A factorization criterion for acyclic directed mixed graphs

Thomas S. Richardson
Department of Statistics
University of Washington
tsr@stat.washington.edu

Abstract

Acyclic directed mixed graphs, also known as semi-Markov models represent the conditional independence structure induced on an observed margin by a DAG model with latent variables. In this paper we present a factorization criterion for these models that is equivalent to the global Markov property given by (the natural extension of) d-separation.

1 Introduction

Acyclic directed mixed graphs, also known as semi-Markov models contain directed ($\rightarrow$) and bi-directed ($\leftrightarrow$) edges subject to the restriction that there are no directed cycles. The use of such graphs to represent a statistical model may be traced to path diagrams introduced by [14], where bi-directed edges are used to represent correlated errors. Such graphs are useful for representing the independence structure arising from a DAG model with hidden variables. More recently they have proved useful in characterizing the set of intervention distributions that are identifiable [4, 13]. Ancestral graphs [12] are a subclass of ADMGs.

The global Markov property for such models is given by the natural extension of d-separation to graphs with bi-directed edges. [11] provided a local Markov property for acyclic directed mixed graphs that was equivalent to the global property for all distributions. [5] introduced a weaker version of the local Markov property which is equivalent to the global property under additional assumptions.

In order to be able to score these models we require a parametrization of the set of distributions obeying the conditional independence relations given by an ADMG. Until now such parametrizations have only been available in the multivariate Gaussian case [2, 12].

In this paper we give a factorization result which leads directly to a parametrization in the multivariate binary case. It is not hard to extend the parameterization to the general discrete case, though we do not do it here for reasons of space.
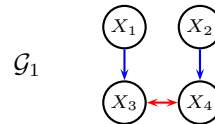
$$\mathcal{G}_1$$

Figure 1: A simple ADMG $\mathcal{G}_1$. (Color is used to distinguish the two edge types.)

We illustrate the main ideas with the graph shown in Figure 1. This graph implies the conditional independence relations $X_1 \perp\!\!\!\perp X_2, X_4$ and $X_2 \perp\!\!\!\perp X_1, X_3$, which cannot be represented by a DAG model (without latents). The theorem given in this paper associates this graph with the following factorizations:

$$
\begin{aligned}
p(x_1, x_2) &= p(x_1)p(x_2), \\
p(x_1, x_3) &= p(x_1)p(x_3 \mid x_1), \\
p(x_2, x_4) &= p(x_2)p(x_4 \mid x_2), \\
p(x_1, x_2, x_3) &= p(x_3 \mid x_1)p(x_1)p(x_2), \\
p(x_1, x_2, x_4) &= p(x_4 \mid x_2)p(x_1)p(x_2), \\
p(x_1, x_2, x_3, x_4) &= p(x_3, x_4 \mid x_1, x_2)p(x_1)p(x_2).
\end{aligned}
$$

Note that the only terms here are: $p(x_1)$, $p(x_2)$, $p(x_3|x_1)$, $p(x_4|x_2)$ and $p(x_3, x_4|x_1, x_2)$. These factorizations lead to a (variation dependent) parametrization in the binary case: if the variables are binary then $p(X_1 = 0)$, $p(X_2 = 0)$, $p(X_3 = 0|x_1)$, $p(X_4 = 0|x_2)$ and $p(X_3 = X_4 = 0|x_1, x_2)$, with $x_1, x_2 \in \{0, 1\}$, are sufficient to specify any multivariate binary distribution obeying the global Markov property for $\mathcal{G}_1$; see [1]. Note that there are $1 + 1 + 2 + 2 + 4 = 10$ quantities here, which corresponds to the known dimension of this Markov model in the binary case.

It is possible to use the local directed Markov property for ADMGs in [11] (see §4.2), together with a given or-

dering (consistent with the directed edges), to obtain factorizations in terms of univariate densities, as proposed in [9]. However, this strategy suffers from the following problem: if only one order is considered then the factorization corresponds to a DAG, and hence, in general, will not be sufficient capture all of the independencies given by the ADMG, e.g. ordering the vertices as they are numbered in $\mathcal{G}_1$ would give:

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2)p(x_3|x_1)p(x_4|x_1, x_2, x_3),$$

which does not imply $X_1 \perp\!\!\!\perp X_4$. On the other hand if all consistent orderings are considered then the set of densities does not form a parametrization as the dimension is too large, e.g. for $\mathcal{G}_1$ we would specify: $p(x_1)$, $p(x_2)$, $p(x_3 \mid x_1)$, $p(x_4 \mid x_2)$, $p(x_3 \mid x_1, x_2)$ and $p(x_4 \mid x_1, x_2)$, which gives 14 parameters in the binary case. Consequently these are not all free: some are constrained to be deterministic functions of others. The factorization in this paper is more complicated as it involves multivariate joint densities, but it has the advantage that it does give a parametrization.

In an ADMG, a set A is said to be *ancestrally closed* if $x \to \cdots \to a$ in $\mathcal{G}$ with $a \in A$ implies that $x \in A$. Our main result is then as follows: a distribution P obeys the global Markov property with respect to an ADMG $\mathcal{G}$ if and only if for every ancestrally closed set A:

$$p(X_A) = \prod_{H_i \in [A]_\mathcal{G}} p(X_{H_i} \mid X_{T_i})$$

where $[A]_\mathcal{G}$ denotes a partition of A into sets $\{H_1, \ldots, H_k\}$, with corresponding conditioning sets $T_1, \ldots, T_k$; for details of the partition and conditioning sets see Theorem 4 below.

This generalizes and unifies three existing results for subclasses of ADMGs: First, the well-known factorization of an ancestrally closed set in a DAG $\mathcal{D}$ [10]:

$$p(X_A) = \prod_{v \in A} p(X_v \mid X_{\mathrm{pa}_\mathcal{D}(v)}).$$

Second, the factorization associated with any set A in a bi-directed graph $\mathcal{B}$ [3]:

$$p(X_A) = \prod_{\substack{C : C \subseteq A; \ C \text{ is connected in } \mathcal{B}, \\ \text{and } C \text{ is inclusion maximal in } A}} p(X_C);$$

where a set C is *connected* if every pair of vertices in C are joined by a path in C, and a connected set C is *inclusion maximal* in A if there is no set of C^*, with $C \subset C^* \subseteq A$, that is connected. (Note that in a bi-directed graph every set is ancestrally closed trivially since there are no directed edges.) Third the factorization associated with a bi-directed chain graph such as $\mathcal{G}_1$; see [1].

The paper is organized as follows: §2 contains basic definitions; §3 contains the factorization result; §4 outlines the proof.

2 Definitions and Basic Concepts

A *directed cycle* is a sequence of edges $x \to \cdots \to x$. If a directed mixed graph $\mathcal{G}$ contains no directed cycles then $\mathcal{G}$ is *acyclic*. There may be two edges between a pair of vertices in an acyclic directed mixed graph (ADMG), but in this case at least one edge must be bi-directed ($x \leftrightarrow y$): otherwise there would be a directed cycle (multiple edges of the same type are not permitted). The induced subgraph of $\mathcal{G}$ given by set A, denoted $\mathcal{G}_A$ consists of the subgraph of $\mathcal{G}$ with vertex set A, and those edges in $\mathcal{G}$ with endpoints in A. The *connected components* of a graph $\mathcal{G}$ are

For a vertex x in a mixed graph $\mathcal{G}$, $\mathrm{pa}_\mathcal{G}(x) \equiv \{v \mid v \to x \text{ in } \mathcal{G}\}$ is the set of *parents* of x; if $y \in \mathrm{pa}_\mathcal{G}(x)$ then $x \in \mathrm{ch}_\mathcal{G}(y)$, the set of *children* of y; $\mathrm{sp}_\mathcal{G}(x) \equiv \{v \mid v \leftrightarrow x \text{ in } \mathcal{G}\}$ is the set of *spouses*[1] of x; $\mathrm{an}_\mathcal{G}(x) \equiv \{v \mid v \to \cdots \to x \text{ in } \mathcal{G} \text{ or } v = x\}$ is the set of *ancestors* of x; similarly $\mathrm{de}_\mathcal{G}(x) \equiv \{v \mid v \leftarrow \cdots \leftarrow x \text{ in } \mathcal{G} \text{ or } v = x\}$ is the set of *descendants* of x. These definitions are applied disjunctively to sets of vertices, so that, for example,

$$\mathrm{pa}_\mathcal{G}(A) \equiv \bigcup_{x \in A} \mathrm{pa}_\mathcal{G}(x), \qquad \mathrm{sp}_\mathcal{G}(A) \equiv \bigcup_{x \in A} \mathrm{sp}_\mathcal{G}(x).$$

(Note that $\mathrm{sp}_\mathcal{G}(A) \cap A$ may be non-empty, and likewise for the other definitions.) A *path* between x and y in $\mathcal{G}$ is a sequence of edges $\langle \epsilon_1, \ldots, \epsilon_n \rangle$, such that there exists a sequence of distinct vertices $\langle x \equiv w_1, \ldots, w_{n+1} \equiv y \rangle$, $(n \geq 0)$, where edge ϵ_i has endpoints w_i, w_{i+1} (paths consisting of a single vertex are permitted for the purpose of simplifying proofs). We denote a *subpath* of a path $\boldsymbol{\pi}$, by $\boldsymbol{\pi}(w_j, w_{k+1}) \equiv \langle \epsilon_j, \ldots, \epsilon_k \rangle$. It is necessary to specify a path as a sequence of edges rather than vertices because the latter does not specify a unique path when there may be two edges between a given pair of vertices. A path of the form $x \to \cdots \to y$ is a *directed path* from x to y.

For a mixed graph $\mathcal{G}$ with vertex set V we consider collections of random variables $(X_v)_{v \in V}$ taking values in probability spaces $(\mathfrak{X}_v)_{v \in V}$, where the probability spaces are either real finite-dimensional vector spaces or finite discrete sets. For $A \subseteq V$ we let $\mathfrak{X}_A \equiv \times_{v \in A}(\mathfrak{X}_v)$, $\mathfrak{X} \equiv \mathfrak{X}_V$ and $X_A \equiv (X_v)_{v \in A}$. We use the usual shorthand notation: v denotes a vertex and a random variable X_v, likewise A denotes a vertex set and X_A.

---

[1]Note that our usage, though established, differs from that of some authors who use the term 'spouse' to denote the other parents of the children of a vertex in a DAG.

2.1 The global Markov property for ADMGs

A non-endpoint vertex z on a path is a *collider on the path* if the edges preceding and succeeding z on the path have an arrowhead at z, i.e. $\rightarrow z \leftarrow$, $\leftrightarrow z \leftrightarrow$, $\leftrightarrow z \leftarrow$, $\rightarrow z \leftrightarrow$. A non-endpoint vertex z on a path which is not a collider is a *non-collider on the path*, i.e. $\leftarrow z \rightarrow$, $\leftarrow z \leftarrow$, $\rightarrow z \rightarrow$, $\leftrightarrow z \rightarrow$, $\leftarrow z \leftrightarrow$. A path between vertices x and y in a mixed graph is said to be *m-connecting given a set* Z if

(i) every non-collider on the path is not in Z, and

(ii) every collider on the path is an ancestor of Z.

If there is no path m-connecting x and y given Z, then x and y are said to be *m-separated* given Z. Sets X and Y are said to be *m-separated* given Z, if for every pair x, y, with $x \in X$ and $y \in Y$, x and y are m-separated given Z.

A probability measure P on $\mathfrak{X}$ is said to satisfy the *global Markov property* for $\mathcal{G}$ if for arbitrary disjoint sets X, Y, Z, (Z may be empty) X is m-separated from Y given Z in $\mathcal{G}$ implies $X \perp\!\!\!\perp Y \mid Z \; [P]$. Note that if $\mathcal{G}$ is a DAG then the above definition is identical to Pearl's d-separation criterion; see [6, 10].

In a given set T with subset A, we will say that B forms the *Markov blanket* for A in T, if $B \subseteq T \setminus A$ and A is m-separated from $T \setminus (B \cup A)$ given B, and there is no subset of B for which this is true.

For an ADMG $\mathcal{G}$ we define:

$$\mathcal{A}(\mathcal{G}) \equiv \{A \mid \operatorname{an}_{\mathcal{G}}(A) = A\}$$

the set of *ancestrally closed sets*. The set $\mathcal{A}(\mathcal{G})$ forms a lattice so that if $A, B \in \mathcal{A}(\mathcal{G})$, then $A \cap B, A \cup B \in \mathcal{A}(\mathcal{G})$.

2.2 Barren subsets

For a given vertex set A we define:

$$barren(A) \equiv \{x \mid x \in A; \deg_{\mathcal{G}}(x) \cap A = \{x\}\}.$$

Note that this definitions is intrinsic to the set A in the sense that a vertex $v \in \operatorname{barren}(A)$ if v has no descendants in $\mathcal{G}$ that are *in A*; v may have descendants in $\mathcal{G}$ that are not in A.

Proposition 1. *In an ADMG,* (i) $\operatorname{barren}(\operatorname{an}(A)) = \operatorname{barren}(A) \subseteq A$; (ii) *if* $A \in \mathcal{A}(\mathcal{G})$ *then* $\operatorname{an}(\operatorname{barren}(A)) = A$; (iii) *if* $B \subseteq A$ *then* $\operatorname{barren}(A) \cap B \subseteq \operatorname{barren}(B)$.

2.3 Districts

For a given ADMG $\mathcal{G}$, the *induced bi-directed graph* $(\mathcal{G})_{\leftrightarrow}$ is the graph formed by removing all directed edges from $\mathcal{G}$. Likewise $(\mathcal{G})_{\rightarrow}$ is the DAG formed by removing all bi-directed edges. The *district* (aka c-component) for a vertex x in $\mathcal{G}$ is the connected component of x in $(\mathcal{G})_{\leftrightarrow}$, or equivalently

$$dis_{\mathcal{G}}(x) \equiv \{y \mid y \leftrightarrow \cdots \leftrightarrow x \text{ in } \mathcal{G} \text{ or } x = y\}.$$

Similarly we let $\operatorname{dis}_A(x) = \operatorname{dis}_{\mathcal{G}_A}(x)$; if $x \notin A$ then we define $\operatorname{dis}_A(x) = \emptyset$. Note that $\operatorname{dis}_A(x) \subseteq \operatorname{dis}(x) \cap A$, with strict inclusion possible: a vertex v is in $\operatorname{dis}_A(x)$ only if $v \in A$ and there is a path from v to x in $(\mathcal{G})_{\leftrightarrow}$ on which every vertex is in A; the condition for membership in $\operatorname{dis}(x) \cap A$ is the same, except that the requirement that every vertex on the path is in A is removed. As usual we apply the definition disjunctively to sets:

$$\operatorname{dis}_A(B) = \bigcup_{x \in B} \operatorname{dis}_A(x).$$

Proposition 2. *If W is a set in an ADMG $\mathcal{G}$, then:* (i) *the set of districts* $\{\operatorname{dis}_W(v) \mid v \in W\}$ *forms a partition of W;* (ii) *for any v,* $\operatorname{dis}_W(v) = \operatorname{dis}_{\operatorname{dis}_W(v)}(v)$; (iii) *vf $w \in W^* \subseteq W$ then* $\operatorname{dis}_{W^*}(w) \subseteq \operatorname{dis}_W(w)$.

A set C is *path-connected* in $(\mathcal{G})_{\leftrightarrow}$ if every pair of vertices in C are connected via a path in $(\mathcal{G})_{\leftrightarrow}$; equivalently, every vertex in C has the same district in $\mathcal{G}$. Note that a set C may be path-connected in $(\mathcal{G})_{\leftrightarrow}$ without forming a connected set in $(\mathcal{G})_{\leftrightarrow}$. (Recall that if a set C is connected then between every pair of vertices in C there is a path on which every vertex is itself in C; equivalently, every vertex in C is in the same district in $\mathcal{G}_C$.)

A district $D = \operatorname{dis}_W(v)$ of a vertex v in a set W in $\mathcal{G}$ is said to be *ancestrally closed* if for every vertex v,

$$D = \operatorname{dis}_{\operatorname{an}_{\mathcal{G}}(D)}(v).$$

Thus the district of v within W would not get any larger if we were to add to W all vertices in $\mathcal{G}$ that are ancestors of vertices that are already in the district of v within W. (Note that if this property holds for one vertex $v \in D$, then it holds for all vertices in D.) By extension, a set W will be said to have *ancestrally closed districts* if every district in W is ancestrally closed.

Proposition 3. *If $W \in \mathcal{A}(\mathcal{G})$ then W has ancestrally closed districts.*

Note however that the converse is false: a set may have ancestrally closed districts but not be ancestral. For example in the graph in Figure 2 all subsets of the vertices have ancestrally closed districts, including those which are not ancestral.

3 Factorization

The main result of the paper is as follows:

Theorem 4. *A probability distribution P obeys the global Markov property for $\mathcal{G}$ if and only if for every $A \in \mathcal{A}(\mathcal{G})$,*

$$p(X_A) = \prod_{H \in [A]_{\mathcal{G}}} p(X_H \mid X_{\mathrm{tail}(H)}) \qquad (1)$$

where $[A]_{\mathcal{G}}$ denotes a partition of A into sets $\{H_1, \ldots, H_k\} \subseteq \mathcal{H}(\mathcal{G})$, defined with $\mathrm{tail}(H)$, below.

The factorization associated with an ADMG is of necessity more involved than in DAGs or chain graphs: in a DAG, one may always find a total ordering under which the parents of a vertex v precede v. In chain graphs this is generalized so that the parents of an 'undirected component' precede the component itself. In the case of an ADMG graph, 'blocks' are *heads* (defined below) which are subsets of connected components of $\mathcal{G}_{\leftrightarrow}$. However, it is easy to find ADMG graphs in which there is no ordering under which every vertex that is a parent of at least one vertex in a given head, occurs prior to any vertex in that head. For example,
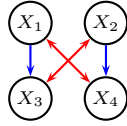

Figure 2: An ADMG in which there is no vertex ordering under which all parents of a head precede any vertex in the head.

in Figure 2, $\{1, 4\}$ and $\{2, 3\}$ form heads. It is easy to see that if 1 is ordered before 2 and 3, then we cannot also have 2 ordered before 1 and 4.

The upshot of this is that the inclusion of a term in the factorization in which vertex i is on the left of the conditioning bar, and j is on the right, does not preclude the inclusion of a term, in the same factorization, in which the reverse holds. For example, we associate with the graph in Figure 2, the factorization:

$$p(x_1, x_2, x_3, x_4) = p(x_1, x_4 \mid x_2)p(x_2, x_3 \mid x_1). \quad (2)$$

It is easy to see that this factorization implies that $X_3 \perp\!\!\!\perp X_4 \mid X_1, X_2$. However, after marginalizing x_3 and x_4 it can be seen that (2) implies

$$p(x_1, x_2) = p(x_1 \mid x_2)p(x_2 \mid x_1),$$

so $X_1 \perp\!\!\!\perp X_2$, which also follows from the global Markov property associated with this graph. In this sense the factorization is 'cyclic' in nature.

3.1 Factorization terms: heads and tails

Each term in the factorization takes the form $p(X_H \mid X_T)$, $H, T \subseteq V$; $H \cap T = \emptyset$. Following [8], we will refer to H as the *head* of the term $p(X_H \mid X_T)$, and T as the *tail*. However, only certain pairs of sets (H, T) will occur in factorizations associated with $\mathcal{G}$. An ordered pair of sets (H, T) form the *head* and *tail* of a term associated with $\mathcal{G}$ if and only if all of the following hold:

(i) $H = \mathrm{barren}(\mathrm{an}_{\mathcal{G}}(H))$,

(ii) H forms a path-connected set in $(\mathcal{G}_{\mathrm{an}(H)})_{\leftrightarrow}$,

and

(iii) $T = (\mathrm{dis}_{\mathrm{an}(H)}(H) \setminus H) \cup \mathrm{pa}(\mathrm{dis}_{\mathrm{an}(H)}(H))$.

Condition (i) may be re-expressed by saying that for every vertex $x \in H$, $\mathrm{an}_{\mathcal{G}}(x) \cap H = \{x\}$. Condition (ii) is equivalent to the assertion that

$$\mathrm{dis}_{\mathrm{an}(H)}(H) = \bigcap_{v \in H} \mathrm{dis}_{\mathrm{an}(H)}(v).$$

For a graph $\mathcal{G}$, the set of *heads* is denoted by $\mathcal{H}(\mathcal{G})$. By condition (iii), if $H \in \mathcal{H}(\mathcal{G})$, then there is a unique tail which we denote $\mathit{tail}(H)$. The tail associated with H consists of the variables in $\mathrm{dis}_{\mathrm{an}(H)}(H)$ that are not in H, and any vertex which is a parent of a vertex in $\mathrm{dis}_{\mathrm{an}(H)}(H)$, which includes all vertices in $\mathrm{pa}(H)$. In what follows it will be useful to distinguish between these two subsets of $\mathrm{tail}(H)$ hence we introduce the following definitions:

$$\begin{aligned} \mathit{dis\text{-}tail}(H) &\equiv \mathrm{dis}_{\mathrm{an}(H)}(H) \setminus H, \\ \mathit{pa\text{-}tail}(H) &\equiv \mathrm{pa}(\mathrm{dis}_{\mathrm{an}(H)}(H)) \end{aligned}$$

(note that these two sets need not be disjoint). All vertices in the tail are connected to a vertex in H by a path on which every non-endpoint vertex is a collider, and every vertex is an ancestor of some vertex in H. Given a set $W \supseteq \mathrm{an}(H)$, with $\mathrm{de}(W) \cap H = H$; $\mathrm{sp}(\mathrm{dis}_{\mathrm{an}(H)}(H)) \cap W \subseteq \mathrm{an}(H)$, the global Markov property for $\mathcal{G}$ implies that

$$H \perp\!\!\!\perp W \setminus (H \cup \mathrm{tail}(H)) \mid \mathrm{tail}(H)$$

or in other words, $\mathrm{tail}(H)$ is the Markov blanket for H in the set W.

For a bi-directed graph $\mathcal{B}$, $\mathcal{H}(\mathcal{B})$ is simply the set of connected subsets of vertices; for a DAG $\mathcal{D}$ it is the set of all singleton sets: $\mathcal{H}(\mathcal{D}) = \{\{v\} | v \in V\}$.

Lemma 5. *If $H \in \mathcal{H}(\mathcal{G})$, then $\mathrm{tail}(H) \cap \mathrm{de}(H) = \emptyset$.*

Proof: Suppose for a contradiction that there is some vertex $v \in \mathrm{tail}(H) \cap \mathrm{de}(H)$. Since, by definition,

$\text{tail}(H) = \left(\text{dis}_{\text{an}(H)}(H) \setminus H\right) \cup \text{pa}(\text{dis}_{\text{an}(H)}(H)) \subseteq \text{an}(H)$, this implies that $H \neq \text{barren}(\text{an}_{\mathcal{G}}(H))$, which is a contradiction. $\quad\square$

Corollary 6. *If $H \in \mathcal{H}(\mathcal{G})$ then $H \cap \text{tail}(H) = \emptyset$.*

Proof: Follows from Lemma 5 since $H \subseteq \text{de}(H)$. $\quad\square$

In a complete ADMG $\mathcal{G}$, for every set $W \in \mathcal{A}(\mathcal{G})$, $\text{barren}(W)$ is path-connected in $(\mathcal{G}_W)_{\leftrightarrow}$, so

$$\mathcal{H}(\mathcal{G}) = \{H \mid H = \text{barren}(W) \text{ for some } W \in \mathcal{A}(\mathcal{G})\}.$$

Proposition 7. *If $A \in \mathcal{A}(\mathcal{G})$, then $\mathcal{H}(\mathcal{G}_A) = \{H \mid H \in \mathcal{H}(\mathcal{G}), H \subseteq A\}$.*

Proposition 8. *The mapping $\mathcal{H}(\mathcal{G}) \to \mathcal{A}(\mathcal{G})$ given by $A = \text{an}_{\mathcal{G}}(H)$ is injective (one-to-one). Further, if $\mathcal{G}$ is complete then it is surjective (onto).*

3.2 Decomposition of a set: $[W]_{\mathcal{G}}$

For any ADMG $\mathcal{G}$ and a subset of vertices W, we define the following functions:

$$\Phi_{\mathcal{G}}(\emptyset) \equiv \emptyset,$$
$$\Phi_{\mathcal{G}}(W) \equiv$$
$$\left\{ H \;\middle|\; H = \bigcap_{x \in H} \text{barren}\left(\text{an}_{\mathcal{G}}\left(\text{dis}_W(x)\right)\right); \; H \neq \emptyset \right\},$$

$$\psi_{\mathcal{G}}(W) \equiv W \setminus \bigcup_{H \in \Phi_{\mathcal{G}}(W)} H,$$

$$\psi_{\mathcal{G}}^{(k)}(W) \equiv \underbrace{\psi_{\mathcal{G}}(\cdots \psi_{\mathcal{G}}(W) \cdots)}_{k\text{-times}}, \qquad \psi_{\mathcal{G}}^{(0)}(W) \equiv W.$$

Finally, we define the *partition induced by $\mathcal{G}$ on W* to be:

$$[W]_{\mathcal{G}} \equiv \bigcup_{k \geq 0} \Phi_{\mathcal{G}}\left(\psi_{\mathcal{G}}^{(k)}(W)\right).$$

As described below, when applied to a set W, the function $\Phi_{\mathcal{G}}(W)$ returns a set of heads, one for each district in W. For a given district $D \subseteq W$, the associated head consists of those vertices in D that are not ancestors (in $\mathcal{G}$) of any other vertices in D. The purpose of the intersection over elements in H is to ensure that H is path-connected in $(\mathcal{G}_W)_{\leftrightarrow}$; an equivalent definition is thus:

$$\Phi_{\mathcal{G}}(W) = \{H \mid \emptyset \neq H = \text{barren}\left(\text{an}_{\mathcal{G}}\left(\text{dis}_W(H)\right)\right);$$
$$H \text{ is path-connected in } (\mathcal{G}_W)_{\leftrightarrow}\}.$$

Thus every pair of distinct vertices in H are connected by a path in $(\mathcal{G}_W)_{\leftrightarrow}$ on which every edge is bi-directed, and is in W. As before it will not be true in general that H itself will form a connected set in $(\mathcal{G}_W)_{\leftrightarrow}$.

The function $\psi_{\mathcal{G}}(W)$ simply removes from W all those vertices which occur in some head in $\Phi_{\mathcal{G}}(W)$.

3.3 Parametrization

Multivariate binary distributions obeying the global Markov property with respect to an ADMG $\mathcal{G}$ may be parametrized by the following set of probabilities:

$$\left\{ p(X_H = 0 \mid X_{\text{tail}(H)} = x_{\text{tail}(H)}) \;\middle|\; H \in \mathcal{H}(\mathcal{G}), \right.$$
$$\left. x_{\text{tail}(H)} \in \{0,1\}^{|\text{tail}(H)|} \right\}.$$

Let $\alpha : V \mapsto \{0,1\}^{|V|}$ be an assignment of values to the variables indexed by V. Define $\alpha(W)$ to be the values assigned to variables indexed by a subset $W \subseteq V$. Let $\alpha^{-1}(0) = \{v \mid v \in V, \alpha(v) = 0\}$. Then:

$$p\left(X_V = \alpha(V)\right) = \sum_{C \,:\, \alpha^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus \alpha^{-1}(0)|} \times$$
$$\prod_{H \in [C]_{\mathcal{G}}} P\left(X_H = 0 \,\middle|\, X_{\text{tail}(H)} = \alpha(\text{tail}(H))\right).$$
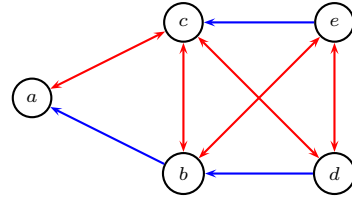
3.4 Examples

Figure 3: An ADMG used to illustrate recursive decompositions.

First consider the graph in Figure 2. We have the following decompositions:

$$
\begin{aligned}
{[\{x_1, x_2\}]}_{\mathcal{G}} &= \{\{x_1\}, \{x_2\}\} \\
{[\{x_1, x_2, x_3\}]}_{\mathcal{G}} &= \{\{x_2, x_3\}, \{x_1\}\}, \\
{[\{x_1, x_2, x_4\}]}_{\mathcal{G}} &= \{\{x_1, x_4\}, \{x_2\}\}, \\
{[\{x_1, x_2, x_3, x_4\}]}_{\mathcal{G}} &= \{\{x_1, x_4\}, \{x_2, x_3\}\}, \\
{[\{x_1, x_3\}]}_{\mathcal{G}} &= \{\{x_1\}, \{x_3\}\}, \\
{[\{x_2, x_4\}]}_{\mathcal{G}} &= \{\{x_2\}, \{x_4\}\}.
\end{aligned}
$$

Together with the singleton sets $\{x_1\}, \{x_2\}$ which have trivial decompositions, these represent all of the ancestral sets in this graph. These decompositions lead to the following factorizations:

$$
\begin{aligned}
p(x_1, x_2) &= p(x_1)p(x_2), \\
p(x_1, x_2, x_3) &= p(x_2, x_3|x_1)p(x_1), \\
p(x_1, x_2, x_4) &= p(x_1, x_4|x_2)p(x_2), \qquad (3) \\
p(x_1, x_2, x_3, x_4) &= p(x_1, x_4|x_2)p(x_2, x_3|x_1), \\
p(x_1, x_3) &= p(x_3|x_1)p(x_1), \\
p(x_2, x_4) &= p(x_4|x_2)p(x_2).
\end{aligned}
$$

Examples of the parametrization are as follows:

$$p(X_{1234}{=}0) = p(X_{14}{=}0|X_2{=}0)p(X_{23}{=}0|X_1{=}0),$$

where we use the obvious shorthand, e.g. $\{X_{14}{=}0\} \equiv \{X_1{=}X_4{=}0\}$;

$$
\begin{aligned}
&p(X_{12}{=}0, X_{34}{=}1)\\
&= \; p(X_1{=}0)p(X_2{=}0) - p(X_{23}{=}0|X_1{=}0)p(X_1{=}0)\\
&\quad - p(X_{14}{=}0|X_2{=}0)p(X_2{=}0)\\
&\quad + p(X_{14}{=}0|X_2{=}0)p(X_{23}{=}0|X_1{=}0);
\end{aligned}
$$

$$
\begin{aligned}
&p(X_{12}{=}1, X_{34}{=}0)\\
&= \; p(X_3{=}0|X_1{=}1)p(X_4{=}0|X_2{=}1)\\
&\quad - p(X_3{=}0|X_1{=}1)p(X_{14}{=}0|X_2{=}1)\\
&\quad - p(X_{23}{=}0|X_1{=}1)p(X_4{=}0|X_2{=}1)\\
&\quad + p(X_{23}{=}0|X_1{=}1)p(X_{14}{=}0|X_2{=}1).
\end{aligned}
$$

For a more complex example, we present five decompositions resulting from the graph in Figure 3:

$$
\begin{aligned}
[\{d,e\}]_{\mathcal{G}} &= \{\{d,e\}\},\\
[\{c,d,e\}]_{\mathcal{G}} &= \{\{c,d\},\{e\}\},\\
[\{b,c,d,e\}]_{\mathcal{G}} &= \{\{b,c\},\{d,e\}\},\\
[\{a,b,c,d,e\}]_{\mathcal{G}} &= \{\{a,c\},\{b,e\},\{d\}\},\\
[\{a,b,d,e\}]_{\mathcal{G}} &= \{\{a\},\{b,e\},\{d\}\},
\end{aligned}
$$

with the corresponding factorizations:

$$
\begin{aligned}
p(d,e) &= p(d,e),\\
p(c,d,e) &= p(c,d|e)p(e),\\
p(b,c,d,e) &= p(b,c|d,e)p(d,e),\\
p(a,b,c,d,e) &= p(a,c|b,d,e)p(b,e|d)p(d),\\
p(a,b,d,e) &= p(a|b)p(b,e|d)p(d).
\end{aligned}
$$

The first four do not imply any constraint on the associated margin. This is a consequence of the fact that there is only one district in each of these ancestral sets. Note that even though $\{a,b,c,d,e\}$ forms a single district, $\{a,b,d,e\}$ contains two districts.

3.5   Properties of $\Phi_{\mathcal{G}}(W)$

Proposition 9. *If $H \in \Phi_{\mathcal{G}}(W)$, then $H \subseteq W$.*

Proof: If for some $x \in H$, $\mathrm{dis}_W(x) = \emptyset$ then $H \subseteq \mathrm{barren}(\mathrm{an}_{\mathcal{G}}(\emptyset)) = \emptyset$, which is a contradiction. But if $\mathrm{dis}_W(x) \neq \emptyset$ then by definition $x \in W$. $\qquad\square$

Lemma 10. *If $x_1, x_2 \in H \in \Phi_{\mathcal{G}}(W)$, then $\mathrm{dis}_W(x_1) = \mathrm{dis}_W(x_2)$, hence for all $x \in H$, $H = \mathrm{barren}(\mathrm{an}_{\mathcal{G}}(\mathrm{dis}_W(x)))$.*

Proof: Let $D_i = \mathrm{dis}_W(x_i)$, $i \in \{1,2\}$. By Proposition 9 $\mathrm{dis}_W(x_i) \neq \emptyset$, $i \in \{1,2\}$. By Proposition 2 (i), if $D_1 \neq$

D_2, then $D_1 \cap D_2 = \emptyset$. Further, by Proposition 1(i) $\mathrm{barren}\,(\mathrm{an}_{\mathcal{G}}(D_i)) \subseteq D_i$ $(i \in \{1,2\})$. Hence $D_1 \cap D_2 = \emptyset$ implies that $H = \emptyset$, which is a contradiction. $\qquad\square$

Corollary 11. *If $H_1, H_2 \in \Phi_{\mathcal{G}}(W)$ and $H_1 \neq H_2$ then $\mathrm{dis}_W(H_1) \cap \mathrm{dis}_W(H_2) = \emptyset$.*

Proof: Suppose for a contradiction that $x \in \mathrm{dis}_W(H_1) \cap \mathrm{dis}_W(H_2)$. This implies that $\mathrm{dis}_W(h_j) = \mathrm{dis}_W(x)$, for any $h_j \in H_j$, $j \in \{1,2\}$. It then follows by Lemma 10 that

$$H_1 = \mathrm{barren}(\mathrm{an}_{\mathcal{G}}(\mathrm{dis}_W(x))) = H_2,$$

which is a contradiction. $\qquad\square$

Corollary 12. *If $H_1, H_2 \in \Phi_{\mathcal{G}}(W)$ and $H_1 \neq H_2$ then $H_1 \cap H_2 = \emptyset$.*

Proof: Immediate from Corollary 11, since $H_i \subseteq \mathrm{dis}_W(H_i)$, $i \in \{1,2\}$. $\qquad\square$

Corollary 13. *For any set A, $[A]_{\mathcal{G}}$ partitions A.*

Proof: By Corollary 12, no two sets in $\Phi_{\mathcal{G}}\left(\psi_{\mathcal{G}}^{(k)}(A)\right)$, for fixed k, overlap. Further, by Proposition 9, and the definition of $\psi_{\mathcal{G}}^{(k)}(A)$, the sets in $\Phi_{\mathcal{G}}\left(\psi_{\mathcal{G}}^{(i)}(A)\right)$ and $\Phi_{\mathcal{G}}\left(\psi_{\mathcal{G}}^{(j)}(A)\right)$ do not overlap. Since $W \neq \emptyset$ implies $\Phi_{\mathcal{G}}(W) \neq \emptyset$, $\{\psi_{\mathcal{G}}^{(i)}(A)\}_i$ is a monotonically decreasing sequence of sets, so the sets in $[A]_{\mathcal{G}}$ exhaust A. $\qquad\square$

Lemma 14. *If $H \in \Phi_{\mathcal{G}}(W)$ then $H \in \mathcal{H}(\mathcal{G})$.*

Proof: We first show that $\mathrm{barren}(\mathrm{an}(H)) = H$. Since for any $h \in H$, $H = \mathrm{barren}(\mathrm{an}_{\mathcal{G}}(\mathrm{dis}_W(h)))$, it follows by Proposition 1 (ii) that $\mathrm{an}_{\mathcal{G}}(H) = \mathrm{an}_{\mathcal{G}}(\mathrm{dis}_W(h))$ hence $\mathrm{barren}(\mathrm{an}_{\mathcal{G}}(H)) = \mathrm{barren}(\mathrm{an}_{\mathcal{G}}(\mathrm{dis}_W(h)))$. Since this holds for any $h \in H$, it then follows that

$$\mathrm{barren}(\mathrm{an}_{\mathcal{G}}(H)) = \bigcap_{h \in H} \mathrm{barren}(\mathrm{an}_{\mathcal{G}}(\mathrm{dis}_W(h))) = H.$$

We now show that $\mathrm{dis}_{\mathrm{an}(H)}(H)$ is path-connected in $(\mathcal{G}_{\mathrm{an}(H)})_{\leftrightarrow}$, or equivalently that

$$\mathrm{dis}_{\mathrm{an}_{\mathcal{G}}(H)}(H) = \bigcap_{v \in H} \mathrm{dis}_{\mathrm{an}_{\mathcal{G}}(H)}(v),$$

and thus $H \in \mathcal{H}(\mathcal{G})$. For any $h \in H$,

$$
\begin{aligned}
\mathrm{dis}_{\mathrm{an}_{\mathcal{G}}(H)}(h) &= \mathrm{dis}_{\mathrm{an}_{\mathcal{G}}(\mathrm{dis}_W(h))}(h)\\
&\supseteq \mathrm{dis}_{\mathrm{dis}_W(h)}(h) = \mathrm{dis}_W(h),
\end{aligned}
$$

where: the first equality is by Proposition 1(ii); the inclusion follows from Proposition 2 (iii); the last equality uses Proposition 2 (ii). Hence

$$\bigcap_{h \in H} \mathrm{dis}_{\mathrm{an}_{\mathcal{G}}(H)}(h) \supseteq \bigcap_{h \in H} \mathrm{dis}_W(h) \supseteq H,$$

where the last inclusion follows from the definition of $\Phi_{\mathcal{G}}(W)$, and Proposition 1(i). By Proposition 2 (i) the LHS is either equal to the empty set or $\mathrm{dis}_{\mathrm{an}_{\mathcal{G}}(H)}(H)$, which concludes the proof. $\square$

For $x \in A$ we define $head(x; A)$ to be the head H containing x in $[A]_{\mathcal{G}}$. For $H \in [A]_{\mathcal{G}}$ we refer to the $k \geq 0$, for which $H \in \Phi_{\mathcal{G}}(\psi^{(k)}(A))$, as the *depth* of H in A, denoted $dep_A(H)$. Similarly the *depth of a vertex* $h \in A$, is the depth in A of the head containing it: $dep_A(x) = dep_A(head(x; A))$.

3.6 Properties of $\psi_{\mathcal{G}}(W)$ when W has ancestrally closed districts

We will show that if W has ancestrally closed districts then so does $\psi_{\mathcal{G}}(W)$. We will also show that if W has ancestrally closed districts then the district-tail of any head $H \in \Phi_{\mathcal{G}}(W)$ is contained in $\psi_{\mathcal{G}}(W)$.

In subsequent proofs we will make use of the following:

Lemma 15. *If W has ancestrally closed districts and $H \in \Phi_{\mathcal{G}}(W)$ then $\mathrm{dis}_W(v) = \mathrm{dis}_{\mathrm{an}(H)}(v)$, for all $v \in \mathrm{dis}_W(H)$.*

Proof: First observe that $\mathrm{an}(H) = \mathrm{an}(\mathrm{barren}(\mathrm{an}(\mathrm{dis}_W(H)))) = \mathrm{an}(\mathrm{dis}_W(H)) \supseteq \mathrm{dis}_W(H)$. Hence $\mathrm{dis}_{\mathrm{an}(H)}(v)$ is not empty, if $v \in \mathrm{dis}_W(H)$. Further,

$$\begin{aligned} \mathrm{dis}_{\mathrm{an}(H)}(v) &= \mathrm{dis}_{\mathrm{an}(\mathrm{dis}_W(H))}(v) \\ &= \mathrm{dis}_{\mathrm{an}(\mathrm{dis}_W(v))}(v) = \mathrm{dis}_W(v), \end{aligned}$$

where the second equality follows since $v \in \mathrm{dis}_W(H)$, the third equality follows since W has ancestrally closed districts. $\square$

Lemma 16. *If W has ancestrally closed districts, then so does $\psi(W)$.*

Corollary 17. *If $A \in \mathcal{A}(\mathcal{G})$ then for $k \geq 0$, $\psi_{\mathcal{G}}^{(k)}(A)$ has ancestrally closed districts.*

Proof: The result follows from Proposition 3 and k-applications of Lemma 16. $\square$

Lemma 18. *If $A \in \mathcal{A}(\mathcal{G})$, and $H \in [A]_{\mathcal{G}}$, with $dep_A(H) = k$, then $\mathrm{dis\text{-}tail}(H) \subseteq \psi^{(k+1)}(A)$.*

Proof: By hypothesis $H \in \Phi_{\mathcal{G}}(\psi^{(k)}(A))$. By Corollary 17, $\psi^{(k)}(A)$ has ancestrally closed districts. Thus by Lemma 15, $\mathrm{dis}_{\mathrm{an}(H)}(H) = \mathrm{dis}_{\mathrm{an}(\psi^{(k)}(A))}(H)$. Further, since $\psi^{(k)}(A)$ has ancestrally closed districts, and $H \subseteq \psi^{(k)}(A)$, $\mathrm{dis}_{\mathrm{an}(\psi^{(k)}(A))}(H) = \mathrm{dis}_{\psi^{(k)}(A)}(H)$. By Corollary 11, for $H' \in \Phi_{\mathcal{G}}(\psi^{(k)}(A))$, with $H' \neq H$, $\mathrm{dis}_{\psi^{(k)}(A)}(H) \cap \mathrm{dis}_{\psi^{(k)}(A)}(H') = \mathrm{dis}_{\psi^{(k)}(A)}(H) \cap H' = \emptyset$.

Hence

$$\begin{aligned} \mathrm{dis\text{-}tail}(H) &= \mathrm{dis}_{\mathrm{an}(H)}(H) \setminus H \\ &= \mathrm{dis}_{\mathrm{an}(H)}(H) \setminus \left(\bigcup_{H' \in \Phi_{\mathcal{G}}(\psi^{(k)}(A))} H' \right) \\ &= \mathrm{dis}_{\psi^{(k)}(A)}(H) \setminus \left(\bigcup_{H' \in \Phi_{\mathcal{G}}(\psi^{(k)}(A))} H' \right) \\ &\subseteq \psi^{(k+1)}(A). \end{aligned}$$

Here the last inclusion follows by definition of $\psi(\cdot)$. $\square$

4 Equivalence of global Markov property and factorization

An inductive argument establishing Theorem 4 faces several challenges: If A is an ancestral set, there may be no head $H \in [A]_{\mathcal{G}}$ such that $A \setminus H$ is an ancestral set: see for example the set $\{x_1, x_2, x_3, x_4\}$ in Figure 2, in which $\{x_1, x_4\}$ and $\{x_2, x_3\}$ are the heads. Since Theorem 4 only makes assertions about factorizations over margins $p(X_A)$ where A is an ancestral set, this means that we cannot apply an inductive argument where we add one term at a time. Further, as the examples in section 3.4 make clear, the heads occurring in the decomposition of an ancestral set A do not necessarily occur in the decomposition of an ancestral superset $A \cup \{x\}$. Thus any inductive argument which hypothesizes that the factorization holds for $p(X_A)$ and then extends this to $p(X_{A \cup \{x\}})$ must also transform the terms in the factorization of $p(X_A)$.
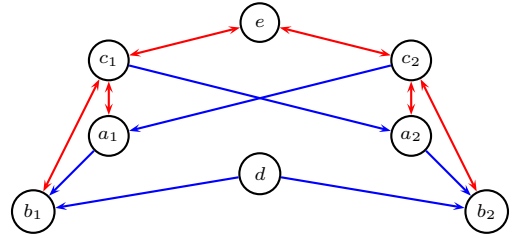

Figure 4: An ADMG used to illustrate the proof strategy for Theorem 4.

We first sketch our proof strategy by considering the example in Figure 4. Suppose, as an inductive hypothesis that, as required by Theorem 4,

$$\begin{aligned} &p(a_1, a_2, b_1, b_2, c_1, c_2, d) \\ &= p(b_1, c_1 | a_1, c_2, d) p(a_1 | c_2) p(b_2, c_2 | a_2, c_1, d) \\ &\quad \times p(a_2 | c_1) p(d). \end{aligned} \tag{4}$$

We wish to show that the global Markov property implies that

$$
\begin{aligned}
p(a_1, a_2, b_1, b_2, c_1, c_2, d, e) \\
= \quad & p(e, b_1, b_2 | a_1, a_2, c_1, c_2, d) \\
& \times p(d) p(a_1, c_1 | c_2) p(a_2, c_2 | c_1). \quad (5)
\end{aligned}
$$

To this end we use the chain rule to further factorize the terms in equation (4), as follows:

$$
\begin{aligned}
p(b_1, c_1 | a_1, c_2, d) &= p(b_1 | a_1, c_1, c_2, d) p(c_1 | a_1, c_2, d), \\
p(b_2, c_2 | a_2, c_1, d) &= p(b_2 | a_2, c_1, c_2, d) p(c_2 | a_2, c_1, d).
\end{aligned}
$$

Since c_1 is m-separated from d given a_1 and c_2, $p(c_1 | a_1, c_2, d) = p(c_1 | a_1, c_2)$. Likewise $p(c_2 | a_2, c_1, d) = p(c_2 | a_2, c_1)$. Also since b_1 is m-separated from a_2 given $\{a_1, c_1, c_2, d\}$ we have $p(b_1 | a_1, c_1, c_2, d) = p(b_1 | a_1, a_2, c_1, c_2, d)$. Finally, we note that b_2 is m-separated from $\{a_1, b_1\}$ given $\{a_2, c_1, c_2, d\}$, hence $p(b_2 | a_2, c_1, c_2, d) = p(b_2 | a_1, a_2, b_1, c_1, c_2, d)$. We now obtain:

$$
\begin{aligned}
& p(e | a_1, a_2, b_1, b_2, c_1, c_2, d) p(a_1, a_2, b_1, b_2, c_1, c_2, d) \\
= \quad & \Big(p(e | a_1, a_2, b_1, b_2, c_1, c_2, d) \\
& \quad \times p(b_1 | a_1, a_2, c_1, c_2, d) p(b_2 | a_1, a_2, b_1, c_1, c_2, d) \Big) \\
& \times p(d) \Big(p(c_1 | a_1, c_2) p(a_1 | c_2) \Big) \\
& \quad \times \Big(p(c_2 | a_2, c_1) p(a_2 | c_1) \Big)
\end{aligned}
$$

which establishes the required factorization (5). A key step in this argument was the ordering used to decompose $p(b_1, c_1 | a_1, c_2, d)$ into univariate marginals. Had this been decomposed into $p(c_1 | a_1, b_1, c_2, d) p(b_1 | a_1, c_2, d)$, the argument would not have gone through, as the global Markov property could not have been applied, and the terms could not have been re-arranged in the required manner.

We first consider the relation between the depth of a vertex in A, and in $A \cup \{x\}$, (in the case where both sets are ancestral). We show that there exists a (not necessarily unique) total order on the vertices in $A \cup \{x\}$, which is compatible with both orderings on the heads, in the sense that a precedes b if the head containing a precedes the head containing b in $[A]_{\mathcal{G}}$, or in $[A \cup \{x\}]_{\mathcal{G}}$. We use this total order to factor the head terms in the factorization of $p(x_A)$, given by the inductive hypothesis, into univariate terms. We will then apply the global Markov property to add or remove terms to the conditioning sets as required. Finally, re-grouping terms will yield the required factorization of $p(x_{A \cup \{x\}})$.

Lemma 19. *If $A, A \cup \{x\} \in \mathcal{A}(\mathcal{G})$, $(x \notin A)$ then for $w \in A$, $\mathrm{dep}_A(w) \leq \mathrm{dep}_{A \cup \{x\}}(w)$.*

Proof: By induction on $k = \mathrm{dep}_{A \cup \{x\}}(w)$. $\qquad \square$

Lemma 20. *If $A, A \cup \{x\} \in \mathcal{A}(\mathcal{G})$, $(x \notin A)$ then for $w \in A$, $\mathrm{dep}_{A \cup \{x\}}(w) \leq \mathrm{dep}_A(w) + 1$.*

A total ordering $<$ on the vertices in a set $A \in \mathcal{A}(\mathcal{G})$ will be said to be *consistent with the depth ordering induced by A*, if whenever $\mathrm{dep}_A(b) < \mathrm{dep}_A(c)$, then $b < c$.

Corollary 21. *If $A, A \cup \{x\} \in \mathcal{A}(\mathcal{G})$, $(x \notin A)$ then there exists a total ordering on the vertices in A that is consistent with the depth ordering induced by A **and** the depth ordering induced by $A \cup \{x\}$.*

We note in passing that this Corollary cannot, in general, be extended to ancestral sets differing by more than one vertex. Consider the sets $A = \{c, d, e\}$, and $A^* = \{a, b, c, d, e\}$ in Figure 3. $\mathrm{dep}_A(e) = 1 > \mathrm{dep}_A(d) = 0$, but $\mathrm{dep}_{A^*}(d) = 2 > \mathrm{dep}_{A^*}(e) = 1$.

If $<$ is a total ordering, define $suc_<(x) = \{v \mid v \geq x\}$.

4.1 Markov blankets derived from a depth ordering

Lemma 22. *Let $H \in [A]_{\mathcal{G}}$ and $H = \{h_{i_1}, \ldots, h_{i_k}\}$ where the vertices are ordered according to a depth consistent total ordering $<$. Let $T_j = \{h_{i_j}, \ldots, h_{i_k}\} \cup \mathrm{tail}(H)$. Then*

$$
B_j = \Big(\mathrm{dis}_{suc_<(h_{i_j})}(h_{i_j}) \setminus \{h_{i_j}\} \Big) \cup \mathrm{pa}(\mathrm{dis}_{suc_<(h_{i_j})}(h_{i_j}))
$$

forms the Markov blanket of $\{h_{i_j}\}$ in T_j.

Note that B_j is solely determined by h_{i_j}, the ordering, and the graph $\mathcal{G}$. Thus, if A and $A \cup \{x\}$ are ancestral sets $(x \notin A)$ and for some vertex $v \in A$, $\mathrm{head}(v; A) = H$, $\mathrm{head}(v; A \cup \{x\}) = H^*$, with associated tails T and T^*, then, under the ordering given in Corollary 21, the Markov blankets of $\{v\}$ in the sets $(H \cup T) \cap suc_<(h)$ and $(H^* \cup T^*) \cap suc_<(h)$ are the same.

4.2 The ordered local Markov property for ADMGs

A total ordering $(\prec)$ on the vertices of $\mathcal{G}$, which satisfies $x \prec y \Rightarrow y \notin \mathrm{an}(x)$ is said to be *consistent with the ancestor relations in $\mathcal{G}$*. Let $\mathrm{pre}_{\mathcal{G}, \prec}(x) \equiv \{v \mid v \prec x \text{ or } v = x\}$. If $A \in \mathcal{A}(\mathcal{G})$ and $x \in \mathrm{barren}(A)$ then the Markov blanket of $\{x\}$ with respect to the induced subgraph on A is

$$
\mathrm{mb}(x, A) \equiv \mathrm{pa}_{\mathcal{G}} \Big(\mathrm{dis}_A(x) \Big) \cup \Big(\mathrm{dis}_A(x) \setminus \{x\} \Big).
$$

A probability measure P satisfies the *ordered local Markov property* for $\mathcal{G}$, with respect to the ancestor

consistent ordering $\prec$, if for any x, and ancestral set A such that $x \in A \subseteq \mathrm{pre}_{\mathcal{G},\prec}(x)$,

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

Note that $\mathrm{ch}_{\mathcal{G}}(x) \cap \mathrm{pre}_{\mathcal{G},\prec}(x) = \emptyset$, so $\mathrm{ch}_{\mathcal{G}}(x) \cap A = \emptyset$. Since in a DAG, for an arbitrary ancestral set A containing no children of x, $\mathrm{mb}(x, A) = \mathrm{pa}(x)$, the ordered local Markov property given above reduces to the local well-numbering Markov property introduced by [7]. In [11] it is shown that a probability measure obeys the global Markov property for an ADMG $\mathcal{G}$ if and only if it obeys the ordered local Markov property for $\mathcal{G}$.

Lemma 23. *If P factorizes according to $\mathcal{G}$, then P obeys the ordered local Markov property for $\mathcal{G}$.*

Proof: Let $x \in A \in \mathcal{A}(\mathcal{G})$, with $A \subseteq \mathrm{pre}_{\mathcal{G},\prec}(x)$. Let $H = \mathrm{head}(x; A)$. It is sufficient to note that $\mathrm{mb}(x; A) = \mathrm{tail}(H) \cup (H \setminus \{x\})$. $\quad\square$

4.3 Proof of Theorem 4

Proof: It follows directly from Lemma 23, and the results in [11] that if a distribution factorizes according to (1) then it obeys the global Markov property for $\mathcal{G}$.

To prove that if $\mathcal{G}$ obeys the global Markov property then it factorizes according to (1) we proceed by induction on the size of the ancestral set $A \in \mathcal{A}(\mathcal{G})$.

If $|A| = 1$, then the claim is clearly trivial.

If $|A| > 1$, then $\mathrm{barren}(A) \neq \emptyset$. Let w be a vertex in $\mathrm{barren}(A)$, hence $A \setminus \{w\} \in \mathcal{A}(\mathcal{G})$. Let $A^\dagger = A \setminus \{w\}$. By the induction hypothesis, (1) holds for $p(x_{A^\dagger})$. By Corollary 21 there exists a total ordering ($<$) on the vertices in $A^\dagger$ that is consistent with the depth orderings induced by $A^\dagger$ and A. Let v be an arbitrary vertex in $A^\dagger$. Let $H = \mathrm{head}(v; A)$ and $H^\dagger = \mathrm{head}(v; A^\dagger)$. Further let $H = \{h_1, \ldots h_k\}$, and $H^\dagger = \{h_1^\dagger, \ldots h_{k^\dagger}^\dagger\}$, where the vertices are sequenced according to $<$. (The vertices in H and $H^\dagger$ will, in general, be subsequences of the sequence of vertices totally ordered by $<$. However, we refrain from using double subscripts in order to keep the notation simple.) Hence $v = h_j = h_{j^\dagger}^\dagger$, for some $j, j^\dagger$. By Lemma 22, the Markov blankets of $\{v\}$ in the sets $T_j = \{h_j, \ldots, h_k\} \cup \mathrm{tail}(H)$ and $T_j^\dagger = \{h_j^\dagger, \ldots, h_{k^\dagger}^\dagger\} \cup \mathrm{tail}(H^\dagger)$ are the same. Hence

$$p(v \mid h_{j+1}, \ldots, h_k, \mathrm{tail}(H))$$
$$= \ p\left(v \mid h_{j^\dagger+1}^\dagger, \ldots, h_{k^\dagger}^\dagger, \mathrm{tail}(H^\dagger)\right)$$

(where the h's are omitted if $v = h_k$ and the $h^\dagger$'s are omitted if $v = h_k^\dagger$). Thus the density over the $A^\dagger$ margin may be factorized into univariate terms of the

appropriate form. Finally, we observe that since $w \in \mathrm{barren}(A)$, w is m-separated from $A \setminus (\mathrm{head}(w; A) \cup \mathrm{tail}(w; A))$ given $(\mathrm{head}(w; A) \cup \mathrm{tail}(w; A)) \setminus \{w\}$, so

$$p(w \mid A \setminus \{w\})$$
$$= \ p\left(w \mid (\mathrm{head}(w; A) \cup \mathrm{tail}(w; A)) \setminus \{w\}\right).$$

This completes the proof. $\quad\square$

References

[1] M. Drton. Discrete chain graph models. *Bernoulli*, To appear, 2009.

[2] M. Drton, M. Eichler, and T. S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *JMLR*, To appear, 2009.

[3] M. Drton and T. S. Richardson. Binary models for marginal independence. *J. Roy. Statist. Soc. Ser. B*, 70(2):287–309, 2008.

[4] Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In *UAI-06*.

[5] C. Kang and J. Tian. Local markov property for models satisfying composition axiom. In *UAI-05*.

[6] S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, UK, 1996.

[7] S. L. Lauritzen, A. P. Dawid, B. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20:491–505, 1990.

[8] S. L. Lauritzen and F. V. Jensen. Local computation with valuations from a commutative semigroup. *Ann. Math. & Artif. Intell.*, 21(1):51–69, 1997.

[9] S. Meganck, P. Leray, and B. Manderick. Causal graphical models with latent variables: learning and inference. In Khaled Mellouli, editor, *Proceedings, ECSQARU 2007*, pages 5–16. Springer, 2007.

[10] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[11] T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scand. J. Statist.*, 30(1):145–157, 2003.

[12] T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Ann. Statist.*, 30:962–1030, 2002.

[13] I. Shpitser and J. Pearl. Identification of conditional interventional distributions. In *UAI-06*.

[14] S. Wright. The method of path coefficients. *Ann. Math. Statist.*, 5:161–215, 1934.