent of the randomly rotated bivariate normal, and more generally the connection between symmetries and independence (Dawid, 1985), we have the following lemma.

**Lemma 32** *Consider the SCM with associated graph* $2 \longrightarrow 1$, *so that 2 is the only parent of 1. Suppose that* $X_2 = \epsilon_2$, $X_1 = \phi_1(X_2, \epsilon_1)$ *and that* $\epsilon_2$ *and* $\epsilon_1$ *are independent. Then* $X_2$ *is independent of* $X_1$ *if and only*

$$\phi_1(x_2, \epsilon_1) \overset{d}{=} X_1 \tag{3}$$

*for almost every* $x_2 \in \mathcal{X}_2$.

Note that (3) is easily satisfied if $\phi_1$ is actually a function of $\epsilon_1$ alone. In (3), it may be helpful to think of $x_2$ as indexing (distinct) ways of generating a random variable with law $X_1$, given the randomization $\epsilon_1$.

**Proof of Lemma 32**     Let $S_2 \subset \mathcal{X}_2$ and $S_1 \subset \mathcal{X}_2$ be measurable subsets. Let $Q_{X_2}$ and $Q_{\epsilon_1}$ be the laws of $X_2$ and $\epsilon_1$, respectively. Since $2 \longrightarrow 1$, and $X_2$ and $\epsilon_1$ are independent, we have

$$
\begin{aligned}
\mathbb{P}(X_1 \in S_1, X_2 \in S_2) &= \mathbb{P}(\phi_1(X_2, \epsilon_1) \in S_1, X_2 \in S_2) \\
&= \int \int \mathbf{1}[\phi_1(x_2, e_1) \in S_1)]\mathbf{1}[x_2 \in S_2]dQ_{X_2}(x_2)dQ_{\epsilon_1}(e_1) \\
&= \int_{S_2} \mathbb{P}(\phi_1(x_2, \epsilon_1) \in S_1)dQ_{X_2}(x_2).
\end{aligned}
$$

Thus if (3) holds, we have

$$
\begin{aligned}
\mathbb{P}(X_1 \in S_1, X_2 \in S_2) &= \int_{S_2} \mathbb{P}(X_1 \in S_1)dQ_{X_2}(x_2) \\
&= \mathbb{P}(X_1 \in S_1)Q_{X_2}(S_2) \\
&= \mathbb{P}(X_1 \in S_1)\mathbb{P}(X_2 \in S_2)
\end{aligned}
$$

and the desired independence follows.

If $X_2$ and $X_1$ are independent, then

$$\mathbb{P}(X_1 \in S_1) = \frac{1}{\mathbb{P}(X_2 \in S_2)} \int_{S_2} \mathbb{P}(\phi_1(x_2, \epsilon_1)) \in S_1)dQ_{X_2}(x_2)$$

for all measurable subsets $S_2$ with $Q_{X_2}(S_2) > 0$. By setting

$$S_2 = \{x_2 \in \mathcal{X}_2 : \mathbb{P}(\phi_1(x_2, \epsilon_1) \in S_1) > \mathbb{P}(X_1 \in S)\},$$

we see that $Q_{X_2}(S_2) = 0$. Hence together with reversing the inequality in the definition of $S_2$, we obtain (3). ∎

In fact, the joint distributions of SCMs may not satisfy any of the axioms for faithfulness, presented in Proposition 4, and discussed in the Sections 2.2 and 2.3. For a counter-example and a thorough discussion on the intersection property, see Peters (2015). For an example of an SCM not satisfying singleton-transitivity, see Example 18. For counter-examples for other axioms, we provide the following examples:

**Example 33 (Failure of composition and ordered upward-stability)** *Consider the following SCM associated to the graph $G_0$ in Figure 6:*

$$X_3 = \epsilon_3; \quad X_2 = \epsilon_2; \quad X_1 = X_2 \oplus X_3 \oplus \epsilon_1 \bmod 2;$$

*where $\epsilon_1, \epsilon_2$, and $\epsilon_3$ are independent Bernoulli random variables, $\epsilon_2$ and $\epsilon_3$ with parameter $\frac{1}{2}$, and $\epsilon_1$ with parameter $\frac{1}{3}$.*

*It is easy to verify that $X_1 \perp\!\!\!\perp X_2$, and $X_1 \perp\!\!\!\perp X_3$, but $X_1 \not\perp\!\!\!\perp (X_2, X_3)$ and $X_1 \not\perp\!\!\!\perp X_2 \mid X_3$. Hence this is an example for an SCM not satisfying the composition property or ordered upward-stability.*
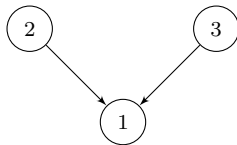


Figure 6: A directed acyclic graph $G_0$.

**Example 34 (Failure of ordered downward-stability)** *Consider a complete DAG with nodes $(1, 2, 3)$. Let $\epsilon_3 = (N_3, U_3)$ and $\epsilon_2 = (U_2, V_2)$, where $N_3, U_3, U_2$, and $V_2$ are all independent, and $N_3$ is Poisson distributed, and $U_3, U_2$ and $V_2$ are uniformly distributed on $[0, 1]$. Let $X_3 = \epsilon_3$. Let $X_2 = \phi_2(X_3, \epsilon_2) = (\psi_2(N_3, U_2), V_2)$, where $\psi_2$ is a deterministic function such that $\psi_2(n_3, U_2)$ is uniformly distributed on the integers $\{1, \ldots, n_3 + 1\}$. Let $X_1 = \phi_1(X_3, X_2) = (N_3, V_2)$. Clearly, $X_2$ is not independent of $X_3$. However, conditional on $X_1$, we have that $X_2$ is independent of $X_3$.*

An important property used in the results of previous section is converse pairwise Markov property. Here, we are mainly concerned with providing conditions for verifying this property. For an SCM $\mathcal{C}$, we provide an assumption on the joint distribution of $\mathcal{C}$ that implies the converse pairwise Markov property. Consider the following extension of Lemma 32.

**Proposition 35** *Consider an SCM, where $j \longrightarrow i$. Write $X_i = \phi_i(X_{\mathrm{pa}(i) \setminus \{j\}}, X_j, \epsilon_i)$. Suppose that $\mathrm{an}(i, j) \neq \emptyset$. Then $i \perp\!\!\!\perp j \mid \mathrm{an}(i, j)$ if and only if for almost all $x_{\mathrm{an}(i,j)} \in \mathcal{X}_{\mathrm{an}(i,j)}$ the conditional law of $X_i$ given $X_{\mathrm{an}(i,j)} = x_{\mathrm{an}(i,j)}$ is the same as the law of $\phi_i(x_{\mathrm{pa}(i) \setminus \{j\}}, x_j, \epsilon_i)$ for almost all $x_j \in \mathcal{X}_j$ in the support of the conditional distribution of $X_j$ given $x_{\mathrm{an}(i,j)}$.*

23

**Proof** Let $\mathcal{F}_j$ and $\mathcal{F}_i$ be the $\sigma$-algebras for $\mathcal{X}_j$ and $\mathcal{X}_i$. Let $D : \mathcal{X}_{\mathrm{an}(i,j)} \times (\mathcal{F}_j, \mathcal{F}_i) \to [0,1]$ be a regular conditional distribution for $(X_j, X_i)$ given $X_{\mathrm{an}(i,j)}$. Fix $x_{\mathrm{an}(i,j)} \in \mathcal{X}_{\mathrm{an}(i,j)}$. All the parents of $i$ are included in $\mathrm{an}(i,j)$, except $j$. Let $Q_i$ and $Q_j$ be marginal distributions of $X_i$ and $X_j$ under $D_{x_{\mathrm{an}(i,j)}} = D(x_{\mathrm{an}}(i,j), \cdot, \cdot)$. Following the proof of Lemma 32, since $j \not\rightarrow i$ and the graph is ancestral, we have that $\epsilon_i$, with law $Q_{\epsilon_i}$, is independent of $(X_{\mathrm{an}(i,j)}, X_j)$. Thus for $S_j \in \mathcal{F}_j$ and $S_i \in \mathcal{F}_i$, we have

$$
\begin{aligned}
D_{x_{\mathrm{an}(i,j)}}(S_j, S_i) &= \int \int_{S_j} \mathbf{1}[\phi_i(x_{\mathrm{pa}(i)\setminus\{j\}}, x_j, e_i) \in S_i)] dQ_j(x_j) dQ_{\epsilon_i}(e_i) \\
&= \int_{S_j} \mathbb{P}(\phi_i(x_{\mathrm{pa}(i)\setminus\{j\}}), x_j, \epsilon_i) \in S_i) dQ_j(x_j).
\end{aligned}
$$

Thus, if

$$
Q_i(S_i) = \mathbb{P}(\phi_i(x_{\mathrm{pa}(i)\setminus\{j\}}), x_j, \epsilon_i) \in S_i) \tag{4}
$$

for all $Q_j$-almost all $x_j \in \mathcal{X}_j$, then

$$
D_{x_{\mathrm{an}(i,j)}}(S_j, S_i) = Q_j(S_j)Q_i(S_i)
$$

as desired.

A routine variation of the rest of the proof of Lemma 32, gives that the condition on $\phi_i$ expressed in (4) is also necessary for the conditional independence of $X_j$ and $X_i$. ∎

Proposition 35 includes Lemma 32 by making the suitable interpretations in the case that $\mathrm{an}(i,j) = \emptyset$.

Proposition 35 suggests the following condition that implies the converse pairwise Markov property, in the case that true causal graph is a DAG. We say that the (joint distribution) of an SCM satisfies *positivity* if $\mathbb{P}(X_B \in S | X_A = x_A) > 0$ almost surely, for every set of nodes $B$ disjoint from $A$ and every measurable $S$ for which $\mathbb{P}(X_B \in S) > 0$, for almost every $x_A \in \mathcal{X}_A$.

Observe that each function $\phi$ in an SCM has the form

$$
\phi : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \times \mathcal{E} \to \mathcal{X}_{n+1},
$$

which is endowed with a product probability measure $\mu \otimes Q_\epsilon$, the joint law of the parents and exogenous variable on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n \times \mathcal{E}$, and a probability measure $\nu = (\mu \otimes Q_\epsilon) \circ \phi^{-1}$ on $\mathcal{X}_{n+1}$, the law of the child. Let $\mu^j$ be the projection of $\mu$ to space

$$
\mathcal{X}^j := \mathcal{X}_1 \times \cdots \times \mathcal{X}_{j-1} \times \mathcal{X}_{j+1} \times \cdots \times \mathcal{X}_n.
$$

Let $\epsilon$, $X = (X_1, \ldots, X_n)$, and $X_j$ have laws $Q_\epsilon$, $\mu$, and $\mu_j$ respectively. We say that $\phi$ has *non-constant fibers* if for all $1 \le j \le n$ there is some $F \subset \mathcal{X}^j$ with positive $\mu^j$-measure for each

$$
x^j = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \in F
$$

and there is some $S \subset \mathcal{X}_{n+1}$ such that the function $k : \mathcal{X}_j \to [0,1]$ given by

$$
k(x_j) = \mathbb{P}(\phi(x_1, \ldots, x_j, \ldots, x_n, \epsilon) \in S)
$$

is non-constant $\mu_j$-almost surely. We say that an SCM has *non-constant fibers* if every function in the SCM has non-constant fibers.

**Remark 36** *Note that in the proof of Proposition 35, equality* (4) *fails if $\phi_i$ has non-constant fibers and the SCM satisfies positivity.*

**Remark 37** *It is easy to see that additive noise models with the form*

$$\phi(x, \epsilon) = h(x) + \epsilon$$

*will have non-constant fibers under the mild condition that for any $1 \leq j \leq n$, and any fixed $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n$, the function*

$$g(x_j) = h(x_1, \ldots, x_j, \ldots x_n)$$

*is non-constant almost surely; assume for simplicity that all the relevant conditional distributions are equivalent to Lebesgue measure. Hence, by Remark 36, these additive noise models will satisfy the converse pairwise Markov property, provided that the true causal graph is a DAG.*

Recall that if $i \leftarrowtail\rightarrowtail j$, then $\epsilon_i$ and $\epsilon_j$ are dependent. In order to deal with the possibility of arcs, we have the following restriction that allows us to preserve dependence. We say that $\phi : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \times \mathcal{E} \to \mathcal{X}_{n+1}$ in an SCM is *noise injective* if for each fixed $x_1, \ldots, x_n$, the function $e \mapsto \phi(x_1, \ldots, x_n, e)$ is injective.

**Lemma 38** *Consider an SCM, where $j \leftarrowtail\rightarrowtail i$. If both $\phi_i$ and $\phi_j$ are noise injective, then $i \not\perp\!\!\!\perp j \,|\, \mathrm{an}(i,j)$.*

**Proof** Fix $x_{\mathrm{an}(i,j)}$. Since $j \leftarrowtail\rightarrowtail i$, all the parents of $i$ and $i$ are included in $\mathrm{an}(i,j)$ and $\epsilon_i$ and $\epsilon_j$ are dependent. The conditional distribution of $(X_i, X_j)$ given $X_{\mathrm{an}(i,j)} = x_{\mathrm{an}(i,j)}$ is the law of

$$\big( (\phi_i(x_{\mathrm{pa}(i)}, \epsilon_i), \phi_j(x_{\mathrm{pa}(j)}, \epsilon_j) \big),$$

since the graph is ancestral and the joint law of $(\epsilon_i, \epsilon_j)$ remains the same under this conditioning. Since $\phi_i$ and $\phi_j$ are noise injective, it follows that $\phi_i(x_{\mathrm{pa}(i)}, \epsilon_i)$ and $\phi_j(x_{\mathrm{pa}(j)}, \epsilon_j)$ are dependent if and only if $\epsilon_i$ and $\epsilon_j$ are dependent. ∎

From the proof of Lemma 38, we see that noise injectivity is a condition which allows us to preserve the dependencies of the noises; in general if the random variables $\epsilon_i$ and $\epsilon_j$ are dependent, we cannot conclude (like independence) that functions of $\epsilon_i$ and $\epsilon_j$ will remain dependent. It is easy to verify that noise injectivity is satisfied by the additive noise models.

**Corollary 39** *The additive noise models of Remark 37 satisfy the converse pairwise assumption, without the assumption that the true causal graph is a DAG.*

**Proof** Immediate from Remark 37 and Lemma 38. ∎

We also say that an SCM is *noise injective* if every function in the SCM is noise injective. Notice that these conditions do not depend on the structure of the causal graph.

**Corollary 40** *Any SCM satisfying positivity, with non-constant fibers, and noise injectivity has the converse pairwise Markvov property; furthermore, if the true causal graph is a DAG, then the noise injectivity assumption may be dropped.*

**Proof** Immediate from Proposition 35, Lemma 38, and Remark 36. ∎

Hence, under some additional assumptions, the SCMs are not only Markovian, but minimal Markovian to their true causal graph:

**Corollary 41** *Consider an SCM with positivity, non-constant fibers, noise injectivity, and ordered upward- and downward-stability w.r.t. $G_0$. Then $P$ is minimally Markovian to $G_0$; furthermore, if the true causal graph is a DAG, then the noise injectivity assumption may be dropped.*

**Proof** The proof follows from the Corollary 15 together with Theorem 27 and Corollary 40. ∎

We can now present our main result for structure learning for SCMs:

**Theorem 42** *Consider an SCM with positivity, non-constant fibers, and a distribution $P$, which satisfies ordered upward- and downward-stability w.r.t. the causal graph $G_0$. Consider natural structure-learning algorithms.*

1. *If $P$ is path-stable, and the SCM is noise injective, then $G(P)$, an output of a natural structure-learning algorithm, is Markov equivalent to $G_0$.*

2. *If $G_0$ is a DAG, and $P$ is V-stable, then a DAG-output of a natural structure-learning algorithm is Markov equivalent to $G_0$.*

**Proof** The proof follows from Theorem 25, Theorem 27, and Corollary 40. ∎

In Theorem 42, the most difficult condition to verify are ordered stabilities, which may not be directly testable.

### 6.3 Specialization to specific SCMs

We consider some more restrictive assumptions on an SCM that will imply the converse pairwise Markov property.

#### 6.3.1 Discrete random variables

For discrete distributions, we have the following simple sufficient condition for an SCM to satisfy the converse pairwise Markov property. We say that an SCM is *discrete* if both the endogenous and exogenous random variables are discrete.

**Proposition 43** *Consider an SCM that is discrete and satisfies positivity. Suppose that for all the nodes $i$, the cardinality of the support of $\epsilon_i$ is strictly less than the support of $X_i$.*

1. *If $P$ is path-stable, and the SCM is noise injective, then $G(P)$, an output of a natural structure-learning algorithm, is Markov equivalent to $G_0$.*

2. *If $G_0$ is a DAG, and P is V-stable, then a DAG-output of a natural structure-learning algorithm is Markov equivalent to $G_0$.*

**Proof** From Theorems 25 and 27 it suffices to verify the converse pairwise Markov property for a pair of adjacent nodes $i$ and $j$. Lemma 38 takes the care of the case where $i \leftrightarrow j$. Suppose that $i \longrightarrow j$. If the support of $\epsilon_i$ is strictly less than the support of $X_i$, then positivity implies that (4) is not satisfied so that Proposition 35 gives that the conditional independence statement $i \perp\!\!\!\perp j \,|\, \mathrm{an}(i,j)$ does not hold. ■

The assumption on the cardinality of the supports has the interpretation that information at a node is *not* mostly noise.

In the next lemma, we show that under some assumptions, the failure of the converse pairwise Markov property for $j \longrightarrow i$, forces the uniform distribution on both $X_i$ and $\epsilon_i$. Let $\phi : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \times \mathcal{E} \to \mathcal{X}_{n+1}$ be a noise injective function in a discrete SCM. Given $x_1, \ldots, x_j, \ldots, x_n$, let $\Phi_{x_j}^{x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n} = \Phi_{x_j}^{x^j}$ denote the inverse of $\phi$. Suppose that for all choices of $1 \le j \le n$ and $x^j$ there exists $x_{n+1}^* \in \mathcal{X}_{n+1}$ such that

$$\{\Phi_{x_j}^{x^j}(x_{n+1}^*) : x_j \in \mathcal{X}_j\} = \mathcal{E}.$$

Then we say that $\phi$ is *noise surjective*.

**Proposition 44** *Consider the following SCM, where $j \longrightarrow i$ and*

$$X_i = \phi_i(X_{\mathrm{pa}(i)\setminus\{j\}}, X_j, \epsilon_i).$$

*Suppose that the joint distribution is discrete and satisfies positivity. Suppose also that $\phi_i$ is noise injective and surjective. If $i \perp\!\!\!\perp j \,|\, \mathrm{an}(i,j)$, then $\epsilon_i$ and $X_j$ are both uniformly distributed and $\mathcal{E}_i$ and $\mathcal{X}_i$ have the same cardinality.*

**Proof** Fix $x_{\mathrm{an}(i,j)}$. Note that all the parents of $i$, save $j$, are included in $\mathrm{an}(i,j)$, By Proposition 35, we have

$$
\begin{aligned}
\mathbb{P}(X_i = x_i | X_{\mathrm{an}(i,j)} = x_{\mathrm{an}(i,j)}) &= \mathbb{P}(\phi_i(x_{\mathrm{pa}(i)\setminus\{j\}}), x_j, \epsilon_i) = x_i) \\
&= \mathbb{P}(\epsilon_i = \Phi_{x_j}^{x_{\mathrm{pa}(i)\setminus\{j\}}}(x_i))
\end{aligned}
$$

for all $x_i \in \mathcal{X}_i$ and all $x_j \in \mathcal{X}_j$, by positivity. By noise surjectivity, substituting $x_i = x_i^*$, we deduce that $\epsilon_i$ is uniformly distributed, with finite support $\mathcal{E}_i$. Thus

$$\mathbb{P}(X_i = x_i | X_{\mathrm{an}(i,j)} = x_{\mathrm{an}(i,j)}) = \frac{1}{\#\mathcal{E}_i}$$

for all $x_i$. Hence it follows that $X_i$ is also uniformly distributed. ■

**Corollary 45** *Consider a discrete SCM satisfying positivity. Suppose that all the functions are noise injective and noise surjective. If each of the exogenous random variables are not uniformly distributed, then the converse pairwise Markov property is satisfied.*

**Proof** Immediate from Proposition 44 and Lemma 38. ∎

**Corollary 46** *Consider an SCM that is discrete and satisfies positivity. Suppose that all the functions are noise injective and noise surjective. Assume each of the exogenous random variables are not uniformly distributed.*

1. *If $P$ is path-stable, then $G(P)$, an output of a natural structure-learning algorithm, is Markov equivalent to $G_0$.*

2. *If $G_0$ is a DAG, and $P$ is V-stable, then a DAG-output of a natural structure-learning algorithm is Markov equivalent to $G_0$.*

**Proof** The previous corollary assures us that the converse pairwise Markov property is satisfied, so that the result follows from Theorems 25 and 27. ∎

Returning to Example 31, it is easy to verify that the simple SCM is both noise injective and surjective. Notice that if $\epsilon_1$ is chosen with Bernoulli parameter $p \neq \frac{1}{2}$, and all other choices are left the same, then $X_2$ is no longer independent of $X_1$. Corollaries 45 and 46 suggest that if it is possible to avoid or exclude the possibility that the exogenous random variables are uniformly distributed, then a natural structure-learning algorithm is more likely to work.

### 6.3.2 Continuous random variables

We note that the modular addition in Example 31 is not continuous. We show that under some regularity conditions, up to a change of coordinates, only trivial functions can result in independence. Given $\phi : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \times \mathcal{E} \to \mathcal{X}_{n+1}$, we say that $\psi$ is a *j-change of coordinates* if there exists $g : [-1/2, 1/2] \to \mathcal{E}$ such that for all $x^j \in \mathcal{X}^j$, there exists $h_{x^j} : \mathcal{X}_{n+1} \to [-1/2, 1/2]$ such that if $U$ is uniformly distributed on $[-1/2, 1/2]$ and

$$\psi(x_j, u) := h_{x^j} \circ \phi(x_1, \ldots, x_j, \ldots, x_n, g(u)),$$

then $g(U)$ has the law of $\epsilon$ and $\psi(x_j, \epsilon)$ is uniformly distributed on $[-1/2, 1/2]$ for all $x_j \in \mathcal{X}_j$. We say that $\psi$ is *regular* if for fixed $x_j$, the function $\zeta(u) = \psi(x_j, u)$ is injective and $\zeta(u)$ is monotone; note that under these conditions an easy argument given in Lemma 47 shows that $\psi$ is a *reflection* of the form $\psi(x_j, u) = u$ or $\psi(x_j, u) = -u$ for all $x_j \in \mathcal{X}_j$ and all $u \in [-1/2, 1/2]$. We say that $\phi$ is *non-trivial* if for each $j$, there does not exist a regular $j$-change of coordinates. We say that an SCM is *non-trivial* if every function in the SCM is non-trivial. We say that a function $\phi$ in an SCM is *noise monotone* if for all $x_1, \ldots, x_n$, the function $e \mapsto \phi(x_1, \ldots, x_n, e) \in \mathcal{X}_{n+1} \subset \mathbb{R}$ is monotone in $e \in \mathcal{E} \subset \mathbb{R}$.

**Proposition 47** *Consider the following SCM, where $j \dashrightarrow i$ and*

$$X_i = \phi_i(X_{\mathrm{pa}(i) \setminus \{j\}}, X_j, \epsilon_i).$$

*Suppose that $\phi_i$ is noise injective and noise monotone. Also assume that $\epsilon_i$ is a continuous random variable, and that for almost all $x_{\mathrm{an}(i,j)}$, the conditional distribution law of $X_i$ given*

$X_{\mathrm{an}(i,j)}$ *is continuous with respect to Lebesgue measure. If $i \perp\!\!\!\perp j \mid \mathrm{an}(i,j)$, then $\phi_i$ admits a* *$j$-regular change of coordinates and any such change of coordinates is a reflection.*

**Proof** Recall that if $F$ is the cumulative distribution function for a continuous random variable $Z$, then $F$ is non-decreasing and $F(Z) - \frac{1}{2}$ is uniformly distributed on $[-1/2, 1/2]$. In addition, the inverse transform method tells us that $F^{-1}(U + \frac{1}{2})$ has law $Z$ if $U$ is uniformly distributed on $[-1/2, 1/2]$. Hence by Proposition 35, and the absolute continuity assumptions, it follows that $\phi_i$ has a $j$-regular change of coordinates $\psi$. Fix $x_j \in \mathcal{X}_j$ and write $\psi(x_j, u) = \psi(u)$. If $\psi$ is increasing in $u$, then for $z \in [-1/2, 1/2]$, we have for $U$ uniformly distributed on $[-1/2, 1/2]$ that

$$z + 1/2 = \mathbb{P}(\psi(U) \leq z) = \mathbb{P}(U \leq \psi^{-1}(z)) = \psi^{-1}(z) + 1/2,$$

from which it follows that $\psi(x_j, z) = z$. If $\psi$ is decreasing, then a similar argument gives that $\psi(x_j, z) = -z$. ∎

**Corollary 48** *If the law of an SCM is given by a jointly continuous distribution, and the SCM is noise injective and monotone, and is non-trivial, then it satisfies the converse pairwise Markov property. Moreover,*

1. *If $P$ is path-stable, then $G(P)$, an output of a natural structure-learning algorithm, is Markov equivalent to $G_0$.*

2. *If $G_0$ is a DAG, and $P$ is V-stable, then a DAG-output of a natural structure-learning algorithm is Markov equivalent to $G_0$.*

**Proof** The converse pairwise Markov property is immediate from Proposition 47 and Lemma 38. Hence the latter assertions that a natural structure-learning algorithm works follow from Theorems 25 and 27. ∎

## 7. Summary and discussion

We formalized the problem of structure learning from independence statements to find a graph that is Markov equivalent to the "true causal graph," which is the unknown graph to which the distribution is assumed to be Markovian. In fact, all the results presented here are also correct for an abstract independence model $\mathcal{J}$ instead of a probability distribution $P$.

A main message of this paper is that an important property of constraint-based structure-learning algorithms is to direct the edges of the skeleton of the probability distribution $P$ so that ordered upward- and downward-stability are satisfied by $P$ w.r.t. the generated graph. We call such algorithms a natural constraint-based structure-learning algorithm. Among the conditions that constitute the widely-used assumption of faithfulness, provided in Proposition 4, ordered upward- and downward-stability are the only ones that induce an ordering, which is what is used for directing the edges in structure learning. Since,

under faithfulness, ordered upward- and downward-stability are necessary for the output of the algorithm, all exact structure-learning algorithms that assume faithfulness in the literature are members of the class of natural structure-learning algorithms. Therefore, for any (newly) proposed algorithm, to prove that the algorithm works, it is sufficient to test whether $P$ satisfies ordered upward- and downward-stability w.r.t. its output.

Following this line of thinking, this result could lead to finding the "optimal" natural constraint-based structure-learning algorithm. The idea is that such an algorithm would be designed to solely direct the edges of the skeleton to ensure an ordering w.r.t. which ordered upward- and downward-stability are satisfied. Hence, for certain recorded independence statements, one should perform some additional tests on whether independence is preserved by certain additions to the conditioning set of those statements. The results of such tests determine the direction of the edges of the output. This is a direction for future work.

Moreover, we provided two assumptions, both of which are satisfied by faithfulness, in order to ensure that the skeleton of the output of a natural structure-learning algorithm is the same as the skeleton of the causal graph. One condition is the converse pairwise Markov property and the other one is satisfied by the assumption that the distribution satisfies ordered upward- and downward-stability w.r.t. the causal graph. In order to ensure that directing the edges of the skeleton of the output will result to a Markov equivalent graph to the causal graph, a sufficient condition is path-stability in the general case of directed ancestral graphs, and V-stability in the case where we only deal with DAGs.

In contrast to well-known conditions in the literature such as orientation-faithfulness, path- and V-stability are only properties of the distribution and *not* the causal graph. V-stability is also weaker than singleton-transitivity, and consequently faithfulness. However, there are cases where faithfulness is satisfied, but not path-stability. Path- and V-stability are easier to test than singleton-transitivity as one needs to only test independence for the endpoints of specific paths or V-configurations in $\mathrm{sk}(P)$ respectively. In addition to singleton-transitivity, out of the equivalent conditions to faithfulness, provided in Sadeghi (2017), intersection and composition properties are not needed. In the case of a Gaussian distribution, however, all these assumptions are automatically satisfied, and therefore, our assumptions imply faithfulness.

It is commonplace to perform structure learning for SCMs. We have specialized the theory for SCMs. For SCMs that we associate with directed ancestral graphs, we showed that the main assumption of $P$ being Markovian to the causal graph is always satisfied.

In addition, we provided simple sufficient conditions on the SCM that imply the converse pairwise Markov property. Although converse pairwise Markov property is indeed related to the causal graph, these sufficient conditions are not. Our key condition of having non-constant fibers has a simple interpretation when the function of an SCM has only one discrete endogenous variable and one exogenous variable: there is a choice of two values for the endogenous variable, whereby the distributions of the random variables given by evaluating the function at these two fixed values and the exogenous random variable are distinct. In the general case where the SCM may not be a DAG, we need to impose a further strong injectivity condition on the functions.

The assumption of distributions (of the SCM) satisfying ordered upward- and downward-stability w.r.t. the causal graph is, in principle, not directly testable. However, in our primary study, we observe that it is difficult to come up with natural examples of distributions

that do not satisfy these conditions when dealing with SCMs. Nevertheless, a thorough study of these conditions for SCMs and other causal models is needed, and is subject of future work.

## Acknowledgments

## References

R. A. Ali, T. S. Richardson, and P. Spirtes. Markov equivalence for ancestral graphs. *Ann. Statist.*, 37(5B):2808–2837, 2009.

S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of markov equivalence classes for acyclic digraphs. *Ann. Statist.*, 25(2):505–541, 1997.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *Ann. Statist.*, 49(5):2885–2915, 2021. doi: 10.1214/21-aos2064.

J. T. Chang and D. Pollard. Conditioning as disintegration. *Statist. Neerlandica*, 51(3):287–317, 1997. doi: 10.1111/1467-9574.00056.

D. M. Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, 2003. doi: 10.1162/153244303321897717.

T. Claassen and T. Heskes. A logical characterization of constraint-based causal discovery. *CoRR*, abs/1202.3711, 2012. URL http://arxiv.org/abs/1202.3711.

D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15:3741–3782, 2014.

D. R. Cox and N. Wermuth. *Multivariate Dependencies : models, analysis and interpretation*. Chapman & Hall, London, United Kingdom, 1996.

A. P. Dawid. Conditional independence in statistical theory (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 41(1):1–31, 1979.

A. P. Dawid. Invariance and independence in multivariate distribution theory. *J. Multivariate Anal.*, 17(3):304–315, 1985. doi: 10.1016/0047-259X(85)90086-7.

A. P. Dawid. Causal inference without counterfactuals. *J. Amer. Statist. Assoc.*, 95(450):407–424, 2000.

S. Fallat, S. Lauritzen, K. Sadeghi, C. Uhler, N. Wermuth, and P. Zwiernik. Total positivity in markov structures. *Annals of Statistics*, 45(3):1152–1184, 2017.

P. Forre and J. M. Mooij. Markov properties for graphical models with cycles and latent variables. *ArXiv preprint:1710.08775*, 2017.

M. Forster, G. Raskutti, R. Stern, and N. Weinberger. The frugal inference of causal relations. *The British Journal for the Philosophy of Science*, 69:821 – 848, 2018.

D. Geiger, T. S. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20:507–534, 1990.

C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering Causal Structure*. Academic Press, 1987.

C. Hitchcock. Causal models. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018.

P. W. Holland. Statistics and causal inference. *J. Amer. Statist. Assoc.*, 81(396):945–970, 1986. With discussion and a reply by the author.

H. Kiiveri, T. P. Speed, and J. B. Carlin. Recursive causal models. *J. Aust. Math. Soc., Ser. A*, 36:30–52, 1984.

J. T. A. Koster. On the validity of the markov interpretation of path diagrams of gaussian structural equations systems with correlated errors. *Scand. J. Stat.*, 26(3):413–431, 1999.

S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, United Kingdom, 1996.

S. L. Lauritzen and K. Sadeghi. Unifying Markov properties for graphical models. *Ann. Statist.*, 46(5):2251–2278, 2018.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 50:157–224, 1988.

R. Lněnička and F. Matúš. On Gaussian conditional independence structures. *Kybernetika*, 43(3):327–342, 2007.

R. E. Neapolitan. *Learning Bayesian networks*. Prentice Hall, Upper Saddle River, NJ, USA, 2004.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems : networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, CA, USA, 1988.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 052189560X, 9780521895606.

J. Pearl and A. Paz. *Graphoids: A Graph-based Logic for Reasoning about Relevance Relations*. University of California (Los Angeles). Computer Science Department, 1985.

J. Peters. On the intersection property of conditional independence and its application to causal discovery. *J. Causal Inference*, 3:97–108, 2015.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, MA, USA, 2017.

J. Ramsey, P. Spirtes, and J. Zhang. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, page 401–408, Arlington, Virginia, USA, 2006. AUAI Press. ISBN 0974903922.

G. Raskutti and C. Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018. doi: https://doi.org/10.1002/sta4.183. e183 sta4.183.

T. Richardson. Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.*, 30(1): 145–157, 2003.

T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Ann. Statist.*, 30(4): 962–1030, 2002.

K. Sadeghi. Stable mixed graphs. *Bernoulli*, 19(5(B)):2330–2358, 2013.

K. Sadeghi. Marginalization and conditioning for lwf chain graphs. *Ann. Statist.*, 44(4): 1792–1816, 2016.

K. Sadeghi. Faithfulness of probability distributions and graphs. *J. Mach. Learn. Res.*, 18 (148):1–29, 2017.

K. Sadeghi and S. L. Lauritzen. Markov properties for mixed graphs. *Bernoulli*, 20(2): 676–696, 2014.

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

D. Steel. Homogeneity, selection, and the faithfulness condition. *Minds and Machines*, 16 (3):303–317, 2006.

M. Studený. Multiinformation and the problem of characterization of conditional independence relations. *Problems of Control and Information Theory*, 18:3–16, 1989.

M. Studený. *Probabilistic Conditional Independence Structures*. Springer-Verlag, London, United Kingdom, 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.

T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence (UAI)*, 4:352–359, 03 1988.

T. Verma and J. Pearl. On the equivalence of causal models. In *Proceedings of the Proceedings of the Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-90)*, pages 220–227, New York, NY, 1990. Elsevier Science.

N. Wermuth. Probability distributions with summary graph structure. *Bernoulli*, 17(3): 845–879, 2011.

N. Wermuth and K. Sadeghi. Sequences of regressions and their independences. *TEST*, 21 (2):215–252 and 274–279, 2012.

J. Woodward. Causal independence and faithfulness. *Multivariate Behavioral Research*, 33 (1):129–148, 1998.

J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, 172(16–17):1873–1896, 2008.

J. Zhang. A comparison of three Occam's razors for Markovian causal models. *British Journal for the Philosophy of Science*, 64(2):423–448, June 2012. ISSN 0007-0882. doi: 10.1093/bjps/axs005.

J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, pages 632–639, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, 2008.

H. Zhao, Z. Zheng, and B. Liu. On the Markov equivalence of maximal ancestral graphs. *Science in China Ser. A, Mathematics*, 48:548–562, 2004.