Submitted to the Annals of Statistics

MARGINS OF DISCRETE BAYESIAN NETWORKS

By Robin J. Evans

University of Oxford

Bayesian network models with latent variables are widely used in statistics and machine learning. In this paper we provide a complete algebraic characterization of these models when the observed variables are discrete and no assumption is made about the state-space of the latent variables. We show that it is algebraically equivalent to the so-called nested Markov model, meaning that the two are the same up to inequality constraints on the joint probabilities. In particular these two models have the same dimension, differing only by inequality constraints for which there is no general description. The nested Markov model is therefore the closest possible description of the latent variable model that avoids consideration of inequalities. A consequence of this is that the constraint finding algorithm of Tian and Pearl [2002] is complete for finding equality constraints.

Latent variable models suffer from difficulties of unidentifiable parameters and non-regular asymptotics; in contrast the nested Markov model is fully identifiable, represents a curved exponential family of known dimension, and can easily be fitted using an explicit parameterization.

1. Introduction.

Directed acyclic graph (DAG) models, also known as Bayesian network models, are widely used multivariate models in probabilistic reasoning, machine learning and causal inference [Bishop, 2007, Darwiche, 2009, Pearl, 2009]. These models are defined by simple factorizations of the joint distribution, and in the case of discrete or jointly Gaussian random variables, are curved exponential families of known dimension. The inclusion of latent variables within Bayesian network models can greatly increase their flexibility, and also account for unobserved confounding. However, this flexibility comes at the cost of creating models that are not easy to explicitly describe when considered as marginal models over the observed variables. Latent variable models generally do not have fully identifiable parameterizations [Allman et al., 2009], and contain 'singularities' that lead to non-regular asymptotics [Drton, 2009]. In addition, using them may force a modeller to specify a parametric structure over the latent variables, introducing additional assumptions that are generally difficult to test and may

Primary: graphical models, Bayesian networks; secondary: algebraic statistics

Keywords and phrases: algebraic statistics, Bayesian network, latent variable model, nested Markov model, Verma constraint

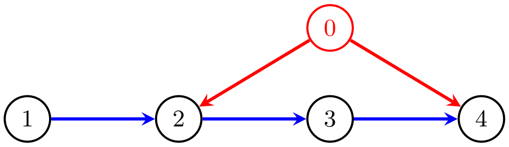


Fig 1. *A directed acyclic graph on five vertices.*

be unreasonable.

In order to avoid potentially erroneous assumptions about the parametric structure of the latent variables or their state-space, we can use an implicitly defined *marginal model*. However, no explicit characterization of this model is available, nor is there any obvious method for fitting it to data.

Example 1.1. Consider the DAG on five vertices shown in Figure 1. The graph represents a multivariate model over five random variables X_0, X_1, X_2, X_3 and X_4, with the restriction that the joint density factorizes as

$$p(x_0, x_1, x_2, x_3, x_4) = p(x_0) \cdot p(x_1) \cdot p(x_2 \,|\, x_0, x_1) \cdot p(x_3 \,|\, x_2) \cdot p(x_4 \,|\, x_0, x_3);$$

here, for example, $p(x_3 \,|\, x_2)$ represents the conditional density of X_3 given X_2. This model arises naturally in the context of dynamic treatment regimes and longitudinal exposures [Robins, 1986]: X_1 and X_3 represent treatments and X_2 and X_4 some outcome of interest. The treatments are randomized, though the second treatment X_3 may depend upon the first outcome X_2, for example a dose may be dynamically adjusted. Since the outcomes are measured on the same patient, they are assumed to be correlated due to a common cause X_0, which might represent an underlying health status, as well as genetic and lifestyle factors.

If we treat X_0 as a latent variable, the *marginal model* over the remaining observed variables (X_1, X_2, X_3, X_4) is the collection of probability distributions that can be written in the form

$$\begin{aligned} &p(x_1, x_2, x_3, x_4) \\ &\quad = \int_{\mathfrak{X}_0} p(x_0) \cdot p(x_1) \cdot p(x_2 \,|\, x_0, x_1) \cdot p(x_3 \,|\, x_2) \cdot p(x_4 \,|\, x_0, x_3) \, dx_0. \end{aligned} \tag{1}$$

That is, the model consists of any (X_1, X_2, X_3, X_4)-margin of a distribution which factorizes according to the DAG over all five variables, for any state-space or distribution[1] of X_0. From (1) we can deduce that the conditional

[1] In general it is sufficient to assume hidden variables are uniform on $(0, 1)$ [see, for

independence $X_3 \perp\!\!\!\perp X_1 \,|\, X_2$ holds in the marginal model; i.e.

$$p(x_3 \,|\, x_1, x_2) = p(x_3 \,|\, x_2). \tag{2}$$

In addition this model satisfies the so-called *Verma constraint*, originally due to Robins [1986] [see also Verma and Pearl, 1990], because the expression

$$q(x_4 \,|\, x_3) \equiv \sum_{x_2} p(x_2 \,|\, x_1) \cdot p(x_4 \,|\, x_1, x_2, x_3) \tag{3}$$

does not depend upon x_1 (see Example 3.2).

The set of distributions satisfying both (2) and (3) is a so-called *nested Markov model* [Richardson et al., 2017]. If the four observed variables are binary these equations represent four independent constraints, and the nested model is an 11-dimensional subset of the 15-dimensional probability simplex.

It is not immediately clear whether or not this nested model is the same as the marginal model defined by (1): in principle the marginal model might impose additional restrictions beyond (2) and (3). This begs the question, is the set of distributions that satisfy (1) characterized by (2) and (3)?

The answer turns out to be 'almost', in the sense that the set of distributions that can be written in the form (1) is a full-dimensional subset of the set that satisfy (2) and (3), though there are additional inequality constraints. This situation is represented by Figure 2, which shows the marginal model ($\mathcal{M}$, in blue) lying strictly within the nested model ($\mathcal{N}$, in red), but the two having the same dimension.

This paper shows that this near-equivalence between the marginal and nested models holds generally for all graphs of this kind. Nested models in general are defined by conditional independences such as (2), and Verma-type constraints such as (3). These latter constraints may always be interpreted as a conditional independence that holds under a different experimental regime to the one observed: in the example above it implies that if we intervene to perform an experiment that sets $\{X_1 = x_1, X_3 = x_3\}$ then the resulting distribution of the final outcome X_4 does not causally depend upon the value of the first treatment, x_1.

1.1. *Other Approaches.*

Alternative approaches to the problem of describing Bayesian network models with hidden variables either make use of

example, Evans, 2016]; for this particular graph, it is a consequence of Theorem 4.7 that we can choose X_0 to be finite and discrete without loss of generality provided it has a sufficiently large number of states.

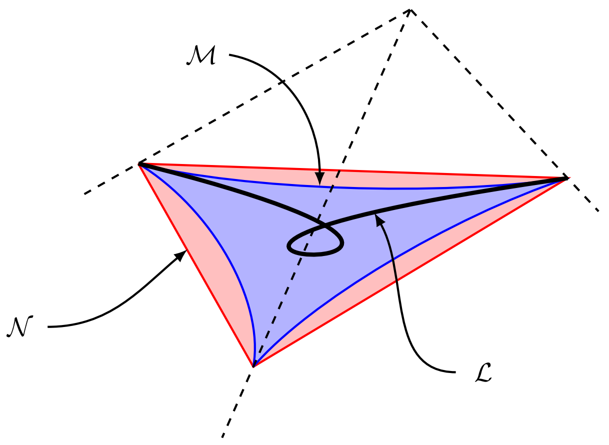


Fig 2. *Diagrammatic representation of the probability simplex (dashed outline) and a marginal model ($\mathcal{M}$, in blue) sitting strictly within the associated nested model ($\mathcal{N}$, in red); note the two models have the same dimension. The boundary of $\mathcal{N}$ is in the simplex boundary, while that of $\mathcal{M}$ is given by inequalities which are generally unknown. Any parametric latent variable model will be contained strictly within $\mathcal{M}$, but it may have a smaller dimension (an example is shown as $\mathcal{L}$). The 'ordinary Markov model' is not shown, but contains $\mathcal{N}$ and would generally have larger dimension.*

parametric structure on the latent variables [for example, Silva and Ghahramani, 2009, Anandkumar et al., 2013], or are restricted to testing conditional independences and do not consider constraints such as (3). This latter category includes the ancestral graph models of Richardson and Spirtes [2002] and the equivalent[2] models on acyclic directed mixed graphs (ADMGs) of Richardson [2003]; these pure conditional independence models, which we refer to as the *ordinary Markov models*, generally have a larger dimension than the observable part of any latent variable model, so using them as a proxy leads to a loss of power to distinguish between certain kinds of model.

On the other hand parametric hidden variable models suffer from various problems caused by the choice of state-space. They may be 'too large', in the sense that the dimension of the parameter space is greater than the dimension of the set of probability distributions in the induced model, thereby introducing identifiability problems. They may also be 'too small', in that unwanted additional restrictions are implied by the parametric structure, and therefore the models have a smaller dimension than the marginal model: this is depicted by the curve labelled $\mathcal{L}$ in Figure 2.

[2] The models are equivalent if selection variables are not present, which is the case throughout this paper.

Paradoxically, it may even be the case that a hidden variable model is 'too large' and 'too small' at the same time! For example, take a latent variable model in Example 1.1 with the simplest possible state-space in which everything is binary: the full model over all five variables has dimension 12; however, we have already established that the dimension of the marginal model over the observed variables is at most 11, so the model is clearly over-parameterized. In fact, it can be shown that the dimension of this latent variable model over the observed variables is only 10, so an additional—and perhaps unwelcome—restriction is present due to the choice of a binary latent variable model [see Appendix A, Evans, 2017].

If X_0 is given enough states, the latent variable model and the marginal model coincide for graphs such as the one in Figure 1, a fact we will exploit in our proofs. However, such a latent variable model is less useful for statistical inference because it is generally massively over-parameterized. See Example 6.2 for a demonstration of this.

None of this should be construed as suggesting that the marginal model supersedes all latent variable models, since sometimes the additional parametric assumptions made in a latent variable model are crucial to their utility. For example, hidden Markov models and phylogenetic tree models are important and widely used latent variable models, but their corresponding marginal models are saturated. Using the marginal model would therefore be statistically uninteresting and likely lead only to trivial inferences. However, as noted above for Example 1.1 and as we will see again in Example 6.2, in some examples marginal models are more suitable than any latent variable model. Marginal models are also of interest in the Quantum Information literature, because they enable comparison between 'classical' latent variable models and the more general quantum entangled states [Henson et al., 2014]. We discuss the implications of our results for quantum models in Section 6.2.

1.2. *A Short Algebra Tutorial.*

This paper makes use of some results from real algebraic geometry, which provides powerful tools for analysing these complicated sets of distributions. All our statistical models are collections of distributions within the probability simplex that satisfy certain constraints. The constraints on a Bayesian network model are conditional independences, and can be represented as the requirement that certain polynomials in the probabilities are equal to zero; for example the conditional independence $X_1 \perp\!\!\!\perp X_3 \mid X_2$ is equivalent to

$$p(x_2) \cdot p(x_1, x_2, x_3) - p(x_1, x_2) \cdot p(x_2, x_3) = 0 \qquad \forall x_1, x_2, x_3.$$

The set of points at which a collection of polynomials are all zero is called an *algebraic variety*, or sometimes an *algebraic set*. This perspective is explored in depth for Bayesian network models by Garcia et al. [2005]. In addition to equality constraints, these models will satisfy polynomial inequalities; i.e. $p(x_V) \geq 0$. A set defined by a combination of polynomial equalities and inequalities is said to be *semi-algebraic*; this category includes many common finite-dimensional statistical models. Semi-algebraic sets have the nice property that their images are semi-algebraic under any polynomial map, which includes elimination of variables or projection onto a linear subspace. A consequence of this is that the margin of any model defined by a semi-algebraic set is also defined by a semi-algebraic set.

The *Zariski closure* of a set is the smallest algebraic variety that contains it; the fact that this is well-defined is a significant result in algebraic geometry. For a semi-algebraic set one can informally think of its Zariski closure as the set obtained by keeping the equality constraints and 'throwing away' the inequality constraints. Semi-algebraic sets have many interesting properties, but they are not necessarily 'nice' from a statistical perspective, in the sense of leading to regular asymptotics. For this we need our set to be a *manifold*, i.e. to be locally Euclidean.

1.3. *Contribution.* In this paper we show that marginal models with finite discrete observed variables are algebraically equivalent to the appropriate nested Markov model, in the sense that the Zariski closures of the marginal model and the nested model are the same. A consequence of this is that a margin of a DAG model and its nested counterpart have the same dimension, and differ only by inequality constraints. The marginal model defined by (1) in Example 1.1 is indeed 11-dimensional, and is algebraically defined by (2) and (3); however, the marginal model also satisfies polynomial inequality constraints that the nested model does not. The result can be interpreted as showing that the constraint finding algorithm of Tian and Pearl [2002] is 'complete', in the sense that no other equality constraints are necessary to describe the marginal model.

THEOREM 1.2. *Let $\mathcal{G}$ be a Bayesian network model with vertices $V \cup H$, where X_V are discrete random variables and X_H have an arbitrary state-space. The resulting model over the margin of X_V has the same Zariski closure as the set of distributions satisfying the constraints listed in Tian and Pearl [2002].*

This means that we have, for the first time, a full algebraic characterization of margins of Bayesian network models. It also shows that the nested model represents a sensible and pragmatic approximation to the marginal model: we currently have no way to derive inequality constraints efficiently, so the nested model—which has a factorization criterion, separation criteria, and a discrete parameterization [Richardson et al., 2017]—is much easier to work with, and can easily be fitted with existing algorithms [Evans and Richardson, 2010]. In addition, the nested model inside the probability simplex is a manifold and therefore regular whenever the joint distribution is positive, whereas the marginal model may have a boundary that lies strictly inside the simplex. The nested model therefore has better statistical properties than the marginal model, in the sense that data generated from any strictly positive distribution will lead to regular asymptotics.

Causal discovery methods such as the FCI algorithm [Spirtes et al., 2000] that use conditional independence constraints could, in principle, be extended to use the constraints implied by nested models; our main result shows that is 'as good as it gets', in the sense that there are no other equality constraints to test without making further (e.g. parametric) assumptions. Thus, this paper probes the limits of what it is possible to learn about causal models with hidden variables from observational data when we have no further knowledge about the latent state-space.

We work with a class of hyper-graphs called mDAGs, with which we associate marginals of DAG models [Evans, 2016]. That paper also shows that these mDAGs are a sufficiently rich class of graphs as to represent all the marginal models we will consider. Nested models are introduced in detail by Richardson et al. [2017], and a parameterization of them in the discrete case given by Evans and Richardson [2015]. The existence of this parameterization will allow us to prove our main results.

The remainder of the paper is organized as follows: Section 2 reviews DAG models, their margins and mDAGs, and carefully defines the problem of interest. Section 3 defines the nested Markov model, and recalls relevant properties from Richardson et al. [2017] and Evans and Richardson [2015]. The remaining sections contain entirely new material: in Section 4 we introduce latent variable models with specific state-spaces, and show that they can be used to represent some marginal models without loss of generality; Section 5 contains the main results of the paper, including the proof of Theorem 1.2. Finally, in Section 6 we show that a large class of marginal models represent smooth manifolds, and provide some discussion.

2. Directed Graphical Models.

We begin with some elementary graphical definitions.

DEFINITION 2.1. A *directed graph*, $\mathcal{G}(V, \mathcal{E})$, consists of a finite set of

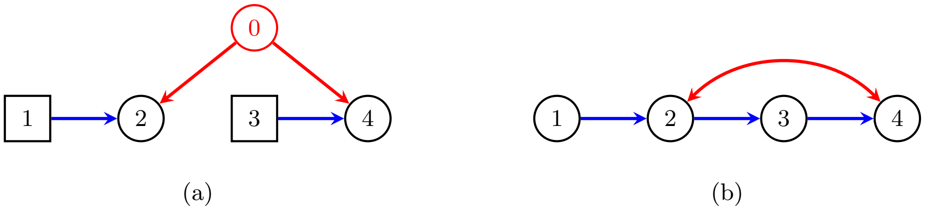


FIG 3. *(a) A conditional directed acyclic graph with three random vertices (0,2,4) and two fixed vertices (1,3). (b) An mDAG representing the DAG in (a), with the vertex 0 treated as unobserved.*

vertices, V, and a collection of edges, $\mathcal{E}$, which are ordered pairs of distinct elements of V. If $(v, w) \in \mathcal{E}$ we denote this by $v \to w$, and say that v is a *parent* of w; the set of parents of w is denoted by $\mathrm{pa}_{\mathcal{G}}(w)$. Similarly w is a *child* of v, and the child set is denoted by $\mathrm{ch}_{\mathcal{G}}(v)$.

A directed graph is *acyclic* if there is no sequence of edges $v_1 \to v_2 \to \cdots \to v_k \to v_1$ for $k > 1$. We call such a graph a *directed acyclic graph*, or DAG.

Graphs are best understood visually: an example of a DAG with five vertices and five edges is given in Figure 1. We will require the following generalization of a DAG that allows for two separate types of vertex.

DEFINITION 2.2. A *conditional DAG* $\mathcal{G}(V, W, \mathcal{E})$ is a DAG with vertices[3] $V \dot{\cup} W$ and edge set $\mathcal{E}$, with the restriction that no vertex in W may have any parents. The elements of V are the *random vertices*, and W the *fixed vertices*; these two sets are disjoint.

If $W = \emptyset$, this reduces to the ordinary definition of a DAG. We depict fixed vertices with square nodes, and random ones with round nodes: see the example in Figure 3(a).

2.1. *Graphical Models.* A graphical model arises from the identification of a graph with a collection of multivariate probability distributions; see Lauritzen [1996] for an introduction. Each vertex $v \in V$ represents a random variable X_v taking values in a finite state-space $\mathfrak{X}_v$, and a model for their joint distribution is determined by the structure of the graph. With a conditional DAG $\mathcal{G}$ we associate a collection of probability measures $P(\cdot \,|\, x_W)$ on $\mathfrak{X}_V \equiv \times_{v \in V} \mathfrak{X}_v$, indexed by $x_W \in \mathfrak{X}_W$. Mathematically, fixed nodes play a similar role to the 'parameter nodes' used by Dawid [2002].

[3] Here and throughout $\dot{\cup}$ denotes a disjoint union of sets.

Following Lauritzen [1996], we say a *probability kernel* over $\mathfrak{X}_A$ given $\mathfrak{X}_B$ is a non-negative function $q : \mathfrak{X}_A \times \mathfrak{X}_B \to \mathbb{R}$ such that $\sum_{x_A} q(x_A \,|\, x_B) = 1$ for all $x_B \in \mathfrak{X}_B$. A kernel behaves much like a conditional probability distribution, but no assumption is made about any distribution over the indexing set $\mathfrak{X}_B$. We apply the usual definitions for marginalizing and conditioning in kernels:

$$q(x_A \,|\, x_B) \equiv \sum_{x_C} q(x_A, x_C \,|\, x_B), \qquad q(x_A \,|\, x_B, x_C) \equiv \frac{q(x_A, x_C \,|\, x_B)}{q(x_C \,|\, x_B)}.$$

If $q(x_A \,|\, x_B, x_C)$ does not depend upon x_B then we will denote it $q(x_A \,|\, x_C)$, and say that $X_A \perp\!\!\!\perp X_B \mid X_C \; [q]$. Here, and elsewhere, we use the shorthand VW for $V \cup W$ in subscripts.

Definition 2.3. Let $p(x_V \,|\, x_W)$ be a probability kernel over $\mathfrak{X}_V$ indexed by $\mathfrak{X}_W$. We say that p obeys the *factorization criterion* with respect to a DAG $\mathcal{G}$ if it factorizes into univariate kernels as

$$p(x_V \,|\, x_W) = \prod_{v \in V} p(x_v \,|\, x_{\mathrm{pa}(v)}), \qquad x_{VW} \in \mathfrak{X}_{VW}. \tag{4}$$

Note that if $\mathcal{G}(V \cup W, \mathcal{E})$ is a causally interpreted DAG, then (4) gives the usual formula for the distribution $p(x_V \,|\, do(x_W))$, the distribution of X_V after intervening to set $X_W = x_W$.

The definition reduces to the familiar factorization criterion for DAGs if $W = \emptyset$. The extra generality will be useful for discussing Markov properties which involve factorization of the distribution into conditional pieces. The fixed vertices are analogous to variables that have been conditioned upon.

A Bayesian network model can also be defined by insisting that each random variable X_v can be written as a measurable function of $X_{\mathrm{pa}(v)}$ and an independent noise variable; we call this the *structural equation property*; for discrete variables in particular, these two criteria are equivalent. Although the factorization property is often simpler to work with for practical purposes such as modelling and fitting, the structural equation property is useful in proofs.

Example 2.4. A distribution P with density p obeys the factorization criterion for the graph in Figure 1 if the density has the form

$$p(x_0, x_1, x_2, x_3, x_4) = p(x_0) \cdot p(x_1) \cdot p(x_2 \,|\, x_0, x_1) \cdot p(x_3 \,|\, x_2) \cdot p(x_4 \,|\, x_0, x_3).$$

Such distributions are precisely those which satisfy the conditional independences

$$X_1 \perp\!\!\!\perp X_0, \qquad X_3 \perp\!\!\!\perp X_0, X_1 \,|\, X_2, \qquad X_4 \perp\!\!\!\perp X_1, X_2 \,|\, X_0, X_3.$$

EXAMPLE 2.5. A kernel p obeys the factorization criterion for the conditional DAG in Figure 3(a) if it can be written as

$$p(x_0, x_2, x_4 \,|\, x_1, x_3) = p(x_0) \cdot p(x_2 \,|\, x_0, x_1) \cdot p(x_4 \,|\, x_0, x_3).$$

2.2. *Latent Variables and mDAGs.* We now introduce the possibility that some of the random variables are unobserved or *latent*, leaving the marginal distribution over the remaining *observed* variables. We represent the collection of margins of DAG models using a larger class of hyper-graphs called mDAGs ('marginal DAGs'). These avoid dealing with latent variables directly, by instead introducing additional edges to represent them. For example, the DAG in Figure 1, with the vertex 0 treated as a latent variable, is represented by the mDAG in Figure 3(b).

Define an *abstract simplicial complex* $\mathcal{B}$ over V as a collection of nonempty subsets of V such that (i) $\{v\} \in \mathcal{B}$ for every $v \in V$, and (ii) if $A \in \mathcal{B}$ and $B \subseteq A$ with $B \neq \emptyset$, then $B \in \mathcal{B}$.

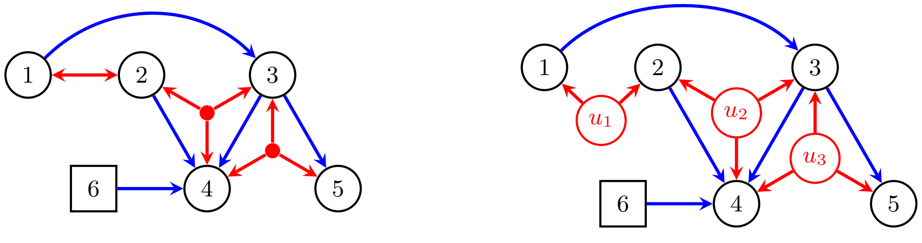
DEFINITION 2.6. An *mDAG*, $\mathcal{G}(V, W, \mathcal{E}, \mathcal{B})$, is a hyper-graph consisting of a conditional DAG with random vertices V, fixed vertices W and directed edge set $\mathcal{E}$, together with an abstract simplicial complex $\mathcal{B}$ over V, called the *bidirected faces*.

We say that $\mathcal{G}'(V', W', \mathcal{E}', \mathcal{B}')$ is a *subgraph* of $\mathcal{G}$ if $V' \subseteq V$, $\mathcal{E}' \subseteq \mathcal{E}$, $\mathcal{B}' \subseteq \mathcal{B}$, and $W' \subseteq V \cup W$: that is, each component is contained within the previous one, but random vertices may become fixed.

The mDAG was introduced by Evans [2016], without the additional generality of fixed vertices. This aspect changes very little about the theory of these graphs, but is necessary for understanding the nested Markov model; note that bidirected faces only involve the random vertices. As with conditional DAGs, when representing mDAGs graphically the fixed vertices are drawn as square nodes and random vertices as circles.

The bidirected simplicial complex is represented by its maximal nontrivial elements (i.e. those of size at least 2), called the *bidirected hyperedges*, or just *edges*. These are drawn in red, as in Figure 4(a); in this case $W = \{6\}$ and the maximal sets of $\mathcal{B}$ are $\{1, 2\}$, $\{2, 3, 4\}$, and $\{3, 4, 5\}$.

With each mDAG, $\mathcal{G}$, we can associate a conditional DAG $\bar{\mathcal{G}}$ by replacing each maximal element $B \in \mathcal{B}$ (of size at least 2) with a new random vertex u, such that the children of u are precisely the vertices in B. The new vertex u becomes the 'unobserved' variable represented by the bidirected edge B. We call $\bar{\mathcal{G}}$ the *canonical DAG* associated with $\mathcal{G}$. The mDAG in Figure 4(a) is thus associated with the canonical DAG in Figure 4(b).



(a) (b)

FIG 4. *(a) An mDAG, $\mathcal{G}$, and (b) a DAG with hidden variables, $\bar{\mathcal{G}}$, representing the same model (the canonical DAG).*

Our interest in mDAGs lies in their representation of the margin of the associated canonical DAG, and so we define our model in this spirit. From the definitions it may seem as though the set of models is restricted to cases where the latent variables have no parents; in fact this does not cause any loss of generality since—if we make no assumption about state-space of the latents—all marginal DAG models can be represented in this way [see Evans, 2016, Theorem 2].

DEFINITION 2.7. Let $\mathcal{G}$ be an mDAG with vertices $V \dot{\cup} W$, and let $\bar{\mathcal{G}}$ be the canonical DAG with vertices $V \dot{\cup} U \dot{\cup} W$. A kernel p over $\mathfrak{X}_V$ indexed by $\mathfrak{X}_W$ is said to be in the *marginal model* for $\mathcal{G}$ if there exists a kernel q that factorizes according to $\bar{\mathcal{G}}$, and

$$p(x_V \,|\, x_W) = \int_{\mathfrak{X}_U} q(x_V,\, x_U \,|\, x_W)\, dx_U.$$

That is, the margin of q over X_V is p. Denote the collection of such kernels by $\mathcal{M}(\mathcal{G})$.

In other words, the marginal model is the collection of kernels that could be constructed as the margin of a Bayesian network with latent variables replacing the bidirected edges. If $\mathcal{G}$ is a DAG then the marginal model is just the usual model defined by the factorization.

A latent variable model corresponding to a canonical DAG $\bar{\mathcal{G}}$ (i.e. possibly with parametric or distributional assumptions on the latent variables) always lies within the marginal model corresponding to the mDAG $\mathcal{G}$.

2.3. *Districts and Sterile Vertices.*

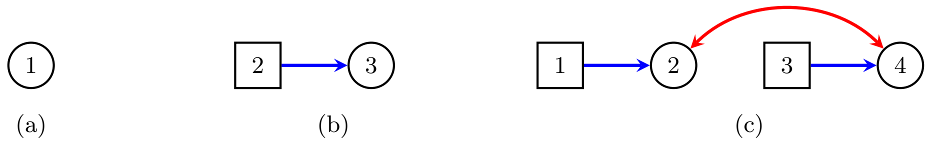


FIG 5. *Subgraphs corresponding to factorization of the graph in Figure 3(b) into districts. Parent nodes of the district are drawn as squares.*

DEFINITION 2.8. A collection of random vertices $C \subseteq V$ in an mDAG $\mathcal{G}$ is *bidirected-connected* if for any distinct $v, w \in C$, there is a sequence of vertices $v = v_0, v_1, \ldots, v_k = w$ all in C such that, for each $i = 1, \ldots, k$, the pair $\{v_{i-1}, v_i\} \in \mathcal{B}$. A *district* of an mDAG is an inclusion maximal bidirected-connected set of random vertices.

More informally, a district is a maximal set of random vertices joined by the red edges in an mDAG. It is easy to see from the definition that districts form a partition of the random vertices in an mDAG. The mDAG in Figure 3(b), for example, contains three districts, $\{1\}$, $\{3\}$ and $\{2, 4\}$. Districts inspire a useful reduction of mDAGs, via the following special subgraph.

DEFINITION 2.9. Let $\mathcal{G}$ be an mDAG containing random vertices $C \subseteq V$. Then $\mathcal{G}[C]$ is the subgraph of $\mathcal{G}$ with

(i) random vertices C and fixed vertices $\mathrm{pa}_{\mathcal{G}}(C) \setminus C$;
(ii) those directed edges $w \to v$ such that $v \in C$ (and $w \in \mathrm{pa}_{\mathcal{G}}(C)$);
(iii) the bidirected simplicial complex $\mathcal{B}_C \equiv \{B \cap C : B \in \mathcal{B}(\mathcal{G})\}$.

$\mathcal{G}[C]$ is therefore the subgraph induced over C, together with parents of C and edges directed towards C. Any edges (whether directed or bidirected) between the newly fixed vertices are removed.

For the graph in Figure 3(b) the subgraphs $\mathcal{G}[\{1\}]$, $\mathcal{G}[\{3\}]$ and $\mathcal{G}[\{2, 4\}]$ are shown in Figures 5(a), (b) and (c) respectively. Note in particular that the edge $2 \to 3$ is not in the subgraph $\mathcal{G}[\{2, 4\}]$.

DEFINITION 2.10. Let $\mathcal{G}$ be an mDAG with random vertices V. For an arbitrary set $C \subseteq V$, define $\mathrm{sterile}_{\mathcal{G}}(C) \equiv C \setminus \mathrm{pa}_{\mathcal{G}}(C)$. In words $\mathrm{sterile}_{\mathcal{G}}(C)$ is the subset of C whose elements have no children in C. We say a set C is *sterile* if $C = \mathrm{sterile}_{\mathcal{G}}(C)$.

3. Nested Markov Property.

The nested Markov property imposes constraints on a joint distribution that mimic those satisfied by the marginal model, including conditional independences and the Verma constraint in Example 1.1 [Richardson et al., 2017]. It is defined in the following recursive way, which is a modification of the algorithm of Tian and Pearl [2002].

DEFINITION 3.1 (Nested Markov Property). A kernel p over $\mathfrak{X}_V$ indexed by $\mathfrak{X}_W$ obeys the *nested Markov property* for an mDAG $\mathcal{G}(V, W)$ if $V = \emptyset$, or both:

1. p factorizes over the districts $D_1, \ldots, D_l$ of $\mathcal{G}$:
$$p(x_V \,|\, x_W) = \prod_{i=1}^{l} g_i(x_{D_i} \,|\, x_{\mathrm{pa}(D_i) \setminus D_i})$$
where each g_i is a kernel which (if $l \geq 2$ or $W \setminus \mathrm{pa}_{\mathcal{G}}(V) \neq \emptyset$) obeys the nested Markov property with respect to $\mathcal{G}[D_i]$; and
2. for each $v \in V$ such that $\mathrm{ch}_{\mathcal{G}}(v) = \emptyset$, the marginal kernel
$$p(x_{V \setminus v} \,|\, x_W) = \sum_{x_v} p(x_V \,|\, x_W)$$
obeys the nested Markov property with respect to $\mathcal{G}[V \setminus \{v\}]$.

The set of kernels that obey the nested Markov property for $\mathcal{G}$ is the *nested Markov model*, denoted by $\mathcal{N}(\mathcal{G})$.

The condition that $l \geq 2$ or $W \setminus \mathrm{pa}_{\mathcal{G}}(V) \neq \emptyset$ in the first criterion of this definition is simply to prevent an infinite recursion of the definition: all the graphs invoked recursively have either fewer random vertices or fewer vertices overall than their predecessor in the recursion. When we reach a graph with a single random vertex v such that all fixed vertices are parents of v, then any kernel $p(x_v \,|\, x_{\mathrm{pa}(v)})$ satisfies the nested Markov property.

The discrete nested model is equivalently defined by the constraints above and the parameterization in Evans and Richardson [2015] (as well the nested Markov properties described in Richardson et al. [2017]). We will make use of these equivalent definitions throughout.

EXAMPLE 3.2. Consider again the mDAG in Figure 3(b). Applying criterion 1 to this graph implies that
$$p(x_1, x_2, x_3, x_4) = g_1(x_1) \cdot g_{24}(x_2, x_4 \,|\, x_1, x_3) \cdot g_3(x_3 \,|\, x_2)$$
for some g_1, g_3 and g_{24} obeying the nested Markov property with respect to the mDAGs in Figures 5(a), (b) and (c) respectively. Applying the second

criterion to g_{24} and the now childless vertex 2 (see Figure 5(c)) gives

$$\sum_{x_2} g_{24}(x_2, x_4 \,|\, x_1, x_3) = h(x_4 \,|\, x_3),$$

for some function h independent of x_1 (by a further application of the first criterion); this is precisely the Verma constraint.

The marginal model implies additional conditions on joint distributions because, although it satisfies the properties used to define the nested model, these properties are not sufficient to describe it. In particular, for p to be in the marginal model, the kernel g_{24} must satisfy Bell's inequalities [see, for example, ver Steeg and Galstyan, 2011, Section 4.1].

The nested Markov property is 'sound' with respect to marginal models, in the sense that all constraints represented by the former also hold in the latter. The following theorem is a consequence of the results in Tian and Pearl [2002].

THEOREM 3.3. *For any mDAG $\mathcal{G}$ we have $\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G})$.*

3.1. *Parameterizing Sets.*

DEFINITION 3.4. Let $\mathcal{G}$ be an mDAG. A subset of random vertices $S \subseteq V$ is called *intrinsic* if S is a district in any graph which can be obtained by iteratively applying graphical operations of the form 1 and 2 in Definition 3.1 (i.e. taking the graph $\mathcal{G}[D]$ for a district D, or $\mathcal{G}[V \setminus \{v\}]$ for a sterile vertex v).

Given an intrinsic set, S, define $H = \text{sterile}_{\mathcal{G}}(S)$ to be the *recursive head*, and $T = \text{pa}_{\mathcal{G}}(S)$ the *tail*, associated with S (note that H and T are disjoint). The collection of all recursive heads in $\mathcal{G}$ is denoted by $\mathcal{H}(\mathcal{G})$. There is a one-to-one correspondence between intrinsic sets and recursive heads [Evans and Richardson, 2015]. Throughout we will use H and T to indicate recursive heads and tails respectively, with the context making it clear which intrinsic set is being referred to. We will sometimes write $T(H)$ to make clear that the head determines the tail.

The definitions above also appear in Evans and Richardson [2015]. We introduce a new definition: let

$$\mathcal{A}(\mathcal{G}) \equiv \{H \cup A \,|\, H \in \mathcal{H}(\mathcal{G}), A \subseteq T(H)\}$$

be the *parameterizing sets* of $\mathcal{G}$. This collection of sets is so-called because it (locally) describes the set of distributions (or kernels) contained in the nested and marginal models, as we will prove in Section 5.

Example 3.5. The mDAG in Figure 3(b) has districts $\{1\}$, $\{3\}$ and $\{2, 4\}$, so these are all intrinsic sets. Further, in the subgraph $\mathcal{G}[\{2, 4\}]$ the vertices 2 and 4 have no children, so we can marginalize either to see that respectively $\{4\}$ and $\{2\}$ are intrinsic sets. The corresponding recursive heads and tails are then:

S	H	T	$\mathcal{A}$
$\{1\}$	$\{1\}$	$\emptyset$	$\{1\}$
$\{2\}$	$\{2\}$	$\{1\}$	$\{2\}$, $\{1,2\}$
$\{3\}$	$\{3\}$	$\{2\}$	$\{3\}$, $\{2,3\}$
$\{4\}$	$\{4\}$	$\{3\}$	$\{4\}$, $\{3,4\}$
$\{2,4\}$	$\{2,4\}$	$\{1,3\}$	$\{2,4\}$, $\{1,2,4\}$, $\{2,3,4\}$, $\{1,2,3,4\}$

.

Note that every non-empty subset of V is represented in $\mathcal{A}$ except for $\{1, 3\}$, $\{1, 2, 3\}$, $\{1, 4\}$ and $\{1, 3, 4\}$. The first two of these correspond to the conditional independence $X_1 \perp\!\!\!\perp X_3 \mid X_2$ in (2), and the others to the Verma constraint (3).

We use the $\triangle$ operator to denote the symmetric difference of two sets: $A \triangle B \equiv (A \setminus B) \cup (B \setminus A)$. Given a finite collection A_i, $i = 1, \dots, k$, let

$$\mathop{\triangle}_{i=1}^{k} A_i \equiv A_1 \triangle A_2 \triangle \cdots \triangle A_k.$$

denote the symmetric difference of all the A_i. That is, it is the set containing precisely those elements a which appear in an odd number of the sets A_i.

The following result gives a characterization of the parameterizing sets in terms of symmetric differences which will be fundamental to our proof of the main results in this paper.

Lemma 3.6. *A set $A \in \mathcal{A}(\mathcal{G})$ if and only if there exists a bidirected-connected set $C = \{v_1, \dots, v_k\}$ in $\mathcal{G}$, and sets A_i, $i = 1, \dots, k$, satisfying $\{v_i\} \subseteq A_i \subseteq \{v_i\} \cup \mathrm{pa}_{\mathcal{G}}(v_i)$, such that*

$$A = \mathop{\triangle}_{i=1}^{k} A_i = A_1 \triangle \cdots \triangle A_k. \tag{5}$$

The proof is found in Section B.1 of the supplement [Evans, 2017].

3.2. *Parameterization of the nested model.* The nested Markov model can be parameterized with parameters indexed by head-tail sets [Evans and

Richardson, 2015], and the parameterization defines a smooth bijection between an open subset of a real vector space (i.e. the parameter space) and the model (the set of probability distributions). This has some nice consequences that we now state [for proofs see Evans and Richardson, 2015].

In particular, for a fixed state-space $\mathfrak{X}_{VW}$ the set $\mathcal{N}(\mathcal{G})$ is a smooth manifold within the strictly positive probability simplex, and has dimension[4]

$$d(\mathcal{G}, \mathfrak{X}_{VW}) \equiv \sum_{H \in \mathcal{H}(\mathcal{G})} |\mathfrak{X}_{T(H)}| \prod_{h \in H} (|\mathfrak{X}_h| - 1).$$

In the all-binary case this reduces to $d(\mathcal{G}, \mathfrak{X}_{VW}) \equiv \sum_{H \in \mathcal{H}(\mathcal{G})} 2^{|T(H)|}$. Our main result will show that $\mathcal{M}(\mathcal{G})$ always has the same dimension as $\mathcal{N}(\mathcal{G})$. Indeed, the parameterization of $\mathcal{N}(\mathcal{G})$ will in principle also serve as a parameterization of $\mathcal{M}(\mathcal{G})$, except that one would also have to restrict the parameter space in order to enforce the inequality constraints; of course, this is currently impractical since the inequality constraints are not generally known.

3.3. *Relationship between mDAGs and ADMGs.* Previous papers considering marginal and nested models for DAGs have used *acyclic directed mixed graphs*, which are the restriction of mDAGs with random vertices so that each bidirected edge has size two [Richardson, 2003, Evans and Richardson, 2014, Richardson et al., 2017]. From the perspective of the nested Markov model this distinction is unimportant: if we replace any bidirected simplicial complex with all its subsets of size 2, we obtain a conditional ADMG that represents the same model under the nested Markov property.

It therefore follows from Theorem 1.2 that there is no difference in equality constraints between graphs that differ only in this manner; *algebraically* the model defined by having a single latent parent for several variables is the same as having separate parents for each pair of vertices. Note that the marginal models are not always equal, as the restriction to pairwise independent latent parents will sometimes introduce additional inequality constraints. See Evans [2016] for a more detailed discussion.

4. Geared mDAGs.

In this section we introduce a special class of mDAGs which we term ‘geared’. For marginal models relating to such graphs, the state-space of the hidden vertices can be restricted without loss of generality, making proofs considerably easier. In Section 5 we prove our main result first for geared graphs, and then extend the result to the general case.

[4] Note that we use the convention that $|\mathfrak{X}_\emptyset| = 1$.

Definition 4.1. Let $\mathcal{G}$ be an mDAG with bidirected simplicial complex $\mathcal{B}$. We say that $\mathcal{G}$ is *geared* if the maximal elements of $\mathcal{B}$ satisfy the running intersection property. That is, there is an ordering of the edges $B_1, \dots, B_k$ such that for each $j > 1$, there exists $s(j) < j$ with

$$B_j \cap \bigcup_{i<j} B_i = B_j \cap B_{s(j)}.$$

In other words, the vertices that are contained in both B_j and any previous edge are all contained within one such edge $B_{s(j)}$.

A particular ordering of the elements of $\mathcal{B}$ which satisfies running intersection is called a *gearing* of $\mathcal{G}$.[5]

Example 4.2. The simplest non-geared mDAG is the bidirected 3-cycle, which has bidirected edge sets $\{1, 2\}$, $\{2, 3\}$, $\{1, 3\}$. We cannot order these in a way that satisfies the running intersection property, since whichever edge is placed last in the ordering shares a different vertex with each of the two other edges.

The following fact about geared subgraphs of mDAGs will allow us to generalize our later results to graphs which are not geared.

Lemma 4.3. *Let $\mathcal{G}$ be an mDAG with parameterizing sets $\mathcal{A}(\mathcal{G})$. For any $A \in \mathcal{A}(\mathcal{G})$ there exists a geared mDAG $\mathcal{G}' \subseteq \mathcal{G}$, such that $A \in \mathcal{A}(\mathcal{G}')$.*

Proof. By Lemma 3.6, A is of the form (5) for some bidirected-connected set C. Let $\mathcal{G}'$ have the same vertices (random and fixed) and directed edges as $\mathcal{G}$, but be such that the set C is *singly connected* by bidirected edges (i.e. the edges are all of size 2 and removing any of them will cause C to be disconnected) chosen to be a subgraph of $\mathcal{G}$. Then $\mathcal{G}'$ is geared by standard properties of trees and running intersection, and using Lemma 3.6 again we have $A \in \mathcal{A}(\mathcal{G}')$. □

[5] The term 'geared' is chosen because a collection of bidirected edges which satisfies running intersection may appear rather like 'cogs' in a set of gears: see Figure 4. The definition is equivalent to the requirement that the simplicial complex $\mathcal{B}$ is vertex decomposable [Provan and Billera, 1980], and is also closely related to the notion of decomposability in an undirected or directed graph. Indeed the term 'decomposable' is used by Fox et al. [2014] to describe the same idea. We avoid using this terminology because of its existing meaning in connection with undirected and directed graphical models: for example, ordinary DAGs are trivially geared, but they may or may not be decomposable in the original sense [Lauritzen, 1996].

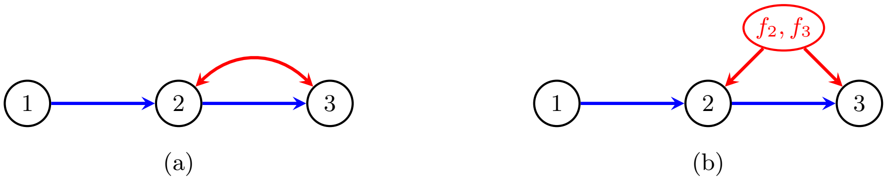
4.1. *Functional Models.* The key property of geared graphical models is that we can find a finite discrete latent variable model that is the same (over the observed variables) as the marginal model; that is, if the latent variables have a sufficiently large state-space then they do not impose additional restrictions on the observed distribution. This is achieved by letting each observed variable be a deterministic function of its latent and observed parents. We illustrate this with an example.

EXAMPLE 4.4. Consider the mDAG in Figure 6(a) representing the *instrumental variables* model, used to model non-compliance in clinical trials; here, for example, X_1 represents a randomized treatment, X_2 the treatment actually taken, and X_3 a patient's outcome or response, such as survival. Suppose that each of these quantities is binary, taking values in $\{0, 1\}$. Conceptually, it can be useful to posit the existence of two different *potential outcomes* $X_3(0), X_3(1)$ for the survival response, one for each level of the treatment; $X_3(0)$ is the patient's outcome given that they choose not to take the treatment (i.e. when $X_2 = 0$) and $X_3(1)$ is their outcome given that they do ($X_2 = 1$). For example, if $X_3(0) = 0$ and $X_3(1) = 1$ then the patient survives if they take the treatment but dies if they do not. This pair of values is known as a patient's *response type*. Of course, we can only ever observe one of these outcomes in a given patient, the one corresponding to the observed value of X_2.

Similarly, we can conceive of two versions of the treatment $X_2(0), X_2(1)$ depending upon the assigned value of X_1, this pair being called the patient's *compliance type*. For example, $X_2(0) = X_2(1) = 0$ means that the patient will not take the treatment, regardless of whether or not they are assigned to the treatment group. These concepts have proved fruitful in causal inference, as they enable discussion of whether treatments have effects at the level of individual patients, rather than just over the entire population on average [Neyman, 1923, Rubin, 1974, Richardson et al., 2011].

Now, since the latent variable (say U) with children $\{2, 3\}$ can take any value, we can—without loss of generality—assume that it includes the pair $(X_3(0), X_3(1))$, or equivalently a function $f_3 : \mathfrak{X}_2 \to \mathfrak{X}_3$ that determines, given the observed X_2, which value X_3 will take. In this case X_3 is still a measurable function of its parents U and X_2. Similarly we can assume U includes a function $f_2 : \mathfrak{X}_1 \to \mathfrak{X}_2$ that determines X_2 given an observed X_1.

An observation for a particular patient can be obtained by drawing a random treatment assignment X_1, a random compliance type for the patient f_2, and a random response type f_3, and then evaluating $(X_1, X_2, X_3) = (X_1, f_2(X_1), f_3(f_2(X_1)))$. The key point is that one can place a distribu-



(a)

(b)

FIG 6. *(a) An mDAG representing the instrumental variables model; (b) a DAG with functional latent variables equivalent to the potential outcomes model of instrumental variables.*

tion over (X_1, f_2, f_3) and obtain a distribution over the observed variables (X_1, X_2, X_3). The only requirement for the distribution to be Markov with respect to this particular graph is that $X_1 \perp\!\!\!\perp \{f_2, f_3\}$, as depicted in Figure 6(b).

The functional construction outlined above is mathematically equivalent to potential outcomes, and provides a model that is somewhat simpler to study than the general latent variable model. In fact, any geared mDAG can be reduced to a latent variable model in the way described above, something we will proceed to show in Theorem 4.7.

The nested model in the case of Figure 6(b) is saturated, and therefore not particularly interesting. For causal modelling the potential outcomes framework is likely to be substantially more useful in this example. However, it is important to note that potential outcomes do not always give a practical alternative to the nested model; see Example 6.2.

4.2. *Remainder Sets.* Given a single-district, geared mDAG with at least one bidirected edge and a gearing $B_1, \dots, B_k$, define

$$R_j \equiv B_j \setminus \bigcup_{i<j} B_i$$

(taking $R_1 \equiv B_1$) to be the *remainder set* associated with B_j. Remainder sets partition V, so for a random vertex $v \in V$, define $r(v)$ to be the unique j such that $v \in R_j$.

Now say that an ordering $<$ on the vertices in V *respects the gearing* if for $v \in R_i$ and $w \in R_j$, we have $v < w$ whenever $i > j$; in other words, all the vertices in R_k precede all those in R_{k-1}, etc; such an ordering always exists. For each $v \in V$ with $r(v) = j$, let

$$\pi(v) = \bigcup_{\substack{i>j \\ v \in B_i}} R_i;$$

that is, the remainders associated with all bidirected edges which contain v and are later than j in the ordering. Then define a collection of functions

$$\mathcal{F}_v \equiv \{f : \mathfrak{X}_{\mathrm{pa}(v)} \times \mathcal{F}_{\pi(v)} \to \mathfrak{X}_v\},$$

where $\mathcal{F}_A = \times_{a \in A} \mathcal{F}_a$ and $\mathcal{F}_\emptyset = \mathfrak{X}_\emptyset = \{1\}$. This is well-defined, since all the vertices in $\pi(v)$ precede v in an ordering which respects the gearing.

Example 4.5. The mDAG in Figure 6(a) has only one bidirected edge and therefore is trivially geared with $R_1 = B_1 = \{2, 3\}$. This leads to the sets $\mathcal{F}_2 = \{f_2 : \mathfrak{X}_1 \to \mathfrak{X}_2\}$ and $\mathcal{F}_3 = \{f_3 : \mathfrak{X}_2 \to \mathfrak{X}_3\}$, which are precisely the sets of functions for compliance type and response type respectively.

Example 4.6. Consider the mDAG in Figure 4, and order the bidirected edges as $B_1 = \{1, 2\}$, $B_2 = \{2, 3, 4\}$ and $B_3 = \{3, 4, 5\}$, giving respective remainder sets $R_1 = \{1, 2\}$, $R_2 = \{3, 4\}$ and $R_3 = \{5\}$. The ordering $5 < 4 < 3 < 2 < 1$ of the random vertices respects the gearing, and we have

$$\pi(1) = \pi(5) = \emptyset, \qquad \pi(3) = \pi(4) = \{5\}, \qquad \pi(2) = \{3, 4\}.$$

In this case then

$$\begin{aligned} \mathcal{F}_5 &= \{f : \mathfrak{X}_3 \to \mathfrak{X}_5\} & \mathcal{F}_4 &= \{f : \mathfrak{X}_{2,3,6} \times \mathcal{F}_5 \to \mathfrak{X}_4\} \\ \mathcal{F}_3 &= \{f : \mathfrak{X}_1 \times \mathcal{F}_5 \to \mathfrak{X}_3\} & \mathcal{F}_2 &= \{f : \mathcal{F}_{3,4} \to \mathfrak{X}_2\} \\ \mathcal{F}_1 &= \{f : \{1\} \to \mathfrak{X}_1\} \end{aligned}$$

Alternatively, if we order the bidirected edges as $\{2, 3, 4\}$, $\{1, 2\}$, $\{3, 4, 5\}$, then we could take $5 < 1 < 2 < 3 < 4$, and

$$\pi(1) = \pi(5) = \emptyset, \qquad \pi(3) = \pi(4) = \{5\}, \qquad \pi(2) = \{1\};$$

this yields $\mathcal{F}_2 = \{f : \mathcal{F}_1 \to \mathfrak{X}_2\}$, with other collections $\mathcal{F}_v$ unchanged.

4.3. *Functional Models for Geared Graphs.* If a vertex v is contained within exactly one bidirected edge, B, then without loss of generality we can assume that the latent variable corresponding to B contains all the residual information about how X_v should behave given the values of its visible parents, $X_{\mathrm{pa}(v)}$. In other words, the latent variable associated with B includes a (random) function $f_v : \mathfrak{X}_{\mathrm{pa}(v)} \to \mathfrak{X}_v$ which, once instantiated, 'tells' $X_v = f_v(X_{\mathrm{pa}(v)})$ which value it should take for each value of its other parents, exactly as in Example 4.5.[6] All the randomness of X_v is collapsed into f_v and $X_{\mathrm{pa}(v)}$.

[6] Equivalently, one could take a deterministic function f_v and introduce an 'error term' E_v so that $X_v = f_v(X_{\mathrm{pa}(v)}, E_v)$, as in the non-parametric structural equation models of Pearl [2009].

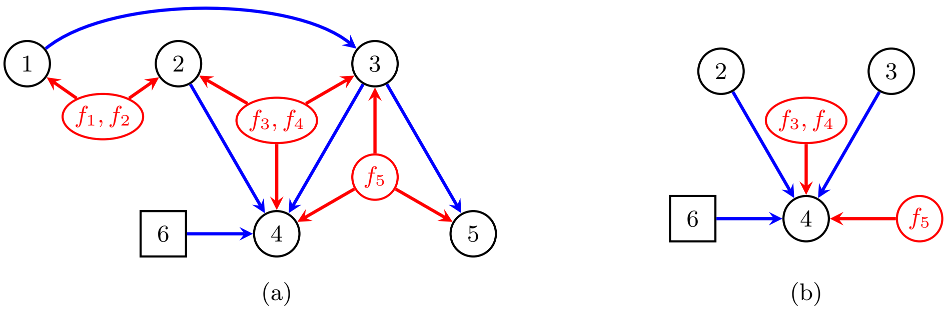


FIG 7. *(a) A DAG with functional latent variables, associated with a gearing of the mDAG in Figure 4(a). (b) Subgraph of the DAG in (a), containing the vertex 4 and its parents.*

If v is contained within two or more bidirected edges, say B_i and B_j, we might say that B_i tells X_v what value to take for every value of its visible parents and the other latent variables. However, it is not clear how to define such a function until the state-space associated with the other latent parents (i.e. B_j) has already been fixed. The decomposable structure of geared graphs makes it possible to iteratively fix state-spaces for latent variables without loss of generality.

To see this, suppose we have a single-district, geared mDAG $\mathcal{G}$ with remainder sets $R_1, \ldots, R_k$, and form the canonical DAG $\bar{\mathcal{G}}$ by replacing each bidirected edge B_i in $\mathcal{G}$ with a new vertex u_i, such that $\mathrm{ch}_{\bar{\mathcal{G}}}(u_i) = B_i$. Compare, for example, the structure of the graphs in Figures 4(a) and (b).

Note that each vertex $v \in R_k$ has a single latent parent u_k in $\bar{\mathcal{G}}$. Then, without loss of generality, incorporate the function $f_v : \mathfrak{X}_{\mathrm{pa}_{\mathcal{G}}(v)} \to \mathfrak{X}_v$ into the latent variable U_k. We 'replace' U_k with the collection of such functions $f_{R_k} \in \mathcal{F}_{R_k}$.

Each vertex $v \in R_{k-1}$ has latent parent u_{k-1} and possibly also u_k; but since the state-space of U_k has been fixed as $\mathcal{F}_{R_k}$, we can define $f_v : \mathfrak{X}_{\mathrm{pa}(v)} \times \mathcal{F}_{R_k} \to \mathfrak{X}_v$ for those v with latent parents u_{k-1} and u_k, and just $f_v : \mathfrak{X}_{\mathrm{pa}(v)} \to \mathfrak{X}_v$ otherwise. These functions $f_{R_{k-1}}$ can be integrated into U_{k-1}, and the process repeated for $i = k-2, \ldots, 1$.

We end up with latent variables U_i taking values in $\mathcal{F}_{R_i}$ for $i = 1, \ldots, k$. For example, with the first gearing given in Example 4.6 for the graph in Figure 4(a), we would have $U_1 = (f_1, f_2)$, $U_2 = (f_3, f_4)$, and $U_3 = (f_5)$. Associating each variable U_i with the vertex u_i leads to the DAG in Figure 7(a). Notice that, for each $v \in V$, the function f_v is contained within a parent variable of v. In addition, all the arguments of the function f_v are also parents of v. For example, take $v = 4$, whose parents are drawn separately in Figure 7(b). The function $f_4 \in \mathcal{F}_4$ is generated as part of the latent variable $U_2 = (f_3, f_4)$, and the associated vertex u_2 is indeed a parent of 4. In addition, $\mathcal{F}_4 = \{f : \mathfrak{X}_{2,3,6} \times \mathcal{F}_5 \to \mathfrak{X}_4\}$, so the arguments of the function f_4, namely X_2, X_3, X_6 and f_5, all correspond to vertices which are also parents of 4. Thus, in setting $X_4 = f_4(X_2, X_3, X_6, f_5)$ we ensure that X_4 is a well defined function of its parent variables.

In fact using this construction we can set $X_v := f_v(X_{\mathrm{pa}(v)}, f_{\pi(v)})$ for every $v \in V$, which is well defined because the directed part of the original mDAG is acyclic. The following result shows that the resulting conditional distribution over X_V given X_W is in the marginal model for the original mDAG.

THEOREM 4.7. *Let $\mathcal{G}$ be a geared mDAG, and $R_i, i = 1, \ldots, k$ be the remainder sets corresponding to some gearing of $\mathcal{G}$. Suppose we generate functions $f_v \in \mathcal{F}_v$ according to a distribution in which*

$$(f_v \,|\, v \in R_i) \perp\!\!\!\perp (f_w \,|\, w \in V \setminus R_i), \qquad\qquad i = 1, \ldots, k,$$

and then define $X_v = f_v(X_{\mathrm{pa}(v)}, f_{\pi(v)})$ for each $v \in V$. Then the induced conditional distribution on X_V given X_W is in the marginal model for $\mathcal{G}$.

Conversely, any distribution in the marginal model for $\mathcal{G}$ can be generated by such a scheme.

PROOF. For each bidirected edge B_i, define the random variable $U_i = (f_v \,|\, v \in R_i)$. The U_is are represented by exogenous variables on the DAG $\bar{\mathcal{G}}$, and the conditions given in the statement of the theorem ensures they are all independent. The structural equation property for $\bar{\mathcal{G}}$ will therefore be satisfied if each X_v is a well defined function of its parents in the graph.

In other words, the three components f_v, $f_{\pi(v)}$ and $X_{\mathrm{pa}(v)}$ must all be determined from random variables which are parents of v in $\bar{\mathcal{G}}$. This holds for $X_{\mathrm{pa}(v)}$ by definition. Additionally $v \in R_i$ implies that $v \in B_i$, and that therefore the variable $U_i \equiv (f_v : v \in R_i)$ is a parent variable of X_v.

Lastly suppose $w \in \pi(v)$; this happens if and only if $w, v \in B_j$ for some $j > i$, in which case $w \in R_j$ for the minimal such j by the running intersection property of the gearing. Then f_w is contained in U_j, which is also a parent variable of X_v.

For the converse suppose that $p \in \mathcal{M}(\mathcal{G})$ satisfies the structural equation property and let R_k be the final remainder set with associated random variable U_k. Each X_v for $v \in R_k$ is a measurable function of its parents $X_{\mathrm{pa}(v)}$ and U_k. Define the random function $f_v : \mathfrak{X}_{\mathrm{pa}(v)} \to \mathfrak{X}_v$ by $f_v(\cdot) = X_v(\cdot, U_k)$, and incorporate it into the latent variable U_k. Repeating this for all $v \in R_k$

gives us $U'_k = (U_k, f_{R_k})$. By Theorem 2.2 of Čencov [2000] we can rewrite $U'_k = (g(f_{R_k}, E), f_{R_k})$ for some measurable function g and random variable E, independent of f_{R_k}, whilst keeping the distribution of U'_k unchanged.

Since $\mathcal{G}$ is geared, all children of U_k that are also children of any other latent variable all share a latent parent, say U_j, $j < k$. If we then augment U_j with E, and replace U'_k with $U''_k \equiv f_{R_k}$, then all variables remain as measurable functions of their parents because $f_w(X_{\mathrm{pa}(w)}, U_k, U_j)$ can be replaced with $f_w(X_{\mathrm{pa}(w)}, g(U''_k, E), U_j)$. Now that U''_k has a fixed, finite state-space, we can apply this process again to U_{k-1}. A simple induction gives the result. □

Since each of these latent variables takes values in a finite collection of functions, this means that the marginal model of a geared graph is equivalent to a latent variable model in which all the random variables (latent and observed) are finite and discrete. It follows from this that marginal models for geared mDAGs are semi-algebraic sets by the Tarski-Seidenberg theorem [Basu et al., 1996, Chapter 2].

EXAMPLE 4.8. Consider the marginal model for the graph in Figure 3(b). In this case the vertices 2 and 4 are each contained in only one bidirected edge, so without loss of generality this edge could be replaced in the canonical DAG (Figure 1) with a latent variable taking values in $\mathcal{F}_2 \times \mathcal{F}_4$ where

$$\mathcal{F}_2 \equiv \{f : \mathfrak{X}_1 \to \mathfrak{X}_2\}, \qquad\qquad \mathcal{F}_4 \equiv \{f : \mathfrak{X}_3 \to \mathfrak{X}_4\}.$$

That is, the latent variable may be assumed to be $U = (f_2, f_4)$, where f_2 and f_4 respectively assign values to X_2 and X_4 given particular values of X_1 and X_3.

For non-geared graphs such as that in Figure 8(a), there is no clear way to write the marginal model as a latent variable model without possible loss of generality. We therefore cannot use this approach to prove that marginal models corresponding to non-geared mDAGs are semi-algebraic. However, it has recently been proven that any marginal model can be written as a latent variable model with finite discrete latent states, provided the state-space is sufficiently large (Denis Rosset, personal communication). It follows that all marginal models are semi-algebraic.

4.4. *Generating Distributions for Geared mDAGs.*

Let $\mathcal{G}$ be a single-district, geared mDAG, with gearing given by remainder sets $R_1, \ldots, R_k$; assign a probability distribution ρ_i to each collection of functions $U_i \equiv$

$(f_v \,|\, v \in R_i)$. Suppose we draw values for variables $U_i = (f_v)_{v \in R_i}$ independently according to ρ_i, and use them to generate values for the observed variables X_V for each possible value of the fixed vertices X_W. The resulting (conditional) distribution over X_V given X_W is, by Theorem 4.7, in the marginal model for $\mathcal{G}$.

Let $\pi(R_i) \equiv \bigcup_{v \in R_i} \pi(v)$ and $f_A \equiv (f_v \,|\, v \in A)$. Define

$$p[\rho_k, \ldots, \rho_1](x_V \,|\, x_W) = \sum_{f_{R_k} \in \Phi_k(x_{VW})} \rho_k(f_{R_k}) \cdots \sum_{f_{R_1} \in \Phi_1(f_{\pi(R_1)}, x_{VW})} \rho_1(f_{R_1}), \tag{6}$$

where

$$\Phi_i(f_{\pi(R_i)}, x_{VW}) = \{f_{R_i} \,|\, f_v(x_{\mathrm{pa}(v)}, f_{\pi(v)}) = x_v \text{ for each } v \in R_i\}; \tag{7}$$

that is, $\Phi_i(f_{\pi(R_i)}, x_{VW})$ is precisely the set of functions f_{R_i} that, given the indicated values of parents variables, jointly evaluate to x_{R_i}. Hence (6) is a sum over all the combinations of functions f_V that, given the input $X_W = x_W$, recursively evaluate to x_V.

The function $p[\cdot]$ takes distributions over the functions f_V and returns a kernel over $\mathfrak{X}_V$ indexed by $\mathfrak{X}_W$. For brevity we will generally denote this by $p[\rho_k, \ldots, \rho_1] = \sum_{\Phi_k} \rho_k \cdots \sum_{\Phi_1} \rho_1$, with the dependence upon x_{VW} left implicit. It may be helpful to think of this as an over-parameterized family of kernels for X_V given X_W, with parameters $\rho_1, \ldots, \rho_k$.

The mapping $p[\cdot]$ is clearly smooth (infinitely differentiable), and its image defines the marginal model. Hence we will be able to deduce various aspects of the model's geometry by studying $p[\cdot]$ and its derivatives. Choosing $\rho_i(f_{R_i}) = 1$ for each i (up to a constant of proportionality which, for simplicity, we do not write explicitly) induces the uniform distribution on $\mathfrak{X}_V$ for each $x_W \in \mathfrak{X}_W$; we denote this kernel by $p_0 \equiv p[1, \ldots, 1]$. Clearly p_0 is contained within $\mathcal{M}(\mathcal{G})$ for any mDAG $\mathcal{G}$—as, in fact, is any distribution corresponding to all variables being independent.

Example 4.9. For the instrumental variables model in Figure 6 (if we consider X_1 to be fixed), we have

$$p[\rho](x_2, x_3 \,|\, x_1) = \sum_{\Phi(x_{123})} \rho(f_2, f_3)$$

where $\Phi(x_{123}) = \{(f_2, f_3) : f_2(x_1) = x_2, f_3(x_2) = x_3\}$.

Example 4.10. In the case of the mDAG in Figure 4(a) we have three bidirected edges and remainder sets, and the gearing used in Figure 7(a) gives

$$p[\rho_3, \rho_2, \rho_1] = \sum_{\Phi_3} \rho_3(f_5) \sum_{\Phi_2} \rho_2(f_3, f_4) \sum_{\Phi_1} \rho_1(f_1, f_2),$$

where

$$\begin{aligned} \Phi_1 &= \{(f_1, f_2) \mid f_1 = x_1,\ f_2(f_3, f_4) = x_2\} \\ \Phi_2 &= \{(f_3, f_4) \mid f_3(x_1) = x_3,\ f_4(x_2, x_3, x_6, f_5) = x_4\} \\ \Phi_3 &= \{f_5 \mid f_5(x_3) = x_5\}. \end{aligned}$$

5. Main Results.

In this section we prove our main result, by showing that the marginal model $\mathcal{M}(\mathcal{G})$ has the same dimension as the nested model. This is done first for geared mDAGs, and the result is then extended to general graphs. For geared graphs, the marginal model is just the image of the infinitely differentiable function $p[\cdot]$ described in the previous section. Such functions can be locally approximated at a particular point, say $p_0 = p[1, \ldots, 1]$, by the linear map given by the derivative of $p[\cdot]$.

This column space of this linear map (also called the *pushforward* map) gives the linear space that approximates the model at p_0, also known as the tangent space[7]. We will show that the tangent space to the marginal model at p_0 is equal to the tangent space of the nested model $\mathcal{N}(\mathcal{G})$ at p_0. To do this we take a basis of the tangent space of $\mathcal{N}(\mathcal{G})$, and for every vector λ in the basis we explicitly construct a vector δ such that the directional derivative of $p[\cdot]$ with respect to δ is equal to λ. This shows that each λ is also contained in the tangent space of $\mathcal{M}(\mathcal{G})$. Since the marginal model is contained within the nested model, it will then follow from results in algebraic geometry that the two models coincide in a neighbourhood of p_0.

For non-geared graphs we have do slightly more work, showing that we can combine maps from different geared sub-graphs to obtain the same result.

5.1. *Vector Spaces and Tangent Cones.*

A probability kernel $p(x_V \mid x_W)$ can be thought of equally as a vector with entries indexed by $\mathfrak{X}_{VW}$, or a real function with domain $\mathfrak{X}_{VW}$. The following decomposition of the vector space $\mathbb{R}^{|\mathfrak{X}_V|}$ will prove useful.

Definition 5.1. For any $A \subseteq V$, let Λ_A be the subspace of $\mathbb{R}^{|\mathfrak{X}_V|}$ consisting of vectors p such that

(i) $\sum_{y_a \in \mathfrak{X}_a} p(y_a, x_{V \setminus a}) = 0$ for each $a \in A$ and $x_{V \setminus \{a\}} \in \mathfrak{X}_{V \setminus \{a\}}$;
(ii) $p(x_V) = p(y_V)$ whenever $x_A = y_A$.

[7] In general the column space could be a subspace of the tangent space, but it is a consequence of Theorem 5.3 that they are equal in this case.

In other words, considered as a function $p : \mathfrak{X}_V \to \mathbb{R}$, the value of $p \in \Lambda_A$ only depends upon x_A, and its sum over x_a for $a \in A$ (keeping the other arguments fixed) is 0. In particular $\Lambda_\emptyset$ is the subspace spanned by the vector of 1s. The dimension of Λ_A is $\prod_{a \in A}(|\mathfrak{X}_a| - 1)$; in the case where all the variables are binary, each Λ_A has dimension one and is the same as the space spanned by the corresponding column of a log-linear design matrix.

It is simple to check that the spaces Λ_A are all orthogonal, and that the real vector space $\mathbb{R}^{|\mathfrak{X}_V|}$ can be decomposed as the direct sum

$$\mathbb{R}^{|\mathfrak{X}_V|} = \bigoplus_{A \subseteq V} \Lambda_A.$$

Definition 5.2. Let $\mathfrak{A}$ be a subset of $\mathbb{R}^k$ containing a point $\boldsymbol{x}$. The *tangent cone* of $\mathfrak{A}$ at $\boldsymbol{x}$ is the set of vectors of the form $\boldsymbol{v} = \lim_{n \to \infty} \alpha_n(\boldsymbol{v}_n - \boldsymbol{x})$ where $\alpha_n \to \infty$ and each $\boldsymbol{v}_n \in \mathfrak{A}$.

A tangent cone is a cone, but may or may not be a vector space, depending upon whether the set $\mathfrak{A}$ is regular at $\boldsymbol{x}$. If $\mathfrak{A}$ is defined by the image of a differentiable bijective map then the tangent cone is a vector space, and the same as the image of the pushforward map. This is the case with the nested model $\mathcal{N}(\mathcal{G})$, which has an explicit and smooth parameterization [Evans and Richardson, 2015]. Its tangent cone at the uniform distribution p_0 is

$$\mathrm{TS}_0^n \equiv \bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A, \tag{8}$$

where $\mathcal{A}(\mathcal{G})$ are the parameterizing sets; this can be deduced by looking directly at the parameterization.

As noted in Section 4, any marginal model $\mathcal{M}(\mathcal{G})$ also contains the uniform distribution $p_0(x_V \,|\, x_W) \equiv |\mathfrak{X}_V|^{-1}$, for all $x_V \in \mathfrak{X}_V, x_W \in \mathfrak{X}_W$, at which point all variables are jointly independent. The tangent cone of the marginal model $\mathcal{M}(\mathcal{G})$ at p_0 is also the vector space (8), which forms the main result of this section.

Theorem 5.3. *The tangent cone of $\mathcal{M}(\mathcal{G})$ at p_0, denoted* TC_0*, is the vector space* $\mathrm{TC}_0 = \mathrm{TS}_0^n \equiv \bigoplus_{A \in \mathcal{A}} \Lambda_A$.

That $\mathrm{TC}_0 \subseteq \mathrm{TS}_0^n$ follows from the fact that $\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G})$. The proof of the reverse inclusion is the subject of this section. We first show this for geared graphs in Section 5.2 then extend to general mDAGs in Section 5.3, culminating in the proof of Theorem 1.2.

5.2. *Results for Geared Graphs.*

DEFINITION 5.4. Let $\lambda : \mathfrak{X}_A \to \mathbb{R}$; we say that λ is A-*degenerate* (or just degenerate) if for each $a \in A$, and $x_{A\setminus a} \in \mathfrak{X}_{A\setminus a}$,

$$\sum_{y_a} \lambda(y_a, x_{A\setminus a}) = 0.$$

It is not hard to see that the set of A-degenerate functions is isomorphic to the vector space Λ_A; both formulations will be useful.

DEFINITION 5.5. Given a degenerate function $\varepsilon_i : \mathcal{F}_{R_i} \to \mathbb{R}$, define

$$D_i(\varepsilon_i) = \lim_{\eta \downarrow 0} \eta^{-1} \left\{ p[1, \dots, 1 + \eta\varepsilon_i, \dots, 1] - p[1, \dots, 1, \dots, 1] \right\},$$

so that $D_i(\varepsilon_i)$ is a vector in $\mathbb{R}^{|\mathfrak{X}_{VW}|}$, the directional derivative of the ith component of $p[\cdot]$ with respect to ε_i. For sufficiently small $\eta > 0$, the vector $1 + \eta\varepsilon_i$ is non-negative and therefore a valid distribution over $\mathcal{F}_{R_i}$ (up to the normalizing constant); it follows that $D_i(\varepsilon_i) \in \mathrm{TC}_0$, the tangent cone of $\mathcal{M}(\mathcal{G})$ at p_0.

Let $T_i = \{D_i(\varepsilon_i) \,|\, \varepsilon_i \text{ degenerate}\}$. Since the function $p[\cdot]$ is differentiable at $[1, \dots, 1]$ it follows that T_i is a vector space, and also that the vector space $T_1 + \cdots + T_k$ is contained within the tangent cone of $\mathcal{M}$ at the uniform distribution. We will show that $T_1 + \cdots + T_k$ is in fact the same as (8).

It will be useful to define the following collection of supersets of Φ_i, for $B \subseteq V$:

$$\Phi_i^B(f_{\pi(R_i)}, x_{VW}) \equiv \{f_{R_i} \,|\, f_v(x_{\mathrm{pa}(v)}, f_{\pi(v)}) = x_v \text{ for each } v \in R_i \cap B\}. \tag{9}$$

In words, this is the collection of functions f_{R_i} such that, given inputs $f_{\pi(R_i)}$ and $x_{\mathrm{pa}(R_i)\setminus R_i}$, the values of $f_{B\cap R_i}$ jointly evaluate to $x_{B\cap R_i}$. Note that $\Phi_i^B = \Phi_i$ for any $B \supseteq R_i$.

LEMMA 5.6. *Let $C \subseteq R_i$, with* $\mathrm{sterile}_{\mathcal{G}}(C) \subseteq A \subseteq C \cup \mathrm{pa}_{\mathcal{G}}(C)$ *and $E \subseteq \pi(C)$. Then for every degenerate function $\lambda : \mathfrak{X}_A \times \mathcal{F}_E \to \mathbb{R}$, there exists a degenerate function $\delta : \mathcal{F}_C \to \mathbb{R}$ such that*

$$\sum_{f_{R_i} \in \Phi_i} \delta(f_C) = \lambda(x_A, f_E),$$

where Φ_i is given by (7). In addition,

$$\sum_{f_{R_i} \in \Phi_i^B} \delta(f_C) = \begin{cases} |\mathfrak{X}_{R_i\setminus B}|\lambda(x_A, f_E) & \textit{if } C \subseteq B \\ 0 & \textit{otherwise.} \end{cases}$$

The proof is in the Supplementary Material, Section B.3 [Evans, 2017].

Remark 5.7. Note that if we set $E = \emptyset$, the above result shows that for any $\lambda \in \Lambda_A$ there exists a δ such that

$$\begin{aligned}&\eta^{-1}\{p[1,\dots,1+\eta\delta,\dots,1] - p[1,\dots,1,\dots,1]\}\\ &= \eta^{-1}\left\{\sum_{\Phi_k}\cdots\sum_{\Phi_i}\eta\delta(f_C)\sum_{\Phi_{i-1}}\cdots\sum_{\Phi_1}1\right\}\\ &= \lambda.\end{aligned}$$

Hence $\Lambda_A \leq T_i$ (that is, Λ_A is a subspace of the vector space T_i) for any A such that $\mathrm{sterile}_{\mathcal{G}}(C) \subseteq A \subseteq C \cup \mathrm{pa}_{\mathcal{G}}(C)$ and $C \subseteq R_i$.

The next result, also proved in the supplement [Evans, 2017, Section B.3], forms the backbone for proving Theorem 5.3: it extends Lemma 5.6 to sets C that are not contained within a single remainder set.

Lemma 5.8. *Let C be a bidirected-connected set, and for each i define $C_i \equiv C \cap R_i$; let $I \equiv \{i \,|\, C_i \neq \emptyset\}$. For $\mathrm{sterile}_{\mathcal{G}}(C_i) \subseteq A_i \subseteq C_i \cup \mathrm{pa}_{\mathcal{G}}(C_i)$, let $A = \triangle_{i\in I} A_i$. Then Λ_A is a subspace of T_l, where l is the minimal element of I.*

Corollary 5.9. *For a geared mDAG $\mathcal{G}$ with $k \geq 1$ bidirected edges,*

$$\bigoplus_{A\in\mathcal{A}(\mathcal{G})} \Lambda_A \leq T_1 + \cdots + T_k.$$

Proof. By Lemma 3.6, there is some bidirected-connected set C such that $A = \triangle_{v\in C} A_v$ for sets $\{v\} \subseteq A_v \subseteq \{v\} \cup \mathrm{pa}_{\mathcal{G}}(v)$. Take $C_i \equiv C \cap R_i = \{v_{i1}, \dots, v_{ik_i}\}$, so A is of the form

$$A = \bigtriangleup_{i,j} A_i^j = \bigtriangleup_{i}\left(\bigtriangleup_{j} A_i^j\right)$$

where $A_i^j \equiv A_{v_{ij}}$ (here we have changed nothing other than to label the vertices v_{ij} by which remainder set they are contained in).

Applying Lemma 3.6 in reverse to the bidirected-connected set C_i shows that $A_i \equiv \triangle_j A_i^j$ is in $\mathcal{A}(\mathcal{G})$, and therefore satisfies $\mathrm{sterile}_{\mathcal{G}}(C_i) \subseteq A_i \subseteq C_i \cup \mathrm{pa}_{\mathcal{G}}(C_i)$. Then by Lemma 5.8 the space Λ_A is contained in some T_i, $i = 1, \dots, k$. □

EXAMPLE 5.10. The instrumental variables model from Figure 6 (see Examples 4.5 and 4.9) has a saturated nested model with the following parameterizing sets:

head H	tail T	parameterizing sets $\mathcal{A}$
$\{1\}$	$\emptyset$	$\{1\}$
$\{2\}$	$\{1\}$	$\{2\}$, $\{1,2\}$
$\{3\}$	$\{1,2\}$	$\{3\}$, $\{1,3\}$, $\{2,3\}$, $\{1,2,3\}$

Indeed, taking the functional parameterization suggested in Example 4.9, one can see that altering the distribution of the compliance functions f_2 will affect the distribution of X_2 conditional on X_1, which is why Λ_2 and Λ_{12} are contained in TC_0. For example, to introduce a correlation between X_1 and X_2 whilst keeping the marginal distributions fixed, we can increase the proportion of 'compliers' (that is the people for whom $f_2(0) = 0, f_2(1) = 1$) and decrease the proportion of 'defiers' ($f_2(0) = 1, f_2(1) = 0$).

Similarly, modifying the distribution of f_3 gives us Λ_3 and Λ_{23}. Obtaining the directional derivatives in Λ_{13} and Λ_{123} requires modifying the distribution of f_2, f_3 jointly.

None of the examples given in this paper require the full generality of Lemma 5.8 to prove that Theorem 5.3 applies to them, however an example in which this is necessary may be found in the Supplementary Material, Section B.4 [Evans, 2017].

5.3. *Extension to non-geared graphs.* Corollary 5.9 puts us in a position to prove Theorem 5.3 for geared graphs; however it does not so far extend to the general case, because we cannot fix the state-spaces of the latent variables without a gearing. In this section we will show that the tangent cone of a general marginal model at the uniform distribution is the vector space spanned by the tangent cones of its geared subgraphs, and that therefore the problem can be reduced to geared graphs.

PROPOSITION 5.11. *Let $\mathcal{G}$ be an arbitrary mDAG containing geared subgraphs $\mathcal{G}_1, \ldots, \mathcal{G}_k$. Suppose that, for each subgraph and a suitable gearing $\Lambda_{A_i} \leq \mathrm{TC}_0(\mathcal{G}_i)$. Then $\Lambda_{A_1} + \cdots + \Lambda_{A_k} \leq \mathrm{TC}_0(\mathcal{G})$.*

In other words, the tangent cone of $\mathcal{G}$ includes the vector space spanned by all the tangent cones of the subgraphs. The proof is found in the Supplementary Material, Section B.5 [Evans, 2017].

EXAMPLE 5.12. The bidirected 4-cycle in Figure 8(a) is not geared, and therefore we cannot apply our earlier results to it directly. The nested model

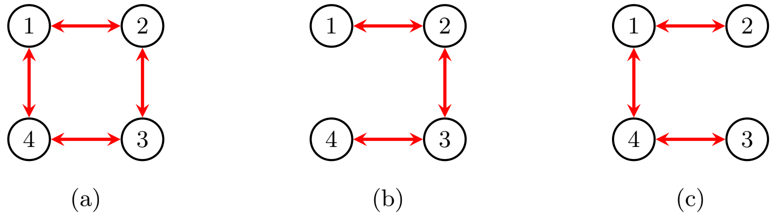


FIG 8. *(a) the bidirected 4-cycle, and (b), (c) two geared subgraphs.*

for this graph is equivalent to the model defined by the constraints $X_1 \perp\!\!\!\perp X_3$ and $X_2 \perp\!\!\!\perp X_4$, and has parameterizing sets

$$\begin{aligned}\mathcal{A}(\mathcal{G}) = \{&\{1\}, \{2\}, \{1,2\}, \{3\}, \{2,3\}, \{1,2,3\},\\ &\{4\}, \{1,4\}, \{1,2,4\}, \{3,4\}, \{1,3,4\}, \{2,3,4\}, \{1,2,3,4\}\},\end{aligned}$$

which are also the bidirected-connected sets of vertices. The two subgraphs in Figures 8(b) and (c), say $\mathcal{G}_1$ and $\mathcal{G}_2$, *are* geared, however, and their parameterizing sets combined include all sets in $\mathcal{A}(\mathcal{G})$. Hence, by applying Proposition 5.11 with these graphs, we find that

$$\bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A = \bigoplus_{A \in \mathcal{A}(\mathcal{G}_1)} \Lambda_A + \bigoplus_{A \in \mathcal{A}(\mathcal{G}_2)} \Lambda_A \leq \mathrm{TC}_0(\mathcal{G}).$$

It follows that the marginal model is also defined by the independences $X_1 \perp\!\!\!\perp X_3$ and $X_2 \perp\!\!\!\perp X_4$, possibly with some additional inequality constraints.

We are now in a position to put together these ideas and prove the main results for general mDAGs.

PROOF OF THEOREM 5.3. Suppose first that $\mathcal{G}$ is geared. $p[1, \dots, 1 + \eta\varepsilon_i, \dots, 1]$ obeys the nested Markov property for any degenerate function ε_i and η sufficiently small that $1 + \eta\varepsilon_i$ is positive; it follows that $T_i \leq \mathrm{TC}_0$ for each i, and that therefore using Corollary 5.9,

$$\bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A \leq T_1 + \cdots + T_k$$

is also contained in TC_0, by the differentiability of $p[\cdot]$ at $(1, \dots, 1)$.

Now for general $\mathcal{G}$, and each $A \in \mathcal{A}(\mathcal{G})$, there exists a geared subgraph $\mathcal{G}'$ of $\mathcal{G}$ such that $\Lambda_A \leq \mathrm{TC}_0(\mathcal{G}')$ by Lemma 4.3. Then applying Proposition 5.11, we see that the space spanned by these subspaces is contained within the tangent cone for $\mathcal{G}$: i.e. $\bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A \leq \mathrm{TC}_0(\mathcal{G})$. If a distribution is in the marginal model then it is also in the nested model, and therefore TC_0 is contained within the tangent space TS_0^n of $\mathcal{N}(\mathcal{G})$ at p_0, which has dimension

$$\dim(\mathrm{TS}_0^n) = \sum_{H \in \mathcal{H}(\mathcal{G})} |\mathfrak{X}_{T(H)}| \prod_{h \in H} (|\mathfrak{X}_h| - 1) = \sum_{A \in \mathcal{A}(\mathcal{G})} \dim(\Lambda_A);$$

the second equality here follows from $\dim(\Lambda_A) = \prod_{h \in A}(|\mathfrak{X}_h| - 1)$ and

$$\sum_{H \subseteq A \subseteq H \cup T} \dim(\Lambda_A) = \sum_{H \subseteq A \subseteq H \cup T} \prod_{h \in A} (|\mathfrak{X}_h| - 1) = |\mathfrak{X}_{T(H)}| \prod_{h \in H} (|\mathfrak{X}_h| - 1).$$

Combining $\bigoplus_{A \in \mathcal{A}(\mathcal{G})} \Lambda_A \leq \mathrm{TC}_0 \subseteq \mathrm{TS}_0^n$ with the dimension of TS_0^n gives the result. □

Theorem 1.2 is now a corollary of this result.

PROOF OF THEOREM 1.2. Since $\mathcal{N}(\mathcal{G})$ is parametrically defined via polynomials, its Zariski closure is an irreducible variety [see, e.g. Cox et al., 2007, Proposition 4.5.5]. The Zariski closure of $\mathcal{M}(\mathcal{G})$ is, by definition, also an algebraic variety. For algebraic varieties V_1, V_2, if $V_1 \subseteq V_2$ and V_2 is irreducible, then either V_1 has a strictly smaller dimension than V_2, or they are identical. By Theorem 5.3, the Zariski closures of $\mathcal{M}$ and $\mathcal{N}$ have the same dimension and therefore they coincide. This means that, in a neighbourhood of p_0, the models themselves are also the same. □

6. Smoothness of the marginal model.

The results of Section 5, together with the smoothness of the nested model, allow us to show that for geared graphs, the interior of the marginal model is a smooth manifold.

THEOREM 6.1. *For any mDAG $\mathcal{G}$ and state-space $\mathfrak{X}_{VW}$, the relative interior of the marginal model $\mathcal{M}(\mathcal{G})$ is a manifold of dimension $d(\mathcal{G}, \mathfrak{X}_{VW})$, and is described by a finite number of semi-algebraic constraints.*

PROOF. The nested Markov model is parametrically defined (with a polynomial parameterization), and therefore its Zariski closure is an irreducible variety [see, e.g. Cox et al., 2007, Proposition 4.5.5]. Furthermore Evans and Richardson [2015] give a diffeomorphism between the set of strictly positive distributions obeying the nested Markov property, and an open parameter set. It follows that $\mathcal{N}(\mathcal{G})$ is a manifold on the interior of the simplex [see, for example, Kass and Vos, 1997, Appendix A].

As noted in Section 4, the marginal model is a semi-algebraic set. Since $\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G})$ and these two sets have the same Zariski closure, it follows

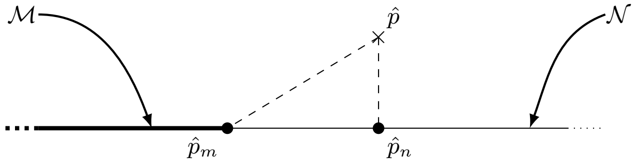


FIG 9. *Diagramatic representation of estimation with the marginal model. The thicker line represents the marginal model, and its thinner extension the nested model. The unconstrainted MLE is shown as $\hat{p}$, and its projection to MLEs under the marginal and nested models as $\hat{p}_m$ and $\hat{p}_n$ respectively. Note that $\hat{p}_m$ is on the boundary of $\mathcal{M}$; if the true data generating distribution is on the boundary this generally leads to irregular asymptotics.*

that $\mathcal{M}(\mathcal{G})$ is defined from $\mathcal{N}(\mathcal{G})$ by a finite number of additional polynomial inequalities. It further follows that it is also a manifold at any point these inequality constraints are not active. □

It follows from Theorem 6.1 that the interior of the marginal model for an mDAG is a curved exponential family of dimension $d(\mathcal{G}, \mathfrak{X}_{VW})$, and that therefore the nice statistical properties of these models can be applied. For example, the maximum likelihood estimator (MLE) of a distribution within the model will be asymptotically normal and unbiased, and the likelihood ratio statistic for testing this model has an asymptotic χ^2-distribution.

For a point on the boundary defined by an active inequality constraint, the asymptotic distribution of the likelihood ratio statistic may be much more complicated than for a point on the relative interior [Drton, 2009]; in general it is a mixture of χ^2-distributions, and this mixture will vary depending upon the unknown truth. A possible advantage of the nested model is that we can guarantee that the true distribution does not lie on the boundary of $\mathcal{N}$ if the MLE consists of strictly positive probabilities, because the boundary only consists of distributions with at least some zero probabilities; the same cannot be said for $\mathcal{M}$. This is depicted in Figure 9, in which the MLE under the nested model ($\hat{p}_n$) is in the interior of $\mathcal{N}$, but the MLE for the marginal model $\hat{p}_m$ lies on the boundary of $\mathcal{M}$.

Inequality constraints are generally much more complicated than equality constraints, and efforts to characterize them fully in DAGs with latent variable models have been limited by computational challenges. See, for example, the discussion in ver Steeg and Galstyan [2011].

6.1. *Why marginal models?.* Theorem 4.7 tells us that we can use an explicit latent variable model with potential outcomes to obtain the marginal model for any geared graph; in this event, one might ask why we need

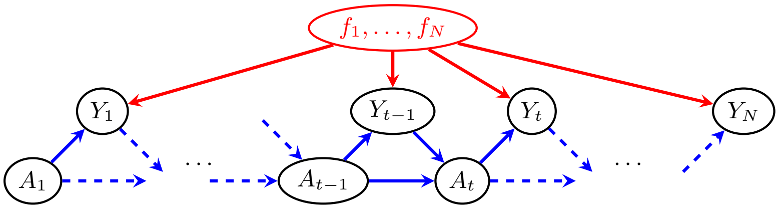


FIG 10. *Bayesian network representing the dynamic treatment model in Example 6.2.*

the constraint-based representation afforded by the nested model. Unfortunately, whilst often useful for causal inference, potential outcomes are impractical to use as a full parameterization in all but the smallest models, because their state-space quickly becomes infeasibly large if there are several treatments or several possible outcomes. In the simplest case of two variables $A \to Y$ with with respectively m and n states, there are n^m possible types for Y, compared to only mn distinct, observable outcomes. In reality m is often quite large, because Y may have multiple parents.

EXAMPLE 6.2. Consider a longitudinal treatment program where at each stage $t = 1, \dots, N$, patients are given treatment A_t and an outcome Y_t is measured. Treatments are chosen by clinicians to depend only on the treatment and outcome at the previous time point, but outcomes are correlated due to unobserved confounding. For $N = 2$ this is similar to Example 1.1.

Following the same approach as in Example 4.4 leads to a latent variable consisting of random functions f_t that map values of treatments A_t onto potential outcomes Y_t; see Figure 10 for a graph of the relevant latent variable model. For the simplest case of binary variables each of these functions would have four states; the full latent variable would consist of all N such functions, so would have 4^N possible states. This latent variable identifies the quantity $P(Y_1, \dots, Y_N \,|\, do(A_1, \dots, A_N))$, which is the distribution of $Y_1, \dots, Y_N$ after intervening on $A_1, \dots, A_N$; this has a dimension of 'only' $\frac{2}{3}(4^N - 1)$ under this model, and the difference between these two models grows exponentially in N. If Y_t depends on $k > 1$ treatments then the problem becomes much worse, as f_t requires 2^{2^k} states. This leads to a latent variable of dimension $O(2^{2^k T})$ on a contingency table of dimension just 2^{2T}.

How would alternative approaches deal with this model? An ancestral graph model [Richardson and Spirtes, 2002] would replace the latent variable with directed edges $Y_i \to Y_j$ for each $i < j$ and and $A_i \to Y_j$ for each $i \leq j$. As such it would throw away most of the structural information about causal relationships, and give a model of significantly larger dimension.

An ordinary latent variable model (i.e. without explicit potential outcomes) could reduce the dimension by using fewer states, and would eventually lead to an identified model. However, as we have already seen in the Introduction, there is no way to achieve identifiability in this example without imposing additional equality constraints; in many contexts, including epidemiological examples such as the one above, there is typically no domain knowledge about the hidden variables that would justify such assumptions. They are also difficult to test because the additional constraints are not generally explicitly available, and goodness of fit statistics such as likelihood ratio tests do not have the same asymptotic distribution at all points in the parameter space [Drton, 2009].

By contrast, the marginal model (and therefore the nested model) has the correct dimension, which is never larger than the relevant contingency table, and does not impose any constraints not explicitly implied by the Bayesian network model. The discrete nested model is identified everywhere and has a smooth parameterization, and can easily be fitted by maximum likelihood using the algorithm in Evans and Richardson [2010]. In addition, the parameterization is made up of precisely of the identifiable causal quantities from the model. For large N the marginal model may still result in a model that is too large for a particular dataset; in this case further parametric constraints or simplifying assumptions such as additivity, sparsity or symmetry can easily be placed on parameters that are identifiable [see Shpitser et al., 2013, for an example of this]. Unlike a latent variable model, these additional assumptions would lead to a transparent reduction in dimension, do not lead to new questions about identifiability, and can be tested directly.

6.2. *Quantum Causal Models.* In the Quantum Information literature there is interest in models where the latent variables are replaced with quantum states or other, even more general objects [see Henson et al., 2014, and references therein]. This can result in larger models, most famously as quantum violations of Bell's inequalities [see Gill, 2014, for a statistical introduction]. However, as a consequence of results by Henson et al. [2014], nested constraints also apply to quantum models, and hence the quantum model is also of the same dimension as the nested and marginal models.

Acknowledgements. We thank Angelos Armen for a very close reading and substantial comments, as well as the associate editor and several anonymous referees for excellent suggestions to improve the clarity of the paper.

References.

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A):3099–3132, 2009.

A. Anandkumar, D. Hsu, A. Javanmard, and S. Kakade. Learning linear Bayesian networks with latent variables. In *Proceedings of The 30th International Conference on Machine Learning*, volume 28, pages 249–257, 2013.

S. Basu, R. Pollack, and M.-F. Roy. *Algorithms in real algebraic geometry*. Springer, 1996.

C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2007.

D. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms*. Springer, 2007. Third Edition.

A. Darwiche. *Modeling and reasoning with Bayesian networks*. CUP, 2009.

A. P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002.

M. Drton. Likelihood ratio tests and singularities. *Annals of Statistics*, 37(2):979–1012, 2009.

R. J. Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

R. J. Evans. Supplement to "Margins of discrete Bayesian networks. 2017.

R. J. Evans and T. S. Richardson. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the 26th conference on Uncertainty in Artificial Intelligence*, 2010.

R. J. Evans and T. S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, 42:1452–1482, 2014.

R. J. Evans and T. S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. arXiv:1511.06813, 2015.

C. J. Fox, A. Käufl, and M. Drton. On the causal interpretation of acyclic mixed graphs under multivariate normality. *Linear Algebra and its Applications*, 2014.

L. D. Garcia, M. Stillman, and B. Sturmfels. Algebraic geometry of Bayesian networks. *Journal of Symbolic Computation*, 39(3-4):331–355, 2005.

R. D. Gill. Statistics, causality and Bell's theorem. *Statist. Sci.*, 29(4):512–528, 2014.

J. Henson, R. Lal, and M. F. Pusey. Theory-independent limits on correlations from generalized bayesian networksayesian networks. *New Journal of Physics*, 16(11):113043, 2014.

R. E. Kass and P. W. Vos. *Geometrical foundations of asymptotic inference*. John Wiley & Sons, 1997.

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

J. Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923. In Polish; English translation by D. Dabrowska and T. Speed in *Statist. Science* 5 463–472, 1990.

J. Pearl. *Causality: Models, Reasoning, and Inference*. CUP, second edition, 2009.

J. Provan and L. Billera. Decompositions of simplicial complexes related to diameters of convex polyhedra. *Mathematics of Operations Research*, 5(4):576–594, 1980.

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

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

T. S. Richardson, R. J. Evans, and J. M. Robins. Transparent parameterizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610, 2011.

T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. arXiv:1701.06686, 2017.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

I. Shpitser, R. J. Evans, T. S. Richardson, and J. M. Robins. Sparse nested markov models with log-linear parameters. In *29th conference on Uncertainty in Artificial Intelligence*, pages 576–585, 2013.

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

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

J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 519–527, 2002.

N. N. Čencov. *Statistical decision rules and optimal inference.* Number 53 in Translations of Mathematical Monographs. American Mathematical Society, 2000. Translated from Russian, 1982.

G. ver Steeg and A. Galstyan. A sequence of relaxations constraining hidden variable models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 2011.

T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pages 255–268, 1990.

SUPPLEMENTARY MATERIAL

Supplement to :"Margins of discrete Bayesian networks"
(doi: 10.1214/00-AOASXXXXSUPP; .pdf). Technical proofs and some additional examples are contained in the supplement.