Smooth, identifiable supermodels of discrete DAG models with latent variables

ROBIN J. EVANS[1] and THOMAS S. RICHARDSON[2]

[1]*Department of Statistics, University of Oxford, 24–29 St Giles', Oxford, OX1 3LB, UK.*
E-mail: evans@stats.ox.ac.uk
[2]*Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195, USA.*
E-mail: thomasr@u.washington.edu

We provide a parameterization of the discrete nested Markov model, which is a supermodel that approximates DAG models (Bayesian network models) with latent variables. Such models are widely used in causal inference and machine learning. We explicitly evaluate their dimension, show that they are curved exponential families of distributions, and fit them to data. The parameterization avoids the irregularities and unidentifiability of latent variable models. The parameters used are all fully identifiable and causally-interpretable quantities.

Keywords: Bayesian network; DAG; nested Markov model; parameterization

1. Introduction

Directed acyclic graph (DAG) models, also known as Bayesian networks, are a widely used class of multivariate models in probabilistic reasoning, machine learning and causal inference (Bishop [1], Darwiche [2], Pearl [15]). The inclusion of latent variables within Bayesian network models can greatly increase their flexibility, and also account for unobserved confounding; however, latent variable models are typically non-regular, their dimension can be hard to calculate, and they generally do not have fully identifiable parameterizations. In this paper, we will present an alternative approach which overcomes these difficulties, and does not require any parametric assumptions to be made about the latent variables.

Example 1.1. Suppose we are interested in the relationship between family income during childhood X, an individual's education level E, their military service M, and their later income Y. We might propose the model shown in Figure 1(a), which includes a hidden variable U representing motivation or intelligence. Let the four observed variables be binary, but make no assumption about U.

One can check using Pearl's d-separation criterion (Pearl [15]) that $M \perp\!\!\!\perp X \mid E$ under this model, in other words there is no relationship between military service and family income after controlling for level of education; this places two independent constraints on the variables' joint distribution $p(x, e, m, y)$ (one for each level of E). In addition, let $q_{EY}(e, y \mid x, m) \equiv p(e \mid$

(a)   (b)
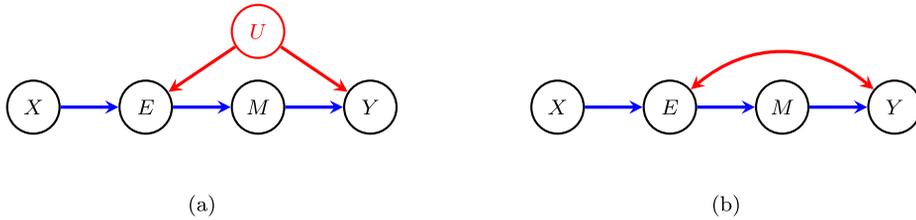
Figure 1. (a) A directed acyclic graph with the latent variable U; (b) a (conditional) acyclic directed mixed graph (the *Verma graph*) representing the observed distribution in (a).

$x) \cdot p(y \mid x, m, e)$; then the quantity

$$q_{EY}(y \mid x, m) \equiv \sum_e q_{EY}(e, y \mid x, m)$$
$$= \sum_e p(e \mid x) \cdot p(y \mid x, m, e) \tag{1}$$

does not depend upon x (Robins [18]); a short proof of this is given in Appendix A. If the graph is interpreted causally, then $q_{EY}(y \mid x, m) = p(y \mid \mathrm{do}(x, m))$, that is, the distribution of Y in an experiment that externally sets $\{X = x, M = m\}$. Note that generally $q_{EY}(y \mid x, m) \neq p(y \mid x, m)$.

The restriction that (1) does not depend on x corresponds to two further independent constraints on p, one for each level of m. The set of distributions that satisfy all four constraints is the *nested Markov model* associated with the graph(s) in Figure 1; the number of free parameters is $15 - 2 - 2 = 11$.

The distributions in the model all factorize as

$$p(x, e, m, y) = p(x) \cdot p(m \mid e) \cdot q_{EY}(e, y \mid x, m),$$

and each of the three factors can be parameterized separately. The model can therefore be described using the following 11 free parameters:

$$p(x = 0), \qquad p(m = 0 \mid e), \qquad q_{EY}(e = 0 \mid x),$$
$$q_{EY}(y = 0 \mid m), \qquad q_{EY}(e = y = 0 \mid x, m).$$

If we interpret the model causally these are, respectively, the quantities

$$P(X = 0), \qquad P\big(M = 0 \mid \mathrm{do}(E = e)\big), \qquad P\big(E = 0 \mid \mathrm{do}(X = x)\big),$$
$$P\big(Y = 0 \mid \mathrm{do}(M = m)\big), \qquad P\big(E = 0, Y = 0 \mid \mathrm{do}(X = x, M = m)\big).$$

The map from the set of positive probability distributions that satisfy the 4 constraints to these 11 parameters is smooth and bijective, and the parameters are fully identifiable. It follows that the model is a curved exponential family of distributions, and that it can be fitted using standard numerical methods.

An alternative modelling approach would be to include a latent variable U explicitly in the model, but this leads to some parameters being unidentifiable. For example, with a binary U the model implied by Figure 1(a) has 12 parameters. We know that the true marginal distribution has at most dimension 11, so at least one of these 12 parameters is unidentifiable. Even though the model is not identified, this latent variable model is still 'too small', in the sense that the model over the observed margin only has dimension 10, whereas dimension 11 can be obtained if U is allowed to have more than two states. As U is not observed, it is undesirable to make specific assumptions about U's state-space because one may unwittingly impose restrictions on the observable distribution. Further, latent variable models are not statistically regular, so standard statistical theory for likelihood ratio tests and asymptotic normality of parameter estimates does not apply (Mond et al. [13], Drton [5]).

1.1. Other work and this paper's contribution

Models of conditional independence associated with margins of DAG models (we refer to these as 'ordinary Markov models') have been studied by Richardson and Spirtes [25]; see also Wermuth [30]. These models were parameterized and shown to be smooth by Evans and Richardson [10]. Other approaches using probit models (Silva and Ghahramani [24]) and cumulative distribution networks (Huang and Frey [12], Silva et al. [23]) are more parsimonious than ordinary Markov models, but impose additional constraints due to their parametric structure.

None of the models mentioned in the previous paragraph can account for constraints of the kind in (1), which were first identified by Robins [18] and separately by Verma and Pearl [29]. Such constraints are attractive because they allow finer distinctions between different causal models from purely observational data: for example, going by conditional independence alone the graph in Figure 1(b) is Markov equivalent to the DAGs in Figure 2, and these causal models are therefore indistinguishable without using other constraints; however the DAGs do not imply the Verma constraint (1), so under the nested Markov model one *can* distinguish between these models.

An algorithm for finding such constraints was given by Tian and Pearl [28], and developed into a fully nonparametric statistical model (the nested Markov model) by Richardson et al. [17]. In this paper, we provide a smooth, statistically regular and fully identifiable parameterization of the discrete version of nested Markov models. As a result, discrete nested Markov models are shown to be curved exponential families of distributions of known dimension. All the parameters we derive are interpretable as straightforward causal quantities. Evans [7] shows that the discrete nested Markov model that we describe here is the best possible algebraic approximation to DAG

Figure 2. DAGs that represent the same conditional independence model as Figure 1(b), but which do not imply the Verma constraint.

models with latent variables, in the sense that the models have the same dimension over the observed variables. An earlier review paper (Shpitser et al. [20]) mentions the parameterization given here, but no proofs are provided.

The conditional independence constraints we consider here include the constraints described in Tian and Pearl [28]. They are also a special case of the *dormant independences*. However, not all dormant independences lead to constraints on the observed distribution – some impose restrictions (solely) on intervention distributions; see Shpitser et al. [20]. A complete algorithm for generating dormant constraints is given in Shpitser and Pearl [21].

The remainder of the paper is organized as follows. In Section 2, we introduce Conditional Acyclic Directed Mixed Graphs, the class of graphs we use to represent our models; those models are formally introduced in Section 3. Some graphical theory is given in Section 4, before the main results in Section 5. Section 6 applies the method to data from a panel study.
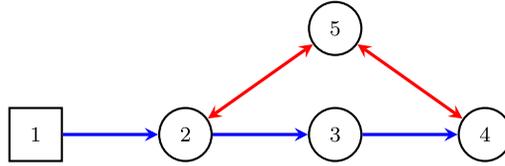
2. Conditional acyclic directed mixed graphs

A directed acyclic graph (DAG) contains vertices representing random variables, and edges (arrows) that imply some structure on the joint probability distribution. A DAG with latent vertices can be transformed into an acyclic directed mixed graph (ADMG) over just its observed vertices via an operation called *latent projection* (Pearl and Verma [14]). In the simplest case this just involves replacing latent variables with bidirected edges ($\leftrightarrow$), as illustrated by the transformation from Figure 1(a) to (b); the transformed graph represents the marginal distribution over the observed random variables X_V.

For technical reasons, we work with a slightly larger class of graphs, called *conditional* acyclic directed mixed graphs (CADMGs). These have two sets of vertices, fixed (W) and random (V), and are used to represent the structure of a set of distributions for X_V indexed by possible values of X_W.

Definition 2.1. A *conditional acyclic directed mixed graph* (CADMG) $\mathcal{G}$ is a quadruple $(V, W, \mathcal{E}, \mathcal{B})$. There are two disjoint sets of vertices: *random*, V, and *fixed*, W. The *directed edges* $\mathcal{E} \subseteq (V \cup W) \times V$ are ordered pairs of vertices; if $(a, b) \in \mathcal{E}$ we write $a \to b$. Loops $a \to a$ and directed cycles $a \to \cdots \to a$ are not allowed (hence 'acyclic'). The *bidirected edges*, $\mathcal{B}$, are unordered pairs of distinct random vertices, and if $\{a, b\} \in \mathcal{B}$ we write $a \leftrightarrow b$.

For convenience, throughout this paper we will only consider CADMGs in which for every fixed vertex w there is at least one edge $w \to v$. (Note that it follows from the definition of a CADMG that v will be random.)

These graphical concepts are most easily understood by example: see the CADMG in Figure 3. We depict random vertices with round nodes, and fixed vertices with square nodes. CADMGs are not generally simple graphs, because it is possible to have up to two edges between each pair of vertices in V (one directed and one bidirected); see Figure 6 for two examples. CADMGs are a slight generalization of ADMGs (Richardson [16]), which correspond to the special case $W = \varnothing$. Note that no arrowheads can be adjacent to any fixed vertex: so neither $a \to w$ nor $a \leftrightarrow w$ is allowed for any $w \in W$. This reflects the fact that fixed vertices cannot depend on

Figure 3. A conditional acyclic directed mixed graph $\mathcal{L}$, with random vertices $V = \{2, 3, 4, 5\}$ and fixed vertices $W = \{1\}$.

other variables, observed or unobserved, but that random vertices may depend upon fixed ones. Mathematically, fixed nodes play a similar role to the 'parameter nodes' used by Dawid [3].

We make use of the following standard familial terminology for directed graphs.

Definition 2.2. If $a \rightarrow b$, we say that a is a *parent* of b, and b a *child* of a. The set of parents of b is denoted $\mathrm{pa}_{\mathcal{G}}(b)$. We say that w is an *ancestor* of v if either $v = w$ or there is a sequence of directed edges $w \rightarrow \cdots \rightarrow v$. The set of ancestors of v is denoted $\mathrm{an}_{\mathcal{G}}(v)$. These definitions are applied disjunctively to sets of vertices so that, for example, $\mathrm{pa}_{\mathcal{G}}(A) \equiv \bigcup_{a \in A} \mathrm{pa}_{\mathcal{G}}(a)$. An *ancestral* set is one that contains all its own ancestors: $\mathrm{an}_{\mathcal{G}}(A) = A$.

Note that the definitions of parents, children and ancestors do not distinguish between random and fixed vertices. A *random-ancestral* set, $A' \subseteq V$, is a set of random vertices such that $\mathrm{an}_{\mathcal{G}}(A') \subseteq A' \cup W$; that is, all the random ancestors of A' are contained in A' itself.

A set of vertices B is said to be *sterile* if it does not contain any of its children: equivalently $\mathrm{pa}_{\mathcal{G}}(B) \cap B = \varnothing$. The *sterile subset* of a set $C \subseteq V$ is $\mathrm{sterile}_{\mathcal{G}}(C) \equiv C \setminus \mathrm{pa}_{\mathcal{G}}(C)$ (sometimes called the set of 'sink nodes' in the induced subgraph on C).

Example 2.3. Consider the CADMG $\mathcal{L}$ in Figure 3. The set of parents of the vertex 3 is $\mathrm{pa}_{\mathcal{L}}(3) = \{2\}$, and the set of ancestors is $\mathrm{an}_{\mathcal{L}}(3) = \{1, 2, 3\}$; hence $\{1, 2, 3\}$ is ancestral, and $\{2, 3\}$ is random-ancestral. The set $\{2, 4, 5\}$ is sterile, but $\{2, 3, 5\}$ is not.

Definition 2.4. A set of random vertices $B \subseteq V$ is *bidirected-connected* if for each $a, b \in B$ there is a sequence of edges $a \leftrightarrow \cdots \leftrightarrow b$ with all intermediate vertices in B. A maximal bidirected-connected set is a *district* of the graph $\mathcal{G}$ (sometimes called a *c-component*). The set of districts is a partition of the random vertices of a graph; the district containing $v \in V$ is denoted $\mathrm{dis}_{\mathcal{G}}(v)$.

We draw bidirected edges in red, which makes it easy to identify districts as the maximal sets connected by red edges. In Figure 1(b) for example, there are three districts: $\{X\}$, $\{M\}$, and $\{E, Y\}$. In Figure 3, there are two: $\{3\}$ and $\{2, 4, 5\}$.

2.1. Transformations

We now introduce two operations that transform CADMGs by removing vertices: the first separates into districts and the second one forms ancestral subgraphs. We will use these transformations to define our Markov property (and thereby our statistical model) in Section 3.
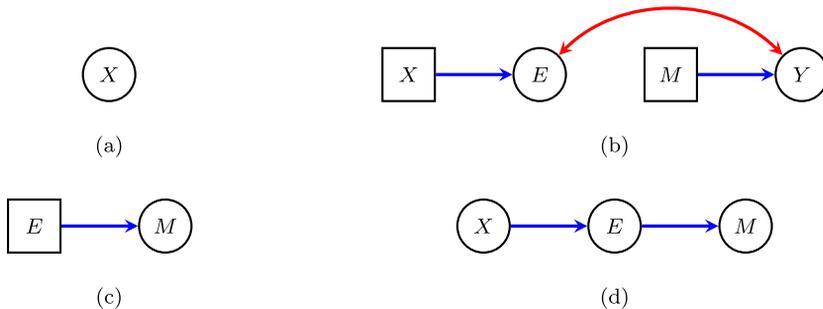
Figure 4. Four reachable subgraphs of the graph in Figure 1(b). Graphs in (a), (b) and (c) correspond to factorization into the districts $\{X\}$, $\{E, Y\}$ and $\{M\}$ respectively. Graph (d) corresponds to marginalizing the childless node Y.

Definition 2.5. Let $\mathcal{G}$ be a CADMG containing a district D. Define $\mathfrak{d}_D(\mathcal{G})$ to be the CADMG with: the set of random vertices D; the set of fixed vertices $\text{pa}_{\mathcal{G}}(D) \setminus D$; the set of bidirected edges whose endpoints are both in D in $\mathcal{G}$; the set of directed edges from $\mathcal{G}$ pointing to a vertex in D (including directed edges between vertices in D).

Let A be a random-ancestral set in $\mathcal{G}$. Define $\mathfrak{m}_A(\mathcal{G})$ to be the graph with the set of random vertices A, the set of fixed vertices $\text{pa}_{\mathcal{G}}(A) \setminus A$, and all edges between these vertices that are in $\mathcal{G}$. Note that, since A is random-ancestral, by definition the vertices in $\text{pa}_{\mathcal{G}}(A) \setminus A$ are already fixed vertices in $\mathcal{G}$.

If a graph $\mathcal{G}'$ can be obtained from $\mathcal{G}$ by iteratively applying operations of the form $\mathfrak{d}$ and $\mathfrak{m}$, we say that $\mathcal{G}'$ is *reachable* from $\mathcal{G}$.

Note that if we start with a graph $\mathcal{G}$ in which all the fixed vertices $w \in W$ have at least one child, then this is also true of the graph obtained after applying either $\mathfrak{m}_A$ or $\mathfrak{d}_D$.

Example 2.6. The graph in Figure 1(b) contains the districts $\{X\}$, $\{E, Y\}$ and $\{M\}$. The corresponding graphs $\mathfrak{d}_D(\mathcal{G})$ are given in Figure 4(a), (b) and (c), respectively. The sets $\{X, E, M\}$, $\{X, E\}$ and $\{X\}$ are ancestral in $\mathcal{G}$, and the graphs $\mathfrak{m}_{\{X\}}(\mathcal{G})$ and $\mathfrak{m}_{\{X,E,M\}}(\mathcal{G})$ are shown in Figure 4(a) and (d), respectively.

Example 2.7. The graph in Figure 3 contains the district $\{2, 4, 5\}$, and $\mathfrak{d}_{\{2,4,5\}}(\mathcal{L})$ gives us the graph in Figure 5(a). The sets $\{2, 4\}$ and $\{4, 5\}$ are both random-ancestral in $\mathfrak{d}_{\{2,4,5\}}(\mathcal{L})$, so we can apply either $\mathfrak{m}_{\{2,4\}}$ or $\mathfrak{m}_{\{4,5\}}$ to obtain the CADMGs in Figure 5(b) and (c), respectively.

As we will see in the next section, both of these graphical operations correspond to an operation on a probability distribution we associate with the graph: $\mathfrak{m}_A$ to marginalization, and $\mathfrak{d}_D$ to a factorization. The 'fixing' operation described in Richardson et al. [17] unifies and generalizes $\mathfrak{m}$ and $\mathfrak{d}$, but the statistical model we will describe is ultimately the same. For the purposes of defining a parameterization, it is more convenient to use the formulation given here.

It is important to note that sets may become districts or random ancestral sets after several iterations of $\mathfrak{m}$ and $\mathfrak{d}$. For example, $\{2, 4\}$ is not random-ancestral in $\mathcal{L}$, but it is in
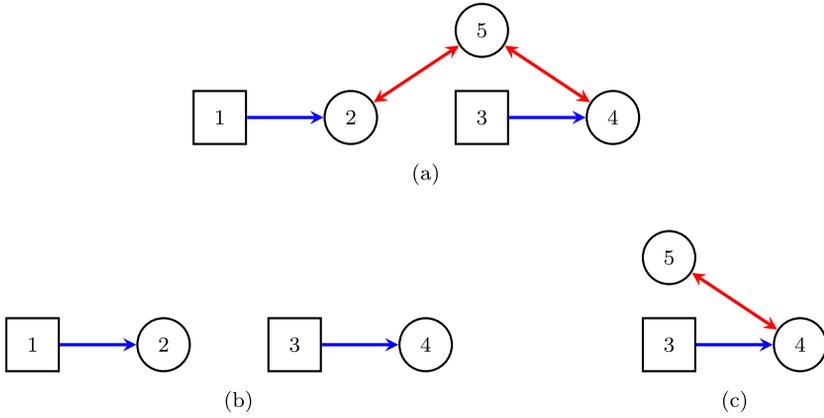
Figure 5. Three CADMGs reachable from the graph in Figure 3.

$\mathfrak{d}_{\{2,4,5\}}(\mathcal{L})$. Similarly, $\{4\}$ is not a district in $\mathfrak{d}_{\{2,4,5\}}(\mathcal{L})$, but it is in $\mathfrak{m}_{\{2,4\}}(\mathfrak{d}_{\{2,4,5\}}(\mathcal{L}))$; see Figure 5(b).

We now give a characterization of what reachable graphs look like.

Definition 2.8. Let $\mathcal{G}$ be a CADMG with random vertex set V. Given $C \subseteq V$ the graph $\mathcal{G}[C]$ is defined to be the CADMG with the set of random vertices C, fixed vertices $\text{pa}_{\mathcal{G}}(C) \setminus C$, those bidirected edges in $\mathcal{G}$ with both endpoints in C, and those directed edges that are directed from $C \cup \text{pa}_{\mathcal{G}}(C)$ to C.

In other words, $\mathcal{G}[C]$ is the subgraph containing precisely the edges whose arrowheads are all in C. For example, if $\mathcal{G}$ is the graph in Figure 1(b), then Figure 4(a)–(d) corresponds to $\mathcal{G}[\{X\}]$, $\mathcal{G}[\{E, Y\}]$, $\mathcal{G}[\{M\}]$ and $\mathcal{G}[\{X, E, M\}]$, respectively.

Lemma 2.9. *Suppose that the graph $\mathcal{G}'$ is reachable from $\mathcal{G}$ and has set of random vertices C. Then $\mathcal{G}' = \mathcal{G}[C]$.*

Proof. Since we assume all fixed vertices have at least one child, then $\mathcal{G} = \mathcal{G}[V]$. In addition, it is clear from the definitions of $\mathfrak{d}$ and $\mathfrak{m}$ that precisely the edges and fixed vertices mentioned are preserved at each step.   □

In the rest of this paper, we will only refer to $\mathcal{G}[C]$ if C is a reachable set, though Definition 2.8 in principle applies to any $C \subseteq V$. Unfortunately, there is generally no simple way of characterizing which sets C correspond to reachable subgraphs without iteratively applying $\mathfrak{d}$ and $\mathfrak{m}$ as defined above. If a set A is random-ancestral, then clearly $\mathcal{G}[A]$ is reachable just by applying $\mathfrak{m}$. Note that Richardson et al. [17] use a more general definition of reachable sets.

3. Nested Markov property

Graphical models relate the structure of a graph to a collection of joint probability distributions over a set of random variables. We will work with the nested Markov property, which relates a (C)ADMG and each of its reachable subgraphs to a collection of probability distributions over random vertices, indexed by fixed vertices.

Suppose we are interested in random variables X_v taking values in a finite discrete set $\mathfrak{X}_v$. For a set of vertices C, let $\mathfrak{X}_C \equiv \times_{v \in C} \mathfrak{X}_v$. A *probability kernel for V given W* (or simply a kernel) is a function $p_{V|W} : \mathfrak{X}_V \times \mathfrak{X}_W \to [0, 1]$ such that for each $x_W \in \mathfrak{X}_W$,

$$\sum_{x_V \in \mathfrak{X}_V} p_{V|W}(x_V \mid x_W) = 1.$$

In other words, a kernel behaves like a conditional probability distribution for X_V given X_W. We use the word 'kernel' to emphasize that some of the conditional distributions we obtain are not equal to the usual conditional distribution obtained from elementary definitions, but instead correspond to certain interventional quantities.

In what follows, $\dot{\cup}$ is used to denote a union of disjoint sets.

Definition 3.1. Let $p_{V|W}$ be a kernel, and let $A \mathbin{\dot{\cup}} B \mathbin{\dot{\cup}} C = V$. The *marginal kernel* over $A, B \mid W$ is defined to be:

$$p_{AB|W}(x_A, x_B \mid x_W) \equiv \sum_{x_C} p_{V|W}(x_V \mid x_W).$$

It is easy to check that $p_{AB|W}$ is also a kernel. A (version of the) *conditional kernel* of $A|B, W$ is any kernel $p_{A|BW}$ satisfying

$$p_{A|BW}(x_A \mid x_B, x_W) \cdot p_{B|W}(x_B \mid x_W) \equiv p_{AB|W}(x_A, x_B \mid x_W).$$

This is uniquely defined precisely for x_B, x_W such that $p_{B|W}(x_B \mid x_W) > 0$.

Remark 3.2. Note that, for convenience, if some of the fixed variables $W^* \subseteq W$ in a kernel $p_{V|W}$ are entirely irrelevant, (i.e. if the functions $p_{V|W}(\cdot \mid \cdot, y_{W^*})$ are identical for all $y_{W^*} \in \mathfrak{X}_{W^*}$) we will describe it interchangably as a kernel of V given W, and as a kernel of V given $W \setminus W^*$, since in this case these objects are isomorphic: $p_{V|W} = p_{V|W\setminus W^*}$.

We are now in a position to define the nested model. The definition is recursive, and works by reference to the model applied iteratively to smaller and smaller graphs. The model is introduced in Richardson et al. [17], and is based on the constraint finding algorithm of Tian and Pearl [28], which follows a similar recursive structure.

Definition 3.3. Let $\mathcal{G}$ be a CADMG and $p_{V|W}$ a probability kernel. Say that $p_{V|W}$ *recursively factorizes* according to $\mathcal{G}$, and write $p_{V|W} \in \mathcal{M}_{rf}(\mathcal{G})$ if either $|V| = 1$, or both:

(i) if $\mathcal{G}$ has districts $D_1, \ldots, D_k$, $k \geq 1$, then

$$p_{V|W}(x_V \mid x_W) = \prod_i r_i(x_{D_i} \mid x_{\mathrm{pa}(D_i)\setminus D_i}) \tag{2}$$

and where, if $k \geq 2$, each r_i recursively factorizes according to $\mathcal{G}[D_i] = \mathfrak{d}_{D_i}(\mathcal{G})$; and

(ii) for each ancestral set A with $V \setminus A \neq \varnothing$, the marginal distribution

$$p_{A\cap V|W}(x_{A\cap V} \mid x_W) = \sum_{x_{V\setminus A}} p_{V|W}(x_V \mid x_W)$$

does not depend upon $x_{W\setminus A}$ (so we denote it by $p_{A\cap V|A\cap W}$ in line with Remark 3.2), and this recursively factorizes according to $\mathcal{G}[V \cap A] = \mathfrak{m}_{V\cap A}(\mathcal{G})$.

Given a graph $\mathcal{G}$, we shall refer to $\mathcal{M}_{rf}(\mathcal{G})$ as the *nested model* associated with $\mathcal{G}$, and say that distributions in that set satisfy the *nested Markov property* with respect to $\mathcal{G}$. There are other, equivalent definitions: see Richardson et al. [17].

Remark 3.4. It is important to note that, in terms of the factors r_i whose existence is implied by condition (i), the definition of recursive factorization 'starts from scratch' each time we perform the recursion. For example, we make no claim (yet) about the connection between a factor r_i obtained from (i) and any such factors which arise after first applying (ii) and then later (i): see Example 3.5.

In the base case $V = \{v\}$ the definition places no restriction on the distribution of X_v given X_W (recall that, by assumption, all fixed vertices have at least one random child). The observed distribution obtained from a directed acyclic graph model with latent variables will satisfy conditions (i) and (ii) with respect to the ADMG that is the latent projection of that DAG (Tian and Pearl [27,28]). The models defined by the Markov properties for ADMGs introduced by Richardson [16] and parameterized by Evans and Richardson [9,10] can be defined by replacing (i) with the weaker requirement:

(i′) $\mathcal{G}$ has districts $D_1, \ldots, D_k$, $k \geq 1$, and $p_{V|W} = \prod_i r_i$ where each r_i is a kernel for D_i given $\mathrm{pa}_{\mathcal{G}}(D_i) \setminus D_i$.

In other words, although the distribution must satisfy the ancestrality condition (ii) and then factorize, no further conditions are imposed on those factors: they are not required to obey any additional constraints implied by the graph $\mathcal{G}[D_i]$. This leads to a model defined entirely by conditional independence relations on the original joint distribution $p_{V|W}$.

As a consequence of this, the m-separation criterion for ordinary Markov models (as well as the other Markov properties described by Richardson [16]) can be applied correctly to the initial ADMG $\mathcal{G}$ to derive conditional independences in $p(x_V)$; however, applied solely to $\mathcal{G}$, this does not completely describe the nested model.

Example 3.5. Consider the CADMG in Figure 1(b). Criterion (i) of recursive factorization requires that

$$p(x, e, m, y) = r_X(x) \cdot r_{EY}(e, y \mid x, m) \cdot r_M(m \mid e)$$

for distributions r_X, r_{EY} and r_M which recursively factorize according to the graphs in Figure 4(a), (b) and (c), respectively.

On the other hand, if we apply condition (ii) to the childless node Y we see that the margin $p(x, e, m)$ must satisfy recursive factorization with respect to the DAG in Figure 4(d), so

$$p(x, m, e) = \tilde{r}_X(x) \cdot \tilde{r}_E(e \mid x) \cdot \tilde{r}_M(m \mid e)$$

for some kernels $\tilde{r}_X$, $\tilde{r}_M$ and $\tilde{r}_E$. This factorization implies the conditional independence $X \perp\!\!\!\perp M \mid E$, which can also be deduced using m-separation. We add a tilde to the kernels to emphasise that the *definition* starts afresh at each iteration, and makes no claim of any relationship between this factorization and the factorization of $p = r_X r_{EY} r_M$. However, it is not hard to verify that in this case

$$r_X(x) = \tilde{r}_X(x) = p(x),$$

$$r_M(m \mid e) = \tilde{r}_M(m \mid e) = p(m \mid e),$$

$$\sum_y r_{EY}(e, y \mid x, m) = \tilde{r}_E(e \mid x) = p(e \mid x).$$

In fact, it will follow from Theorem 5.4 that, in general, kernels such as r_X and $\tilde{r}_X$ that have the same random vertex set but are derived in different ways are equal under the model. Note that

$$r_{EY}(e, y \mid x, m) = p(e \mid x) \cdot p(y \mid x, m, e)$$

$$\neq p(e \mid x, m) \cdot p(y \mid x, m, e)$$

$$= p(e, y \mid x, m),$$

and so r_{EY} is *not* the usual conditional distribution of E, Y given X, M.

3.1. Properties of the recursive kernels

Here we show that the kernels r_i from (2) in Definition 3.3 are products of conditional distributions derived from $p_{V|W}$ at the current level of the recursion, and that they are uniquely defined up to versions of those conditional distributions.

A *topological ordering* of the random vertices of a CADMG is a total ordering $<$ on V such that every vertex precedes its children. We denote by $\mathrm{pre}_<(v)$ the set of (random) vertices which precede v under $<$.

The following proposition shows that the factors in the definition of recursive factorization are unique up to versions of conditional distributions.

Proposition 3.6. *Let $\mathcal{G}$ be a CADMG with districts $D_1, \ldots, D_k$, and let $<$ be any topological ordering of V. Let $p_{V|W} = \prod_i r_i$, where each r_i recursively factorizes with respect to $\mathcal{G}[D_i]$. Then*

$$r_i(x_{D_i} \mid x_{\mathrm{pa}_{\mathcal{G}}(D_i) \setminus D_i}) = \prod_{v \in D_i} p_{v| \mathrm{pre}_<(v) \cup W}(x_v \mid x_{\mathrm{pre}_<(v)}, x_W), \tag{3}$$

where $p_{v| \mathrm{pre}_<(v) \cup W}$ is any $p_{V|W}$-version of the conditional distribution of $X_v | X_{\mathrm{pre}_<(v)}, X_W$.

Remark 3.7. The equation in (3) is an instance of the *g-formula* of Robins [18]. The result also appears as Corollary 1 in Tian [26], Section 4.3, in the case of latent variable models.

Proof. For the purposes of induction, we generalize the result slightly to allow D_i to be collections of several districts. Let $E_i \equiv \mathrm{pa}_\mathcal{G}(D_i) \setminus D_i$. We proceed by induction on $|V|$: if $|V| \leq 1$ there is nothing to show. Otherwise, let $t \in D_k$ be the last vertex in the ordering $<$, so that x_t only appears as a variable in the factor r_k. Then

$$p_{V\setminus\{t\}|W}(x_{V\setminus\{t\}} \mid x_W) \equiv \sum_{x_t} p_{V|W}(x_V \mid x_W)$$

$$= \sum_{x_t} \prod_{i=1}^{k} r_i(x_{D_i} \mid x_{E_i})$$

$$= \left(\sum_{x_t} r_k(x_{D_k} \mid x_{E_k}) \right) \prod_{i=1}^{k-1} r_i(x_{D_i} \mid x_{E_i})$$

$$= \tilde{r}_k(x_{D_k\setminus\{t\}} \mid x_{E_k}) \prod_{i=1}^{k-1} r_i(x_{D_i} \mid x_{E_i}),$$

where, by property 1 of recursive factorization, the kernel $\tilde{r}_k$ recursively factorizes with respect to the graph $\mathcal{G}[D_k \setminus \{t\}]$. Similarly, all the factors r_i for $i = 1, \ldots, k-1$ recursively factorize with respect to $\mathcal{G}[D_i]$, so by the induction hypothesis each such r_i is of the required form (3), and

$$\tilde{r}_k(x_{D_k\setminus\{t\}} \mid x_{E_k}) = \prod_{v \in D_k \setminus \{t\}} p_{v|\mathrm{pre}_<(v)\cup W}(x_v \mid x_{\mathrm{pre}_<(v)}, x_W).$$

But then

$$\prod_i r_i = p_{V|W} = p_{t|W,V\setminus\{t\}} \cdot p_{V\setminus\{t\}|W} = p_{t|W,V\setminus\{t\}} \cdot \tilde{r}_k \cdot \prod_{i=1}^{k-1} r_i;$$

therefore whenever $p_{V\setminus\{t\}|W} > 0$

$$r_k(x_{D_k} \mid x_{E_k}, x_W) = p_{t|W,V\setminus\{t\}}(x_t \mid x_W, x_{V\setminus\{t\}}) \cdot \tilde{r}_k. \tag{4}$$

Hence $p_{t|W,V\setminus\{t\}}$ satisfies (4) if and only if it is a version of the relevant conditional distribution, as required. $\qquad\square$

The next result shows that the positivity of $p_{V|W}$ is preserved in any derived kernels.

Lemma 3.8. *Let* $p_{V|W}(x_V \mid x_W)$ *be a probability distribution,* $<$ *some total ordering on* V*, and let* $A \subseteq V$ *and* $B \equiv W \cup \mathrm{pre}_<(A) \setminus A$*. Define*

$$r_{A|B}(x_A \mid x_B) \equiv \prod_{v \in A} p_{v|\mathrm{pre}_<(v),W}(x_v \mid x_{\mathrm{pre}_<(v)}, x_W),$$

for some versions $p_{v|\mathrm{pre}_<(v),W}$ *of the conditional distributions of* $X_v \mid X_W, X_{\mathrm{pre}_<(v)}$.

Then:

(a) $r_{A|B}$ is a kernel for $X_A \mid X_B$;
(b) *for any* $T \subseteq V$, $x_T \in \mathfrak{X}_T$ *and* $x_W \in \mathfrak{X}_W$, *if* $p_{T|W}(x_T \mid x_W) > 0$ *then*

$$r_{T \cap A|B}(x_{T \cap A} \mid x_B) \equiv \sum_{y_{A \setminus T}} r_{A|B}(y_{A \setminus T}, x_{T \cap A} \mid x_B) > 0$$

and all versions of $r_{T \cap A|B}(x_{T \cap A} \mid x_B)$ *are the same*;
(c) *if* $p_{T|W}(x_T \mid x_W) = 0$ *then there exists* $t \in T$ *such that* (*every version of*)

$$p_{t|\text{pre}_<(t),W}(x_t \mid x_{\text{pre}_<(t)}, x_W) = 0.$$

Proof. (a) Clearly $r_{A|B} \geq 0$ since it is a product of conditional distributions, which are themselves non-negative. In addition, by summing the expression above in reverse order of $<$ it is easy to see that $\sum_{x_A} r_{A|B}(x_A \mid x_B) = 1$ for any $x_B \in \mathfrak{X}_B$. Hence, $r_{A|B}$ is a kernel.

For (b), note that if $p_{T|W}(x_T \mid x_W) > 0$, then there exists some $x_{V \setminus T} \in \mathfrak{X}_{V \setminus T}$ such that $p_{V|W}(x_V \mid x_W) > 0$. Then

$$p_{V|W}(x_V \mid x_W) = \prod_{v \in V} p_{v|\text{pre}_<(v),W}(x_v \mid x_{\text{pre}_<(v)}, x_W)$$

$$= r_{A|B}(x_A \mid x_B) \prod_{v \in V \setminus A} p_{v|\text{pre}_<(v),W}(x_v \mid x_{\text{pre}_<(v)}, x_W),$$

so if the left-hand side is positive then so is $r_{A|B}(x_A \mid x_B) > 0$. Since all the events in this expression have positive $p_{V|W}$ probability, all versions of each conditional probability are equal.

Lastly, if $p_{T|W}(x_T \mid x_W) = 0$ then clearly some factor of

$$0 = p_{T|W}(x_T \mid x_W) = \prod_{t \in T} p_{t|\text{pre}_<(t),W}(x_t \mid x_{\text{pre}_<(t)}, x_W)$$

is also zero. Pick the $<$-minimal t such that this holds, so that $p_{\text{pre}_<(t)|W}(x_{\text{pre}_<(t)} \mid x_W) > 0$. Then (c) holds. □

A corollary of this lemma is the following.

Corollary 3.9. *Let* $p_{V|W} \in \mathcal{M}_{rf}(\mathcal{G})$ *be a strictly positive kernel. Then any kernel derived from* $p_{V|W}$ *by repeated applications of Definition* 3.3 (*using* $\mathcal{G}$) *is uniquely defined.*

Proof. Clearly applying (ii) is always unique, since it only involves summing. By Proposition 3.6, application of (i) is a factorization into univariate conditional distributions, each of which is uniquely defined when the joint distribution is positive. In addition, by Lemma 3.8 each such conditional distribution is also strictly positive, so following the recursion with each unique factor gives the result. □

4. Intrinsic sets and partitions

In this section, we provide the necessary theory to link the graphical notions of Section 3 to the parameterization in Section 5. The parameterization uses factorizations of the distribution into pieces that correspond to special subsets of vertices in the graph; these subsets are themselves derived from the idea of the 'reachable' sets already introduced.

Definition 4.1. Let $\mathcal{G}$ be a CADMG. A non-empty set S of random vertices is *intrinsic* if it is bidirected-connected and the graph $\mathcal{G}[S]$ is reachable from $\mathcal{G}$.

For each intrinsic set S, define the associated *recursive head* by $\mathrm{rh}_{\mathcal{G}}(S) = \mathrm{sterile}_{\mathcal{G}}(S)$; that is, it is the set of sink nodes in the induced subgraph over S. The set of recursive heads is denoted by $\mathcal{H}(\mathcal{G})$, or simply $\mathcal{H}$.[1]

The *tail* associated with a recursive head H (and the relevant intrinsic set S) is $T(H) \equiv \mathrm{pa}_{\mathcal{G}}(S)$. We will denote a tail by T if it is unambiguous which recursive head it is derived from.

Intrinsic sets are central to the nested Markov property as they are the sets of variables over which the kernels r_i in Definition 3.3 specify distributions. Intrinsic sets do not appear to be easily characterized in terms of the presence of a path in the original graph; Definition 4.1 implicitly considers a sequence of graphs generated via repeated applications of the two operations $\mathfrak{d}$ and $\mathfrak{m}$. The set of intrinsic sets may be found in polynomial time; see Shpitser et al. [22].

Example 4.2. For the graph $\mathcal{L}$ in Figure 3, $\{2, 4, 5\}$ and $\{3\}$ are districts and therefore intrinsic sets. The graph $\mathcal{L}[\{2, 4, 5\}]$ is shown in Figure 5(a); applying $\mathfrak{m}$ appropriately to random-ancestral sets yields all the other intrinsic sets: $\{2, 5\}$, $\{4, 5\}$, $\{2\}$, $\{4\}$ and $\{5\}$. Each recursive head is equal to the associated intrinsic set.

Definition 4.3. Let $B \subseteq V$ be a set of random vertices in $\mathcal{G}$. Suppose we alternately marginalize vertices that are not ancestors of B, and remove those which are not in the same district as some element of B:

$$\mathcal{G} \mapsto \mathfrak{m}_{\mathrm{an}_{\mathcal{G}}(B)}(\mathcal{G}), \qquad \mathcal{G} \mapsto \mathfrak{d}_{\mathrm{dis}_{\mathcal{G}}(B)}(\mathcal{G}). \qquad (5)$$

If these two operations change anything at all then they reduce the size of the set of random vertices; consequently repeatedly applying both these operations successively will eventually reach some stable point, which is a graph whose set of random vertices we denote by $I_{\mathcal{G}}(B)$. Note that at each step of (5) the random vertices in the resulting graph always include B, so $B \subseteq I_{\mathcal{G}}(B)$.

If $I_{\mathcal{G}}(B)$ is bidirected-connected, then it is an intrinsic set by definition, and we call $I_{\mathcal{G}}(B)$ the *intrinsic closure* of B.

---

[1]Note that the definition of a recursive head differs from the *head* used in Evans and Richardson [10] for ADMGs. We will see in Example 4.12 that $\{E, Y\}$ is a recursive head in the graph in Figure 1(b), but one can check that it is not a head in the Evans and Richardson [10] sense.

(a)   (b)

Figure 6. (a) A (C)ADMG $\mathcal{G}$ and (b) $\mathcal{G}_1 \equiv \eth_{\mathrm{dis}(Y)}(\mathcal{G})$.

Proposition 4.4. *If $\mathcal{G}' = \mathcal{G}[C]$ is reachable from $\mathcal{G}$ for some set $C \supseteq B$, then*

$$\mathfrak{m}_{\mathrm{an}_{\mathcal{G}'}(B)}\big(\mathcal{G}'\big) \subseteq \mathfrak{m}_{\mathrm{an}_{\mathcal{G}}(B)}(\mathcal{G}), \qquad \eth_{\mathrm{dis}_{\mathcal{G}'}(B)}\big(\mathcal{G}'\big) \subseteq \eth_{\mathrm{dis}_{\mathcal{G}}(B)}(\mathcal{G}).$$

Note that here and in what follows we use $\subseteq$ as a subgraph relation when applied to graphs.

Proof. From Lemma 2.9, $\mathcal{G}' = \mathcal{G}[C]$; We have $\mathfrak{m}_{\mathrm{an}_{\mathcal{G}'}(B)}(\mathcal{G}') = \mathcal{G}[\mathrm{an}_{\mathcal{G}'}(B)]$ and $\mathfrak{m}_{\mathrm{an}_{\mathcal{G}}(B)}(\mathcal{G}) = \mathcal{G}[\mathrm{an}_{\mathcal{G}}(B)]$. Any ancestor of B in the subgraph $\mathcal{G}' = \mathcal{G}[C]$ must be also be an ancestor in $\mathcal{G}$, so clearly $\mathcal{G}[\mathrm{an}_{\mathcal{G}'}(B)] \subseteq \mathcal{G}[\mathrm{an}_{\mathcal{G}}(B)]$. A similar argument holds for $\eth$. □

Both of the operators in (5) are idempotent; in addition, since the sets $\mathrm{an}(B)$ and $\mathrm{dis}(B)$ only get smaller through repeated iterations, it follows from Proposition 4.4 that the stable point does not depend upon which operation is applied first. Hence, $I_{\mathcal{G}}(B)$ is well-defined.

Example 4.5. Let $\mathcal{G}$ be the graph in Figure 6(a) and consider the intrinsic closure of the bidirected-connected set $\{Y\}$. The graph $\mathfrak{m}_{\mathrm{an}(Y)}(\mathcal{G})$ is just $\mathcal{G}$, since everything is an ancestor of Y. However $\mathcal{G}_1 \equiv \eth_{\mathrm{dis}(Y)}(\mathcal{G})$ gives the graph $\mathcal{G}[\{X, Y\}]$ shown in Figure 6(b) in which Z is fixed, but the edges are all unchanged. It then becomes clear that repeatedly applying $\mathfrak{m}$ and $\eth$ will not result in any further changes to the graph. Hence, the intrinsic closure is just the set of random vertices in this graph: $I_{\mathcal{G}}(\{Y\}) = \{X, Y\}$.

On the other hand, consider the graph $\mathcal{L}$ in Figure 3 and the intrinsic closure of the set $\{4, 5\}$. Again $\mathfrak{m}_{\mathrm{an}(\{4,5\})}(\mathcal{L}) = \mathcal{L}$, and then $\eth_{\mathrm{dis}(\{4,5\})}(\mathcal{L})$ gives the graph in Figure 5(a). Applying $\mathfrak{m}_{\mathrm{an}(\{4,5\})}(\cdot)$ to this graph yields the graph in Figure 5(c), whose only random vertices are $\{4, 5\}$. Hence, the procedure terminates and, since it forms a district in this graph, $\{4, 5\}$ is an intrinsic set and its own intrinsic closure.

One consequence of the next result is that, as we would hope, every intrinsic set is its own intrinsic closure.

Lemma 4.6. *Let S be an intrinsic set with recursive head H in a graph $\mathcal{G}$. Then for any set A such that $H \subseteq A \subseteq S$ we have $I_{\mathcal{G}}(A) = S$.*

Proof. By the definition of H, every vertex in S is either in H or is a parent of some other element of S. Since S is bidirected-connected, the operations $\eth_A$, $\mathfrak{m}_A$ therefore cannot remove any element of S without also having removed an element of H, but this is not allowed since $H \subseteq A$. Hence, no element of S is ever removed, and $I_{\mathcal{G}}(A) \supseteq S$.

Suppose that $I_{\mathcal{G}}(A) \supset S$ and so $B \equiv I_{\mathcal{G}}(A) \setminus S$ is non-empty. Every element of B is an ancestor of some other entry in $I_{\mathcal{G}}(A)$. In addition, every element of $I_{\mathcal{G}}(A)$ is connected to $A \subseteq S$ by sequences of bidirected edges through $I_{\mathcal{G}}(A)$, so $I_{\mathcal{G}}(A)$ is, like S, a bidirected-connected set. Thus, we cannot remove any element of B via operations of the form $\mathfrak{m}, \mathfrak{d}$ without first removing some element of $A \subseteq S$. If B is non-empty, then this implies S is not reachable, which contradicts the assumption that S is intrinsic. □

Note that a corollary of this result is that recursive heads are in one-to-one correspondence with intrinsic sets: two distinct intrinsic sets may not have the same recursive head.

Proposition 4.7. *If B is a bidirected-connected set with intrinsic closure $I_{\mathcal{G}}(B)$, then the recursive head H associated with the intrinsic set $I_{\mathcal{G}}(B)$ satisfies $H \subseteq B$.*

Proof. By definition of intrinsic closure, every vertex v in $I_{\mathcal{G}}(B)$ is an ancestor of B in $\mathcal{G}[I_{\mathcal{G}}(B)]$. If $v \notin B$, then $v \notin \text{sterile}_{\mathcal{G}}(I_{\mathcal{G}}(B))$, hence $v \notin H$. □

Lemma 4.8. *Every singleton $\{v\}$ for $v \in V$ is a recursive head.*

Proof. Take the intrinsic closure $I_{\mathcal{G}}(\{v\})$ of v. Every element of $I_{\mathcal{G}}(\{v\})$ other than v is a parent of some other element of $I_{\mathcal{G}}(\{v\})$ by definition; therefore $\{v\}$ is the sterile set, and a recursive head. □

Lemma 4.9. *Let $\mathcal{G}$ be a CADMG, and $\mathcal{G}'$ be a CADMG with random vertices V', reachable from $\mathcal{G}$. Then the intrinsic sets of $\mathcal{G}'$ are precisely the intrinsic sets of $\mathcal{G}$ that are contained in V', and their associated recursive heads and tails are the same.*

Proof. Since $\mathcal{G}' = \mathcal{G}[V']$ is reachable from $\mathcal{G}$, any intrinsic set in $\mathcal{G}'$ is also an intrinsic set in $\mathcal{G}$. For the converse, suppose that $D \subseteq V'$ is an intrinsic set in $\mathcal{G}$. Take the intrinsic closure of D in $\mathcal{G}'$, say C; if $C = D$ then we are done.

Suppose not, so that $C \setminus D$ is non-empty. This occurs precisely when C is bidirected-connected in $\mathcal{G}'$, and every vertex in $C \setminus D$ is an ancestor in $\mathcal{G}'$ of some other vertex in C. But if this is true in $\mathcal{G}'$, then it must also be true in $\mathcal{G}$, which contains any edges that $\mathcal{G}'$ does; thus the intrinsic closure of D in $\mathcal{G}$ is a strict superset of D. This contradicts the assumption that D is intrinsic in $\mathcal{G}$.

By Lemma 2.9 the recursive heads and tails associated with each intrinsic set are unchanged, since the parent sets of each random vertex are preserved. □

Corollary 4.10. *Let $\mathcal{G}$ be a CADMG containing random-ancestral sets A_1, A_2. If $H \in \mathcal{H}(\mathcal{G}[A_1])$ and $H \in \mathcal{H}(\mathcal{G}[A_2])$, then $H \in \mathcal{H}(\mathcal{G}[A_1 \cap A_2])$.*

Proof. If A_1 and A_2 are random-ancestral, then so is $A_1 \cap A_2$, so the graph $\mathcal{G}[A_1 \cap A_2]$ is reachable from $\mathcal{G}$. The result follows from Lemma 4.9. □

4.1. Partitions

We follow the approach of Evans and Richardson [10] by defining partitions of sets via appropriate collections of subsets. Define a partial ordering $\prec$ on recursive heads by $H_1 \prec H_2$ whenever $I_{\mathcal{G}}(H_1) \subset I_{\mathcal{G}}(H_2)$.

Definition 4.11. Define a function $\Phi_{\mathcal{G}}$ on sets of random vertices $C \subseteq V$ that 'picks out' the set of $\prec$-maximal recursive heads $H \in \mathcal{H}(\mathcal{G})$ that are subsets of C. That is,

$$\Phi_{\mathcal{G}}(C) \equiv \{H \in \mathcal{H} \mid H \subseteq C \text{ and } H \nprec H' \text{ for all other } H' \subseteq C, H' \in \mathcal{H}\}.$$

Define

$$\psi_{\mathcal{G}}(C) \equiv C \setminus \bigcup_{D \in \Phi_{\mathcal{G}}(C)} D.$$

Now recursively define a function $\llbracket \cdot \rrbracket_{\mathcal{G}}$ that partitions subsets of V: define $\llbracket \varnothing \rrbracket_{\mathcal{G}} = \varnothing$, and

$$\llbracket W \rrbracket_{\mathcal{G}} \equiv \Phi_{\mathcal{G}}(W) \cup \llbracket \psi_{\mathcal{G}}(W) \rrbracket_{\mathcal{G}}.$$

For full details, including a proof that this definition does indeed define a partition, see the Appendix B.

Example 4.12. The recursive heads of the graph in Figure 1(b) are $\{X\}, \{E\}, \{M\}, \{Y\}, \{E, Y\}$, and the ordering requires that $\{E\}$ and $\{Y\}$ precede $\{E, Y\}$. Hence, for example

$$\llbracket \{X, E, Y\} \rrbracket_{\mathcal{G}} = \{\{X\}, \{E, Y\}\},$$

$$\llbracket \{M, Y\} \rrbracket_{\mathcal{G}} = \{\{M\}, \{Y\}\}.$$

The partitioning function $[\cdot]_{\mathcal{G}}$ in Evans and Richardson [10] made use of 'heads' rather than 'recursive heads', and therefore the partition obtained differs from the one here. For example, applied to the same graph as above,

$$[\{X, E, Y\}]_{\mathcal{G}} = \{\{X\}, \{E\}, \{Y\}\}.$$

Lemma 4.13. *If $\mathcal{G}' = \mathcal{G}[D]$ is reachable from $\mathcal{G}$ then $\llbracket C \rrbracket_{\mathcal{G}'} = \llbracket C \rrbracket_{\mathcal{G}}$ for every $C \subseteq D$.*

Proof. By Lemma 4.9, the intrinsic sets of $\mathcal{G}' = \mathcal{G}[D]$ are precisely the intrinsic sets of $\mathcal{G}$ that are subsets of D, with the same associated recursive heads. Hence the result follows from the definition of $\prec$. $\square$

Lemma 4.14. *If $\mathcal{G}$ is such that $V = D_1 \,\dot{\cup}\, D_2$ for sets D_1, D_2 not connected by bidirected edges, then*

$$\llbracket C \rrbracket_{\mathcal{G}} = \llbracket C \cap D_1 \rrbracket_{\mathcal{G}} \cup \llbracket C \cap D_2 \rrbracket_{\mathcal{G}}.$$

Proof. Since every intrinsic set (and therefore recursive head) is a subset of either D_1 or D_2, the result follows from Propositions B.4 and B.5 in the Appendix.   □

5. Parameterization

We are now in a position to introduce the parameterization. Recall that T denotes the tail associated with a recursive head H. We will present the parameterization for binary variables only, that is, those with state-space $\mathfrak{X}_v \equiv \{0, 1\}$, each $v \in V \dot{\cup} W$; the extension to non-binary discrete variables is conceptually simple but notationally cumbersome. Appendix C contains notes on the general case.

Definition 5.1. Let $\mathcal{G}$ be a CADMG with random vertices V and fixed vertices W. We say that $p_{V|W}$ is *parameterized according to* $\mathcal{G}$, and write $p_{V|W} \in \mathcal{M}_p(\mathcal{G})$, if it can be written in the form:

$$p_{V|W}(x_V \mid x_W) = \sum_{C:O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_\mathcal{G}} q_H(x_T), \qquad x_{VW} \in \mathfrak{X}_{VW}, \tag{6}$$

where we define $O \equiv O(x_V) \equiv \{v \in V \mid x_v = 0\}$. Here $q_H(x_T) \in \mathbb{R}$ for each $H \in \mathcal{H}$, $x_T \in \mathfrak{X}_T$, and $T \equiv T(H)$ is the tail associated with the recursive head H.

Note that if $C = \varnothing$ then the product is empty, which we define to be equal to 1. It will be shown in Section 5.3 that if $p_{V|W}$ is of the above form then $q_H(x_T) \in [0, 1]$ for all H and x_T, or can be chosen to be so. In fact, if the graph is interpreted causally, then each $q_H(x_T)$ is the same as $p_{H|T}(0_H \mid x_{T \setminus \tilde{T}}, \mathrm{do}(x_{\tilde{T}}))$, where $\tilde{T}$ is a suitable subset of T (see Theorem 5.5), and 0_H denotes $\bigcup_{h \in H} \{X_h = 0\}$.

5.1. Comparison to other graphical parameterizations

It is worth remarking on some special cases of the parameterization: if $\mathcal{G}$ is a DAG then each H is a singleton $\{h\}$, and (6) is just the familiar parameterization in terms of conditional probability tables using corner-point identifiability constraints: $q_H(x_T) = p_{h|\mathrm{pa}(h)}(0_h \mid x_{\mathrm{pa}(h)})$. If $\mathcal{G}$ has only bidirected edges, then $T = \varnothing$, and (6) reduces to the parameterization given in Drton and Richardson [6]. If $\mathcal{G}$ has a chain graph structure, that is, the districts can be ordered so that $v \to w$ only if v's district is strictly before w's, then the parameterization reduces to that given in Drton [4].

A comparison with the parameterization of Evans and Richardson [10] is more subtle. Since the ordinary Markov models in that paper only use the weaker requirement (i′) (see Section 3) we would expect that they generally have a larger dimension than the nested model for the same graph, and therefore use a different parameterization. If the models are the same, and if each intrinsic set can be obtained from a single marginalization step followed by factorization, then the 'ordinary' heads and tails will be the same as the recursive heads and tails, and hence the parameterization will be identical.
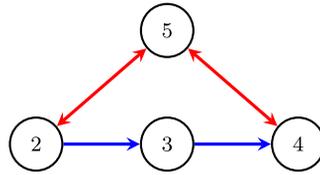
Figure 7. An ADMG whose nested and ordinary Markov models are the same, but for which the parameterizations of Evans and Richardson [10] and this paper are distinct.

However, even if the ordinary and nested models are the same, the parameterizations can be different. Consider the graph in Figure 7 (a modified version of $\mathcal{L}$). In this case, the ordinary and nested models are the same and both represent the distributions for which $X_5 \perp\!\!\!\perp X_3 \mid X_2$ and $X_4 \perp\!\!\!\perp X_2 \mid X_3$; this is the same as the corresponding maximal ancestral graph model. Since the set $\{2, 4, 5\}$ is a recursive head the nested parameterization includes the quantity $q_{245}(x_3) = P(X_2 = X_4 = X_5 = 0 \mid \text{do}(x_3))$ (see Theorem 5.5), whereas the ordinary parameterization does not have such a head, and uses only ordinary conditional probabilities such as $P(X_4 = 0, X_5 = 0 \mid x_2, x_3)$.

In general, the number of parameters in the nested model is no greater than the number in the ordinary Markov model, though this number can be quite large even for sparse graphs if the districts are large. The number of parameters for a particular district will be at least quadratic in the district size, this most parsimonious case occurring if the district is a single chain. The number of parameters may grow exponentially in the number of vertices, even for models with only a linear number of edges: for example, if we have a 'star' graph with all bidirected edges (this is equivalent to a star-shaped DAG with all edges pointing to the central node). Such large models are potentially undesirable, and methods to reduce the parameter count are suggested by Shpitser et al. [19].

5.2. Main results

We will show that distributions are parameterized according to $\mathcal{G}$ precisely when they recursively factorize according to $\mathcal{G}$, so that in fact $\mathcal{M}_{rf}(\mathcal{G}) = \mathcal{M}_p(\mathcal{G})$. In particular, a distribution of the form (6) satisfies properties (i) and (ii) of the recursive factorization. This is shown by the next two lemmas.

Lemma 5.2. *Let $\mathcal{G}$ be a CADMG with random vertices $V = D_1 \,\dot\cup\, \cdots \,\dot\cup\, D_l$, such that for $i \neq j$ there is no bidirected edge in $\mathcal{G}$ from a vertex in D_i to a vertex in D_j. Then for all $x_{VW} \in \mathfrak{X}_{VW}$ and $O \equiv O(x_V)$,*

$$\sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T) = \prod_{i=1}^{l} \sum_{O_i \subseteq C \subseteq D_i} (-1)^{|C \setminus O_i|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T),$$

where $O_i = O \cap D_i$.

Proof. We prove the result for $l = 2$, from which the general result follows by induction. From Lemma 4.14,

$$\prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T) = \prod_{H \in [\![C \cap D_1]\!]_{\mathcal{G}}} q_H(x_T) \times \prod_{H \in [\![C \cap D_2]\!]_{\mathcal{G}}} q_H(x_T).$$

In addition, if $C_i = C \cap D_i$, then $C \setminus O = (C_1 \setminus O_1) \cup (C_2 \setminus O_2)$ and this is the union of two disjoint sets, so $|C \setminus O| = |C_1 \setminus O_1| + |C_2 \setminus O_2|$. Hence,

$$\sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T)$$

$$= \sum_{O \subseteq C \subseteq D_1 \cup D_2} (-1)^{|C \setminus O|} \prod_{H \in [\![C \cap D_1]\!]_{\mathcal{G}}} q_H(x_T) \prod_{H \in [\![C \cap D_2]\!]_{\mathcal{G}}} q_H(x_T)$$

$$= \sum_{O_1 \subseteq C_1 \subseteq D_1} (-1)^{|C_1 \setminus O_1|} \prod_{H \in [\![C_1]\!]_{\mathcal{G}}} q_H(x_T)$$

$$\times \sum_{O_2 \subseteq C_2 \subseteq D_2} (-1)^{|C_2 \setminus O_2|} \prod_{H \in [\![C_2]\!]_{\mathcal{G}}} q_H(x_T). \qquad \square$$

Lemma 5.3. *Let $\mathcal{G}$ be a CADMG with a random vertex v. Then for all $x_{VW} \in \mathfrak{X}_{VW}$ and $O \equiv O(x_V)$,*

$$\sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T)$$

$$= \sum_{O \subseteq C \subseteq V \setminus \{v\}} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T) - \sum_{O \cup \{v\} \subseteq C \subseteq V} (-1)^{|C \setminus (O \cup \{v\})|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T).$$

Proof. Separating the sum into those subsets C that contain v and those which do not gives

$$\sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T)$$

$$= \sum_{O \subseteq C \subseteq V \setminus \{v\}} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T) + \sum_{O \cup \{v\} \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T),$$

which is seen to be the same as the given expression by including a factor of -1 inside and outside the second sum. $\qquad \square$

We now move to the main result of the paper.

Theorem 5.4. *The kernel $p_{V|W}$ recursively factorizes according to $\mathcal{G}$ if and only if it is parameterized according to $\mathcal{G}$.*

Proof. Throughout the proof, we will write the partitions of vertices in a CADMG as $[\![\cdot]\!]_{\mathcal{G}}$ regardless of which graph we are dealing with; since all the graphs we consider are reachable from $\mathcal{G}$, this is justified by Lemma 4.13.

We proceed by induction on the size of V. If $V = \{v\}$ then recursive factorization is by definition, so the condition holds for any distribution. On the other hand, parameterization entails

$$p_{v|W}(0_v \mid x_W) = q_v(x_{\mathrm{pa}(v)}), \qquad p_{v|W}(1_v \mid x_W) = 1 - q_v(x_{\mathrm{pa}(v)}), \tag{7}$$

which follows from setting $q_v(x_{\mathrm{pa}(v)}) = p_{v|W}(0_v \mid x_W)$ and the fact that $p_{v|W}(0_v \mid x_W) + p_{v|W}(1_v \mid x_W) = 1$ because $p_{v|W}$ is a probability distribution; hence parameterization also holds for any distribution with one random variable.

($\Leftarrow$) Now consider a general V and suppose $p_{V|W}$ is parameterized according to $\mathcal{G}$. If $\mathcal{G}$ has multiple districts then, by Lemma 5.2, the kernel factorizes into pieces which are parameterized according to $\mathcal{G}[D_i]$, and so by the induction hypothesis recursively factorize according to $\mathcal{G}[D_i]$.

Otherwise take any $a \in \mathrm{sterile}_{\mathcal{G}}(V)$, and consider a specific $x_{W, V \setminus \{a\}} \in \mathfrak{X}_{W, V \setminus \{a\}}$; let $O = \{v \in V \setminus \{a\} \mid x_v = 0\}$, so then

$$\sum_{x_a} p(x_V \mid x_W) = p(x_{V \setminus a}, 0_a \mid x_W) + p(x_{V \setminus a}, 1_a \mid x_W)$$

$$= \sum_{O \cup \{a\} \subseteq C \subseteq V} (-1)^{|C \setminus (O \cup \{a\})|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T)$$

$$+ \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T)$$

$$= \sum_{O \subseteq C \subseteq V \setminus \{a\}} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T)$$

by Lemma 5.3. By the induction hypothesis, this last expression recursively factorizes according to $\mathcal{G}[V \setminus \{a\}] = \mathfrak{m}_{V \setminus a}(\mathcal{G})$, and this extends easily to any random-ancestral margin $V \setminus B$ by sequentially marginalizing the variables in B. Hence, $p_{V|W}$ obeys properties (i) and (ii) of recursive factorization, and therefore recursively factorizes according to $\mathcal{G}$.

($\Rightarrow$) Conversely, suppose that $p_{V|W}$ recursively factorizes according to $\mathcal{G}$. In this direction, we will strengthen the induction hypothesis slightly and show that if $p_{V|W}$ recursively factorizes according to $\mathcal{G}$ then $p_{V|W}$ is parameterized according to $\mathcal{G}$, *and* that for each parameter $q_H(x_T)$, either: $p_{T \setminus W|W}(x_{T \setminus W} \mid x_{T \cap W}, y_{W \setminus T}) > 0$ for some $y_{W \setminus T}$, in which case $q_H(x_T)$ is uniquely recoverable from $p_{V|W}$; or $p_{T \setminus W|W}(x_{T \setminus W} \mid x_{T \cap W}, y_{W \setminus T}) = 0$ for all $y_{W \setminus T}$, in which case $q_H(x_T)$ can take any value. For the base case with $|V| = 1$, the result follows from the derivation of (7).

If $\mathcal{G}$ has multiple districts then, by definition, $p_{V|W}$ factorizes into pieces which themselves recursively factorize according to the district subgraphs $\mathcal{G}[D_i]$, and by the induction hypothesis each factor is parameterized according to $\mathcal{G}[D_i]$. Applying Lemma 5.2 it follows that $p_{V|W}$ is parameterized according to $\mathcal{G}$, and no parameters are shared between these factors by Lemma 4.14.

For uniqueness of $q_H(x_T)$, note that this parameter only appears in the expansion for probabilities $p_{V|W}(x_V \mid x_W)$ (i.e., those indexed by the same values x_T). If $p_{T \setminus W|W}(x_{T \setminus W} \mid x_W) > 0$,

then the factorization of $p_{V|W}$ is unique for these values by Proposition 3.6, and each factor is also positive for those values of x_T by Lemma 3.8; thus $q_H(x_T)$ is uniquely recoverable from that factor by the strengthened induction hypothesis.

If $p_{T\setminus W|W}(x_{T\setminus W} \mid x_W) = 0$, then by Lemma 3.8 there is some $t \in T \setminus W$ and $x_{V\setminus T}$ such that every version of $p_{t|\operatorname{pre}_<(t),W}(x_t \mid x_{\operatorname{pre}_<(t)}, x_W) = 0$. We split into two cases: either t is in the same district as H, or not; let D_1 be the district containing H, and the associated kernel r_1. If t is in D_1, then it follows from Proposition 3.6 that $r_1(x_{T\cap D_1} \mid x_{T\setminus D_1}) = 0$, and so by the induction hypothesis applied to $\mathcal{G}[D_1]$ we get that $q_H(x_T)$ can take any value. Otherwise if t is in a different district (say D_2), then it follows from Proposition 3.6 that $r_2(x_{T\cap D_2} \mid x_{\operatorname{pa}(D_2)\setminus D_2}) = 0$; so clearly whatever the value of any other factor, including r_1, the product will always be zero.

Now suppose $\mathcal{G}$ has a single district V; it follows from the definitions that V is intrinsic with recursive head $H^* = \operatorname{sterile}_\mathcal{G}(V)$ and tail $T^* = (V \cup W) \setminus H^*$. For any vertex $h \in H^*$ the set $V \setminus \{h\}$ is random-ancestral, so the margin $p_{V\setminus h|W}$ recursively factorizes with respect to $\mathcal{G}[V \setminus \{h\}]$, and therefore (by the induction hypothesis) is also parameterized according to $\mathcal{G}[V \setminus \{h\}]$. Every recursive head H other than H^* is found in at least one random-ancestral margin $V \setminus \{h\}$ of $\mathcal{G}$, so applying the induction hypothesis to $\mathcal{G}[V \setminus \{h\}]$ we obtain either a well defined parameter, or determine that its value is irrelevant.

If two or more random-ancestral margins contain the recursive head H, note that by Corollary 4.10 there is a 'smallest' such margin $p_{\operatorname{an}(H)\setminus W|W}$ containing H; all other random-ancestral margins contain this margin, and therefore by the induction hypothesis they will agree either on a value for $q_H(x_T)$ or agree that it is arbitrary. So for every random-ancestral set $A \subsetneq V$ the margin $\mathcal{G}[A]$ is parameterized according to $p_{A|W}$ and any parameters that two or more of these margins jointly use either are consistent, or can be chosen to be consistent.

The only recursive head not found in a random-ancestral margin is H^*, so the only parameter yet to be defined is $q_{H^*}(x_{T^*})$. We define this to be any version of $p_{H^*|T^*}(0_{H^*} \mid x_{T^*})$; this is well defined if $p(x_{T^*\setminus W} \mid x_{T^*\cap W}) > 0$, and arbitrary otherwise. Then

$$p_{V|W}(0_{H^*}, x_{V\setminus H^*} \mid x_W) = q_{H^*}(x_{T^*}) \cdot p(x_{V\setminus H^*} \mid x_W).$$

Since $V \setminus H^*$ is a random-ancestral margin of $\mathcal{G}$, it follows that $p(x_{V\setminus H^*} \mid x_W)$ is parameterized according to $\mathcal{G}[V \setminus H^*]$, and so

$$p_{V|W}(0_{H^*}, x_{V\setminus H^*} \mid x_W) = p(0_{H^*} \mid x_{V\setminus H^*}, x_W) \cdot \prod_{O \subseteq C \subseteq V\setminus H^*} (-1)^{|C\setminus O|} \prod_{H \in \llbracket C \rrbracket_\mathcal{G}} q_H(x_T)$$

$$= \prod_{O \subseteq C \subseteq V} (-1)^{|C\setminus O|} \prod_{H \in \llbracket C \rrbracket_\mathcal{G}} q_H(x_T).$$

This gives the required result if $x_h = 0$ for all $h \in H^*$. On the other hand, if $x_h = 1_h$ for some $h \in H^*$, then using a second induction on the number of zeros in x_{H^*} we have

$$p(x_{V\setminus h}, 1_h \mid x_W)$$
$$= p(x_{V\setminus h} \mid x_W) - p(x_{V\setminus h}, 0_h \mid x_W)$$

$$= \sum_{O \subseteq C \subseteq V \setminus \{h\}} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T) - \sum_{O \cup \{h\} \subseteq C \subseteq V} (-1)^{|C \setminus (O \cup \{h\})|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T)$$

$$= \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(x_T)$$

using Lemma 5.3. Hence, every probability $p_{V|W}(x_V \mid x_W)$ is of the required form. $\qquad \square$

5.3. Model smoothness

For some ADMGs $\mathcal{G}$, the parameters $q_H(x_T)$ are just (versions of) the ordinary conditional probabilities $P(X_H = 0 \mid X_T = x_T)$, and hence the alternating sum is similar to the Möbius form of the parameterization studied in Evans and Richardson [10] in the context of 'ordinary' Markov models. However, we have already seen that not all of the parameters can be interpreted this way; recall the example in Section 3 for Figure 1(b). In this case, as noted in Example 3.5, $q_{EY}(x, m) = r_{EY}(0, 0 \mid x, m)$ is not an ordinary conditional probability, but if the graph is interpreted causally then it is the conditional probability of $\{E = Y = 0\}$ after intervening to fix $\{X = x, M = m\}$:

$$q_{EY}(x, m) = p_{E|X}(0 \mid x) \cdot p_{Y|XME}(0 \mid x, m, 0)$$
$$= P(Y = E = 0 \mid \mathrm{do}(X = x, M = m)).$$

By the requirement that the graph is 'interpreted causally' we mean that it is the latent projection of a causal DAG in the sense of Pearl [15], Definition 1.3.1. This result holds more generally.

Theorem 5.5. *If $p_{V|W}$ is strictly positive and recursively factorizes according to some CADMG $\mathcal{G}$, then all the parameters $q_H(x_T)$ are unique and can be smoothly recovered from $p_{V|W}$ (i.e., there is an infinitely differentiable function from $p_{V|W}$ to the $q_H(x_T)$).*

In addition, if the graph is interpreted causally, then

$$q_H(x_T) = P(X_H = 0_H \mid X_{T \setminus \tilde{T}} = x_{T \setminus \tilde{T}}, \mathrm{do}(X_{\tilde{T}} = x_{\tilde{T}})),$$

where $\tilde{T} \equiv T \setminus S = \mathrm{pa}_{\mathcal{G}}(S) \setminus S$ is the subset of T that does not intersect S.

Proof. The first claim follows directly from the proof of Theorem 5.4, since the operations involved are just summations and divisions by positive quantities; the fact that $p_{V|W}$ is strictly positive ensures that each parameter is always uniquely defined rather than being arbitrary.

For the second part: recall that the steps (i) and (ii) in Definition 3.3 correspond to the algorithm in Tian and Pearl [27], so it follows from that paper that the conditional distribution obtained when we reach a particular intrinsic set S is $p_V(x_S \mid \mathrm{do}(x_{\mathrm{pa}(S) \setminus S}))$. Then calculating $q_H(x_T)$ just gives $p_V(0_H \mid x_{S \setminus H}, \mathrm{do}(x_{\mathrm{pa}(S) \setminus S}))$, and hence the result. $\qquad \square$

We remark that if the distribution is not strictly positive then it follows from the '$\Rightarrow$' part of the proof of Theorem 5.4 that the parameters $q_H(x_T)$ are uniquely defined if and only if

$p(x_{V \cap T} \mid x_{W \cap T}, y_{W \setminus T}) > 0$ for some $y_{W \setminus T}$. In the case that $W = \varnothing$ and $\mathcal{G}$ is an ADMG, this reduces to $q_H(x_T)$ being uniquely defined if and only if $p(x_T) > 0$.

We now return to the generality of a finite discrete state-space $\mathfrak{X}_v$ for each X_v. Let $\tilde{\mathfrak{X}}_v$ be the same set with some arbitrary entry removed (so that $|\tilde{\mathfrak{X}}_v| = |\mathfrak{X}_v| - 1$). Then for any set C let $\tilde{\mathfrak{X}}_C \equiv \times_{v \in C} \tilde{\mathfrak{X}}_v$. See Appendix C for explicit definitions of the parameters in this case.

Corollary 5.6. *The set of strictly positive distributions obeying the recursive factorization property with respect to a CADMG $\mathcal{G}$ is a curved exponential family of dimension*

$$d(\mathcal{G}) = \sum_{H \in \mathcal{H}(\mathcal{G})} |\tilde{\mathfrak{X}}_H| \cdot |\mathfrak{X}_T|.$$

Proof. Theorem 5.5 shows that there is a smooth (infinitely differentiable) map from positive distributions obeying the recursive factorization to the model parameters; it is clear from the form of the parameterization that the map from parameters to the probabilities is also smooth. The result follows by the same argument as Theorem 6.5 of Evans and Richardson [10]. □

This result allows us to invoke standard statistical theory within this class of models. For example, if $\mathcal{G}'$ is a subgraph of $\mathcal{G}$, then we can perform a hypothesis test of $H_0 : p_{V|W} \in \mathcal{M}_{rf}(\mathcal{G}')$ versus $H_1 : p_{V|W} \in \mathcal{M}_{rf}(\mathcal{G})$ by comparing the likelihood ratio statistic to a χ_k^2 distribution, where $k = d(\mathcal{G}) - d(\mathcal{G}')$.

Fitting these models is relatively straightforward given the explicit maps between parameters and probabilities. Maximum likelihood estimation can be performed using the same method as in Evans and Richardson [8]. The parameters $q_H(x_T)$ are clearly variation dependent, which can cause algorithmic complications and interpretability problems. A (generally variation dependent) log-linear parameterization of the kind given in Evans and Richardson [9] can relatively easily be adapted to nested models; see also Shpitser et al. [19].

6. Examples

The Wisconsin Longitudinal Study (Hauser et al. [11]) is a panel study of over 10 000 people who graduated from Wisconsin High Schools in 1957. We consider males who, when asked in 1975, had either been drafted or had not served in the military at all; after removing missing data this left 1676 respondents. We wish to know whether, after controlling for family income and education, being drafted had a significant effect on future earnings.

The variables measured were:

- X, an indicator of whether family income in 1957 was above \$5k;
- Y, an indicator of whether the respondent's income in 1992 was above \$37k;
- M, an indicator of whether the respondent was drafted into the military;
- E, an indicator of whether the respondent had education beyond high school.

Dichotomizations for X, Y and E were chosen to be close to the median values of the original variables. The data are shown in Table 1; in each case the value 1 corresponds to the statement

Table 1. Data from the Wisconsin Longitudinal Study

$M \backslash Y$	$X = 0, E = 0$		$M \backslash Y$	$X = 1, E = 0$	
	0	1		0	1
0	241	162	0	161	148
1	53	39	1	33	29

| $M \backslash Y$	$X = 0, E = 1$		$M \backslash Y$	$X = 1, E = 1$	
	0	1		0	1
0	82	176	0	113	364
1	13	16	1	16	30

above being true, 0 otherwise. One possible model is that future income is unrelated to family income at the time of graduation after controlling for military service and level of education. This suggests the graph in Figure 8(a), where the directed edge from X to Y is not present. We can fit this model using the parameterization and an algorithm based on the one given by Evans and Richardson [8]; the resulting fit has a deviance of 31.3 on 2 degrees of freedom, strongly suggesting that the model should be rejected. Unsurprisingly, the graph in Figure 1(b) is also rejected for these data.

On the other hand the model shown in Figure 8(b) has a deviance of 5.57 on 6 degrees of freedom, which indicates a good fit. Note that this implies that there is no evidence of a significant effect of being drafted on future income, even though marginally there is a strong negative correlation. Models obtained by removing any additional edges are strongly rejected. Under this model, the probability of having a high income in 1992 is estimated as 0.50 (standard error 0.018) if the family had high income, and 0.36 (0.016) if not.

In other words, we estimate

$$P\big(Y = 1 \mid \mathrm{do}(X = 1)\big) = 0.50, \qquad P\big(Y = 1 \mid \mathrm{do}(X = 0)\big) = 0.36,$$

indicating a strong causal effect.

(a) (b)
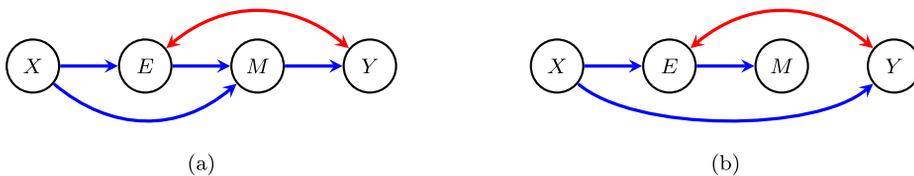
Figure 8. Two models for the Wisconsin miltary service data. (a) A proposed but rejected model; (b) a well-fitting model. See text for discussion.

Appendix A: Proof of the Verma constraint

Note that

$$\sum_e p(e \mid x) \cdot p(y \mid x, m, e) = \sum_e \frac{p(x, m, e, y)}{p(x) \cdot p(m \mid x, e)}$$

$$= \sum_e \frac{\sum_u p(u, x, m, e, y)}{p(x) \cdot p(m \mid x, e)}$$

by elementary laws of conditional probability. Applying the usual factorization of the DAG in Figure 1(a), we obtain

$$= \sum_e \frac{\sum_u p(u) \cdot p(x) \cdot p(e \mid x, u) \cdot p(m \mid e) \cdot p(y \mid m, u)}{p(x) \cdot p(m \mid x, e)}$$

noting that $M \perp\!\!\!\perp X \mid E$, and cancelling, gives

$$= \sum_{e,u} p(u) \cdot p(e \mid x, u) \cdot p(y \mid m, u)$$

$$= \sum_u p(u) \cdot p(y \mid m, u),$$

which does not depend upon x.

Appendix B: Partitions

Let V be an arbitrary finite set, and let $\mathcal{H}$ be an arbitrary collection of non-empty subsets of V, with the restriction that $\{v\} \in \mathcal{H}$ for all $v \in V$ (i.e. all singletons are in $\mathcal{H}$). A partial ordering $\prec$ on the elements of $\mathcal{H}$ will be said to be *partition suitable* if for any $H_1, H_2 \in \mathcal{H}$ with $H_1 \cap H_2 \neq \emptyset$, there exists $H^* \in \mathcal{H}$ such that $H^* \subseteq H_1 \cup H_2$ and $H_i \preceq H^*$ for each $i = 1, 2$. (Here $H_1 \preceq H_2$ means $H_1 \prec H_2$ or $H_1 = H_2$.)

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

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

Define $\psi(W)$ to be the set of vertices not in any set in $\Phi(W)$, that is:

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

Now recursively define a function $[\cdot]$ that partitions subsets of V: define $[\varnothing] = \varnothing$, and

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

It is clear that $\bigcup_{A \in [W]} A = W$.

The next proposition shows that $[W]$ is indeed a partition of W.

Proposition B.1. *If $H_1, H_2 \in \Phi(W)$ with $H_1 \neq H_2$ then $H_1 \cap H_2 = \varnothing$.*

Proof. Suppose $H_1 \cap H_2 \neq \varnothing$. Then by partition suitability, there exists $H^* \subseteq H_1 \cup H_2$ with $H^* \succeq H_1, H_2$, and in particular $H^* \succ H_i$ for at least one of $i = 1, 2$. Hence at least one of the H_i is not maximal in W. $\square$

Proposition B.2. *If $A \subseteq W_1 \subseteq W_2$, and $A \in \Phi(W_2)$ then $A \in \Phi(W_1)$.*

Proof. If A is maximal amongst elements of $\mathcal{H}$ that are subsets of W_2, then it is certainly still maximal amongst those that are subsets of W_1, since there are fewer such sets. $\square$

Proposition B.3. *If $C \in [W]$, then $[W] = \{C\} \cup [W \setminus C]$.*

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

If C is not maximal with respect to $\prec$ among subsets of W, then $\Phi(W) = \Phi(W \setminus C)$, and so

$$[W] = \Phi(W) \cup \big[\psi(W)\big]$$
$$= \Phi(W \setminus C) \cup \big[\psi(W)\big],$$

and the problem reduces to showing that $[\psi(W)] = \{C\} \cup [\psi(W \setminus C)]$, which follows from the induction hypothesis. Thus, suppose $C \in \Phi(W)$.

Now by Proposition B.2, $\Phi(W \setminus C) \cup \{C\} \supseteq \Phi(W)$, and if equality holds we are done. Otherwise let $C_1, \ldots, C_k$ be the sets in $\Phi(W \setminus C)$ but not in $\Phi(W)$. These sets are maximal in $W \setminus C$, so they are in $\Phi(\psi(W))$ by Proposition B.2, since by hypothesis, $\psi(W) \subseteq W \setminus C$. Then the problem reduces to showing that

$$\big[\psi(W)\big] = \{C_1, \ldots, C_k\} \cup \big[\psi(W) \setminus (C_1 \cup \cdots \cup C_k)\big],$$

which follows from repeated application of the induction hypothesis. $\square$

Proposition B.4. *Let $D_1, \ldots, D_k$ be a partition of V, and suppose that each $H \in \mathcal{H}$ is contained within some D_i. Let $\prec$ be a partition-suitable partial ordering. Then*

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

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

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

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

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

From the condition on $\mathcal{H}$ we can write $\Phi(W) = \mathcal{C}_1 \cup \mathcal{C}_2$ where each $H \in \mathcal{C}_i$ is a subset of D_i; since the elements of $\mathcal{C}_i$ are maximal with respect to $\prec$ in W, they are also maximal in $W \cap D_i$. Hence $\mathcal{C}_i \subseteq \Phi(W \cap D_i)$, and then applying Proposition B.3 repeatedly gives

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

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

B.1. Partition suitability of recursive head ordering

The next result, together with Proposition B.1, shows that the function $[\![\cdot]\!]_{\mathcal{G}}$, from Definition 4.11, is indeed a partition.

Proposition B.5. $\prec$ *is partition suitable for* $\mathcal{H}(\mathcal{G})$.

Proof. Lemma 4.8 shows that $\mathcal{H}$ contains the singleton vertices. Now suppose we have two recursive heads H_1, H_2 with $H_1 \cap H_2 \neq \varnothing$. Let the associated intrinsic sets be S_1, S_2. Since S_1, S_2 are bidirected connected sets and they share a common element, $S_1 \cup S_2$ is also bidirected-connected. Let S^* be the intrinsic closure of $S_1 \cup S_2$, with recursive head H^*. Then S^* contains both S_1 and S_2, and therefore $H^* \succeq H_1, H_2$.

By Proposition 4.7, $H^* = \text{sterile}_{\mathcal{G}}(S^*) \subseteq S_1 \cup S_2$; by definition of a recursive head, any $v \in S_1$ is either in H_1 or is a parent of some other element of S_1 (and the same for S_2). Hence $H^* \subseteq H_1 \cup H_2$. $\square$

Appendix C: General discrete state-space

Lemmas 5.2 and 5.3 and Theorem 5.4 are stated and proved for binary variables to avoid cumbersome notation; here we provide some notes on how one would adapt them to the general case.

Suppose that $\mathfrak{X}_{VW}$ is possibly non-binary. For each $v \in V$ pick an arbitrary element $k_v \in \mathfrak{X}_v$ to be a corner-point. Let $\tilde{\mathfrak{X}}_v \equiv \mathfrak{X}_v \setminus \{k_v\}$ and $\tilde{\mathfrak{X}}_C \equiv \times_{v \in C} \tilde{\mathfrak{X}}_v$. In the binary case we took $k_v = 1$, so that $\tilde{\mathfrak{X}}_v = \{0\}$ for each v.

The parameters then become $q_H(x_H \mid x_T)$ for $H \in \mathcal{H}(\mathcal{G})$, $x_H \in \tilde{\mathfrak{X}}_H$ and $x_T \in \mathfrak{X}_T$. The parameterization in (6) becomes:

$$p_{V \mid W}(x_V \mid x_W) = \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \sum_{y_C \in \tilde{\mathfrak{X}}_C : y_O = x_O} \prod_{H \in [\![C]\!]_{\mathcal{G}}} q_H(y_H \mid x_T),$$

where $O \equiv O(x_V) = \{v \mid x_v \in \tilde{\mathfrak{X}}_v\}$. Note that, in the binary case, the inner sum only ever has one term.

Lemma 5.2 goes through as before by splitting the inner sum up as

$$\sum_{y_C \in \tilde{\mathfrak{X}}_C : y_O = x_O} = \sum_{y_{C_1} \in \tilde{\mathfrak{X}}_{C_1} : y_{O_1} = x_{O_1}} \sum_{y_{C_2} \in \tilde{\mathfrak{X}}_{C_2} : y_{O_2} = x_{O_2}}.$$

The proof of Theorem 5.4 is also the same, except that instead of $x_h = 0$ and $x_h = 1$ the important cases become $x_h \in \tilde{\mathfrak{X}}_h$ and $x_h = k_h$.

Acknowledgements

References

[1] Bishop, C.M. (2007). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. New York: Springer. MR2247587

[2] Darwiche, A. (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge: Cambridge Univ. Press. MR2572244

[3] Dawid, A.P. (2002). Influence diagrams for causal modelling and inference. *Int. Stat. Rev.* **70** 161–189.

[4] Drton, M. (2009). Discrete chain graph models. *Bernoulli* **15** 736–753. MR2555197

[5] Drton, M. (2009). Likelihood ratio tests and singularities. *Ann. Statist.* **37** 979–1012. MR2502658

[6] Drton, M. and Richardson, T.S. (2008). Binary models for marginal independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 287–309. MR2424754

[7] Evans, R.J. (2018). Margins of discrete Bayesian networks. *Ann. Statist.* **46** 2623–2656.

[8] Evans, R.J. and Richardson, T.S. (2010). Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the* 26*th Conference on Uncertainty in Artificial Intelligence* 177–184.

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

[10] Evans, R.J. and Richardson, T.S. (2014). Markovian acyclic directed mixed graphs for discrete data. *Ann. Statist.* **42** 1452–1482. MR3262457

[11] Hauser, R.M., Sewell, W.H. and Herd, P. Wisconsin Longitudinal Study (WLS), 1957–2012. Available at http://www.ssc.wisc.edu/wlsresearch/documentation/. Version 13.03, Univ. Wisconsin–Madison, WLS.

[12] Huang, J.C. and Frey, B.J. (2008). Cumulative distribution networks and the derivative-sum-product algorithm. In *Proceedings of the* 24*th Conference on Uncertainty in Artificial Intelligence* 290–297.

[13] Mond, D., Smith, J. and van Straten, D. (2003). Stochastic factorizations, sandwiched simplices and the topology of the space of explanations. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **459** 2821–2845. MR2015992

[14] Pearl, J. and Verma, T.S. (1992). A statistical semantics for causation. *Stat. Comput.* **2** 91–95.

[15] Pearl, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge: Cambridge Univ. Press. MR2548166

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

[17] Richardson, T.S., Evans, R.J., Robins, J.M. and Shpitser, I. (2017). Nested Markov properties for acyclic directed mixed graphs. Preprint. Available at arXiv:1701.06686.

[18] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – Application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. MR0877758

[19] Shpitser, I., Evans, R.J., Richardson, T.S. and Robins, J.M. (2013). Sparse nested Markov models with log-linear parameters. In *Proceedings of the* 29*th Conference on Uncertainty in Artificial Intelligence* 576–585.

[20] Shpitser, I., Evans, R.J., Richardson, T.S. and Robins, J.M. (2014). Introduction to nested Markov models. *Behaviormetrika* **41** 3–39.

[21] Shpitser, I. and Pearl, J. (2008). Dormant independence. Technical Report R-340, Cognitive Systems Laboratory, University of California, Los Angeles.

[22] Shpitser, I., Richardson, T.S., Robins, J.M. and Evans, R.J. (2011). Parameter and structure learning in mixed graph models of post-truncation independence. Draft.

[23] Silva, R., Blundell, C. and Teh, Y.W. (2011). Mixed cumulative distribution networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (*AISTATS*) **15** 670–678.

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

[25] Richardson, T.S. Spirtes, P.L. and (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962-1030. MR1926166

[26] Tian, J. (2002). Studies in causal reasoning and learning. Ph.D. thesis, University of California, Los Angeles.

[27] Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the* 18*th National Conference on Artificial Intelligence*. AAAI.

[28] Tian, J. and Pearl, J. (2002). On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* (*UAI*-02) 519–527. Morgan Kaufmann Publishers Inc.

[29] Verma, T.S. and Pearl, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of the* 7*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-91) 255–268.

[30] Wermuth, N. (2011). Probability distributions with summary graph structure. *Bernoulli* **17** 845–879. MR2817608