ALTERNATIVE MARKOV AND CAUSAL PROPERTIES FOR ACYCLIC DIRECTED MIXED GRAPHS

JOSE M. PEÑA
ADIT, IDA, LINKÖPING UNIVERSITY, SE-58183 LINKÖPING, SWEDEN
JOSE.M.PENA@LIU.SE

Abstract. We extend AMP chain graphs by (i) relaxing the semidirected acyclicity constraint so that only directed cycles are forbidden, and (ii) allowing up to two edges between any pair of nodes. We introduce global, ordered local and pairwise Markov properties for the new models. We show the equivalence of these properties for strictly positive probability distributions. We also describe an exact algorithm for learning the new models via answer set programming. Finally, we show that when the random variables are continuous, the new models can be interpreted as systems of structural equations with correlated errors. This enables us to adapt Pearl’s do-calculus to them.

1. Introduction

Chain graphs (CGs) are graphs with possibly directed and undirected edges but without semidirected cycles. They have been extensively studied as a formalism to represent probabilistic independence models, because they can model symmetric and asymmetric relationships between random variables. Moreover, they are much more expressive than directed acyclic graphs (DAGs) and undirected graphs (UGs) (Sonntag and Peña, 2015b). There are three different interpretations of CGs as independence models: The Lauritzen-Wermuth-Frydenberg (LWF) interpretation (Lauritzen, 1996), the multivariate regression (MVR) interpretation (Cox and Wermuth, 1996), and the Andersson-Madigan-Perlman (AMP) interpretation (Andersson et al., 2001). No interpretation subsumes another (Andersson et al., 2001; Sonntag and Peña, 2015a). However, AMP and MVR CGs rather than LWF CGs are coherent with data generation by block-recursive normal linear regressions (Andersson et al., 2001, Sections 1 and 5).

Richardson (2003) extends MVR CGs by (i) relaxing the semidirected acyclicity constraint so that only directed cycles are forbidden, and (ii) allowing up to two edges between any pair of nodes. The resulting models are called acyclic directed mixed graphs (ADMGs). These are the models in which the do-calculus operate to determine if the causal effect of an intervention is identifiable from observed quantities (Pearl, 1995, 2009). In this paper, we make the same two extensions to AMP CGs. We call our ADMGs alternative as opposed to the ones proposed by Richardson, which we call original. It is worth mentioning that neither the original ADMGs nor any other family of mixed graphical models that we know of (e.g. summary graphs (Cox and Wermuth, 1996), ancestral graphs (Richardson and Spirtes, 2002), MC graphs (Koster, 2002) or loopless mixed graphs (Sadeghi and Lauritzen, 2014)) subsume AMP CGs and hence our alternative ADMGs. To see it, we refer the reader to the works by Richardson and Spirtes (2002, p. 1025) and Sadeghi and Lauritzen (2014, Section 4.1). Therefore, our work complements the existing works.

The rest of the paper is organized as follows. Section 2 introduces some preliminaries. Sections 3 and 4 introduce global, ordered local and pairwise Markov properties, and prove their equivalence. Section 5 describes an exact algorithm for learning ADMGs via answer set programming (Gelfond, 1988; Niemelä, 1999; Simons et al., 2002). When the random variables are continuous, Section 6 offers an intuitive interpretation of ADMGs as systems of structural equations with correlated errors, so that the do-calculus can easily be adapted to them. We close the paper with some discussion in Section 7.
2. Preliminaries

In this section, we introduce some concepts about graphical models. Unless otherwise stated, all the graphs and probability distributions in this paper are defined over a finite set $V$. The elements of $V$ are not distinguished from singletons. An ADMG $G$ is a graph with possibly directed and undirected edges but without directed cycles. There may be up to two edges between any pair of nodes, but in that case the edges must be different and one of them must be undirected to avoid directed cycles. Edges between a node and itself are not allowed. See Figure 1 for two examples of ADMGs.

The parents of $X \in V$ in an ADMG $G$ are $Pa_G(X) = \{A|A \rightarrow B \text{ is in } G \text{ with } B \in X\}$. The children of $X$ in $G$ are $Ch_G(X) = \{A|A \leftarrow B \text{ is in } G \text{ with } B \in X\}$. The neighbours of $X$ in $G$ are $Ne_G(X) = \{A|A-B \text{ is in } G \text{ with } B \in X\}$. The ancestors of $X$ in $G$ are $An_G(X) = \{A|A \rightarrow \ldots \rightarrow B \text{ is in } G \text{ with } B \in X \text{ or } A \in X\}$. The descendants of $X$ in $G$ are $De_G(X) = \{A|A \leftarrow \ldots \leftarrow B \text{ is in } G \text{ with } B \in X \text{ or } A \in X\}$. The connectivity component of $X$ in $G$ is $Cc_G(X) = \{A|A \rightarrow \ldots \rightarrow B \text{ is in } G \text{ with } B \in X \text{ or } A \in X\}$. The connectivity components in $G$ are denoted as $Cc(G)$. A route between a node $V_i$ and a node $V_n$ on $G$ is a sequence of (not necessarily distinct) nodes $V_1, \ldots, V_n$ such that $V_i$ and $V_{i+1}$ are adjacent in $G$ for all $1 \leq i < n$. If the nodes in the route are all distinct, then the route is called a path. We represent with $A \rightarrow B$ that $A \leftarrow B$ or $A - B$ (or both) is in $G$. Finally, the subgraph of $G$ induced by $X \subseteq V$, denoted as $G_X$, is the graph over $X$ that has all and only the edges in $G$ whose both ends are in $X$.

Every probability distribution $p$ satisfies the following four properties, where $X$, $Y$, $W$ and $Z$ disjoint subsets of $V$: Symmetry $X \perp p Y|Z \Rightarrow Y \perp p X|Z$, decomposition $X \perp p Y \cup W|Z \Rightarrow X \perp p Y|Z$, weak union $X \perp p Y \cup W|Z \Rightarrow X \perp p Y|Z \cup W$, and contraction $X \perp p Y|Z \cap W \cap X \perp p W|Z \Rightarrow X \perp p Y \cup W|Z$. If $p$ is strictly positive, then it also satisfies the intersection property $X \perp p Y|Z \cup W \cap X \perp p W|Z \cup Y \Rightarrow X \perp p Y \cup W|Z$. Some (not yet characterized) probability distributions also satisfy the composition property $X \perp p Y|Z \cap X \perp p W|Z \Rightarrow X \perp p Y \cup W|Z$.

3. Global Markov Property for ADMGs

In this section, we introduce four separation criteria for ADMGs. Moreover, we show that they are all equivalent for strictly positive probability distributions. A probability distribution is said to satisfy the global Markov property with respect to an ADMG if every separation in the graph can be interpreted as an independence in the distribution.

3.1. Criterion 1. A node $C$ on a path in an ADMG $G$ is said to be a collider on the path if $A \rightarrow C \leftarrow B$ is a subpath. Moreover, the path is said to be connecting given $Z \subseteq V$ when

- every collider on the path is in $An_G(Z)$, and
- every non-collider $C$ on the path is outside $Z$ unless $A-C-B$ is a subpath and $Pa_G(C) \setminus Z \neq \emptyset$.

Let $X$, $Y$ and $Z$ denote three disjoint subsets of $V$. When there is no path in $G$ connecting a node in $X$ and a node in $Y$ given $Z$, we say that $X$ is separated from $Y$ given $Z$ in $G$ and denote it as $X \perp G Y | Z$.

3.2. Criterion 2. A node $C$ on a route in an ADMG $G$ is said to be a collider on the route if $A \rightarrow C \leftarrow B$ is a subroute. Note that maybe $A = B$. Moreover, the route is said to be connecting given $Z \subseteq V$ when

- every collider on the route is in $Z$, and
- every non-collider $C$ on the route is outside $Z$.

Let $X$, $Y$ and $Z$ denote three disjoint subsets of $V$. When there is no route in $G$ connecting a node in $X$ and a node in $Y$ given $Z$, we say that $X$ is separated from $Y$ given $Z$ in $G$ and denote it as $X \perp G Y | Z$.
3.3. Criterion 3. Let $G^a$ denote the UG over $V$ that contains all and only the undirected edges in $G$. The extended subgraph $G[X]$ with $X \subseteq V$ is defined as $G[X] = G_{\text{Ang}(X)} \cup (G^a)_{\text{Ccg}(\text{Ang}(X))}$. Two nodes $A$ and $B$ in $G$ are said to be collider connected if there is a path between them such that every non-endpoint node is a collider, i.e. $A \leftarrow C \leftarrow B$ or $A \leftarrow C \rightarrow D \rightarrow B$. Such a path is called a collider path. Note that a single edge forms a collider path. The augmented graph $G^a$ is the UG over $V$ such that $A \leftarrow B$ is in $G^a$ if and only if $A$ and $B$ are collider connected in $G$. The edge $A \leftarrow B$ is called augmented if it is in $G^a$ but $A$ and $B$ are not adjacent in $G$. A path in $G^a$ is said to be connecting given $Z \subseteq V$ if no node on the path is in $Z$. Let $X$, $Y$, and $Z$ denote three disjoint subsets of $V$. When there is no path in $G[X \cup Y \cup Z]^a$ connecting a node in $X$ and a node in $Y$ given $Z$, we say that $X$ is separated from $Y$ given $Z$ in $G$ and denote it as $X \perp_G Y | Z$.

3.4. Criterion 4. Given an UG $H$ over $V$ and $X \subseteq V$, we define the marginal graph $H^X$ as the UG over $X$ such that $A \leftarrow B$ is in $H^X$ if and only if $A \leftarrow B$ is in $H$ or $A \leftarrow V_1 \cdots \leftarrow V_n \leftarrow B$ is $H$ with $V_1, \ldots, V_n \notin X$. We define the marginal extended subgraph $G(X)^m$ as $G(X)^m = G_{\text{Ang}(X)} \cup ((G^a)_{\text{Ccg}(\text{Ang}(X))})_{\text{Ang}(X)}$. Let $X$, $Y$, and $Z$ denote three disjoint subsets of $V$. When there is no path in $(G[X \cup Y \cup Z]^m)^a$ connecting a node in $X$ and a node in $Y$ given $Z$, we say that $X$ is separated from $Y$ given $Z$ in $G$ and denote it as $X \perp_G Y | Z$.

The first three separation criteria introduced above coincide with those introduced by Andersson et al. (2001) and Levitz et al. (2001) for AMP CGs. The equivalence for AMP CGs of these three separation criteria has been proven by Levitz et al. (2001) Theorem 4.1). We prove below the equivalence for ADMGs of the four separation criteria introduced above.

**Lemma 1.** If there is a path $\rho$ in an ADMG $G$ between $A \in X$ and $B \in Y$ such that (i) no non-collider $C$ on $\rho$ is in $Z$ unless $A \leftarrow C \leftarrow B$ is a subpath of $\rho$ and $Pa_G(C) \setminus Z \neq \emptyset$, and (ii) every collider on $\rho$ is in $\text{Ang}_G(X \cup Y \cup Z)$, then there is a path in $G$ connecting a node in $X$ and a node in $Y$ given $Z$.

**Proof.** Suppose that $\rho$ has a collider $C$ such that $C \in \text{Ang}_G(D) \setminus \text{Ang}_G(Z)$ with $D \in X$, or $C \in \text{Ang}_G(E) \setminus \text{Ang}_G(Z)$ with $E \in Y$. Assume without loss of generality that $C \in \text{Ang}_G(D) \setminus \text{Ang}_G(Z)$ with $D \in X$ because, otherwise, a symmetric argument applies. Then, replace the subpath of $\rho$ between $A$ and $C$ with $D \leftarrow \ldots \leftarrow C$. Note that the resulting path (i) has no non-collider in $Z$ unless $A \leftarrow C \leftarrow B$ is a subpath of $\rho$ and $Pa_G(C) \setminus Z \neq \emptyset$, and (ii) has every collider in $\text{Ang}_G(X \cup Y \cup Z)$. Note also that the resulting path has fewer colliders than $\rho$ that are not in $\text{Ang}_G(Z)$. Continuing with this process until no such collider $C$ exists produces the desired result. □

**Lemma 2.** Given an ADMG $G$, let $\rho$ denote a shortest path in $G[A \cup B \cup Z]^a$ connecting two nodes $A$ and $B$ given $Z$. Then, a path in $G$ between $A$ and $B$ can be obtained as follows. First, replace every augmented edge on $\rho$ with an associated collider path in $G[A \cup B \cup Z]$. Second, replace every non-augmented edge on $\rho$ with an associated edge in $G[A \cup B \cup Z]$. Third, replace any configuration $C \leftarrow D \leftarrow F \rightarrow D \leftarrow E$ produced in the previous steps with $C \leftarrow D \leftarrow E$.

**Proof.** We start by proving that the collider paths added in the first step of the lemma either do not have any node in common except possibly one of the endpoints, or the third step of the lemma removes the repeated nodes. Suppose for a contradiction that $C \leftarrow D$ and $C' \leftarrow D'$ are two augmented edges on $\rho$ such that their associated collider paths have in common a node which is not an endpoint of these paths. Consider the following two cases.

**Case 1:** Suppose that $D \neq C'$. Then, one of the following configurations must exist in $G[A \cup B \cup Z]$. 

However, the first case implies that $C - D'$ is in $G[A \cup B \cup Z]^a$, which implies that replacing the subpath of $\rho$ between $C$ and $D'$ with $C - D'$ results in a path in $G[A \cup B \cup Z]^a$ connecting $A$ and $B$ given $Z$ that is shorter than $\rho$. This is a contradiction. Similarly for the fourth, sixth, seventh, eighth, ninth and tenth cases. And similarly for the rest of the cases by replacing the subpath of $\rho$ between $D$ and $D'$ with $D - D'$.

**Case 2:** Suppose that $D = C'$. Then, one of the following configurations must exist in $G[A \cup B \cup Z]$.

However, the first case implies that $C - D'$ is in $G[A \cup B \cup Z]^a$, which implies that replacing the subpath of $\rho$ between $C$ and $D'$ with $C - D'$ results in a path in $G[A \cup B \cup Z]^a$ connecting $A$ and $B$ given $Z$ that is shorter than $\rho$. This is a contradiction. Similarly for the second, fourth and seventh cases. For the third, fifth and sixth cases, the third step of the lemma removes the repeated nodes. Specifically, it replaces $C - E \leftarrow D \rightarrow E - D'$ in the third case, $E - F \leftarrow D \rightarrow F - D'$ with $E - F - D'$ in the fifth case, and $E - F \leftarrow D \rightarrow F - H$ with $E - F - H$ in the sixth case.

It only remains to prove that the collider paths added in the first step of the lemma have no nodes in common with $\rho$ except the endpoints. Suppose that $\rho$ has an augmented edge $C - D$. Then, one of the following configurations must exist in $G[A \cup B \cup Z]$.

Consider the first case and suppose for a contradiction that $E$ occurs on $\rho$. Note that $E \notin Z$ because, otherwise, $\rho$ would not be connecting. Assume without loss of generality that $E$ occurs on $\rho$ before $C$ and $D$ because, otherwise, a symmetric argument applies. Then, replacing the subpath of $\rho$ between $E$ and $D$ with $E - D$ results in a path in $G[A \cup B \cup Z]^a$ connecting $A$ and $B$ given $Z$ that is shorter than $\rho$. This is a contradiction. Similarly for the second case. Specifically, assume without loss of generality that $E$ occurs on $\rho$ because, otherwise, a symmetric argument with $F$ applies. Note that $E \notin Z$ because, otherwise, $\rho$ would not be connecting. If $E$ occurs on $\rho$ after
There is a path in an ADMG $\overrightarrow{G}$ connecting two nodes $A$ and $B$ given $Z$. The sequence of non-colliders on $\rho$ forms a path in $G[A \cup B \cup Z]^a$ between $A$ and $B$.

Proof. Consider the maximal undirected subpaths of $\rho$. Note that each endpoint of each subpath is ancestor of a collider or endpoint of $\rho$, because $\rho$ is connecting. Thus, all the nodes on $\rho$ are in $G[A \cup B \cup Z]$. Suppose that $C$ and $D$ are two successive non-colliders on $\rho$. Then, the subpath of $\rho$ between $C$ and $D$ consists entirely of colliders. Specifically, the subpath is of the form $C \leftarrow D$, $C \rightarrow D$, $C \rightarrow E \leftarrow D$ or $C \rightarrow E \rightarrow F \rightarrow D$. Then, $C$ and $D$ are adjacent in $G[A \cup B \cup Z]^a$.

Theorem 1. There is a path in an ADMG $G$ connecting a node in $X$ and a node in $Y$ given $Z$ if and only if there is a path in $G[X \cup Y \cup Z]^a$ connecting a node in $X$ and a node in $Y$ given $Z$.

Proof. We start by proving the only if part. Let $\rho$ denote a path in $G$ connecting $A \in X$ and $B \in Y$ given $Z$. By Lemma 3 the non-colliders on $\rho$ form a path $\rho^a$ between $A$ and $B$ in $G[X \cup Y \cup Z]^a$. Since $\rho$ is connecting, every non-collider $C$ on $\rho$ is outside $Z$ unless $D - C - E$ is a subpath of $\rho$ and $P_{AG}(C) \setminus Z \neq \emptyset$. In the latter case, replace the subpath $D - C - E$ of $\rho^a$ with $D - F - E$ where $F \in P_{AG}(C) \setminus Z$. Then, $\rho^a$ is connecting given $Z$. Note that $D - F$ and $F - E$ are in $G[X \cup Y \cup Z]^a$, because $D - C - E$ and $F \rightarrow C$ are in $G$ with $C \in Z$.

To prove the if part, let $\rho^a$ denote a shortest path in $G[X \cup Y \cup Z]^a$ connecting $A \in X$ and $B \in Y$ given $Z$. We can transform $\rho^a$ into a path $\rho$ in $G$ as described in Lemma 2. Since $\rho^a$ is connecting, no node on $\rho^a$ is in $Z$ and, thus, no non-collider on $\rho$ is in $Z$. Finally, since all the nodes on $\rho$ are in $G[X \cup Y \cup Z]^a$, it follows that every collider on $\rho$ is in $An_G(X \cup Y \cup Z)$. To see it, note that if $C - D$ is an augmented edge in $G[X \cup Y \cup Z]^a$ then the colliders on any collider path associated with $C - D$ are in $An_G(X \cup Y \cup Z)$. Thus, by Lemma 1 there exist a node in $X$ and a node in $Y$ which are connected given $Z$ in $G$.

Theorem 2. There is a path in an ADMG $G$ connecting $A$ and $B$ given $Z$ if and only if there is a route in $G$ connecting $A$ and $B$ given $Z$.

Proof. The only if part is trivial. To prove the if part, let $\rho$ denote a route in $G$ connecting $A$ and $B$ given $Z$. Let $C$ denote a node that occurs more than once in $\rho$. Consider the following cases.

Case 1: Assume that $\rho$ is of the form $A \ldots D \rightarrow C \ldots C \rightarrow E \ldots B$. Then, $C \notin Z$ for $\rho$ to be connecting given $Z$. Then, removing the subroute between the two occurrences of $C$ from $\rho$ results in the route $A \ldots D \rightarrow C \rightarrow E \ldots B$, which is connecting given $Z$.

Case 2: Assume that $\rho$ is of the form $A \ldots D \rightarrow C \ldots C \leftarrow E \ldots B$. Then, $C \in An_G(Z)$ for $\rho$ to be connecting given $Z$. Then, removing the subroute between the two occurrences of $C$ from $\rho$ results in the route $A \ldots D \rightarrow C \leftarrow E \ldots B$, which is connecting given $Z$.

Case 3: Assume that $\rho$ is of the form $A \ldots D \leftarrow C \ldots C \ldots B$. Then, $C \notin Z$ for $\rho$ to be connecting given $Z$. Then, removing the subroute between the two occurrences of $C$ from $\rho$ results in the route $A \ldots D \leftarrow C \ldots B$, which is connecting given $Z$.

Case 4: Assume that $\rho$ is of the form $A \ldots D - C \ldots C \rightarrow E \ldots B$. Then, $C \notin Z$ for $\rho$ to be connecting given $Z$. Then, removing the subroute between the two occurrences of $C$ from $\rho$ results in the route $A \ldots D - C \rightarrow E \ldots B$, which is connecting given $Z$.

Case 5: Assume that $\rho$ is of the form $A \ldots D - C \ldots C \leftarrow E \ldots B$. Then, $C \in An_G(Z)$ for $\rho$ to be connecting given $Z$. Then, removing the subroute between the two occurrences of $C$ from $\rho$ results in the route $A \ldots D - C \leftarrow E \ldots B$, which is connecting given $Z$.

Case 6: Assume that $\rho$ is of the form $A \ldots D - C \ldots C \rightarrow E \ldots B$ and $C \notin Z$. Then, removing the subroute between the two occurrences of $C$ from $\rho$ results in the route $A \ldots D - C \rightarrow E \ldots B$, which is connecting given $Z$.

Case 7: Assume that $\rho$ is of the form $A \ldots D - C \ldots C \rightarrow E \ldots B$ and $C \in Z$. Then, $\rho$ must actually be of the form $A \ldots D - C \leftarrow F \ldots C \rightarrow E \ldots B$ or $A \ldots D - C \rightarrow F \ldots C \leftarrow E \ldots B$. Note that in the former case $F \notin Z$ for $\rho$ to be connecting given $Z$. For the same reason,
Proof. First, if \( A \rightarrow C \leftarrow D \) in an ADMG, then \( A \perp_{G} D \mid Z \) for some \( Z \). This is due
to the following reasons.

Case 1: \( A \rightarrow C \leftarrow D \) is in \( G \). Then, every subpath \( C \rightarrow V_{1} \rightarrow \ldots \rightarrow V_{n} \) in \( G \) is not

Case 2: \( A \rightarrow C \leftarrow D \) is in \( G \). Then, every subpath \( C \rightarrow V_{1} \rightarrow \ldots \rightarrow V_{n} \) in \( G \) is not

Case 3: \( A \rightarrow C \leftarrow D \) is in \( G \). Then, every subpath \( C \rightarrow V_{1} \rightarrow \ldots \rightarrow V_{n} \) in \( G \) is not

Either case implies that there is a path in \( G \) connecting \( A \) and \( B \) in \( G \).

Unlike in AMP CGs, two non-adjacent nodes in an ADMG are not necessarily separated. For example, \( A \perp_{G} D \mid Z \) does not hold for any \( Z \) in the ADMGs in Figure 1. This drawback is shared by the original ADMGs (Evans and Richardson, 2013, p. 752), summary graphs and MC graphs (Richardson and Spirtes, 2002, p. 1023), and ancestral graphs (Richardson and Spirtes, 2002, Section 3.7). For ancestral graphs, the problem can be solved by adding edges to the graph without altering the separations represented until every missing edge corresponds to a separation (Richardson and Spirtes, 2002, Section 5.1). A similar solution does not exist for our ADMGs (we omit the details). In any case, fixing the problem by adding edges may have a negative effect in the parameterization of the model because, typically, the more the edges the more the parameters to specify and estimate. So, one may prefer to acknowledge the problem and live with it rather than to fix it and create a new problem. The following corollary characterizes when two non-adjacent nodes in an ADMG are not separated.

Corollary 1. Given two nodes \( A \) and \( B \) that are not adjacent in an ADMG \( G \), \( A \perp_{G} B \mid S \) for some \( S \subseteq V \setminus (A \cup B) \) if and only if \( G \) has no subgraph of the form

\[
A \rightarrow V_{1} \rightarrow \ldots \rightarrow V_{n} \rightarrow B.
\]

Proof. First, if \( G \) has a subgraph of the form stated in the lemma, then it is clear that \( A \perp_{G} B \mid S \) does not hold for any \( S \subseteq V \setminus (A \cup B) \). If \( V_{1} \rightarrow \ldots \rightarrow V_{n} \) is the path \( A \rightarrow V_{1} \rightarrow \ldots \rightarrow V_{n} \rightarrow B \) is connecting, then the path \( A \rightarrow V_{1} \rightarrow \ldots \rightarrow V_{n} \rightarrow B \) is connecting.

Second, assume that \( G \) has no subgraph of the form stated in the lemma, and that \( A \in \text{Dec}(B) \) and \( B \notin \text{Dec}(A) \). Note that \( A \) and \( B \) are not collider connected in \( G \) since they are not adjacent in \( G \), and the existence of a path \( A \rightarrow C \leftarrow B \) or \( A \rightarrow C \leftarrow D \leftarrow B \) implies that \( C \in \text{An}(B) \), which contradicts that \( B \notin \text{Dec}(A) \). Then, \( A \perp_{G} B \mid \phi \).

Finally, assume that \( G \) has no subgraph of the form stated in the lemma and, without loss of generality, that \( B \in \text{Dec}(A) \). Note that \( A \) and \( B \) are not collider connected in \( G \) since they are not adjacent in \( G \), and the existence of a path \( A \rightarrow C \leftarrow B \) or \( A \rightarrow C \leftarrow D \leftarrow B \) implies that
$C \in \text{An}_G(B)$ and $D \in \text{An}_G(A)$ and, thus, the existence of a directed cycle in $G$ or a subgraph of the form stated in the lemma, which is a contradiction. Then, $A \perp_G B | \emptyset$. 

4. Ordered Local and Pairwise Markov Properties for ADMGs

In this section, we introduce ordered local and pairwise Markov properties for ADMGs. Given an ADMG $G$, the directed acyclicity of $G$ implies that we can specify a total ordering ($<$) of the nodes of $G$ such that $A < B$ only if $B \notin \text{An}_G(A)$. Such an ordering is said to be consistent with $G$. Let the predecessors of $A$ be defined as $\text{Pre}_G(A, <) = \{ B | B < A \text{ or } A = B \}$. Given $S \subseteq V$, we define the Markov blanket of $B \in S$ with respect to $G[S]$ as $M_{b_G}[S](B) = Ch_{G[S]}(B) \cup \text{Ne}_{G[S]}(B \cup Ch_{G[S]}(B))$. We say that a probability distribution $p$ satisfies the ordered local Markov property with respect to $G$ and $< \text{ if for any } A \in V \text{ and } S \subseteq \text{Pre}_G(A, <) \text{ such that } A \in S$

$$B \perp_p S \setminus (B \cup M_{b_G}[S](B)) | M_{b_G}[S](B)$$

for all $B \in S$.

**Theorem 4.** Given a probability distribution $p$ satisfying the intersection property, $p$ satisfies the global Markov property with respect to an ADMG if and only if it satisfies the ordered local Markov property with respect to the ADMG and a consistent ordering of its nodes.

**Proof.** We start by proving the only if part. It suffices to note that every node that is adjacent to $B$ in $G[S]^n$ is in $M_{b_G}[S](B)$, hence $B$ is separated from $S \setminus (B \cup M_{b_G}[S](B))$ given $M_{b_G}[S](B)$ in $G[S]^n$. Thus, $B \perp_p S \setminus (B \cup M_{b_G}[S](B)) | M_{b_G}[S](B)$ by the global Markov property.

To prove the if part, let $A$ be the node in $X \cup Y \cup Z$ that occurs the latest in $<$, and let $S = X \cup Y \cup Z$. Note that for all $B \in S$, the set of nodes that are adjacent to $B$ in $G[S]^n$ is precisely $M_{b_G}[S](B)$. Then, the ordered local Markov property implies the global Markov property by decomposition ([Lauritzen], 1996, Theorem 3.7). 

Similarly, we say that a probability distribution $p$ satisfies the ordered pairwise Markov property with respect to $G$ and $< \text{ if for any } A \in V \text{ and } S \subseteq \text{Pre}_G(A, <) \text{ such that } A \in S$

$$B \perp_p C | V(G[S]) \setminus (B \cup C)$$

for all nodes $B, C \in S$ that are not adjacent in $G[S]^n$, and where $V(G[S])$ denotes the nodes in $G[S]$.

**Theorem 5.** Given a probability distribution $p$ satisfying the intersection property, $p$ satisfies the global Markov property with respect to an ADMG if and only if it satisfies the ordered pairwise Markov property with respect to the ADMG and a consistent ordering of its nodes.

**Proof.** We start by proving the only if part. It suffices to note that if $B$ and $C$ are not adjacent in $G[S]^n$, then they are separated in $G[S]^n$ given $V(G[S]) \setminus (B \cup C)$. Thus, $B \perp_p C | V(G[S]) \setminus (B \cup C)$ by the global Markov property.

To prove the if part, let $A$ be the node in $X \cup Y \cup Z$ that occurs the latest in $<$, and let $S = X \cup Y \cup Z$. Then, the ordered pairwise Markov property implies the global Markov property by decomposition ([Lauritzen], 1996, Theorem 3.7). 

For each $A \in V$ and $S \subseteq \text{Pre}_G(A, <)$ such that $A \in S$, the ordered local Markov property specifies an independence for each $B \in S$. The number of independencies to specify can be further reduced by considering only maximal ancestral sets, i.e., those such that $M_{b_G}[S](B) \subseteq M_{b_G}[T](B)$ for every ancestral set $T$ such that $S \subseteq T \subseteq \text{Pre}_G(A, <)$. The independencies for the non-maximal ancestral sets follow from the independencies for the maximal ancestral sets by decomposition. A characterization of the maximal ancestral sets is possible but notionally cumbersome (we omit the details). All in all, for each node and maximal ancestral set, the ordered local Markov property specifies an independence for each node in the set. This number is greater than for the original ADMGs, where a single independence is specified for each node and maximal ancestral set ([Richardson], 2003, Section 3.1).
Note that [Andersson et al. 2001, Theorem 3] describe local and pairwise Markov properties for AMP CGs that are equivalent to the global one under the assumption of the intersection and composition properties. Our ordered local Markov and pairwise properties above only require assuming the intersection property. This assumption is in line with similar results for UGs (Lauritzen, 1996, Theorem 3.7). For AMP CGs, however, we can do better than just using the ordered local and pairwise Markov properties for ADMGs above. Specifically, we introduce in the next section neater local and pairwise Markov properties for AMP CGs under the intersection property assumption.

4.1. Local and Pairwise Markov Properties for AMP CGs. [Andersson et al. 2001, Theorem 2] introduce the following block-recursive Markov property. A probability distribution $p$ satisfies the global Markov property with respect to an AMP CG $G$ if and only if the following three properties hold for all $C ∈ Cc(G)$:

- **C1:** $C⊥_pNd_G(C) \setminus Cc(G)(Pa_G(C)) \setminus Cc(G)(Pa_G(C))$.
- **C2:** $p(C|Cc(G)(Pa_G(C)))$ satisfies the global Markov property with respect to $G$. C.
- **C3:** $D ⊥_pCc(G)(Pa_G(D)) \setminus Pa_G(D) | Pa_G(D)$ for all $D ⊆ C$.

We simplify the block-recursive Markov property as follows.

**Lemma 4.** $C1$, $C2$ and $C3$ hold if and only if the following two properties hold:

- **C1**: $D ⊥_pNd_G(D) \setminus Pa_G(D) | Pa_G(D)$ for all $D ⊆ C$.
- **C2**: $p(C|Pa_G(C))$ satisfies the global Markov property with respect to $G$.

**Proof.** First, $C1^*$ implies $C3^*$ by decomposition. Second, $C1^*$ implies $C1$ by taking $D = C$ and applying weak union. Third, $C1$ and the fact that $Nd_G(D) = Nd_G(C)$ imply $D ⊥_pNd_G(D) \setminus Cc(G)(Pa_G(C)) \setminus Cc(G)(Pa_G(C))$ by symmetry and decomposition, which together with $C3^*$ imply $C1^*$ by contraction. Finally, $C2$ and $C2^*$ are equivalent because $p(C|Pa_G(C)) = p(C|Cc(G)(Pa_G(C)))$ by $C1^*$ and decomposition.

[Andersson et al. 2001, Theorem 3] also introduce the following local Markov property. A probability distribution $p$ satisfying the intersection and composition properties satisfies the global Markov property with respect to an AMP CG $G$ if and only if the following two properties hold for all $C ∈ Cc(G)$:

- **L1:** $X ⊥_p(C \setminus (X ∪ Ne_G(X)))[Nd_G(C) \cup Ne_G(X)]$ for all $X ∈ C$.
- **L2:** $X ⊥_pNd_G(C) \setminus Pa_G(X) | Pa_G(X)$ for all $X ∈ C$.

We introduce below a local Markov property that is equivalent to the global one under the assumption of the intersection property only.

**Theorem 6.** A probability distribution $p$ satisfying the intersection property satisfies the global Markov property with respect to an AMP CG $G$ if and only if the following two properties hold for all $C ∈ Cc(G)$:

- **L1:** $X ⊥_p(C \setminus (X ∪ Ne_G(X)))[Nd_G(C) \cup Ne_G(X)]$ for all $X ∈ C$.
- **L2**: $X ⊥_pNd_G(C) \setminus Pa_G(X) | Pa_G(X) | Pa_G(X)$ for all $X ∈ C$ and $Z ⊆ C \setminus X$.

**Proof.** To see the only if part, note that $C1^*$ with $D = C$ implies that $p(C|Nd_G(C)) = p(C|Pa_G(C))$. This implies that $p(C|Nd_G(C))$ satisfies the global Markov property with respect to $G$. C. This implies $L1$ [Lauritzen, 1996, Theorem 3.7]. Moreover, let $D = X ∪ Z$ and note that $Nd_G(D) = Nd_G(C)$. Then, $L2^*$ follows from $C1^*$ by symmetry and weak union.

To see the if part, note that $L2^*$ with $Z = Ne_G(X)$ implies that $X ⊥_pNd_G(C) \setminus Pa_G(C) | Ne_G(X) \cup Pa_G(C)$ by weak union. This together with $L1$ imply $X ⊥_p(C \setminus (X ∪ Ne_G(X)))[Nd_G(C) \cup Ne_G(X) | Pa_G(C)]$ by contractions and decomposition. This implies $C2^*$ [Lauritzen, 1996, Theorem 3.7]. Moreover, let $D = \{D_1, \ldots, D_n\}$. Then

1. $D_1 ⊥_pNd_G(C) \setminus Pa_G(D)(D \setminus D_1) \setminus Pa_G(D)$ by $L2^*$ with $X = D_1$ and $Z = D \setminus D_1$.
2. $D_2 ⊥_pNd_G(C) \setminus Pa_G(D)(D \setminus D_2) \setminus Pa_G(D)$ by $L2^*$ with $X = D_2$ and $Z = D \setminus D_2$.
3. $D_1 \cup D_2 ⊥_pNd_G(C) \setminus Pa_G(D)(D \setminus (D_1 \cup D_2)) \setminus Pa_G(D)$ by symmetry and intersection on (1) and (2).
4. $D_3 ⊥_pNd_G(C) \setminus Pa_G(D)(D \setminus D_3) \setminus Pa_G(D)$ by $L2^*$ with $X = D_3$ and $Z = D \setminus D_3$.
5. $D_1 \cup D_2 \cup D_3 ⊥_pNd_G(C) \setminus Pa_G(D)(D \setminus (D_1 \cup D_2 \cup D_3)) \setminus Pa_G(D)$ by symmetry and intersection on (3) and (4).
Continuing with this for $D_4, \ldots, D_n$ leads to $C1^\ast$. □

Finally, Andersson et al. (2001, Theorem 3) also introduce the following pairwise Markov property. A probability distribution $p$ satisfying the intersection and composition properties satisfies the global Markov property with respect to an AMP CG $G$ if and only if the following two properties hold for all $C \in \text{Cc}(G)$:

- P1: $X \indep Y | Nd_G(C) \cup C \setminus (X \cup Y)$ for all $X \in C$ and $Y \in C \setminus (X \cup \text{Ne}_G(X))$.
- P2: $X \indep Y | Nd_G(C) \setminus Y$ for all $X \in C$ and $Y \in Nd_G(C) \setminus \text{Pa}_G(X)$.

We introduce below a pairwise Markov property that is equivalent to the global one under the assumption of the intersection property only.

**Theorem 7.** A probability distribution $p$ satisfying the intersection property satisfies the global Markov property with respect to an AMP CG $G$ if and only if the following two properties hold for all $C \in \text{Cc}(G)$:

- P1: $X \indep Y | Nd_G(C) \cup C \setminus (X \cup Y)$ for all $X \in C$ and $Y \in C \setminus (X \cup \text{Ne}_G(X))$.
- P2*: $X \indep Y | Z \cup Nd_G(C) \setminus Y$ for all $X \in C$, $Z \subseteq C \setminus X$ and $Y \in Nd_G(C) \setminus \text{Pa}_G(X \cup Z)$.

**Proof.** To see the only if part, note that L1 and L2* imply P1 and P2* by weak union. To see the if part, let $Nd_G(C) \setminus \text{Pa}_G(X \cup Z) = \{Y_1, \ldots, Y_n\}$. Then

(1) $X \indep Y_1 | Z \cup Nd_G(C) \setminus Y_1$ by P2* with $Y = Y_1$.
(2) $X \indep Y_2 | Z \cup Nd_G(C) \setminus Y_2$ by P2* with $Y = Y_2$.
(3) $X \indep Y_1 \cup Y_2 | Z \cup Nd_G(C) \setminus (Y_1 \cup Y_2)$ by intersection on (1) and (2).
(4) $X \indep Y_3 | Z \cup Nd_G(C) \setminus Y_3$ by P2* with $Y = Y_3$.
(5) $X \indep Y_1 \cup Y_2 \cup Y_3 | Z \cup Nd_G(C) \setminus (Y_1 \cup Y_2 \cup Y_3)$ by intersection on (3) and (4).

Continuing with this for $Y_4, \ldots, Y_n$ leads to L2*. Finally, let $C \setminus (X \cup \text{Ne}_G(X)) = \{Y_1, \ldots, Y_n\}$. Then

(6) $X \indep Y_1 | Nd_G(C) \setminus C \setminus (X \cup Y_1)$ by P1 with $Y = Y_1$.
(7) $X \indep Y_2 | Nd_G(C) \setminus C \setminus (X \cup Y_2)$ by P1 with $Y = Y_2$.
(8) $X \indep Y_1 \cup Y_2 | Nd_G(C) \setminus C \setminus (X \cup Y_1 \cup Y_2)$ by intersection on (6) and (7).
(9) $X \indep Y_3 | Nd_G(C) \setminus C \setminus (X \cup Y_3)$ by P1 with $Y = Y_3$.
(10) $X \indep Y_1 \cup Y_2 \cup Y_3 | Nd_G(C) \setminus (Y_1 \cup Y_2 \cup Y_3)$ by intersection on (8) and (9).

Continuing with this for $Y_4, \ldots, Y_n$ leads to L1. □

5. Learning ADMGs Via ASP

In this section, we introduce an exact algorithm for learning ADMGs via answer set programming (ASP), which is a declarative constraint satisfaction paradigm that is well-suited for representing and solving computationally hard combinatorial problems (Gelfond, 1988; Niemelä, 1999; Simons et al., 2002). ASP represents constraints in terms of first-order logical rules. Therefore, when using ASP, the first task is to model the problem at hand in terms of rules so that the set of solutions implicitly represented by the rules corresponds to the solutions of the original problem. One or multiple solutions of the original problem can then be obtained by invoking an off-the-shelf ASP solver on the constraint declaration. The algorithms underlying the ASP solver cling (Gebser et al., 2011), which we use in this work, is based on state-of-the-art Boolean satisfiability solving techniques (Biere et al., 2009).

Figure 2 shows the ASP encoding of the learning algorithm. The predicate node($X$) in rule 1 represents that $X$ is a node. The predicates line($X$, $Y$) and arrow($X$, $Y$) represent that there is an undirected and directed edge from $X$ to $Y$. The rules 2-3 encode a non-deterministic guess of the edges, which means that the ASP solver with implicitly consider all possible graphs during search, hence the exactness of the search. The rules 4-5 enforce the fact that undirected edges are symmetric and that there can be at most one directed edge between two nodes. The predicate ancestor($X$, $Y$) represents that $X$ is an ancestor of $Y$. The rules 6-8 enforce that the graph has no directed cycles. The predicates in the rules 9-10 represent whether a node $X$ is or is not in a set of nodes $C$. The rules 11-22 encode the separation criterion 2 in Section 3. The predicate con($X$, $Y$, $C$) in rules 23-26 represents that there is a connecting route between $X$ and $Y$ given $C$. The rule 27 enforces that each dependence in the input must correspond to a connecting route. The rule 28 represents that each independence in the input that is not represented implies a penalty of $W$ units. In our case,
\[ W = 1. \] The rules 29-31 imply a penalty of 1 unit per edge. Other penalty rules can be added similarly.

The left-hand side of Figure 2 shows the ASP encoding of all the (in)dependencies in the probability distribution at hand, e.g. as determined by some available data. In our case, there are only dependencies. Note that it suffices to specify all the (in)dependencies between pair of nodes, because these identify uniquely the rest of the independencies in the probability distribution (Studeny, 2005, Lemma 2.2). Specifically, the predicate \text{dep}(X,Y,C,W)\] represents that there is a dependence between the nodes \(X\) and \(Y\) given the set \(C\). The penalty for failing to represent this dependence is \(W\). The predicate \text{nodes}(3)\] represents that there are three nodes, and the predicate \text{set}(0..7)\] represents that there are eight sets of nodes, indexed from 0 (empty set) to 7 (full set). Note that we do not make any assumption (e.g. faithfulness) about the probability distribution at hand.

By calling an ASP solver with the encodings of the learning algorithm and the (in)dependencies in the domain, the solver will essentially perform an exhaustive search over the space of graphs, and will output the graphs with the smallest penalty. In our case, the learning algorithm finds 37 optimal models. Among them, we have UGs such as line(1,2) line(2,1) line(1,3) line(3,1) line(2,3) line(3,2), DAGs such as arrow(1,2) arrow(1,3) arrow(2,3), AMP CGs such as line(1,2) line(2,1) arrow(1,3) arrow(2,3), and ADMGs such as line(1,2) line(2,1) line(2,3) line(3,2) arrow(1,3) arrow(1,2) arrow(1,3) arrow(2,3). When we constrain the graphs to be consistent with the ordering \(1 \prec 2 \prec 3\) (i.e. we add the rules on the right-hand side of Figure 3), the algorithm finds 12 optimal models when the number of edges is minimized, and 33 otherwise. Both cases include ADMGs.

Finally, the ASP code can easily be extended as shown in Figure 4 to learn not only our ADMGs but also original ADMGs.

6. Causal Interpretation of ADMGs

Let us assume that \(V\) is normally distributed. In this section, we show that an ADMG \(G\) can be interpreted as a system of structural equations with correlated errors. Specifically, the system includes an equation for each \(A \in V\), which is of the form

\[ A = \beta_A Pa_G(A) + \epsilon_A \]

where \(\epsilon_A\) denotes the error term. The error terms are represented implicitly in \(G\). They can be represented explicitly by magnifying \(G\) into the ADMG \(G'\) as follows:

1. Set \(G' = G\)
2. For each node \(A\) in \(G\)
3. Add the node \(\epsilon_A\) and the edge \(\epsilon_A \rightarrow A\) to \(G'\)
4. For each edge \(A - B\) in \(G\)
5. Replace \(A - B\) with the edge \(\epsilon_A - \epsilon_B\) in \(G'\)

The magnification above basically consists in adding the error nodes \(\epsilon_A\) to \(G\) and connect them appropriately. Figure 5 shows an example. Note that every node \(A \in V\) is determined by \(Pa_{G'}(A)\) and that \(\epsilon_A\) is determined by \(A \cup Pa_G(A) \setminus \epsilon_A\). Let \(\epsilon\) denote all the error nodes in \(G'\). Formally, we say that \(A \in V \cup \epsilon\) is determined by \(Z \subseteq V \cup \epsilon\) when \(A \in Z\) or \(A\) is a function of \(Z\). We use \(Dt(Z)\) to denote all the nodes that are determined by \(Z\). From the point of view of the separations, that a node outside the conditioning set of a separation is determined by the conditioning set has the same effect as if the node were actually in the conditioning set. Bearing this in mind, it is not difficult to see that, as desired, \(G\) and \(G'\) represent the same separations over \(V\). Specifically, let \(X, Y\) and \(Z\) denote three disjoint subsets of \(V\), then \(X\) is separated from \(Y\) given \(Z\) in \(G\) if and only if \(X\) is separated from \(Y\) given \(Dt(Z)\) in \(G'\) (Peña, 2014, Theorem 1)\footnote{The theorem proves the result for AMP CGs. However, it also applies to ADMGs because the proof does not make use of the differences between AMP CGs and ADMGs.} Finally, let \(\epsilon \sim \mathcal{N}(0, \Lambda)\) such that
% input predicates:
% indep(X,Y,C): the nodes X and Y are independent given the set of nodes C
% dep(X,Y,C): the nodes X and Y are not independent given the set of nodes C
% nodes(N): number of nodes
% set(X): X is the index of a set of nodes

% nodes
node(X) :- nodes(N), X=1..N. % rule 1

% edges
\{ line(X,Y) \} :- node(X), node(Y), X != Y. % 2
\{ arrow(X,Y) \} :- node(X), node(Y), X != Y. % 3
line(X,Y) :- line(Y,X). % 4
arrow(X,Y) :- arrow(Y,X). % 5

% directed acyclicity
ancestor(X,Y) :- arrow(X,Y). % 6
ancestor(X,Y) :- ancestor(X,Z), ancestor(Z,Y).

% set membership
inside_set(X,C) :- node(X), set(C), 2**(X-1) & C != 0. % 9
outside_set(X,C) :- node(X), set(C), 2**(X-1) & C == 0. % 10

% end_line/head/tail(X,Y,C) means that there is a connecting route
given C from X to Y that ends with an line/arrowhead/arrowtail

% single link route
end_line(X,Y,C) :- line(X,Y), outside_set(X,C). % 11
end_head(X,Y,C) :- arrow(X,Y), outside_set(X,C).
end_tail(X,Y,C) :- arrow(Y,X), outside_set(X,C).

% diverging connection
end_line(X,Y,C) :- end_line(X,Z,C), line(Z,Y), outside_set(Z,C).
end_line(X,Y,C) :- end_tail(X,Z,C), line(Z,Y), outside_set(Z,C).
end_head(X,Y,C) :- end_line(X,Z,C), arrow(Z,Y), outside_set(Z,C).
end_head(X,Y,C) :- end_head(X,Z,C), arrow(Z,Y), outside_set(Z,C).
end_tail(X,Y,C) :- end_tail(X,Z,C), arrow(Z,Y), outside_set(Z,C).
end_tail(X,Y,C) :- end_tail(X,Z,C), arrow(Y,Z), outside_set(Z,C).

% converging connection
end_line(X,Y,C) :- end_head(X,Z,C), line(Z,Y), inside_set(Z,C).
end_tail(X,Y,C) :- end_line(X,Z,C), arrow(Y,Z), inside_set(Z,C).
end_tail(X,Y,C) :- end_head(X,Z,C), arrow(Y,Z), inside_set(Z,C).

% derived non-separations
con(X,Y,C) :- end_line(X,Y,C), X != Y, outside_set(Y,C). % 23
con(X,Y,C) :- end_head(X,Y,C), X != Y, outside_set(Y,C).
con(X,Y,C) :- end_tail(X,Y,C), X != Y, outside_set(Y,C).
con(X,Y,C) :- con(Y,X,C).

% satisfy all non-independencies
:- dep(X,Y,C,W), not con(X,Y,C). % 27

% maximize the number of satisfied independencies

% minimize the number of lines/arrows
:- line(X,Y), X>Y. [1,X,Y,1] % 29
:- arrow(X,Y). [1,X,Y,2]

% show results
#show.
#show line/2.
#show arrow/2.

Figure 2. ASP encoding of the learning algorithm.

nodes(3). % three nodes
set(0..7). % all subsets of three nodes
dep(1,2,0,1). % weight = 1
dep(1,2,4,1).
dep(2,3,0,1).
dep(2,3,1,1).
dep(1,3,0,1).
dep(1,3,2,1).

Figure 3. ASP encoding of the (in)dependencies in the domain.
\{ \text{biarrow}(X,Y) \} :- \text{node}(X), \text{node}(Y), X \neq Y.
\text{biarrow}(X,Y) :- \text{biarrow}(Y,X).
\text{biarrow}(X,Y), \text{line}(Z,W).
\text{end\_head}(X,Y,C) :- \text{biarrow}(X,Y), \text{outside\_set}(X,C).
\text{end\_head}(X,Y,C) :- \text{end\_tail}(X,Z,C), \text{biarrow}(Z,Y), \text{outside\_set}(Z,C).
\text{end\_head}(X,Y,C) :- \text{end\_arrow}(X,Z,C), \text{biarrow}(Y,Z), \text{inside\_set}(Z,C).

\text{#show biarrow/2}.

Figure 4. Additional ASP encoding for learning original and alternative ADGMs.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
\text{G} & \text{G'} \\
\hline
A \rightarrow B & \epsilon_A \rightarrow A \rightarrow B \leftarrow \epsilon_B \\
\hline
C \rightarrow D & \epsilon_C \rightarrow C \leftarrow \epsilon_D \\
\hline
E \leftrightarrow F & \epsilon_E \rightarrow E \rightarrow F \leftarrow \epsilon_F \\
\hline
\end{tabular}
\caption{Example of the magnification of an ADMG.}
\end{table}

Figure 5. Example of the magnification of an ADMG.

\[(\Lambda^{-1})_{\epsilon_A,\epsilon_B} = 0\) if \(\epsilon_A - \epsilon_B\) is not in \(G'\). Then, \(G\) can be interpreted as a system of structural equations with correlated errors as follows. For any \(A \in V\)

\[A = \beta_A \text{Pa}_G(A) + \epsilon_A\]  \hspace{1cm} (1)

and for any other \(B \in V\)

\[
\text{covariance}(\epsilon_A, \epsilon_B) = \Lambda_{\epsilon_A,\epsilon_B}.
\]  \hspace{1cm} (2)

The following two theorems confirm that the interpretation above works as intended.

**Theorem 8.** Every probability distribution \(p(V)\) specified by Equations 1 and 2 is Gaussian.

**Proof.** Modify the equation \(A = \beta_A \text{Pa}_G(A) + \epsilon_A\) by replacing each \(B \in V\) in the right-hand side of the equation with the right-hand side of the equation of \(B\), i.e. \(\beta_B \text{Pa}_G(B) + \epsilon_B\). Since \(G\) is directed acyclic, repeating this process results in a set of equations for the elements of \(V\) whose right-hand sides are linear combinations of the elements of \(\epsilon\). In other words, \(V = \delta \epsilon\) with \(\epsilon \sim \mathcal{N}(0, \Lambda)\). Then, \(V \sim \mathcal{N}(0, \delta \Lambda \delta^T)\). \(\square\)

**Theorem 9.** Every probability distribution \(p(V)\) specified by Equations 1 and 2 satisfies the global Markov property with respect to \(G\).

**Proof.** Let \(A \in V\). Then, Equation 1 implies that

\[A \perp_{p(V,\epsilon)} (V \cup \epsilon) \setminus (A \cup \text{Pa}_{G'}(A)) | \text{Pa}_{G'}(A)\]

and thus

\[A \perp_{p(V,\epsilon)} \text{Nd}_{G'}(A) \setminus \text{Pa}_{G'}(A) | \text{Pa}_{G'}(A)\]  \hspace{1cm} (3)

by decomposition.

Moreover, let \(\epsilon_C \in Cc(G')\). Then, Equation 2 implies that

\[\epsilon_C \perp_{p(V,\epsilon)} \epsilon \setminus \epsilon_C | \emptyset\]

by Lauritzen (1996, Theorem 3.7 and Proposition 5.2), and thus

\[p(\epsilon) = \prod_{C \in Cc(G') \setminus \emptyset} \epsilon_C\]

by Lauritzen (1996, Theorem 3.9). This together with Equation 1 imply that

\[p(\epsilon_C \cup \text{Nd}_{G'}(\epsilon_C)) = p(\epsilon_C)p(\text{Nd}_{G'}(\epsilon_C))\]

and thus

\[\epsilon_C \perp_{p(V,\epsilon)} \text{Nd}_{G'}(\epsilon_C) | \emptyset\]
and thus
\[ \epsilon_A \perp_{p(V \cup e)} Nd_{G'}(\epsilon_C) \mid \epsilon_Z \]
where \( \epsilon_A \in \epsilon_C \) and \( \epsilon_Z \notin \epsilon_C \setminus \epsilon_A \), by decomposition and weak union.

Finally, Equation 2 implies that
\[ \epsilon_A \perp_{p(V \cup e)} \epsilon_C \setminus (\epsilon_A \cup Ne_{G'}(\epsilon_A)) \mid Ne_{G'}(\epsilon_A) \]
by Lauritzen (1996, Theorem 3.7 and Proposition 5.2), and thus
\[ p(\epsilon_C) = h(\epsilon_A \cup Ne_{G'}(\epsilon_A))k(\epsilon_C \setminus \epsilon_A) \]
by Lauritzen (1996, Equation 3.6). This together with Equation 1 imply that
\[ p(\epsilon_C \cup Nd_{G'}(\epsilon_C)) = h(\epsilon_A \cup Ne_{G'}(\epsilon_A))k(\epsilon_C \setminus \epsilon_A)p(Nd_{G'}(\epsilon_C)) \]
and thus
\[ \epsilon_A \perp_{p(V \cup e)} \epsilon_C \setminus (\epsilon_A \cup Ne_{G'}(\epsilon_A)) \mid Nd_{G'}(\epsilon_C)Ne_{G'}(\epsilon_A) \]
by Lauritzen (1996, Equation 3.6).

Consequently, Equations 3–5 imply that \( p(V \cup e) \) satisfies the global Markov property with respect to \( G' \) by Theorem 6 because (i) \( G' \) is actually an AMP CG over \( V \cup e \), (ii) \( A \) is the only node in the connectivity component of \( G' \) that contains \( A \), and (iii) \( \epsilon_C \) has no parents in \( G' \). Then, \( p(V) \) satisfies the global Markov property with respect to \( G \), because \( G \) and \( G' \) represent the same separations over \( V \) [Pená 2014, Theorem 1].

The equations above specify each node as a linear function of its parents with additive normal noise. The equations can be generalized to nonlinear or nonparametric functions as long as the noise remains additive normal. That is, \( A = f(Pa_G(A)) + \epsilon_A \) for all \( A \in V \), with \( \epsilon \sim \mathcal{N}(0, \Lambda) \) such that \((\Lambda^{-1})_{A,A,B} = 0 \) if \( \epsilon_A - \epsilon_B \) is not in \( G' \). That the noise is additive normal ensures that \( \epsilon_A \) is determined by \( A \cup Pa_{G'}(A) \setminus \epsilon_A \), which is needed for Theorem 1 by Pená (2014) to remain valid. Thanks to this, we can prove in the same manner as Theorem 6 that every probability distribution over \( V \) specified by the system of nonlinear equations also satisfies the global Markov property with respect to \( G \).

A less formal but more intuitive alternative interpretation of ADMGs is as follows. We can interpret the parents of each node in an ADMG as its observed causes. Its unobserved causes are grouped into an error node that is represented implicitly in the ADMG. We can interpret the undirected edges in the ADMG as the correlation relationships between the different error nodes. The causal structure is constrained to be a DAG, but the correlation structure can be any UG. This causal interpretation of our ADMGs parallels that of the original ADMGs [Pearl 1995, 2009]. There are however two main differences. First, the noise in the original ADMGs is not necessarily additive normal. Second, the correlation structure of the error nodes in the original ADMGs is represented by a covariance graph, i.e., a graph with only bidirected edges [Pearl and Wermuth 1993]. Therefore, whereas a missing edge between two error nodes in the original ADMGs represents marginal independence, it represents saturated conditional independence in our ADMGs. This means that the original and our ADMGs represent complementary causal models. Consequently, there are scenarios where the identification of the causal effect of an intervention is not possible with the original ADMGs but is possible with ours, and vice versa. We elaborate on this in the next section.

6.1. Do-Calculus for ADGMs. We start by adapting Pearl’s do-calculus, which operates in the original ADMGs, to our ADMGs. The original do-calculus consists of the following three rules, which permit us to determine if the causal effect of an intervention is identifiable from observed quantities:

- Rule 1 (insertion/deletion of observations):
  \[ p(Y|do(X), Z \cup W) = p(Y|do(X), W) \text{ if } Y \perp_{G''} Z \mid X \cup W \mid | X. \]

- Rule 2 (action/observation exchange):
  \[ p(Y|do(X), do(Z), W) = p(Y|do(X), Z \cup W) \text{ if } Y \perp_{G''} I_Z \mid X \cup W \mid Z \mid X. \]

- Rule 3 (insertion/deletion of actions):
  \[ p(Y|do(X), do(Z), W) = p(Y|do(X), W) \text{ if } Y \perp_{G''} I_Z \mid X \cup W \mid X. \]
where $X$, $Y$, $Z$ and $W$ are disjoint subsets of $V$, "$\|X\|$" denotes an intervention on $X$ in $G''$, and $G''$ is the ADMG $G$ augmented with an intervention random variable $IA$ and an edge $IA \rightarrow A$ for every $A \in V$. See Pearl (1995, p. 686) for further details and the proof that the rules are sound. Fortunately, the rules also apply to our ADMGs by simply redefining "$\|X\|$" appropriately. The proof that the rules are still sound is essentially the same as before. Specifically, "$\|X\|$" should be implemented as follows:

- Delete all the directed edges pointing to nodes in $X$,
- for every path $A \leftarrow V_1 \leftarrow \cdots \leftarrow V_n \leftarrow B$ with $A, B \notin X$ and $V_1, \ldots, V_n \in X$, add the edge $A \rightarrow B$, and
- delete all the undirected edges with an endnode in $X$.

The first step is the same as an intervention in an original ADMG. The second and third steps of the intervention are best understood in terms of the magnified ADMG $G'$: They correspond to marginalizing the error nodes associated to the nodes in $X$ out of $G'_e$, the UG that represents the correlation structure of the error nodes. In other words, they replace $G'_e$ with $(G')^e\setminus \epsilon_X$, the marginal graph of $G'_e$ over $\epsilon \setminus \epsilon_X$. This makes sense since $\epsilon_X$ is no longer associated to $X$ due to the intervention and, thus, it should be marginalized out because it is unobserved. This is exactly what the second and third steps of the intervention imply. To see it, note that the ADMG after the intervention and the magnified ADMG after the intervention represent the same separations over $V$ (Peña, 2014, Theorem 1).

Now, we show that the original and our ADMGs allow for complementary causal reasoning. Specifically, we show an example where our ADMGs allow for the identification of the causal effect of an intervention whereas the original ADMGs do not, and vice versa. Consider the DAG in Figure 6, which represents the causal relationships among all the random variables in the domain at hand. However, only $A$, $B$ and $C$ are observed. Moreover, $US$ represents selection bias. Although other definitions may exist, we say that selection bias is present if two unobserved causes have a common effect that is omitted from the study but influences the selection of the samples in the study (Pearl, 2009, p. 163). Therefore, the corresponding unobserved causes are correlated in every sampled selected. Note that this definition excludes the possibility of an intervention affecting the selection because, in a causal model, unobserved causes do not have observed causes. Note also that our goal is not the identification of the causal effect of an intervention in the whole population but in the subgroup that satisfies the selection bias criterion. For causal effect identification in the whole population, see [Bareinboim and Tian, 2017].

The ADGMs in Figure 6 represent the causal model represented by the DAG when only the observed random variables are modeled. According to our interpretation of ADMGs above, our ADMG is derived from the DAG by keeping the directed edges between observed random variables, and adding an undirected edge between two observed random variables if and only if their unobserved causes are not separated in the DAG given the unobserved causes of the rest of the observed random variables. In other words, $U_A \perp U_B | U_C$ holds in the DAG but $U_A \perp U_C | U_B$ and $U_B \perp U_C | U_A$ do not and, thus, the edges $A \rightarrow C$ and $B \rightarrow C$ are added to the ADMG but $A \rightarrow B$ is not. Deriving the original ADMG is less straightforward. The bidirected edges in an original ADMG represent potential marginal dependence due to a common unobserved cause, also known as confounding. Thus, the original ADMGs are not meant to model selection bias. The best we can do is then to use bidirected edges to represent potential marginal dependencies regardless of their origin. This implies that we can derive the original ADGM from the DAG by keeping the directed edges between observed random variables, and adding a bidirected edge between two observed random variables if and only if their unobserved causes are not separated in the DAG given the empty set. Clearly, $p(B|do(A))$ is unidentifiable with the original ADMG but is identifiable with our ADMG. Specifically,

$$p(B|do(A)) = \sum_C p(B|do(A), C)p(C|do(A)) = \sum_C p(B|do(A), C)p(C) = \sum_C p(B|A, C)p(C)$$

\(^2\)For instance, the DAG may correspond to the following fictitious domain: $A =$ Smoking, $B =$ Lung cancer, $C =$ Drinking, $U_A =$ Parents’ smoking, $U_B =$ Parents’ lung cancer, $U_C =$ Parents’ drinking, $U =$ Parents’ genotype that causes smoking and drinking, $U_S =$ Parents’ hospitalization.

\(^3\)For instance, in the fictitious domain in the previous footnote, we are interested in the causal effect that smoking may have on the development of lung cancer for the patients with hospitalized parents.
where the first equality is due to marginalization, the second due to Rule 3, and the third due to Rule 2.

The original ADMGs assume that confounding is always the source of correlation between unobserved causes. In the example, we consider selection bias as an additional source. However, this is not the only possibility. For instance, \( U_B \) and \( U_C \) may be tied by a physical law of the form \( f(U_B, U_C) = \text{constant} \) devoid of causal meaning, much like Boyle’s law relates the pressure and volume of a gas as \( PV = \text{constant} \) if the temperature and amount of gas remain unchanged within a closed system. In such a case, the discussion above still applies and our ADMG allows for causal effect identification but the original does not. For an example where the original ADMGs allow for causal effect identification whereas ours do not, simply replace the subgraph \( U_C \to S \leftarrow U_B \) in Figure 6 with \( U_C \leftarrow W \to U_B \) where \( W \) is an unobserved random variable. Then, our ADMG will contain the same edges as before plus the edge \( A \to B \), making the causal effect unidentifiable. The original ADMG will contain the same edges as before with the exception of the edge \( A \leftrightarrow B \), making the causal effect identifiable.

In summary, the bidirected edges of the original ADMGs have a clear semantics: They represent potential marginal dependence due to a common unobserved cause. This means that we have to know the causal relationships between the unobserved random variables to derive the ADMG. Or at least, we have to know that there is no selection bias or tying law so that marginal dependence can be attributed to a common unobserved cause due to Reichenbach’s principle (Pearl, 2009, p. 30). This knowledge may not be available in some cases. Moreover, the original ADMGs are not meant to represent selection bias or tying laws. To solve these two problems, we may be willing to use the bidirected edges to represent potential marginal dependences regardless of their origin. Our ADMGs are somehow dual to the original ADMGs, since the undirected edges represent potential saturated conditional dependence between unobserved causes. This implies that in some cases, such as in the example above, our ADMGs may allow for causal effect identification whereas the original may not.

7. Discussion

In this work, we have introduced ADMGs as an extension AMP CGs by (i) relaxing the semidirected acyclicity constraint so that only directed cycles are forbidden, and (ii) allowing up to two edges between any pair of nodes. We have introduced and proved the equivalence of global, ordered local and pairwise Markov properties for the new models. We have also described an exact algorithm for learning them. Finally, we have shown that when the random variables are continuous, the new models can be interpreted as systems of structural equations with correlated errors. This has enabled us to adapt the do-calculus to them.

In the future, we plan to unify the original and our ADMGs by allowing directed, undirected and bidirected edges. Moreover, it is clear that the probability distribution over \( V \cup \epsilon \) factorizes according to the corresponding structural equations model. We plan to investigate whether it is possible to derive a similar factorization for the probability distribution over \( V \).
References


