The Annals of Statistics
2014, Vol. 42, No. 4, 1452–1482
DOI: 10.1214/14-AOS1206

MARKOVIAN ACYCLIC DIRECTED MIXED GRAPHS FOR DISCRETE DATA[1]

BY ROBIN J. EVANS AND THOMAS S. RICHARDSON

University of Oxford and University of Washington

Acyclic directed mixed graphs (ADMGs) are graphs that contain directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges, subject to the constraint that there are no cycles of directed edges. Such graphs may be used to represent the conditional independence structure induced by a DAG model containing hidden variables on its observed margin. The Markovian model associated with an ADMG is simply the set of distributions obeying the global Markov property, given via a simple path criterion (m-separation). We first present a factorization criterion characterizing the Markovian model that generalizes the well-known recursive factorization for DAGs. For the case of finite discrete random variables, we also provide a parameterization of the model in terms of simple conditional probabilities, and characterize its variation dependence. We show that the induced models are smooth. Consequently, Markovian ADMG models for discrete variables are curved exponential families of distributions.

1. Introduction. A directed graph is a finite collection of vertices, V, together with a collection of ordered pairs $E \subset V \times V$ such that $(v, v) \notin E$ for any v; if $(v, w) \in E$ we write $v \rightarrow w$. E is the (directed) edge set. We say a directed graph is *acyclic* if it contains no directed cycles; that is, there is no sequence of vertices $v_1 \rightarrow v_2 \rightarrow \cdots \rightarrow v_k \rightarrow v_1$, for any $k > 1$. We call such a graph a *directed acyclic graph* (DAG). Models based on DAGs are popular because of their simple definition in terms of a recursive factorization, easy to determine conditional independence constraints, and potential for causal interpretations [Pearl (1995, 2009) Robins and Richardson (2011), Spirtes, Glymour and Scheines (1993)]. Unfortunately, if some of the variables in a DAG are unobserved, the resulting pattern of conditional independences no longer corresponds to a DAG model (on the observed variables); in this sense, DAGs are not closed under marginalization.

An *acyclic directed mixed graph* (ADMG) consists of a DAG with vertices V and directed edges E, together with a collection B of unordered (distinct) pairs of elements of V; these are the *bidirected edges*. If $\{v, w\} \in B$ we write $v \leftrightarrow w$, and if in addition $(v, w) \in E$ this is denoted $v \rightleftarrows w$. Graphical definitions are best

Received January 2013; revised December 2013.

[1]Supported by the U.S. National Science Foundation Grant CNS-0855230 and U.S. National Institutes of Health Grant R01 AI032475.

MSC2010 subject classifications. 62M45.

Key words and phrases. Acyclic directed mixed graph, curved exponential family, conditional independence, graphical model, m-separation, parameterization.

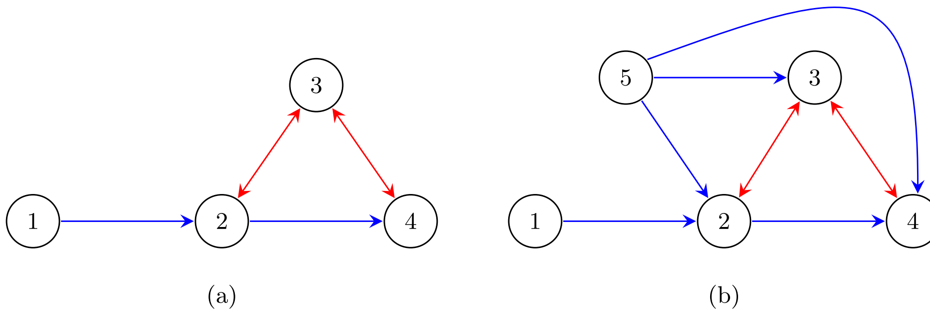

FIG. 1. (a) *An acyclic directed mixed graph*, $\mathcal{L}$. (b) *An ADMG studied by Evans and Richardson* (2013).

understood visually, so we invite the reader to consult the example ADMGs given in Figure 1.

Like DAGs, acyclic directed mixed graphs can be interpreted, via a Markov property, as representing a set of probability distributions defined by conditional independence restrictions; these can be read off the graph using a graphical separation criterion. The advantage of ADMGs is that they are closed under marginalization, in the sense mentioned above [Richardson and Spirtes (2002)]; indeed they represent precisely the conditional independence relations which can be obtained by marginalizing DAGs. Richardson (2003) gave a global Markov property and ordered local Markov property for ADMG models, and showed their equivalence.

The patterns of conditional independence implied by a DAG give rise to curved exponential families in the case of discrete random variables and, therefore, these models have well understood asymptotic statistical properties. However, general models induced by conditional independence constraints do not share this property, and it may be challenging to determine their dimension; for example, certain interpretations of chain graphs are known to lead to non-smooth models [Drton (2009)]. In this paper, we show that discrete ADMG models are curved exponential families, and give a smooth parameterization.

Evans and Richardson (2013) provide a number of applied examples for ADMGs representing discrete distributions—such as using the graph in Figure 1(b) to model an encouragement design for an influenza vaccine—and they discuss the relationship between Markovian ADMG models and marginal log-linear models [Bartolucci, Colombi and Forcina (2007), Bergsma and Rudas (2002)]. ADMGs also arise in studying general conditions for identifying intervention distributions, under the causal interpretation of a DAG model [see Dawid and Didelez (2010), Huang and Valtorta (2006), Pearl and Robins (1995), Shpitser and Pearl (2006a, 2006b), Silva and Ghahramani (2009), Tian and Pearl (2002)].

This paper provides a factorization criterion for joint distributions obeying the global Markov property with respect to an ADMG as well as a parameterization of

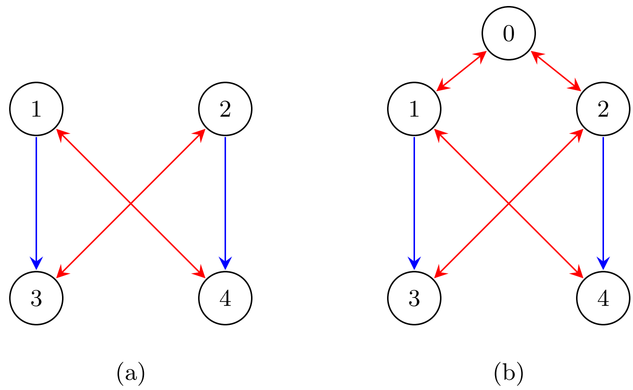

FIG. 2. (a) *An ADMG in which there is no vertex ordering such that all parents of a head precede every vertex in the head*; (b) $\{0, 3, 4\}$ *forms a head in this ADMG*, *but the induced subgraph on* $\{0, 3, 4\}$ *is not connected.*

these models in the discrete case. The factorizations so obtained are unusual: the graph in Figure 2(a), for example, gives

$$f_{1234}(x_1, x_2, x_3, x_4) = f_{23|1}(x_2, x_3 \mid x_1) \cdot f_{14|2}(x_1, x_4 \mid x_2),$$

showing that the joint distribution is a product of two conditional distributions that we would not usually expect to multiply together (see Example 4.13). The factorization criterion generalizes the well known one for DAGs, and is analogous to the Hammersley–Clifford theorem for undirected graphical models [Hammersley and Clifford (1971)]; the parameterization enables model fitting, and is used to prove that the discrete models are curved exponential families of distributions.

ADMGs may be viewed as a subclass of the larger classes of summary graphs [Wermuth (2011)] and ribbonless mixed graphs [Sadeghi (2013), Sadeghi and Lauritzen (2014)], which allow for undirected edges. The factorization and parameterization developed here may be extended to these larger classes without difficulty.

The remainder of the paper is organized as follows: Section 2 introduces basic graphical concepts. In Section 3, we give conditions under which a partial ordering on a class of subsets may be used to define partitions of arbitrary subsets. In Section 4, we use these tools to develop our factorization criterion, which then forms the basis of the simple parameterization introduced in Section 5. In Section 6, we show that the Markov model associated with an ADMG is smooth, and characterize the variation dependence of the parameterization. Finally, Section 7 contains a brief discussion.

2. Graphical definitions and Markov properties. Let $\mathcal{G}$ be an acyclic directed mixed graph with vertices V; the *induced subgraph* of $\mathcal{G}$ over $A \subseteq V$, denoted $\mathcal{G}_A$, is the graph with vertex set A, and all those (directed or bidirected) edges which join two vertices that are both in A.

A *path* in $\mathcal{G}$ is a sequence of adjacent edges, without repetition of a vertex; a path may be empty, or equivalently consist of only one vertex. The first and last vertices on a path are the *endpoints* (these are not distinct if the path is empty); other vertices on the path (if any) are *non-endpoints*. The graph $\mathcal{L}$ in Figure 1(a), for example, contains the path $1 \rightarrow 2 \rightarrow 4 \leftrightarrow 3$, with endpoints 1 and 3, and non-endpoints 2 and 4. A *directed* path is one in which all the edges are directed ($\rightarrow$) and are oriented in the same direction, whereas a *bidirected path* consists entirely of bidirected edges.

We use the usual familial terminology for vertices in a graph. If $w \rightarrow v$, we say that w is a *parent* of v; the set of parents of v is denoted $\mathrm{pa}_{\mathcal{G}}(v)$. More generally, w is an *ancestor* of v if there is a directed path from w to v (note that this includes the case $v = w$); conversely v is a *descendant* of w. The ancestors and descendants of v are denoted $\mathrm{an}_{\mathcal{G}}(v)$ and $\mathrm{de}_{\mathcal{G}}(v)$, respectively. In the graph $\mathcal{L}$ in Figure 1(a), for instance, the ancestors of 4 are the vertices $\mathrm{an}_{\mathcal{L}}(4) = \{1, 2, 4\}$, and

$$\mathrm{pa}_{\mathcal{L}}(4) = \{2\}, \qquad \mathrm{de}_{\mathcal{L}}(4) = \{4\}.$$

The *district* containing v, denoted $\mathrm{dis}_{\mathcal{G}}(v)$, is the set of vertices w such that $v \leftrightarrow \cdots \leftrightarrow w$, including v itself; for example, the district of 4 in $\mathcal{L}$ is $\{2, 3, 4\}$. We apply these functions disjunctively to sets so that, for example,

$$\mathrm{an}_{\mathcal{G}}(W) = \bigcup_{v \in W} \mathrm{an}_{\mathcal{G}}(v).$$

A set of vertices A is *ancestral* if $A = \mathrm{an}_{\mathcal{G}}(A)$; that is, A contains all its own ancestors. Define

$$\mathrm{barren}_{\mathcal{G}}(B) \equiv \{v \in B \,|\, \mathrm{de}_{\mathcal{G}}(v) \cap B = \{v\}\}.$$

We say a set B is *barren* if $B = \mathrm{barren}_{\mathcal{G}}(B)$; that is, it contains none of its nontrivial descendants in $\mathcal{G}$. We will also use the notation $\mathrm{dis}_A(v)$ as a shorthand for $\mathrm{dis}_{\mathcal{G}_A}(v)$, the *district containing v in the induced subgraph of $\mathcal{G}$ on A*.

For an ADMG $\mathcal{G}$ with vertex set V, we consider collections of random variables $(X_v)_{v\in V}$ taking values in probability spaces $(\mathfrak{X}_v)_{v\in V}$; these spaces are either finite discrete sets or finite-dimensional real vector spaces. 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 abuse notation in the usual way: v denotes both a vertex and the random variable X_v, likewise A denotes both a set of vertices and the random vector X_A. For fixed elements of $\mathfrak{X}_v$ and $\mathfrak{X}_A$, we write x_v and x_A, respectively.

The relationship between a graph $\mathcal{G}$ and random variables X_V is governed by Markov properties specified in terms of paths. A non-endpoint vertex c on a path π, is a *collider* on π if the edges preceding and succeeding c on the path both have an arrowhead at c, for example, $\rightarrow c \leftarrow$ or $\leftrightarrow c \leftarrow$; otherwise c is a *non-collider*.

DEFINITION 2.1. A path π in $\mathcal{G}$ between two vertices $v, w \in V(\mathcal{G})$ is said to be *blocked* by a set $C \subseteq V \setminus \{v, w\}$ if either:

(i) there is a non-collider on π, and that non-collider is contained in C; or
(ii) there is a collider on π which is not in $\mathrm{an}_{\mathcal{G}}(C)$.

We say v and w are *m-separated* given C in $\mathcal{G}$ if every path from v to w in $\mathcal{G}$ is blocked by C. Note that C may be empty. Sets $A, B \subseteq V$ are said to be m-separated given $C \subseteq V \setminus (A \cup B)$ if every pair $a \in A$ and $b \in B$ are m-separated given C.

The special case of m-separation for DAGs is the better known d-separation [Lauritzen (1996), Pearl (1988)]. We next relate m-separation to conditional independence, for which we use the now standard notation of Dawid (1979): for random variables X, Y and Z we denote the statement "X is independent of Y conditional on Z" by $X \perp\!\!\!\perp Y | Z$. If Z is empty, we write $X \perp\!\!\!\perp Y$.

DEFINITION 2.2. A probability measure P on $\mathfrak{X}$ is said to satisfy the *global Markov property* (GMP) for an acyclic directed mixed graph $\mathcal{G}$, if for all disjoint sets $A, B, C \subseteq V$ with A and B nonempty, A being m-separated from B given C implies that $X_A \perp\!\!\!\perp X_B | X_C$ under P.

Consider the ADMG $\mathcal{L}$ in Figure 1(a); the vertices 1 and 4 are m-separated conditional on 2, and 1 and 3 are m-separated unconditionally. It is not hard to verify that no other m-separation relations hold for this graph, and that therefore a distribution P obeys the global Markov property with respect to $\mathcal{G}$ if and only if $X_1 \perp\!\!\!\perp X_4 | X_2$ and $X_1 \perp\!\!\!\perp X_3$ under P.

DEFINITION 2.3. Let $\mathcal{G}$ be an ADMG containing an ancestral set A and a vertex $v \in \mathrm{barren}_{\mathcal{G}}(A)$. Define

$$\mathrm{mb}_{\mathcal{G}}(v, A) \equiv \mathrm{pa}_{\mathcal{G}}\big(\mathrm{dis}_A(v)\big) \cup \big(\mathrm{dis}_A(v) \setminus \{v\}\big)$$

to be the *Markov blanket* for v in the induced subgraph $\mathcal{G}_A$. For a set of vertices $W \subseteq \mathrm{barren}_{\mathcal{G}}(A)$, we analogously define the Markov blanket of W to be

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

Let $<$ be a *topological ordering* on the vertices of $\mathcal{G}$, meaning that no vertex appears before any of its ancestors; let $\mathrm{pre}_{\mathcal{G},<}(v)$ be the set of vertices containing v and all vertices preceding v in the ordering. A probability measure P is said to satisfy the *ordered local Markov property* for $\mathcal{G}$ with respect to $<$, if for any v and ancestral set A such that $v \in A \subseteq \mathrm{pre}_{\mathcal{G},<}(v)$,

$$v \perp\!\!\!\perp A \setminus \big(\mathrm{mb}_{\mathcal{G}}(v, A) \cup \{v\}\big) \,|\, \mathrm{mb}_{\mathcal{G}}(v, A)$$

with respect to P.

REMARK 2.4. For $v \in \mathrm{barren}_{\mathcal{G}}(A)$, the Markov blanket for v in A consists of those vertices in $A \setminus \{v\}$ that can be reached from v by paths through A on which all non-endpoints are colliders.

EXAMPLE 2.5. One can easily verify that for the graph in Figure 1(a),

$$\mathrm{mb}_{\mathcal{L}}(4, \{1,2,4\}) = \{2\}, \qquad \mathrm{mb}_{\mathcal{L}}(3, \{1,3\}) = \varnothing,$$

and that therefore under the topological ordering $1, 2, 3, 4$, the ordered local Markov property implies $X_4 \perp\!\!\!\perp X_1 | X_2$ and $X_3 \perp\!\!\!\perp X_1$, just as the global Markov property does.

The following result shows that the two properties are, in fact, always equivalent.

PROPOSITION 2.6 [Richardson (2003), Theorem 2]. *Let $\mathcal{G}$ be an ADMG, and $<$ a topological ordering of its vertices; further let P be a probability measure on $\mathfrak{X}_V$. The following are equivalent*:

(i) *P obeys the global Markov property with respect to $\mathcal{G}$*;
(ii) *P obeys the ordered local Markov property with respect to $\mathcal{G}$ and $<$.*

In particular, this result implies that if the ordered local Markov property is satisfied for some topological ordering $<$, then it is satisfied for all such orderings.

3. Partitions and partial orderings. The global Markov property for DAGs can be equivalently stated in terms of a simple factorization criterion applied to the joint distribution. In order to achieve something similar for ADMGs, we will need to consider partitions of sets of vertices into appropriate blocks. This section develops the necessary mathematical theory on functions that define partitions.

Let V be an arbitrary finite set, and let $\mathcal{H}$ be a collection of nonempty subsets of V, with the restriction that $\{v\} \in \mathcal{H}$ for all $v \in V$ (i.e., all singletons are in $\mathcal{H}$). Let $\prec$ be a partial ordering on the elements of $\mathcal{H}$, and write $H_1 \preceq H_2$ to mean that either $H_1 \prec H_2$ or $H_1 = H_2$.

DEFINITION 3.1. We say that $\prec$ is *partition-suitable* (for $\mathcal{H}$) if for any $H_1, H_2 \in \mathcal{H}$ with $H_1 \cap H_2 \neq \varnothing$, there exists $H^* \in \mathcal{H}$ such that $H^* \subseteq H_1 \cup H_2$ and $H_i \preceq H^*$ for each $i = 1, 2$.

In other words, partition-suitability requires that any two intersecting elements of $\mathcal{H}$ are dominated with respect to $\prec$ by some element of $\mathcal{H}$.

Define a function Φ on subsets of V such that $\Phi(W)$ "picks out" the $\prec$-maximal elements of $\mathcal{H}$ which are subsets of W. That is, it returns the collection of subsets

$$\Phi(W) \equiv \{H \in \mathcal{H} | H \subseteq W \text{ and } H \nprec H' \text{ for all other } H' \subseteq W\}.$$

Partition-suitability ensures that the sets in $\Phi(W)$ are disjoint.

PROPOSITION 3.2. *If $\prec$ is partition-suitable and $H_1, H_2 \in \Phi(A)$ for some set A, then either $H_1 = H_2$ or $H_1 \cap H_2 = \varnothing$.*

PROOF. This is immediate from the definition of partition-suitable. □

Now let

$$\psi(W) \equiv W \setminus \bigcup_{C \in \Phi(W)} C,$$

that is, ψ returns those elements of W which are not contained in any set in $\Phi(W)$. Then recursively define a *partitioning function* $[\cdot]$ on subsets of V by $[\varnothing] = \varnothing$, and

$$[W] \equiv \Phi(W) \cup \bigl[\psi(W)\bigr].$$

The idea is that the function Φ "removes" the maximal sets from W, and the procedure is then applied again to what remains, $\psi(W)$. The following proposition shows that each vertex of W is contained within precisely one set in $[W]$.

PROPOSITION 3.3. *If $\prec$ is partition-suitable, then the function $[\cdot]$ partitions sets. That is, for any $W \subseteq V$,*

$$\bigcup_{H \in [W]} H = W,$$

and if $A, B \in [W]$ then either $A = B$ or $A \cap B = \varnothing$.

PROOF. We proceed by induction on the size of W. If $W = \varnothing$ the result follows from the definition. Also by definition, if $W \neq \varnothing$ then

$$[W] = \Phi(W) \cup \bigl[\psi(W)\bigr],$$

so the induction hypothesis and the definitions of Φ and ψ mean we need only check that $\Phi(W)$ is nonempty and contains disjoint sets.

The first claim follows from the fact that $\prec$ is a partial ordering, and so always contains at least one maximal element (since V is finite); the second is a direct application of Proposition 3.2. □

LEMMA 3.4. *Let $\prec$ be partition-suitable, $A \subseteq V$ and $H \in \Phi(A)$. If $H \subseteq B \subseteq A$ for some subset B, then $H \in \Phi(B)$.*

PROOF. Let $\mathcal{H}_A$ be the set of subsets in $\mathcal{H}$ contained within A. If $H \in \Phi(A) \subseteq \mathcal{H}_A$ then H is maximal with respect to $\prec$ in $\mathcal{H}_A$. It is trivial that $\mathcal{H}_B \subseteq \mathcal{H}_A$, and so H is also maximal in $\mathcal{H}_B$. Thus, $H \in \Phi(B)$. □

We can paraphrase Lemma 3.4 as saying that if a set H is removed from A at the first application of Φ, then H is contained in the partition of any subset B of A (provided B contains H).

The next proposition shows that partitioning functions as we have defined them are stable when some set in the partition is removed. This "stability" is very useful when trying to understand the properties of the partition.

PROPOSITION 3.5. *If* $C \in [W]$, *then* $[W] = \{C\} \cup [W \setminus C]$.

PROOF. We proceed by induction on the size of W. If $[W] = \{C\}$, including any case in which $|W| = 1$, the result is trivial.

If C is not maximal with respect to $\prec$ in W then, by Lemma 3.4, $\Phi(W) = \Phi(W \setminus C)$, so

$$\begin{aligned}[W] &= \Phi(W) \cup \bigl[\psi(W)\bigr] \\ &= \Phi(W \setminus C) \cup \bigl[\psi(W)\bigr],\end{aligned}$$

and the problem reduces to showing that

$$\bigl[\psi(W)\bigr] = \{C\} \cup \bigl[\psi(W \setminus C)\bigr] = \{C\} \cup \bigl[\psi(W) \setminus C\bigr],$$

which holds by the induction hypothesis. Thus, without loss of generality, suppose $C \in \Phi(W)$.

Now, by Lemma 3.4 and the supposition, $\Phi(W \setminus C) \cup \{C\} \supseteq \Phi(W)$, and if equality holds we are done. Otherwise let $C_1, \dots, C_k$ be the sets in $\Phi(W \setminus C)$ but not in $\Phi(W)$. Note that by definition, $C_1, \dots, C_k \subseteq \psi(W)$. Further, these sets are maximal in $W \setminus C$, so by Lemma 3.4 they are also maximal in $\psi(W) \subseteq W \setminus C$. Then the problem reduces to showing that

$$\bigl[\psi(W)\bigr] = \{C_1, \dots, C_k\} \cup \bigl[\psi(W) \setminus (C_1 \cup \cdots \cup C_k)\bigr],$$

which follows from repeated application of the induction hypothesis. □

Lastly, we show that if each set in $\mathcal{H}$ is contained within a piece of some partition of V, then the partitioning function can be applied separately to each piece of this coarser partition.

PROPOSITION 3.6. *Let* $D_1, \dots, D_k$ *be a partition of* V, *and suppose that every* $H \in \mathcal{H}$ *is contained within some* D_i. *Let* $\prec$ *be a partition-suitable partial ordering on* $\mathcal{H}$. *Then for all* $W \subseteq V$,

$$[W] = \bigcup_{i=1}^{k} [W \cap D_i].$$

PROOF. We prove the case $k = 2$, from which the general result follows by repeated applications. If either of $W \cap D_1$ or $W \cap D_2$ are empty, then the result is trivial. By definitions

$$[W] = \Phi(W) \cup \bigl[\psi(W)\bigr];$$

$\psi(W)$ is strictly smaller than W, so by the induction hypothesis

$$[W] = \Phi(W) \cup \bigl[\psi(W) \cap D_1\bigr] \cup \bigl[\psi(W) \cap D_2\bigr].$$

Define $\mathcal{C}_1, \mathcal{C}_2$ so that $\Phi(W) = \mathcal{C}_1 \cup \mathcal{C}_2$ and each $H \in \mathcal{C}_i$ is a subset of D_i only; since the elements of $\mathcal{C}_i$ are maximal with respect to $\prec$ in W, by Lemma 3.4 they are also maximal in $W \cap D_i$. Hence, $\mathcal{C}_i \subseteq \Phi(W \cap D_i)$. Repeatedly applying Proposition 3.5 gives

$$\mathcal{C}_i \cup \left[\psi(W) \cap D_i\right] = [W \cap D_i],$$

because $(\psi(W) \cap D_i) \cup \bigcup_{C \in \mathcal{C}_i} C = W \cap D_i$. Hence the result. □

4. The factorization criterion. Let P be a probability measure having density $f_V : \mathfrak{X}_V \to \mathbb{R}$ with respect to some σ-finite dominating product measure μ on $\mathfrak{X}_V$. For $U, W \subseteq V$, we denote by $f_W : \mathfrak{X}_W \to \mathbb{R}$ the marginal density over W, and by $f_{W|U}(\cdot|u) : \mathfrak{X}_W \to \mathbb{R}$ for $f_U(u) > 0$ the conditional density of W given $U = u$ (more precisely: any member of the equivalence class of such densities). Then P obeys the global Markov property with respect to a DAG if and only if it factorizes as

$$f_V(x_V) = \prod_{v \in V} f_{v|\mathrm{pa}(v)}(x_v | x_{\mathrm{pa}(v)}),$$

for μ-almost all $x_V \in \mathfrak{X}_V$ [see, e.g., Lauritzen (1996)]. In the sequel, all equalities over f are considered to hold almost everywhere with respect to μ.

In this section, we show that factorizations can also be used to characterize Markov models over ADMGs; however, as we shall see, the criterion is more complicated than that for DAGs.

EXAMPLE 4.1. Consider the ADMG in Figure 1(a). A distribution which obeys the global Markov property with respect to this graph satisfies $X_1 \perp\!\!\!\perp X_3$ and $X_1 \perp\!\!\!\perp X_4 | X_2$. It is not possible to specify a factorization on the joint distribution of X_1, X_2, X_3 and X_4 which implies precisely these two independences. Instead, we require factorizations of certain marginal distributions:

$$\begin{aligned} f_{13}(x_1, x_3) &= f_1(x_1) \cdot f_3(x_3), \\ f_{124}(x_1, x_2, x_4) &= f_1(x_1) \cdot f_{2|1}(x_2|x_1) \cdot f_{4|2}(x_4|x_2). \end{aligned}$$

Such marginal factorizations can be used to represent distributions which obey the global Markov property with respect to an ADMG.

DEFINITION 4.2 (Head). A vertex set $H \subseteq V$ is a *head* if it is barren in $\mathcal{G}$ and contained within a single district of $\mathcal{G}_{\mathrm{an}(H)}$. We write $\mathcal{H}(\mathcal{G})$ for the collection of all heads in $\mathcal{G}$.

Note that every singleton vertex $\{v\}$ forms a head.

EXAMPLE 4.3. For the ADMG shown in Figure 2(b), we have the following:

$$\mathcal{H}(\mathcal{G}) = \big\{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}, \{0,1\}, \{0,2\}, \{1,4\}, \{2,3\}, \{0,1,2\}, \{0,1,4\}, \{0,2,3\}, \{0,3,4\}\big\}.$$

Notice that although they are contained within a single district, the sets $\{0,1,2,4\}$, $\{0,1,2,3\}$ and $\{0,1,2,3,4\}$ do not form heads because they are not barren. Also observe that $\{0,3,4\}$ does form a head, even though the induced subgraph $\mathcal{G}_{\{0,3,4\}}$ is not connected [because $\{0,3,4\}$ is a subset of a single district in $\mathcal{G}_{\mathrm{an}(\{0,3,4\})}$, as required].

DEFINITION 4.4 (Tail). For any head H, the *tail* of H is the set

$$\mathrm{tail}_{\mathcal{G}}(H) \equiv \big(\mathrm{dis}_{\mathrm{an}(H)}(H) \setminus H\big) \cup \mathrm{pa}\big(\mathrm{dis}_{\mathrm{an}(H)}(H)\big).$$

If the context makes it clear which head we are referring to, we will sometimes denote a tail simply by T.

Note that the tail is a subset of the ancestors of the head. An intuitive interpretation is that a head H is a set within which no independence relations hold without marginalizing some elements of H, and the tail is the Markov blanket for H within the set $\mathrm{an}_{\mathcal{G}}(H)$. We can therefore factorize ancestral sets into heads conditional upon their tail sets; see Remark 4.14 below.

EXAMPLE 4.5. In the special case of a DAG, the heads are precisely all singleton vertices $\{v\}$, and the tails are the sets of parents $\mathrm{pa}_{\mathcal{G}}(v)$. In a purely bidirected graph, the heads are just the connected sets, and the tails are all empty.

EXAMPLE 4.6. The graph $\mathcal{L}$ in Figure 1(a) has the following head–tail pairs:

H	$\{1\}$	$\{2\}$	$\{3\}$	$\{2,3\}$	$\{4\}$	$\{3,4\}$
T	$\varnothing$	$\{1\}$	$\varnothing$	$\{1\}$	$\{2\}$	$\{1,2\}$

Note that the set $\{2,3,4\}$ is not a head, because it is not barren.

In general, it is not possible to order the vertices in an acyclic directed mixed graph such that, for each head H, all the vertices in $\mathrm{pa}_{\mathcal{G}}(H)$ precede all the vertices in H. A counterexample is given in Figure 2(a), which is taken from Richardson (2009). The head $\{1,4\}$ has parent 2, and the head $\{2,3\}$ has parent 1, so whichever way we order the vertices 1 and 2, the condition will be violated.

However, there is a well-defined partial ordering on heads which will be useful to us, and satisfies the essential property of partition-suitability from Section 3.

DEFINITION 4.7. For two distinct heads H_i and H_j in an ADMG $\mathcal{G}$, say that $H_i \prec H_j$ if $H_i \subseteq \mathrm{an}_{\mathcal{G}}(H_j)$.

LEMMA 4.8. *The (strict) partial ordering $\prec$ is well defined.*

PROOF. We need to verify that $\prec$ is irreflexive, asymmetric and transitive; irreflexivity is by definition. Asymmetry amounts to $H_i \prec H_j \Longrightarrow H_j \not\prec H_i$; suppose not for contradiction, so that there exist distinct heads H_i and H_j with $H_i \prec H_j$ and $H_j \prec H_i$. Since H_i and H_j are distinct, there exists a vertex v which is in one of these heads but not the other; assume without loss of generality that $v \in H_j \setminus H_i$.

Since $H_j \subseteq \mathrm{an}_{\mathcal{G}}(H_i)$, we can find a directed path π_1 from v to some vertex $w \in H_i$; the path is nonempty because $v \notin H_i$. However, since we also have $H_i \subseteq \mathrm{an}_{\mathcal{G}}(H_j)$, we can find a (possibly empty) directed path π_2 from w to some $x \in H_j$. Now, the concatenation of π_1 and π_2 is also a path, because any repeated vertices would imply a directed cycle in the graph. Call this new path π.

But π is a nonempty directed path between two vertices in H_j, which violates the requirement that heads are barren. Hence, asymmetry holds.

For transitivity, if $H_i \prec H_j$ and $H_j \prec H_k$, then clearly we can find a directed path from any element $v \in H_i$ to some element of H_k, simply by concatenating paths from $v \in H_i$ to some $w \in H_j$ and from w to H_k. Hence, $H_i \subseteq \mathrm{an}_{\mathcal{G}}(H_k)$, and so $H_i \prec H_k$. $\square$

LEMMA 4.9. *The partial ordering $\prec$ on the heads $\mathcal{H}(\mathcal{G})$ of an ADMG $\mathcal{G}$ is partition-suitable.*

PROOF SKETCH (SEE THE APPENDIX FOR DETAILS). If two heads H_1, H_2 are distinct and $H_1 \cap H_2 \neq \varnothing$, then $H^* = \mathrm{barren}_{\mathcal{G}}(H_1 \cup H_2)$ is a head, $H_1 \preceq H^*$ and $H_2 \preceq H^*$. $\square$

Note that in general H^* may be a strict subset of $H_1 \cup H_2$. For example, consider the graph shown in Figure 2(b), and let $H_1 = \{0, 1, 4\}$ and $H_2 = \{0, 2, 3\}$ so that $H_1, H_2 \in \mathcal{H}(\mathcal{G})$ and $H_1 \cap H_2 = \{0\}$. However, $H^* = \mathrm{barren}_{\mathcal{G}}(H_1 \cup H_2) = \{0, 3, 4\} \subsetneq H_1 \cup H_2$.

Denote the relevant functions from Section 3 defined by this partial ordering by $\Phi_{\mathcal{G}}$, $\psi_{\mathcal{G}}$ and $[\cdot]_{\mathcal{G}}$, respectively. This partitioning function allows us to factorize probabilities for ADMGs into expressions based upon heads and tails.

EXAMPLE 4.10. For the graph $\mathcal{L}$ in Figure 1(a), we have

H	$\{1\}$	$\{2\}$	$\{3\}$	$\{2, 3\}$	$\{4\}$	$\{3, 4\}$
$\mathrm{an}_{\mathcal{G}}(H)$	$\{1\}$	$\{1, 2\}$	$\{3\}$	$\{1, 2, 3\}$	$\{1, 2, 4\}$	$\{1, 2, 3, 4\}$

so that

$$\{1\} \prec \{2\} \prec \{2, 3\} \prec \{3, 4\},$$

$$\{2\} \prec \{4\} \prec \{3, 4\}, \qquad \{3\} \prec \{2, 3\}.$$

Then, for example, $\Phi_{\mathcal{L}}(\{2,3,4\}) = \{\{3,4\}\}$, and $\Phi_{\mathcal{L}}(\{2\}) = \{\{2\}\}$, giving

$$[\{2,3,4\}]_{\mathcal{L}} = \{\{3,4\},\{2\}\}.$$

EXAMPLE 4.11. For the graph in Figure 2(a), we have

H	$\{1\}$	$\{2\}$	$\{3\}$	$\{4\}$	$\{1,4\}$	$\{2,3\}$
$\mathrm{an}_{\mathcal{G}}(H)$	$\{1\}$	$\{2\}$	$\{1,3\}$	$\{2,4\}$	$\{1,2,4\}$	$\{1,2,3\}$

Thus $\{1\} \prec \{3\} \prec \{2,3\} \succ \{2\}$ and $\{2\} \prec \{4\} \prec \{1,4\} \succ \{1\}$.

Now we can provide a factorization criterion for acyclic directed mixed graphs.

THEOREM 4.12. *Let $\mathcal{G}$ be an ADMG, and P a probability distribution on $\mathfrak{X}_V$ with density f_V. P obeys the global Markov property with respect to $\mathcal{G}$ if and only if for every ancestral set $A \in \mathcal{A}(\mathcal{G})$, and μ-almost all $x_A \in \mathfrak{X}_A$.*

$$f_A(x_A) = \prod_{H \in [A]_{\mathcal{G}}} f_{H|T}(x_H|x_T). \tag{1}$$

A formal proof of this result is given in the Appendix; a sketch proof is given in Richardson (2009), Theorem 4.

EXAMPLE 4.13. For the graph in Figure 2(a), observe that the global Markov property implies precisely that $X_3 \perp\!\!\!\perp X_4 | X_{12}$, and $X_1 \perp\!\!\!\perp X_2$. Applying the partition function to the relevant sets of vertices yields

$$[\{1,2,3,4\}] = \{\{1,4\},\{2,3\}\},$$

so Theorem 4.12 gives us the factorization from the Introduction:

$$f_{1234}(x_1,x_2,x_3,x_4) = f_{23|1}(x_2,x_3 \mid x_1) \cdot f_{14|2}(x_1,x_4 \mid x_2)$$

for all $x_i \in \mathfrak{X}_i$, $i = 1,\dots,4$. The expression may appear slightly strange, since the first factor is the density for $\{X_2, X_3\}$ given X_1, while the second is for $\{X_1, X_4\}$ given X_2; nevertheless this factorization does indeed imply that $X_3 \perp\!\!\!\perp X_4 | X_{12}$. Further, integrating out x_3 and x_4 gives

$$f_{12}(x_1,x_2) = f_{2|1}(x_2|x_1) \cdot f_{1|2}(x_1|x_2), \qquad x_1 \in \mathfrak{X}_1, x_2 \in \mathfrak{X}_2,$$

which implies that $X_1 \perp\!\!\!\perp X_2$.

REMARK 4.14. It follows from Theorem 4.12 that if H is a head, $\mathrm{tail}_{\mathcal{G}}(H)$ is the Markov blanket for H in the set $\mathrm{an}_{\mathcal{G}}(H)$, in the sense that under the global Markov property,

$$H \perp\!\!\!\perp \mathrm{an}_{\mathcal{G}}(H) \setminus \big(H \cup \mathrm{tail}_{\mathcal{G}}(H)\big) \mid \mathrm{tail}_{\mathcal{G}}(H).$$

REMARK 4.15. A different, incorrect definition of $\Phi_{\mathcal{G}}$ (and, therefore, $\psi_{\mathcal{G}}$, $[\cdot]_{\mathcal{G}}$) was given in Richardson (2009) and Evans and Richardson (2010). The erroneous definition coincides with that given here when W is ancestral, so equation (1) holds for both. However, equation (2) below does not hold for the incorrect partitioning function in general.

5. Toward a parameterization of the discrete Markov model for an ADMG. The factorizations in Theorem 4.12 can be used to produce a parameterization of ADMG models when $\mathfrak{X}_V$ is a finite set, and thus the relevant random variables are discrete. For simplicity of exposition, we will henceforth assume that the random variables are binary, so $\mathfrak{X}_V = \{0, 1\}^{|V|}$. Extension to the general finite discrete case is easy but notationally challenging; this is done in the special case of ADMGs with chain graph structure by Drton (2009).

In the following result, and throughout the paper, empty products are assumed to equal 1.

THEOREM 5.1. *Let $\mathcal{G}$ be an ADMG, and P a probability distribution on $\{0, 1\}^{|V|}$. Then P obeys the global Markov property with respect to $\mathcal{G}$ if and only if for every ancestral set A and $x_A \in \mathfrak{X}_A$,*

$$P(X_A = x_A) = \sum_{C : O \subseteq C \subseteq A} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = 0 | X_T = x_T), \tag{2}$$

where $O \equiv \{v \in A | x_v = 0\}$.

Theorem 5.1 shows that conditional probabilities of the form $P(X_H = 0|X_T = x_T)$ are sufficient to form a parameterization of the binary ADMG model; it remains to show that they are nonredundant, which is proved in Section 6.

Note that the sets C in (2) may not be ancestral, which hinders proof by induction. In order to facilitate the proof, we define the following quantity which will be needed in the intermediate steps of the induction.

DEFINITION 5.2. Let A be an ancestral set in an ADMG $\mathcal{G}$, and consider a particular assignment x_A to X_A; write $O \equiv \{v \in A | x_v = 0\}$. For any sets B, W such that $B \subseteq W \subseteq (A \setminus O)$, define the following quantity:

$$g_{x_A}(B, W) \equiv (-1)^{|B|} \prod_{H \in [O \cup W]_{\mathcal{G}}} P(X_{H \cap (O \cup B)} = 0, X_{H \setminus (O \cup B)} = 1 \mid X_T = x_T).$$

Note that if $B = \varnothing$ then the right-hand side has factors of the form $P(X_H = x_H | X_T = x_T)$, and looks much like (1); however, if $B = W$ the expression is a product of the form $P(X_H = 0|X_T = x_T)$, just like each term of (2).

The interpretation is that W is the set of nonzero vertices being partitioned, and which need to have their values on the left-hand side of any conditioning bars "flipped" from 1 to 0 in order to get an expression of the form (2). The set B consists of those vertices for which this "flipping" has already taken place, and those in $W \setminus B$ have yet to be flipped.

The induction starts with the single term $(B, W) = (\varnothing, A \setminus O)$, given via Theorem 4.12. At each step a term is "reduced" into a sum of two further pieces by flipping a single vertex, until the procedure finishes with a sum containing the set of terms

$$\{g_{x_A}(B, W) : (B, W) = (C, C), \text{ where } C \subseteq A \setminus O\},$$

and thus corresponds to an expression of the form (2).

DEFINITION 5.3. Take a triple (x_A, B, W), where $B \subsetneq W \subseteq (A \setminus O)$ for $O \equiv \{v \in A | x_v = 0\}$. We say that (x_A, B, W) is *reducible* if for each $H \in [O \cup W]_{\mathcal{G}}$ such that $H \cap (W \setminus B) \neq \varnothing$, it holds that $\mathrm{dis}_{\mathrm{an}(H)}(H) \setminus H \subseteq O \cup (W \setminus B)$.

In words, given a set W in which not all vertices are flipped, so $W \setminus B \neq \varnothing$, the condition requires that any head H which is in the partition and has not yet been fully "flipped," has the part of its tail which is from the same district [i.e., $\mathrm{dis}_{\mathrm{an}(H)}(H) \setminus H$] consists solely of vertices that are either in W or not yet flipped.

The following technical lemma provides the necessary piece for the induction step.

LEMMA 5.4. *Let A be an ancestral set, and P a distribution obeying the global Markov property with respect to $\mathcal{G}$. If (x_A, B, W) is reducible in $\mathcal{G}_A$, then there is some $w \in W \setminus B$ such that*

$$g_{x_A}(B, W) = g_{x_A}(B, W \setminus \{w\}) + g_{x_A}(B \cup \{w\}, W), \tag{3}$$

and, in addition, either $B = W \setminus \{w\}$ (so $B \cup \{w\} = W$), or both $(x_A, B, W \setminus \{w\})$ and $(x_A, B \cup \{w\}, W)$ are also reducible.

PROOF. See the Appendix. $\square$

Here, w is a vertex that is given the value 1 in every head in $g_{x_V}(B, W)$, but is "flipped" so it is set equal to 0 in $g_{x_V}(B \cup \{w\}, W)$ and is removed from the partition in $g_{x_V}(B, W \setminus \{w\})$. A major difficulty in the overall proof of Theorem 5.1 stems from the fact that, though each g_{x_A} produced after a reduction is itself reducible or of the form $g_{x_A}(C, C)$, we will not generally be able to flip the same vertex in each term.

PROOF OF THEOREM 5.1. By Theorem 4.12, the global Markov property holds if and only if for each ancestral A and x_A,

$$\begin{aligned} P(X_A = x_A) &= \prod_{H \in [A]_{\mathcal{G}}} P(X_H = x_H | X_T = x_T) \\ &= g_{x_A}(\varnothing, A \setminus O) \end{aligned}$$

using the definition of g_{x_V}. It is easy to check from Definition 5.3 that either $A = O$, in which case there is nothing to prove, or $(x_A, \varnothing, A \setminus O)$ is reducible. Then from repeated application of Lemma 5.4 this is just

$$\begin{aligned} &= \sum_{C \subseteq A \setminus O} g_{x_A}(C, C) \\ &= \sum_{C \subseteq A \setminus O} (-1)^{|C|} \prod_{H \in [O \cup C]_{\mathcal{G}}} P(X_H = 0 | X_T = x_T) \end{aligned}$$

which, by inspection, gives the required result. Conversely, suppose (2) holds, and that $v \in \mathrm{barren}_{\mathcal{G}}(A)$ has district $D_1 = \mathrm{dis}_A(v)$; let $D_2 \equiv A \setminus D_1$ and for $C \subseteq A$, let $C_i \equiv C \cap D_i$ and $O_i \equiv O \cap D_i$, $i = 1, 2$. Then $C \setminus O = (C_1 \setminus O_1) \dot{\cup} (C_2 \setminus O_2)$, and from Proposition 3.6 get $[C]_{\mathcal{G}} = [C_1]_{\mathcal{G}} \cup [C_2]_{\mathcal{G}}$. Hence,

$$\begin{aligned} P(X_A = x_A) &= \sum_{C\,:\, O \subseteq C \subseteq A} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = 0 \mid X_T = x_T) \\ &= \sum_{\substack{C_1 : O_1 \subseteq C_1 \subseteq D_1 \\ C_2 : O_2 \subseteq C_2 \subseteq D_2}} (-1)^{|C_1 \setminus O_1| + |C_2 \setminus O_2|} \\ &\quad \times \prod_{H \in [C_1]_{\mathcal{G}} \cup [C_2]_{\mathcal{G}}} P(X_H = 0 \mid X_T = x_T) \\ &= \sum_{C_1 : O_1 \subseteq C_1 \subseteq D_1} (-1)^{|C_1 \setminus O_1|} \prod_{H \in [C_1]_{\mathcal{G}}} P(X_H = 0 \mid X_T = x_T) \\ &\quad \times \sum_{C_2 : O_2 \subseteq C_2 \subseteq D_2} (-1)^{|C_2 \setminus O_2|} \prod_{H \in [C_2]_{\mathcal{G}}} P(X_H = 0 \mid X_T = x_T) \\ &= h(x_{D_1}, x_{\mathrm{pa}_{\mathcal{G}}(D_1)}) \cdot k(x_{D_2}, x_{\mathrm{pa}_{\mathcal{G}}(D_2)}) \end{aligned}$$

for some functions h, k. In particular, k does not involve x_v, so it follows that $v \perp\!\!\!\perp A \setminus (D_1 \cup \mathrm{pa}_{\mathcal{G}}(D_1)) \mid (D_1 \setminus \{v\}) \cup \mathrm{pa}_{\mathcal{G}}(D_1)$ which, by the definition of the Markov blanket of v in A, is equivalent to

$$v \perp\!\!\!\perp A \setminus \bigl(\mathrm{mb}_{\mathcal{G}}(v, A) \cup \{v\}\bigr) \mid \mathrm{mb}_{\mathcal{G}}(v, A).$$

It follows that the ordered local Markov property holds (for any topological ordering); hence, by Proposition 2.6 so does the global Markov property. □

6. Model smoothness. Let $\mathcal{P}_{\mathcal{G}} \subseteq \Delta_{2^n-1}$ denote the set of strictly positive binary probability distributions which obey the global Markov property with respect to an ADMG $\mathcal{G}$, where Δ_k is the strictly positive k-dimensional probability simplex and n is the number of vertices in $\mathcal{G}$. We call $\mathcal{P}_{\mathcal{G}}$ the *model* defined by $\mathcal{G}$ on a binary state-space. In this section, such models are shown to be smooth, in the sense that they are curved exponential families of distributions, and we prove that the conditional probabilities used in Theorem 5.1 constitute a parameterization.

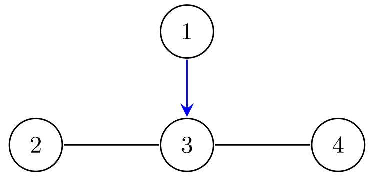

FIG. 3. *A chain graph representing a non-smooth discrete model under the Alternative Markov Property.* (*Note that this is not an acyclic directed mixed graph nor a summary graph.*)

Models induced by patterns of conditional independence may be non-smooth, and determining which are smooth in general is a difficult open problem [Drton and Xiao (2010)]. Non-smoothness can occur even if the conditional independences arise from a Markov property applied to a graph, as in the following example.

EXAMPLE 6.1. Consider the chain graph given in Figure 3, which mixes directed and undirected edges. Under the Alternative Markov Property (AMP) for chain graphs, this graph represents distributions for which $X_2 \perp\!\!\!\perp X_4 | X_1, X_3$ and $X_1 \perp\!\!\!\perp X_2, X_4$ [Andersson, Madigan and Perlman (2001)]. This is shown by Drton (2009) to represent a non-smooth model for discrete random variables.

It follows from Theorem 5.1 that for an ADMG $\mathcal{G}$, the collection of probabilities of the form

$$P(X_H = 0 \mid X_T = x_T), \qquad x_T \in \mathfrak{X}_T, H \in \mathcal{H}(\mathcal{G}),$$

is sufficient to recover the joint distribution under the model $\mathcal{P}_\mathcal{G}$. However, it is not immediately clear that each of these probabilities is necessary, or more specifically that the map in (2) is smooth and of full rank everywhere.

For brevity, we write $q_H(x_T) \equiv P(X_H = 0 | X_T = x_T)$, and the vector of all such probabilities by

$$\mathbf{q} \equiv \bigl(q_H(x_T) | H \in \mathcal{H}(\mathcal{G}), x_T \in \mathfrak{X}_T\bigr). \tag{4}$$

For $\mathbf{p} \in \mathcal{P}_\mathcal{G}$, we—in a mild abuse of notation—let $\mathbf{q}(\mathbf{p})$ be the vector of the form (4) determined by calculating the appropriate conditional probabilities from $\mathbf{p}$. Since this only involves adding and dividing strictly positive numbers, the map $\mathbf{q}$ is smooth (infinitely differentiable). Let $\mathcal{Q}_\mathcal{G} \equiv \mathbf{q}(\mathcal{P}_\mathcal{G})$ be the image of $\mathbf{q}$ over $\mathcal{P}_\mathcal{G}$; we call $\mathcal{Q}_\mathcal{G}$ the set of *derived* parameter values. We will prove that the map in (2) provides a smooth inverse to $\mathbf{q}$. The first result shows that the set of vectors $\mathbf{q}$ that are derived parameters corresponds exactly to those which give strictly positive probabilities under the inverse map.

THEOREM 6.2. *For an ADMG $\mathcal{G}$, a vector $\mathbf{q}$ is derived* (*i.e.*, $\mathbf{q} \in \mathcal{Q}_\mathcal{G}$) *if and only if for each $x_V \in \mathfrak{X}_V$, we have*

$$p_{x_V}(\mathbf{q}) \equiv \sum_{C : x_V^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus x_V^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_H(x_T) > 0, \tag{5}$$

where $x_V^{-1}(0) \equiv \{v \in V \mid x_v = 0\}$.

REMARK 6.3. The boundary of $\mathcal{Q}_\mathcal{G}$ is the set of $\mathbf{q}$ such that $p_{x_V}(\mathbf{q}) \geq 0$ for all $x_V \in \mathfrak{X}_V$, with equality holding in at least one case.

The definition of $p_{x_V}(\mathbf{q})$ in (5) is of the same form as the expression given for $P(X_V = x_V)$ in (2) and so the result might at first seem trivial; clearly probabilities must be nonnegative. However, it is not immediately obvious that this condition is *sufficient* for parameters to be in the image set $\mathcal{Q}_\mathcal{G} \equiv \mathbf{q}(\mathcal{P}_\mathcal{G})$. If we take some $\mathbf{q}^\dagger \notin \mathcal{Q}_\mathcal{G}$ and apply to it the nonlinear functional form in (5) to obtain $\mathbf{p}(\mathbf{q}^\dagger)$, without this result there is no apparent reason why $\mathbf{p}(\mathbf{q}^\dagger)$ should not be a probability distribution, nor indeed in $\mathcal{P}_\mathcal{G}$.

To prove Theorem 6.2, we need the following lemma.

LEMMA 6.4. *Let A be an ancestral set in $\mathcal{G}$, and let $x_A \in \mathfrak{X}_A$. Then for any real vector* $\mathbf{q}$ (*not necessarily in* $\mathcal{Q}_\mathcal{G}$), *the map in* (5) *satisfies*

$$\sum_{y_V\,:\,y_A = x_A} p_{y_V}(\mathbf{q}) = \sum_{C\,:\,x_A^{-1}(0) \subseteq C \subseteq A} (-1)^{|C \setminus x_A^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_H(x_T),$$

where $x_A^{-1}(0) \equiv \{v \in A \mid x_v = 0\}$. *In particular, taking* $A = \varnothing$,

$$\sum_{y_V} p_{y_V}(\mathbf{q}) = 1.$$

Recall that empty products are assumed equal to 1.

PROOF OF LEMMA 6.4. If $A = V$ the result is trivial. If not, pick some $v \in \operatorname{barren}_\mathcal{G}(V) \setminus A$; this is possible because if $A \supseteq \operatorname{barren}_\mathcal{G}(V)$ then $A = V$ by ancestrality of A. So

$$\begin{aligned}
\sum_{\substack{y_V:\\ y_A = x_A}} p_{y_V}(\mathbf{q}) &= \sum_{\substack{y_V:\\ y_A = x_A}} \sum_{y_V^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus y_V^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_H(y_T) \\
&= \sum_{\substack{y_{V\setminus\{v\}}:\\ y_A = x_A}} \sum_{y_v} \sum_{y_V^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus y_V^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_H(y_T) \\
&= \sum_{\substack{y_{V\setminus\{v\}}:\\ y_A = x_A}} \Biggl(\sum_{y_{V\setminus\{v\}}^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus y_{V\setminus\{v\}}^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_H(y_T) \\
&\qquad + \sum_{y_{V\setminus\{v\}}^{-1}(0) \cup \{v\} \subseteq C \subseteq V} (-1)^{|C \setminus (y_{V\setminus\{v\}}^{-1}(0) \cup \{v\})|} \prod_{H \in [C]_\mathcal{G}} q_H(y_T) \Biggr).
\end{aligned}$$

The last equation simply breaks the sum into cases where $y_v = 1$ and $y_v = 0$, respectively, which takes this form because v does not appear in any tail sets. The first inner sum in the last expression can be further divided into the cases where C contains v, and those where it does not, giving

$$\sum_{\substack{y_V:\\ y_A=x_A}} p_{y_V}(\mathbf{q}) = \sum_{\substack{y_{V\setminus\{v\}}:\\ y_A=x_A}} \Bigg(\sum_{y_{V\setminus\{v\}}^{-1}(0)\subseteq C\subseteq V\setminus\{v\}} (-1)^{|C\setminus y_{V\setminus\{v\}}^{-1}(0)|} \prod_{H\in[C]_{\mathcal{G}}} q_H(y_T)$$

$$+ \sum_{y_{V\setminus\{v\}}^{-1}(0)\cup\{v\}\subseteq C\subseteq V} (-1)^{|C\setminus y_{V\setminus\{v\}}^{-1}(0)|} \prod_{H\in[C]_{\mathcal{G}}} q_H(y_T)$$

$$+ \sum_{y_{V\setminus\{v\}}^{-1}(0)\cup\{v\}\subseteq C\subseteq V} (-1)^{|C\setminus (y_{V\setminus\{v\}}^{-1}(0)\cup\{v\})|} \prod_{H\in[C]_{\mathcal{G}}} q_H(y_T)\Bigg).$$

The second and third terms differ only by a factor of -1, and so cancel leaving

$$\sum_{\substack{y_V:\\ y_A=x_A}} p_{y_V}(\mathbf{q}) = \sum_{\substack{y_{V\setminus\{v\}}:\\ y_A=x_A}} \Bigg(\sum_{y_{V\setminus\{v\}}^{-1}(0)\subseteq C\subseteq V\setminus\{v\}} (-1)^{|C\setminus y_{V\setminus\{v\}}^{-1}(0)|} \prod_{H\in[C]_{\mathcal{G}}} q_H(y_T)\Bigg).$$

Repeating this until no vertices outside A are left gives

$$\sum_{\substack{y_V:\\ y_A=x_A}} p_{y_V}(\mathbf{q}) = \sum_{y_A^{-1}(0)\subseteq C\subseteq A} (-1)^{|C\setminus y_A^{-1}(0)|} \prod_{H\in[C]_{\mathcal{G}}} q_H(y_T).$$

In the special case $A = \varnothing$, we end up with an empty product

$$\sum_{y_V} p_{y_V}(\mathbf{q}) = (-1)^{|\varnothing|} \prod_{H\in[\varnothing]_{\mathcal{G}}} q_H(y_T) = 1.$$

□

PROOF OF THEOREM 6.2. The "only if" part of the statement follows from Theorem 5.1 by the fact that if the parameters are derived then $p_{x_V}(\mathbf{q}) = P(X_V = x_V)$, and these are therefore positive by definition of $\mathcal{P}_{\mathcal{G}}$.

For the converse, suppose that the inequalities hold; we will show that we can retrieve the parameters simply by calculating the appropriate conditional probabilities. Lemma 6.4 ensures that $\sum_{x_V} p_{x_V}(\mathbf{q}) = 1$, and that therefore this is a probability distribution.

Choose some $H^* \in \mathcal{H}(\mathcal{G})$, with $T^* = \mathrm{tail}_{\mathcal{G}}(H^*)$ and $A = \mathrm{an}_{\mathcal{G}}(H^*)$; also set $x_{H^*} = 0$ and pick $x_{T^*} \in \{0,1\}^{|T^*|}$. By Lemma 6.4,

$$\sum_{y_V: y_A=x_A} p_{y_V}(\mathbf{q}) = \sum_{y_A^{-1}(0)\subseteq C\subseteq A} (-1)^{|C\setminus y_A^{-1}(0)|} \prod_{H\in[C]_{\mathcal{G}}} q_H(y_T).$$

Now clearly $H^* \in \Phi_{\mathcal{G}}(A)$, so applying Lemma 3.4 and the fact that $H^* \subseteq x_A^{-1}(0) = y_A^{-1}(0)$ shows $H^* \in [C]_{\mathcal{G}}$ for all terms C in the sum and, therefore, we can apply Proposition 3.5 to factor out the parameter associated with H^*:

$$= q_{H^*}(y_{T^*}) \sum_{y_A^{-1}(0)\subseteq C\subseteq A} (-1)^{|C\setminus y_A^{-1}(0)|} \prod_{H\in[C\setminus H^*]_{\mathcal{G}}} q_H(y_T)$$

$$= q_{H^*}(y_{T^*}) \sum_{y_{A\setminus H^*}^{-1}(0)\subseteq C\subseteq A\setminus H^*} (-1)^{|C\setminus y_{A\setminus H^*}^{-1}(0)|} \prod_{H\in[C]_{\mathcal{G}}} q_H(y_T).$$

But note that $A \setminus H^*$ is also an ancestral set, and thus using Lemma 6.4 again,

$$\sum_{y_V : y_{A\setminus H^*}=x_{A\setminus H^*}} p_{y_V}(\mathbf{q}) = \sum_{y_{A\setminus H^*}^{-1}(0)\subseteq C\subseteq A\setminus H^*} (-1)^{|C\setminus y_{A\setminus H^*}^{-1}(0)|} \prod_{H\in[C]_{\mathcal{G}}} q_H(y_T).$$

Hence,

$$\frac{\sum_{y_{V\setminus A}} p_{x_V}(\mathbf{q})}{\sum_{y_{V\setminus(A\setminus H^*)}} p_{x_V}(\mathbf{q})} = q_{H^*}(x_{T^*}),$$

and we can recover the original parameters from the probability distribution $\mathbf{p}$ in the manner we would expect; that $\mathbf{p}$ satisfies the global Markov property for $\mathcal{G}$ then follows from Theorem 5.1. Thus, $\mathbf{p} \in \mathcal{P}_{\mathcal{G}}$ and $\mathbf{q} = \mathbf{q}(\mathbf{p}) \in \mathcal{Q}_{\mathcal{G}}$. $\square$

THEOREM 6.5. *For an ADMG $\mathcal{G}$, the model $\mathcal{P}_{\mathcal{G}}$ of strictly positive binary probability distributions satisfying the global Markov property with respect to $\mathcal{G}$ is smoothly parameterized by $\mathbf{q} \in \mathcal{Q}_{\mathcal{G}}$.*

Consequently, the model $\mathcal{P}_{\mathcal{G}}$ is a curved exponential family of dimension

$$d = \sum_{H\in\mathcal{H}(\mathcal{G})} |\mathfrak{X}_{\mathrm{tail}(H)}| = \sum_{H\in\mathcal{H}(\mathcal{G})} 2^{|\mathrm{tail}(H)|}.$$

PROOF. By Theorem 6.2, the set $\mathcal{Q}_{\mathcal{G}} \subseteq \mathbb{R}^d$ is open. The map $\mathbf{p}(\mathbf{q}) : \mathcal{Q}_{\mathcal{G}} \to \mathcal{P}_{\mathcal{G}}$ is multilinear and, therefore, infinitely differentiable. Its inverse $\mathbf{q} : \mathcal{P}_{\mathcal{G}} \to \mathcal{Q}_{\mathcal{G}}$ is also infinitely differentiable on $\mathcal{P}_{\mathcal{G}}$.

The composition $\mathbf{q} \circ \mathbf{p}$ is the identity function on $\mathcal{Q}_{\mathcal{G}}$ and, therefore, its Jacobian is the identity matrix I_d. However, the Jacobian of a composition of differentiable functions is the product of the Jacobians, so

$$I_d = \frac{\partial \mathbf{q}}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \mathbf{q}}.$$

But this implies that each of the Jacobians has full rank d and, therefore, the map $\mathbf{q}$ is a smooth parameterization of $\mathcal{P}_{\mathcal{G}}$. See Kass and Vos (1997), Corollary A.3. $\square$

7. Discussion. We remark that it is easy to extend the results of Sections 5 and 6 from the binary case to a general finite discrete state-space; we have avoided this only for notational simplicity. It is also a simple matter to extend the results from ADMGs to the *summary graphs* of Wermuth (2011) which incorporate three types of edge: directed ($\rightarrow$), undirected ($—$), and dashed (- - -); the dashed edges are equivalent to bidirected ($\leftrightarrow$) edges [Sadeghi and Lauritzen (2014)]. The undirected component of a summary graph can be dealt with using standard methods for undirected graphs [Lauritzen (1996)], and the remaining parameterization done as for an ADMG, conditional on the undirected component.

APPENDIX: TECHNICAL PROOFS

PROOF OF LEMMA 4.9. Suppose that two heads H_1, H_2 are distinct and $H_1 \cap H_2 \neq \emptyset$. We will show that they are dominated by $H^* \equiv \text{barren}_{\mathcal{G}}(H_1 \cup H_2)$; clearly $H^* \subseteq H_1 \cup H_2$ and $H_1, H_2 \subseteq \text{an}_{\mathcal{G}}(H^*)$, so if H^* is a head then $\prec$ satisfies the requirements for partition-suitability.

Clearly H^* is barren, so we need to prove that it is contained within a single district in $\text{an}_{\mathcal{G}}(H^*)$. By definition, $\text{an}_{\mathcal{G}}(H^*) \supseteq H_1 \cup H_2$; we need to find a bidirected path between any $v, w \in H^* \subseteq H_1 \cup H_2$. If v and w are either both in H_1 or both in H_2, then the existence of such a path follows from the fact that these are heads. If $v \in H_1$ and $w \in H_2$, then construct a bidirected path in $\text{an}_{\mathcal{G}}(H_1)$ from v to some vertex $x \in H_1 \cap H_2$, and a bidirected path in $\text{an}_{\mathcal{G}}(H_2)$ from x to w; these paths can then be concatenated into a new path meeting the requirements, shortening the resulting sequence of edges if necessary to avoid repetition of vertices. Hence, H^* is a head.

Since $H_1, H_2 \subseteq \text{an}_{\mathcal{G}}(H^*)$ we have $H_i \preceq H^*$ for each $i = 1, 2$, and therefore $\prec$ is partition-suitable. $\square$

Proof of factorization.

PROPOSITION A.1. *Let $\prec$ and $\prec'$ be two partition-suitable partial orderings for $\mathcal{H}$, such that for every $H \in \mathcal{H}$ and $W \subseteq V$, H is maximal in W under $\prec$ whenever this is so under $\prec'$. Then $[\cdot]^{\prec} = [\cdot]^{\prec'}$.*

PROOF. We again proceed by induction on the size of W. Recall that for all $v \in V$, we have $\{v\} \in \mathcal{H}$ by the definition of partition-suitability, so $[\{v\}]^{\prec} = [\{v\}]^{\prec'} = \{\{v\}\}$. Now take a general $W \subseteq V$, and suppose that H is maximal under $\prec'$ in W; then by Proposition 3.5

$$\begin{aligned}
[W]^{\prec'} &= \{H\} \cup [W \setminus H]^{\prec'} \\
&= \{H\} \cup [W \setminus H]^{\prec} \\
&= [W]^{\prec}
\end{aligned}$$

by applying the induction hypothesis to $W \setminus H$, and using the fact that $H \in [W]^{\prec}$ because it is also maximal under $\prec$ in W. $\square$

Define a partial ordering $\prec^*$ on heads in an ADMG by $H_1 \prec^* H_2$ if and only if both $H_1 \prec H_2$, and H_1 and H_2 are contained in the same district in $\mathrm{an}_{\mathcal{G}}(H_1 \cup H_2)$; note that this is a weaker ordering than $\prec$, since strictly fewer pairs of sets are comparable. It is easy to see that $\prec^*$ is partition-suitable for heads $\mathcal{H}(\mathcal{G})$ by repeating the proof of Lemma 4.9. In addition, sets which are maximal under $\prec$ will also be maximal under $\prec^*$, so the partitions defined by these two orderings are the same by Proposition A.1.

This weaker partial ordering leads us to a class of sets which play a role similar to that of ancestral set: a set with "ancestrally closed districts" is one whose districts are ancestrally closed (rather than the whole set).

DEFINITION A.2. Let $\mathcal{G}$ be an ADMG, and W be a subset of its vertices. We say W *has ancestrally closed districts* if $\mathrm{dis}_{\mathrm{an}(W)}(W) = W$.

Equivalently, W has ancestrally closed districts if W is not connected to $\mathrm{an}_{\mathcal{G}}(W) \setminus W$ by any bidirected edges. This definition is important because the partitioning function $[\cdot]_{\mathcal{G}}$ will act upon sets with ancestrally closed districts "separately" within the relevant ancestral set: that is, for such sets,

$$[\mathrm{an}_{\mathcal{G}}(W)]_{\mathcal{G}} = [W]_{\mathcal{G}} \cup [\mathrm{an}_{\mathcal{G}}(W) \setminus W]_{\mathcal{G}}.$$

Note that if $D = D_1 \dot{\cup} D_2$ has ancestrally closed districts, and D_1 and D_2 are not joined by any bidirected edges, then D_1 and D_2 themselves have ancestrally closed districts (here $\dot{\cup}$ indicates a disjoint union). If for every $v, w \in D$ there is a bidirected path from v to w such that all the vertices on the path are contained within D, then D cannot be partitioned in this manner, and we say it is *bidirected-connected*.

DEFINITION A.3. Let $C \subseteq V$. We say that an ordering $<$ on the vertices of C is $(C, \prec^*)$-*consistent* if for any $H_1, H_2 \in [C]_{\mathcal{G}}$ such that $H_1 \prec^* H_2$, we have $v_1 < v_2$ for all $v_1 \in H_1$, $v_2 \in H_2$.

LEMMA A.4. *Let $D = D_1 \dot{\cup} D_2$ have ancestrally closed districts and be such that D_1 is not connected to D_2 by any bidirected edges. Let $<_1$ and $<_2$ be orderings on D_1 and D_2 (resp.). If for $i = 1, 2$, $<_i$ is $(D_i, \prec^*)$-consistent, then every extension of $<_1$ and $<_2$ to an ordering $<$ on D is also a $(D, \prec^*)$-consistent ordering.*

PROOF. Orderings between vertices $v_1, v_2 \in D_i$ are specified by $<_i$. Further, if $v_1 \in D_1$ and $v_2 \in D_2$ then since v_1 and v_2 are in different districts in $\mathrm{an}_{\mathcal{G}}(D)$, it follows from the definition of $\prec^*$ that v_1 and v_2 can be ordered in any way to achieve a $(D, \prec^*)$ consistent ordering. $\square$

A total ordering $<_i$ on a set D_i will be said to be *topological in* $\mathcal{G}$ if no vertex $d \in D_i$ precedes any of its proper ancestors in $\mathcal{G}$ that are in D_i.

LEMMA A.5. *Let D_1 and D_2 be disjoint subsets in $\mathcal{G}$. Let $<_1$ and $<_2$ be topological orderings on D_1 and D_2 (resp.). Then there exists an extension of $<_1$ and $<_2$ to a topological ordering $<$ on $D_1 \cup D_2$.*

PROOF. We construct a topological ordering iteratively as follows: let $\langle d_1, \dots, d_{k-1}\rangle$ be the first $k-1$ vertices in $D_1 \cup D_2$ already ordered under $<$; let $E_k = (D_1 \cup D_2) \setminus \{d_1, \dots, d_{k-1}\}$ be the set of vertices remaining to be ordered. Further, let $Q_k = \{d \mid d \in E_k, \mathrm{an}_{\mathcal{G}}(d) \cap E_k = \{d\}\}$ be those vertices in E_k that have no proper ancestors in E_k; $Q_k \neq \varnothing$ since V is finite and $\mathcal{G}$ is acyclic. Finally, if $Q_k \cap D_1 \neq \varnothing$, define d_k to be the first element in Q_k under $<_1$, otherwise define d_k to be the first element in Q_k under $<_2$. That the ordering is topological follows from the definition of Q_k. □

LEMMA A.6. *Let D have ancestrally closed districts, and suppose $C \subseteq \mathrm{barren}_{\mathcal{G}}(D)$. Then $D \setminus C$ has ancestrally closed districts.*

PROOF. Let $D' \equiv D \setminus C$. Since $C \subseteq \mathrm{barren}_{\mathcal{G}}(D)$, $\mathrm{an}_{\mathcal{G}}(D \setminus C) \subseteq \mathrm{an}_{\mathcal{G}}(D) \setminus C$, so

$$\mathrm{dis}_{\mathrm{an}(D')}(D') \subseteq \mathrm{dis}_{\mathrm{an}(D)}(D') \setminus C \subseteq \mathrm{dis}_{\mathrm{an}(D)}(D) \setminus C = D \setminus C = D'.$$

Since $D' \subseteq \mathrm{dis}_{\mathrm{an}(D')}(D')$, the result holds. □

LEMMA A.7. *Let $C \cup W$ have ancestrally closed districts, with $W \subseteq \mathrm{barren}_{\mathcal{G}}(C \cup W)$ and $W \cap C = \varnothing$. Then any ordering on W may be extended to a topological ordering of the vertices in $C \cup W$ which is both $(C, \prec^*)$ and $(C \cup W, \prec^*)$-consistent.*

PROOF. Note that C has ancestrally closed districts by Lemma A.6. We proceed by induction on the size of $C \cup W$; if $|C \cup W| = 0$ or 1 then the result is trivial.

If $C \cup W$ contains two components which are not connected by bidirected edges, then we can split it into two smaller sets $C_1 \cup W_1$ and $C_2 \cup W_2$, each with ancestrally closed districts, where $C = C_1 \dot{\cup} C_2$ and $W = W_1 \dot{\cup} W_2$. Clearly, $W_i \in \mathrm{barren}(C_i \cup W_i)$ for each i, so using the induction hypothesis, we can find topological orderings $<_i$ on the vertices of $C_i \cup W_i$ which are both $(C_i \cup W_i, \prec^*)$ and $(C_i, \prec^*)$ consistent. It then follows from Lemma A.5, taking $D_i = (C_i \cup W_i)$, that there exists a topological ordering $<$ on $C \cup W$ that extends $<_1$ and $<_2$. It further follows from two applications of Lemma A.4 that $<$ is both $(C, \prec^*)$ and $(C \cup W, \prec^*)$-consistent.

Since, by assumption, $C \cup W$ has ancestrally closed districts, if this set does not contain two components then $C \cup W$ is a single district in $\mathrm{an}_{\mathcal{G}}(C \cup W)$. Let $H = \mathrm{barren}_{\mathcal{G}}(C \cup W)$; this is clearly a head and maximal under $\prec^*$ in $C \cup W$. Further, $W \subseteq H$ so applying Proposition 3.5 gives

$$\begin{aligned}[C \cup W]_{\mathcal{G}} &= \{H\} \cup \bigl[(C \cup W) \setminus H\bigr]_{\mathcal{G}} \\ &= \{H\} \cup \bigl[C \setminus (H \setminus W)\bigr]_{\mathcal{G}},\end{aligned}$$

since $W \cap C = \varnothing$. Since $H \setminus W \subseteq \mathrm{barren}_{\mathcal{G}}(C)$, Lemma A.6 shows that $C \setminus (H \setminus W)$ also has ancestrally closed districts; applying the induction hypothesis, we can find a topological ordering of C which is both $(C \setminus (H \setminus W), \prec^*)$ and $(C, \prec^*)$-consistent [possibly $C \setminus (H \setminus W) = C$ in which case this is trivial]. This ordering may be combined with an arbitrary ordering on W by simply concatenating the orderings so that everything in W comes after everything in C. This gives an ordering which is $(C \cup W, \prec^*)$-consistent, because $H \supseteq W$ is maximal; since W is barren in $C \cup W$, the ordering is also topological. □

COROLLARY A.8. *If $D \cup \{w\}$ has ancestrally closed districts with $w \in \mathrm{barren}_{\mathcal{G}}(D \cup \{w\})$, then there exists an ordering $<$ which is both $(D, \prec^*)$ and $(D \cup \{w\}, \prec^*)$-consistent, and such that w is the maximal vertex under $<$.*

PROOF. The claim is trivial if $w \in D$. Otherwise, $\{w\}$ is barren in $D \cup \{w\}$, so we apply the previous lemma. □

Note that the previous lemma and this corollary do not generalize to adding two vertices: there exist graphs with ancestral sets A, $A \cup \{w_1\}$ and $A \cup \{w_1, w_2\}$, such that no topological ordering is $(A, \prec^*)$-, $(A \cup \{w_1\}, \prec^*)$- and $(A \cup \{w_1, w_2\}, \prec^*)$-consistent. See Richardson (2009) for such an example.

Given a path, π, and two vertices v, w on π, the *subpath* $\pi(v, w)$ is the sequence of edges which lie between v and w on π. As with a path, we allow a single vertex (and no edges) to be a degenerate case of a subpath.

LEMMA A.9. *Suppose π is a path from a to b, and is not blocked by C. Then every vertex v on π is contained in $\mathrm{an}_{\mathcal{G}}(\{a, b\} \cup C)$.*

PROOF. Suppose w is on π and is an ancestor of neither a nor b. Then on each of the subpaths $\pi(a, w)$ and $\pi(w, b)$, there is at least one edge with an arrowhead pointing towards w along the subpath. Let v_{aw} and v_{wb} be the vertices at which such arrowheads occur that are closest to w on the respective subpaths. There are now three cases: (1) if $w \neq v_{wb}$ then $\pi(w, v_{wb})$ is a directed path from w to v_{wb}. It further follows that v_{wb} is a collider on π, and since the path is not blocked by v_{wb}, it is an ancestor of C. Hence, $w \in \mathrm{an}_{\mathcal{G}}(C)$. (2) If $w \neq v_{aw}$, then a symmetric argument holds. (3) If $v_{aw} = w = v_{wb}$, then w is a collider on π, hence again an ancestor of C. $\square$

The next two lemmas are used to establish necessary results about Markov blankets:

LEMMA A.10. *Let $H_1, H_2 \in [D]_{\mathcal{G}}$ with $H_1 \neq H_2$. Then at least one of the following holds*:

(i) $H_1 \prec H_2$;
(ii) $H_2 \prec H_1$; *or*
(iii) *there is no bidirected path between any $h_1 \in H_1$ and $h_2 \in H_2$ contained within* $\mathrm{an}_{\mathcal{G}}(H_1 \cup H_2)$.

PROOF. Suppose $H_1, H_2 \in [D]_{\mathcal{G}}$, and that (iii) fails. Then let $H^* \equiv \mathrm{barren}_{\mathcal{G}}(H_1 \cup H_2)$. Since, H_1, H_2 are heads and since (iii) fails, H^* is a barren set which is connected by bidirected paths in $\mathrm{an}_{\mathcal{G}}(H^*) = \mathrm{an}_{\mathcal{G}}(H_1 \cup H_2)$; hence, it is a head. In addition, $H^* \subseteq H_1 \cup H_2 \subseteq D$, and $H^* \succeq H_1, H_2$.

It follows that $H^* \in [D]_{\mathcal{G}}$, which means that either $H^* = H_1$, in which case (ii) holds, or $H^* = H_2$, in which case (i) holds. $\square$

LEMMA A.11. *Let D be bidirected-connected with ancestrally closed districts, and let $D' \equiv D \setminus \{w\}$ for some $w \in \mathrm{barren}_{\mathcal{G}}(D)$. Let $<$ be a total order that is $(D, \prec^*)$- and $(D', \prec^*)$-consistent, and under which w is maximal. For a given $v \in D'$ define H, H' to be the heads such that $v \in H \in [D]_{\mathcal{G}}$ and $v \in H' \in [D']_{\mathcal{G}}$, respectively, and T, T' the corresponding tails. Let*

$$\begin{aligned}
B &\equiv \bigl(\mathrm{dis}_{\mathrm{pre}_<(v)}(v) \setminus \{v\}\bigr) \cup \mathrm{pa}_{\mathcal{G}}\bigl(\mathrm{dis}_{\mathrm{pre}_<(v)}(v)\bigr),\\
C &\equiv \bigl(H \cap \mathrm{pre}_<(v)\bigr) \cup T \quad \textit{and}\\
C' &\equiv \bigl(H' \cap \mathrm{pre}_<(v)\bigr) \cup T'.
\end{aligned}$$

Then $B \subseteq C$ and $B \subseteq C'$, and B m-separates v from both $C \setminus B$ and $C' \setminus B$.

PROOF. Let $S \equiv \mathrm{dis}_{\mathrm{pre}_<(v)}(v) \subseteq D$; we claim that $S \subseteq \mathrm{an}_{\mathcal{G}}(H)$. If not then there is a bidirected path π from v to some $s \in S \setminus \mathrm{an}_{\mathcal{G}}(H)$; let this path be minimal, so that s is adjacent on π to some $t \in \mathrm{an}_{\mathcal{G}}(H)$. Then s lies in some different head $H^* \in [D]_{\mathcal{G}}$, and we have constructed a bidirected path from H to H^* within $\mathrm{an}_{\mathcal{G}}(H \cup H^*)$; it follows from Lemma A.10 that either $H \prec H^*$ or $H^* \prec H$, but the former is ruled out by the existence of π and the $(D, \prec^*)$-consistency of $<$. Hence, $H^* \subseteq \mathrm{an}_{\mathcal{G}}(H)$, and in particular $s \in \mathrm{an}_{\mathcal{G}}(H)$, so we reach a contradiction.

Thus, $S \subseteq \mathrm{an}_{\mathcal{G}}(H)$ and, therefore, $S \equiv \mathrm{dis}_{\mathrm{pre}_<(v)}(v) \subseteq \mathrm{dis}_{\mathrm{an}(H)}(v)$, so

$$S \cup \mathrm{pa}_{\mathcal{G}}(S) \subseteq \mathrm{dis}_{\mathrm{an}(H)}(v) \cup \mathrm{pa}\bigl(\mathrm{dis}_{\mathrm{an}(H)}(v)\bigr) \subseteq H \cup T.$$

Finally, using $S \subseteq \text{pre}_<(v)$, we have

$$B = (S \setminus \{v\}) \cup \text{pa}_{\mathcal{G}}(S) \subseteq (H \cap \text{pre}_<(v)) \cup T = C.$$

It follows from Lemma A.6 and the fact that $w \in \text{barren}_{\mathcal{G}}(D)$, that D' also has ancestrally closed districts, and the same argument as above shows that $B \subseteq C'$.

Now, let π be a path from v to some $c \in C \setminus B$, and assume without loss of generality that π does not intersect $C \setminus B$ other than at c. We will show that π is blocked by B.

Note that $B \subseteq C \subseteq \text{pre}_<(v)$; thus if π includes any vertex $s > v$ then it is blocked by Lemma A.9, because s is not an ancestor of any element of C. Consequently, we may assume that the edge on π adjacent to v is of the form $v \leftrightarrow$ or $v \leftarrow$.

We claim that π contains at least one non-collider; suppose not for a contradiction: then π is of the form

$$v \leftrightarrow t_1 \leftrightarrow \cdots \leftrightarrow t_p \leftrightarrow c, \qquad v \leftrightarrow t_1 \leftrightarrow \cdots \leftrightarrow t_p \leftarrow c \quad \text{or} \quad v \leftarrow c,$$

with every node t_i an ancestor of B and hence of D. Since D has ancestrally closed districts, it follows that every $t_i \in D$ and hence $t_i \in \text{dis}_{\text{pre}_<(v)}(v) \setminus \{v\}$, so $t_i \in B$. But then $c \in B$, which is a contradiction, since we assumed $c \in C \setminus B$.

It follows that π contains at least one non-collider; let d be the non-collider closest to v on the path. But then repeating the argument above (replacing c with d) shows that $d \in B$ and, therefore, π is blocked by B.

Similarly, all paths π' from v to some c' in $C' \setminus B$ are blocked by B. $\square$

The next lemma is the crux of the induction used in the proof of Theorem 4.12.

LEMMA A.12. *Let D have ancestrally closed districts, and $w \in \text{barren}_{\mathcal{G}}(D)$. Then for any f_V obeying the global Markov property with respect to $\mathcal{G}$, we have*

$$\prod_{H \in [D]_{\mathcal{G}}} f_{H|T}(x_H | x_T) = f_{w|\text{an}(D)\setminus\{w\}}(x_w | x_{\text{an}(D)\setminus\{w\}}) \prod_{H \in [D\setminus\{w\}]_{\mathcal{G}}} f_{H|T}(x_H | x_T)$$

μ-almost everywhere.

PROOF. Note that we need only prove the case where D forms a single district, from which the general result will follow because by Proposition 3.6 the factors not involving $\text{dis}_D(w)$ are the same on both sides. Assume therefore that $D = \text{dis}_D(w)$, and thus D is bidirected-connected.

Define $D' = D \setminus \{w\}$, and let $<$ be a topological total ordering which is $(D, \prec^*)$ and $(D', \prec^*)$ consistent, which exists by Corollary A.8. Further, we can choose w to be the maximal element in D.

For any $v \in H \in [D]_{\mathcal{G}}$, let $H_v = H \cap \text{pre}_<(v)$, and similarly for $v \in H' \in [D']_{\mathcal{G}}$, let $H'_v = H' \cap \text{pre}_<(v)$. In addition, let

$$B_v \equiv (\text{dis}_{\text{pre}_<(v)}(v) \setminus \{v\}) \cup \text{pa}_{\mathcal{G}}(\text{dis}_{\text{pre}_<(v)}(v)).$$

Then

$$\prod_{H \in [D]_{\mathcal{G}}} f_{H|T}(x_H|x_T) = \prod_{H \in [D]_{\mathcal{G}}} \prod_{v \in H} f_{v|H_v \cup T}(x_v|x_{H_v}, x_T)$$
$$= \prod_{H \in [D]_{\mathcal{G}}} \prod_{v \in H} f_{v|B_v}(x_v|x_{B_v})$$
$$= \prod_{v \in D} f_{v|B_v}(x_v|x_{B_v}),$$

where the first equality follows from the elementary properties of conditional probabilities, and the second from applying Lemma A.11 to see that B_v m-separates v from $(H_v \cup T) \setminus B_v$.

But B_v also m-separates v from $(H'_v \cup T') \setminus B_v$, so reversing the argument gives

$$\prod_{v \in D} f_{v|B_v}(x_v|x_{B_v}) = f_{w|B_w}(x_w|x_{B_w}) \prod_{v \in D \setminus \{w\}} f_{v|B_v}(x_v|x_{B_v})$$
$$= f_{w|B_w}(x_w|x_{B_w}) \prod_{H' \in [D \setminus \{w\}]_{\mathcal{G}}} \prod_{v \in H'} f_{v|B_v}(x_v|x_{B_v})$$
$$= f_{w|B_w}(x_w|x_{B_w}) \prod_{H' \in [D \setminus \{w\}]_{\mathcal{G}}} \prod_{v \in H'} f_{v|H'_v \cup T'}(x_v|x_{H'_v}, x_{T'})$$
$$= f_{w|B_w}(x_w|x_{B_w}) \prod_{H' \in [D \setminus \{w\}]_{\mathcal{G}}} f_{H'|T'}(x_{H'}|x_{T'}).$$

In addition, note that $B_w = H_w \cup T$, so it is the Markov blanket for w in $\mathrm{an}_{\mathcal{G}}(D)$ using the ordered local Markov property. Thus,

$$f_{w|B_w}(x_w|x_{B_w}) = f_{w|\,\mathrm{an}(D) \setminus \{w\}}(x_w|x_{\mathrm{an}(D) \setminus \{w\}}),$$

which gives the result. $\square$

PROOF OF THEOREM 4.12. We proceed by induction on $|A|$. Clearly, the result holds if $|A| = 1$.

If $|A| > 1$, then let $w \in \mathrm{barren}_{\mathcal{G}}(A)$; thus $A' \equiv A \setminus \{w\}$ is also ancestral. Suppose that the global Markov property holds; then by elementary laws of probability and the induction hypothesis,

$$f_A(x_A) = f_{w|A'}(x_w|x_{A'}) \cdot f_{A'}(x_{A'})$$
$$= f_{w|A'}(x_w|x_{A'}) \prod_{H' \in [A']_{\mathcal{G}}} f_{H'|T'}(x_{H'}|x_{T'})$$

and by Lemma A.12, this is just

$$= \prod_{H \in [A]_{\mathcal{G}}} f_{H|T}(x_H|x_T).$$

Conversely, suppose that (1) holds and let $<$ be a topological ordering of the ancestral set A. By the induction hypothesis, the ordered local Markov property is satisfied for $<$ and all suitable pairs (v, A') such that $A' \subset A$. Let $w \in \mathrm{barren}_{\mathcal{G}}(A)$ be the maximal vertex under $<$ in A, with H such that $w \in H \in [A]_{\mathcal{G}}$; the factorization implies that $H \perp\!\!\!\perp A \setminus (H \cup T) \mid T$. Note that $H = \mathrm{barren}_{\mathcal{G}}(\mathrm{dis}_A(w))$, so

$$\begin{aligned} \mathrm{mb}_{\mathcal{G}}(w, A) &\equiv \big(\mathrm{dis}_A(w) \setminus \{w\}\big) \cup \mathrm{pa}_{\mathcal{G}}\big(\mathrm{dis}_A(w)\big) \\ &= \big(H \setminus \{w\}\big) \cup T. \end{aligned}$$

This then implies $w \perp\!\!\!\perp A \setminus (\mathrm{mb}_{\mathcal{G}}(w) \cup \{w\}) \mid \mathrm{mb}_{\mathcal{G}}(w)$ by the weak union property of conditional independence. Hence, the ordered local Markov property is satisfied. □

Proof of parameterization.

PROPOSITION A.13. *If $H \in [W]_{\mathcal{G}}$ and $D = \mathrm{dis}_{\mathrm{an}(H)}(H) \cap W$ then $[W]_{\mathcal{G}} = [W \setminus D]_{\mathcal{G}} \cup [D]_{\mathcal{G}}$.*

PROOF. Note that since $H \in [W]_{\mathcal{G}}$, $H \subseteq \mathrm{dis}_{\mathrm{an}(H)}(H) \cap W = D$. The proof is by induction on $|W \setminus D|$. If $W \setminus D = \varnothing$, the claim is trivial. Suppose $H^* \in [W]_{\mathcal{G}}$ and $H^* \cap D \neq \varnothing$. Applying Lemma A.10 to $H, H^* \in [W]_{\mathcal{G}}$ we see that either $H^* = H$ or $H^* \prec H$, so $H^* \subseteq D$. Thus, every head in $[W]_{\mathcal{G}}$ is either a subset of D or $W \setminus D$. Consequently, there exists $H^\dagger \in [W]_{\mathcal{G}}$ with $H^\dagger \subseteq W \setminus D$; let $W^\dagger \equiv W \setminus H^\dagger$. By Proposition 3.5, $[W]_{\mathcal{G}} = \{H^\dagger\} \cup [W^\dagger]_{\mathcal{G}}$. Since $D \subseteq W^\dagger$ and $H \in [W^\dagger]_{\mathcal{G}}$, the conclusion follows from the inductive hypothesis applied to $W^\dagger$. □

PROOF OF LEMMA 5.4. It suffices to prove the result for $A = V$, from which the general case follows by applying it to the subgraphs $\mathcal{G}_A$.

Since (x_V, B, W) is reducible, $W \setminus B \neq \varnothing$; let H^* be a maximal head such that both $H^* \in [O \cup W]_{\mathcal{G}}$ and $H^* \cap (W \setminus B) \neq \varnothing$, further take $w \in H^* \cap (W \setminus B)$. Let $D^* \equiv \mathrm{dis}_{\mathrm{an}(H^*)}(H^*)$ be the associated district within the ancestors of H^*. By construction, D^* has ancestrally closed districts and is bidirected-connected.

Define $y_B \equiv 0$, $y_{V \setminus B} \equiv x_{V \setminus B}$; then

$$\begin{aligned} &g_{x_V}(B, W) \\ &\quad \equiv (-1)^{|B|} \prod_{H \in [O \cup W]_{\mathcal{G}}} P(X_H = y_H | X_T = x_T) \\ &\quad = (-1)^{|B|} \prod_{H \in [(O \cup W) \setminus H^*]_{\mathcal{G}}} P(X_H = y_H | X_T = x_T) \\ &\qquad\qquad \times \big\{P(X_w = 1, X_{H^* \setminus \{w\}} = y_{H^* \setminus \{w\}} | X_{T^*} = x_{T^*}) \\ &\qquad\qquad\quad + P(X_w = 0, X_{H^* \setminus \{w\}} = y_{H^* \setminus \{w\}} | X_{T^*} = x_{T^*}) \\ &\qquad\qquad\quad - P(X_w = 0, X_{H^* \setminus \{w\}} = y_{H^* \setminus \{w\}} | X_{T^*} = x_{T^*})\big\}. \end{aligned}$$

The last term after distributing the product is just $g_{x_V}(B \cup \{w\}, W)$, so to prove (3) we need to show that

$$\begin{aligned} & g_{x_V}\bigl(B, W \setminus \{w\}\bigr) \\ & \quad \equiv (-1)^{|B|} \prod_{H \in [(O \cup W) \setminus \{w\}]_{\mathcal{G}}} P(X_H = y_H | X_T = x_T) \\ & \quad = (-1)^{|B|} \\ & \qquad \times \prod_{H \in [(O \cup W) \setminus H^*]_{\mathcal{G}}} P(X_H = y_H | X_T = x_T) \\ & \qquad \qquad \times \bigl\{P(X_w = 1, X_{H^* \setminus \{w\}} = y_{H^* \setminus \{w\}} | X_{T^*} = x_{T^*}) \\ & \qquad \qquad \quad + P(X_w = 0, X_{H^* \setminus \{w\}} = y_{H^* \setminus \{w\}} | X_{T^*} = x_{T^*})\bigr\}. \end{aligned} \tag{6}$$

Note that by the definition of reducibility, $D^* \setminus H^* \subseteq O \cup (W \setminus B)$, so $D^* \setminus H^*$ does not contain any "flipped" vertices; hence, $D^* \cap B \subseteq H^*$. Further, $D^* \subseteq O \cup W$.

By Proposition A.13, applied to H^*, D^* and $O \cup W$, $[O \cup W]_{\mathcal{G}} = [(O \cup W) \setminus D^*]_{\mathcal{G}} \cup [D^*]_{\mathcal{G}}$. Thus, every head $H^\dagger \in [O \cup W]_{\mathcal{G}}$ which contains a vertex in $D^* \setminus H^*$ is such that $H^\dagger \subseteq D^*$. Hence, by applying Lemma A.10 to D^*, it follows that $H^\dagger \prec H^*$ [since $H^\dagger \subseteq D^* \setminus H^* \subseteq \mathrm{an}_{\mathcal{G}}(H^*)$ rules out $H^* \prec H^\dagger$, while $H^*, H^\dagger \subseteq D^*$ rules out (iii)]. Thus, D^* is made up of H^* and the heads which precede it under $\prec$, and hence also under $\prec^*$.

Suppose we replace $[O \cup W]_{\mathcal{G}}$ with $[(O \cup W) \setminus S]_{\mathcal{G}}$ for some $S \subseteq H^*$; from Lemma 3.4, it is clear that only heads which precede H^* under $\prec^*$ will be affected, so in particular:

$$\begin{aligned} & \bigl[(O \cup W) \setminus H^*\bigr]_{\mathcal{G}} = \bigl[(O \cup W) \setminus D^*\bigr]_{\mathcal{G}} \cup \bigl[D^* \setminus H^*\bigr]_{\mathcal{G}} \quad \text{and} \\ & \bigl[(O \cup W) \setminus \{w\}\bigr]_{\mathcal{G}} = \bigl[(O \cup W) \setminus D^*\bigr]_{\mathcal{G}} \cup \bigl[D^* \setminus \{w\}\bigr]_{\mathcal{G}}. \end{aligned} \tag{7}$$

It follows that to establish (6) it suffices to show:

$$\begin{aligned} & \prod_{H \in [D^* \setminus \{w\}]_{\mathcal{G}}} P(X_H = y_H | X_T = x_T) \\ & \quad = \bigl\{P(X_w = 0, X_{H^* \setminus \{w\}} = y_{H^* \setminus \{w\}} | X_{T^*} = x_{T^*}) \\ & \qquad + P(X_w = 1, X_{H^* \setminus \{w\}} = y_{H^* \setminus \{w\}} | X_{T^*} = x_{T^*})\bigr\} \\ & \qquad \times \prod_{H \in [D^* \setminus H^*]_{\mathcal{G}}} P(X_H = y_H | X_T = x_T). \end{aligned} \tag{8}$$

Let $z_{D^*\setminus\{w\}} \equiv y_{D^*\setminus\{w\}}$ and $z_{V\setminus D^*} \equiv x_{V\setminus D^*}$ (with z_w remaining free). Since $D^* \cap B \subseteq H^*$, applying Lemma A.12 to D^* and w using the values of z_V gives

$$P(X_w = z_w | X_{(H^*\cup T^*)\setminus\{w\}} = z_{(H^*\cup T^*)\setminus\{w\}}) \prod_{H\in[D^*\setminus\{w\}]_{\mathcal{G}}} P(X_H = z_H | X_T = z_T)$$

$$= \prod_{H\in[D^*]_{\mathcal{G}}} P(X_H = z_H | X_T = z_T)$$

$$= P(X_{H^*} = z_{H^*} | X_{T^*} = z_{T^*}) \prod_{H\in[D^*\setminus H^*]_{\mathcal{G}}} P(X_H = z_H | X_T = z_T).$$

Summing both sides of the equation over z_w yields (8). Thus, (3) holds.

It remains to demonstrate that if $B \cup \{w\} \neq W$, the triples $(x_V, B \cup \{w\}, W)$ and $(x_V, B, W \setminus \{w\})$ are also reducible.

For the first, consider $H \in [O \cup W]_{\mathcal{G}}$ with $H \cap (W \setminus (B \cup \{w\})) \neq \varnothing$. Let $D \equiv \mathrm{dis}_{\mathrm{an}(H)}(H) \subseteq O \cup W$; by construction D has ancestrally closed districts. Since $H \cap (W \setminus B) \supseteq H \cap (W \setminus (B \cup \{w\})) \neq \varnothing$, by the reducibility of (x_V, B, W), $D \setminus H \subseteq O \cup (W \setminus B)$. It is sufficient to show that $w \notin D \setminus H$. Since by Proposition A.13, $[O \cup W]_{\mathcal{G}} = [(O \cup W) \setminus D]_{\mathcal{G}} \cup [D]_{\mathcal{G}}$, if $w \in D \cap H^*$ then $H^* \in [D]_{\mathcal{G}}$. If $H = H^*$, then $w \notin D \setminus H$. If $H \neq H^*$, then applying Lemma A.10 we have $H^* \prec H$ (by the same argument as above). But this contradicts that H^* is a maximal head in $[O \cup W]_{\mathcal{G}}$ such that $H^* \cap (W \setminus B) \neq \varnothing$. Hence, $(x_V, B \cup \{w\}, W)$ is reducible.

We now consider $(x_V, B, W \setminus \{w\})$. Let $H \in [O \cup (W \setminus \{w\})]$, with $H \cap ((W \setminus \{w\}) \setminus B) \neq \varnothing$. Again, let $D \equiv \mathrm{dis}_{\mathrm{an}(H)}(H)$.

First suppose $H \in [(O \cup W) \setminus D^*]_{\mathcal{G}}$ then, by (7), $H \in [O \cup W]_{\mathcal{G}}$. We showed above that if $H \in [O \cup W]_{\mathcal{G}}$ and $H \cap (W \setminus (B \cup \{w\})) \neq \varnothing$ then $D \setminus H \subseteq (W \setminus (B \cup \{w\}))$. This is sufficient since $W \setminus (B \cup \{w\}) = (W \setminus \{w\}) \setminus B$.

If $H \notin [(O \cup W) \setminus D^*]_{\mathcal{G}}$ then (7) implies $H \in [D^* \setminus \{w\}]_{\mathcal{G}}$. Lemma A.6 applied to D^* implies that $D^* \setminus \{w\}$ has ancestrally closed districts, so $D \subseteq D^* \setminus \{w\}$. Since $D \subseteq D^*$, if a vertex v is not barren in D then $v \notin \mathrm{barren}_{\mathcal{G}}(D^*) = H^*$. Hence, $H^* \cap D \subseteq \mathrm{barren}_{\mathcal{G}}(D) = H$. Thus,

$$D \setminus H \subseteq D \setminus H^* \subseteq D^* \setminus H^* \subseteq O \cup (W \setminus B),$$

where the third inclusion follows from the reducibility of (x_V, B, W) and the choice of H^*. But since $D \subseteq D^* \setminus \{w\}$, we have $D \setminus H \subseteq O \cup ((W \setminus \{w\}) \setminus B)$ as required. □

REFERENCES

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Stat.* **28** 33–85. MR1844349

BARTOLUCCI, F., COLOMBI, R. and FORCINA, A. (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statist. Sinica* **17** 691–711. MR2398430

BERGSMA, W. P. and RUDAS, T. (2002). Marginal models for categorical data. *Ann. Statist.* **30** 140–159. MR1892659

DAWID, A. P. (1979). Conditional independence in statistical theory. *J. Roy. Statist. Soc. Ser. B* **41** 1–31. MR0535541

DAWID, A. P. and DIDELEZ, V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Stat. Surv.* **4** 184–231. MR2740837

DRTON, M. (2009). Discrete chain graph models. *Bernoulli* **15** 736–753. MR2555197

DRTON, M. and XIAO, H. (2010). Smoothness of Gaussian conditional independence models. In *Algebraic Methods in Statistics and Probability II. Contemp. Math.* **516** 155–177. Amer. Math. Soc., Providence, RI. MR2730747

EVANS, R. J. and RICHARDSON, T. S. (2010). Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the* 26*th Conference on Uncertainty in Artificial Intelligence* 177–184. AUAI Press, Corvallis, OR.

EVANS, R. J. and RICHARDSON, T. S. (2013). Marginal log-linear parameters for graphical Markov models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75** 743–768. MR3091657

HAMMERSLEY, J. M. and CLIFFORD, P. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

HUANG, Y. and VALTORTA, M. (2006). Pearl's calculus of interventions is complete. In *Proceedings of the* 22*nd Conference On Uncertainty in Artificial Intelligence*. AUAI Press, Arlington, VA.

KASS, R. E. and VOS, P. W. (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley, New York. MR1461540

LAURITZEN, S. L. (1996). *Graphical Models*. Oxford Univ. Press, New York. MR1419991

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA. MR0965765

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710. MR1380809

PEARL, J. (2009). *Causality*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166

PEARL, J. and ROBINS, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence* (*Montreal*, *PQ*, 1995) 444–453. Morgan Kaufmann, San Francisco, CA. MR1615028

RICHARDSON, T. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.* **30** 145–157. MR1963898

RICHARDSON, T. S. (2009). A factorization criterion for acyclic directed mixed graphs. In *Proceedings of the* 25*th Conference on Uncertainty in Artificial Intelligence* 462–470. AUAI Press, Arlington, VA.

RICHARDSON, T. and SPIRTES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. MR1926166

ROBINS, J. M. and RICHARDSON, T. S. (2011). Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology*: *Finding the Determinants of Disorders and Their Cures* (P. Shrout, K. Keyes and K. Ornstein, eds.) **6** 1–52. Oxford Univ. Press, London.

SADEGHI, K. (2013). Stable mixed graphs. *Bernoulli* **19** 2330–2358. MR3160556

SADEGHI, K. and LAURITZEN, S. (2014). Markov properties for mixed graphs. *Bernoulli* **20** 676–696. MR3178514

SHPITSER, I. and PEARL, J. (2006a). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the* 21*st National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA.

SHPITSER, I. and PEARL, J. (2006b). Identification of conditional interventional distributions. In *Proceedings of the* 22*nd Conference on Uncertainty in Artificial Intelligence* 437–444. AUAI Press, Arlington, VA.

SILVA, R. and GHAHRAMANI, Z. (2009). The hidden life of latent variables: Bayesian learning with mixed graph models. *J. Mach. Learn. Res.* **10** 1187–1238. MR2520804

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation*, *Prediction*, *and Search*. *Lecture Notes in Statistics* **81**. Springer, New York. MR1227558

TIAN, J. and PEARL, J. (2002). A general identification condition for causal effects. In *Proceedings of the* 18*th National Conference on Artificial Intelligence*. AAAI Press, Menlo Park.

WERMUTH, N. (2011). Probability distributions with summary graph structure. *Bernoulli* **17** 845–879. MR2817608

DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
1–2 SOUTH PARKS ROAD
OXFORD OX1 3TG
UNITED KINGDOM
E-MAIL: robin.evans@stats.ox.ac.uk

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE WASHINGTON 98195
USA
E-MAIL: thomasr@u.washington.edu