Behaviormetrika
Vol.41, No.1, 2014, 3–39

Invited paper

INTRODUCTION TO NESTED MARKOV MODELS

Ilya Shpitser*, Robin J. Evans**, Thomas S. Richardson***, and James M. Robins****

Graphical models provide a principled way to take advantage of independence constraints for probabilistic and causal modeling, while giving an intuitive graphical description of "qualitative features" useful for these tasks. A popular graphical model, known as a Bayesian network, represents joint distributions by means of a directed acyclic graph (DAG). DAGs provide a natural representation of *conditional independence* constraints, and also have a simple causal interpretation. When all variables are observed, the associated statistical models have many attractive properties. However, in many practical data analyses unobserved variables may be present. In general, the set of marginal distributions obtained from a DAG model with hidden variables is a much more complicated statistical model: the likelihood of the marginal is often intractable; the model may contain singularities. There are also an infinite number of such models to consider.

It is possible to avoid these difficulties by modeling the observed marginal directly. One strategy is to define a model by means of conditional independence constraints induced on the observed marginal by the hidden variable DAG; we call this the *ordinary Markov model.* This model will be a supermodel that contains the set of marginal distributions obtained from the original DAG. Richardson and Spirtes (2002) and Evans and Richardson (2013a) gave parametrizations of this model in the Gaussian and discrete case, respectively.

However, it has long been known that hidden variable DAG models also imply nonparametric constraints which generalize conditional independences; these are sometimes called "Verma Constraints". In this paper we describe a natural extension of the ordinary Markov approach, whereby both conditional independences and these generalized constraints are used to define a *nested Markov model.* The binary nested Markov model may be parametrized via a simple extension of the binary parametrization of the ordinary Markov model of Evans and Richardson (2013a). We also give evidence for a characterization of nested Markov equivalence for models with four observed variables. A consequence of this characterization is that, in some instances, most structural features of hidden variable DAGs can be recovered exactly when a single generalized independence constraint holds under the distribution of the observed variables.

1. Introduction

Graphical models provide a principled way to take advantage of independence constraints for probabilistic and causal modeling, while giving an intuitive graphical description of "qualitative features" useful for these tasks. A popular graphical model represents a joint distribution by means of a directed acyclic graph (DAG), where each random variable over which the joint is defined corresponds to a vertex in the graph. Such a model represents all conditional distributions where a random variable

Key Words and Phrases:

* University of Southampton. Email: i.shpitser@soton.ac.uk

** University of Oxford. Email: robin.evans@stats.ox.ac.uk

*** University of Washington, Email: thomasr@u.washington.edu

****Harvard University, Email: robins@hsph.harvard.edu

is independent of random variables corresponding to its non-descendants in the graph conditioned on the variables corresponding to its parents in the graph. The popularity of DAG models stems from their well understood theory, and from the fact that such models have an intuitive causal interpretation. Informally, if all common causes of variables on the graph are themselves on the graph, then an arrow from a variable X to a variable Y in a DAG model can be interpreted to mean that X is a "direct cause" of Y. See Section 2.3, Spirtes et al. (1993), or Pearl (2000) for further discussion.

Frequently, causal and probabilistic inference problems contain latent variables. While existing theoretical machinery based on DAGs can be applied to such settings (Cooper and Herskovits, 1992; Friedman, 1997; Beal and Ghahramani, 2004), this creates a number of problems. First, a particular marginal distribution can, in general, be obtained from an infinite number of joint distributions associated with an infinite number of hidden variable DAGs. Second, prior knowledge about latent variables is often scarce, so any modeling assumptions which explicitly represent latents may leave one open to model misspecification bias. Third, evaluating the likelihood of the observed marginal distribution under a hidden variable DAG model is often computationally intractable. Finally, the models defined by marginals of DAGs do not possess many of the nice statistical properties enjoyed by their fully observed counterparts; This lack of regularity may complicate inference procedures (see for instance Drton et al. (2009)).

An alternative approach is to define a larger statistical model using constraints that the latent variable DAG induces on the observed margin. One such model uses precisely the conditional independence constraints on the observed marginal; we call this the *ordinary Markov model.* This model will be a supermodel that contains the set of marginal distributions obtained from the original DAG. Richardson and Spirtes (2002) and Evans and Richardson (2013a) gave parametrizations of this model in the Gaussian and discrete case, respectively.

A natural extension of this ordinary Markov approach is to use a larger set of (equality) constraints implied by the latent variable model to define a *nested Markov model.* Specifically, we augment the ordinary conditional independence constraints with generalized conditional independences. A number of recent papers (Shpitser et al., 2011; Richardson et al., 2012; Shpitser et al., 2013) have developed this idea.

In more detail, the constraints defining the nested Markov model are encoded in an acyclic directed mixed graph (ADMG) – called a latent projection, and first described in Verma and Pearl (1990) – derived from the DAG with latents. As an illustration, this approach associates the ADMG in Fig. 1(c) with the hidden variable DAGs shown in Fig. 1(a) and (b). As an example of a generalized constraint, in any marginal distribution $P(x_1, x_2, x_3, x_4)$ consistent with the hidden variable DAGs in Fig. 1(a) and (b), it is the case that

$$\frac{\partial}{\partial x_1} \sum_{x_2} P(x_4 \mid x_1, x_2, x_3) P(x_2 \mid x_1) = 0 \tag{1}$$

Note that (1) can be viewed as an equality constraint, because it states that the

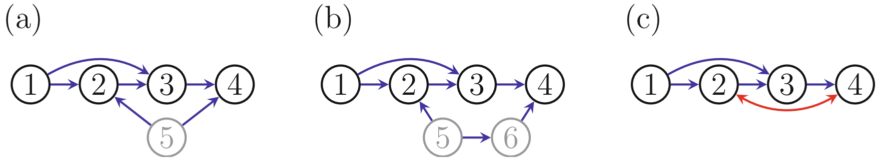

Figure 1: (a) A DAG representing distributions that factorize as $P(x_1)$ $P(x_2|x_1, x_5)$ $P(x_3|x_2, x_1)$ $\times$ $P(x_4|x_3, x_5)$ $P(x_5)$. (b) A DAG representing distributions that factorize as $P(x_5)$ $\times$ $P(x_6|x_5)$ $P(x_1)$ $P(x_2|x_1, x_5)$ $P(x_3|x_2, x_1)$ $P(x_4|x_3, x_6)$. (c) A mixed graph representing the marginal distribution $P(x_1, x_2, x_3, x_4)$ of a joint that factorizes as in (a) or in (b).

expression $\sum_{x_2} P(x_4 \mid x_1, x_2, x_3)P(x_2 \mid x_1)$ is not a function of x_1. There are no ordinary conditional independences that hold in the marginal distribution: in fact, (1) is the only equality constraint that holds. Robins (1999) presented a method for finding such constraints; a more comprehensive algorithm that finds additional constraints was given by Tian and Pearl (2002).

Nested Markov models avoid the non-identifiability, misspecification and intractability issues of hidden variable DAG models by not referring to hidden variables at all. For example, as discussed later, the nested Markov model associated with the mixed graph in Fig. 1(c) is the set of densities over X_1, X_2, X_3 and X_4 in which (1) holds. This constraint on $P(x_1, x_2, x_3, x_4)$ is implied by any distribution $P(x_1, x_2, x_3, x_4, x_5)$ that is in the DAG model of Fig. 1(a) and any distribution $P(x_1, x_2, x_3, x_4, x_5, x_6)$ that is in the DAG model of Fig. 1(b).

Nested Markov models have theoretical and practical advantages when applied to the problem of learning causal structure from data. A common strategy for structure learning for graphical models with a known likelihood is to score via criteria such as BIC, and select graphs corresponding to high scoring models. The rationale for this procedure, given in (Schwarz, 1978), is that the BIC score is an approximation of the model posterior. The proof of this relies crucially on the smoothness of the parameter map of the model. This approximation holds for nested Markov models, since they form curved exponential families, but does not hold for hidden variable DAGs (Rusakov and Geiger, 2005; Drton, 2009; Drton and Plummer, 2013). In other words, there is no obvious Bayesian reason to adopt the BIC score for structure learning with hidden variable DAG models. Furthermore, since nested Markov models are defined in terms of generalized independence constraints, structure learning algorithms that work with these models can recover more causal features than algorithms based solely on conditional independences, such as scoring algorithms that search over classes of models defined solely by conditional independence constraints (Evans and Richardson, 2010; Spirtes et al., 1993). See Section 4.4 for examples. Nested Markov models also offer an attractive approach to latent variable causal modeling. See Section 3.3.4 for more on this. Finally, we describe log-linear nested Markov models which offer a well-behaved, low dimensional submodel of the nested Markov model. See Section 4.3 for details.

This manuscript serves as an introduction to nested Markov models, and their connections to existing theory on statistical and causal graphical models. We have chosen to prioritize examples and connections to existing work over technical rigor and completeness. The nested Markov model is defined formally in the Appendix; we also outline the binary parametrization[1)].

The remainder of the paper is structured as follows. Section 2 contains an overview of causal and statistical DAG models, and a short history of generalized independence constraints, such as the "Verma constraint." Section 3 describes mixed graphs, defines the ordinary and nested Markov models associated with DAGs with latents, and explains the connections between such models and earlier work on generalized independences constraints. Section 4 compares parameterizations of the ordinary and nested Markov models in the binary case, and describes preliminary results on equivalence in nested Markov models. Section 5 contains concluding remarks and describes avenues for future work.

2. Overview of DAG Models

We review causal and statistical DAG models, and give a brief history of generalized independence constraints that motivated the development of nested Markov models. The theory of both statistical (Pearl, 1988; Koller and Friedman, 2009) and causal (Spirtes et al., 1993; Pearl, 2000) DAG models is very well developed, and there are many applications (Shwe et al., 1991; Friedman et al., 2000; Pourret et al., 2008).

2.1 Notation

To avoid confusion arising from overloaded notation, we will need to distinguish between vertices in graphs, corresponding random variables, and values those random variables attain. We will denote a singleton vertex as lowercase roman, e.g. b, and a set of vertices by uppercase roman, e.g. B. A single random variable corresponding to a vertex b will be denoted as X_b, and a particular value of this variable as x_b. Similarly, a set of random variables corresponding to a vertex set B will be X_B and the corresponding set of values will be x_B.

We will employ the standard "genealogy metaphor" when describing relationships among vertices in a DAG. Let a, b and c be vertices in a DAG $\mathcal{G}$. If an edge $b \rightarrow a$ exists in $\mathcal{G}$, then we say that b is a *parent* of a, and a is a *child* of b. A *path* π between a and b in $\mathcal{G}$ is a sequence of edges $\langle \epsilon_1, \ldots, \epsilon_n \rangle$, such that there exists a sequence of distinct vertices $\langle a \equiv w_1, \ldots, w_{n+1} \equiv b \rangle$, $(n \geq 0)$, where an edge ϵ_i has endpoints w_i, w_{i+1}. A vertex a is said to be an *ancestor* of a vertex d if *either* there is a directed path $a \rightarrow \cdots \rightarrow d$ from a to d, *or* $a = d$; similarly d is said to be a *descendant* of a. The set of parents, children, ancestors, descendants and non-descendants of a in $\mathcal{G}$ are written $\text{pa}_{\mathcal{G}}(a)$, $\text{ch}_{\mathcal{G}}(a)$, $\text{an}_{\mathcal{G}}(a)$, $\text{de}_{\mathcal{G}}(a)$, $\text{nd}_{\mathcal{G}}(a)$ respectively. By convention, a is always both an ancestor and a descendant of a, but is not a non-descendant of a. We

[1)] Technical proofs will be provided elsewhere.

apply these definitions disjunctively to sets, e.g. $\text{an}_{\mathcal{G}}(A) = \bigcup_{a \in A} \text{an}_{\mathcal{G}}(a)$.

A set of vertices A in $\mathcal{G}$ is called *ancestral* if $a \in A \Rightarrow \text{an}_{\mathcal{G}}(a) \subseteq A$. An ordering $\prec$ of nodes in $\mathcal{G}$ is said to be *topological* if for any vertex pair $a, b \in \mathcal{G}$, if $a \prec b$, then $a \notin \text{de}_{\mathcal{G}}(b)$.

2.2 Statistical DAG Models

A directed acyclic graph (DAG) $\mathcal{G}(V, E)$ is a graph over a vertex set V, containing a set E of directed edges ($\rightarrow$) subject to the restriction that there are no directed cycles $v \rightarrow \cdots \rightarrow v$.

Definition 1 *A distribution $P(x_V)$ is said to be* Markov relative *to a DAG $\mathcal{G}$ if*

$$P(x_V) = \prod_{v \in V} P(x_v \mid x_{\text{pa}_{\mathcal{G}}(v)}). \tag{2}$$

For example, a distribution $P(x_1, x_2, x_3, x_4, x_5)$ is Markov relative to the DAG shown in Fig. 1(a) if:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_5)P(x_2 \mid x_1, x_5)P(x_3 \mid x_1, x_2)P(x_4 \mid x_5, x_3).$$

We refer to the set of distributions $P(x_V)$ which are Markov relative to a DAG $\mathcal{G}$ as the *(statistical) DAG model* associated with $\mathcal{G}$, or $\mathcal{P}^d(\mathcal{G})$. Statistical DAG models are sometimes called *Bayesian networks* (Pearl, 1988).

An equivalent way to define these models is as the set of distributions $P(x_V)$ satisfying the conditional independences

$$\Big(X_v \perp\!\!\!\perp X_{\text{nd}_{\mathcal{G}}(v) \setminus \text{pa}_{\mathcal{G}}(v)} \mid X_{\text{pa}_{\mathcal{G}}(v)}\Big)_{[P(x_V)]} \quad \text{for all } v \in V, \tag{3}$$

where $(X_A \perp\!\!\!\perp X_B \mid X_C)_{[P(x_V)]}$ means "X_A is conditionally independent of X_B given X_C in $P(x_V)$"[2]. Equation (2) is called the *Markov factorization*, and (3) defines the so-called *local Markov property.*

2.2.1 d-separation

Conditional independences in (3) imply an additional set of independences that may be read off directly from the associated DAG using the *d-separation criterion* (Pearl, 1988), defined on paths in the DAG.

A non-endpoint vertex z on a path is a *collider on the path* if the edges preceding and succeeding z on the path have an arrowhead at z, i.e. $\rightarrow z \leftarrow$. A non-endpoint vertex z on a path which is not a collider is a *non-collider on the path*, i.e. $\leftarrow z \rightarrow$, $\leftarrow z \leftarrow$, $\rightarrow z \rightarrow$. A path between vertices a and b in a DAG $\mathcal{G}$ is said to be *d-connecting given a set* C in $\mathcal{G}$ if every non-collider on the path is not in C, and every collider on the path is an ancestor of C in $\mathcal{G}$. If there is no path d-connecting a and b given C, then a and b are said to be *d-separated* given C. Sets A and B are said to

[2] By convention (3) is taken to be satisfied trivially for any v for which $\text{nd}_{\mathcal{G}}(v) \setminus \text{pa}_{\mathcal{G}}(v)$ is empty.

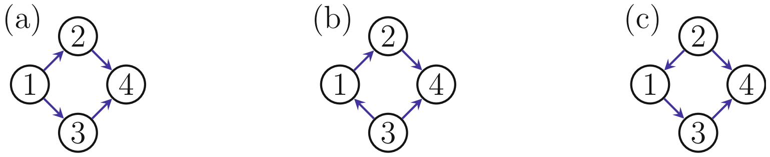

Figure 2: Three Markov equivalent DAGs.

be *d-separated* given C, if for all a, b, with $a \in A$ and $b \in B$, a and b are d-separated given C. As a shorthand we write this as

$$(A \perp\!\!\!\perp B \mid C)_{[\mathcal{G}]} \,,$$

where we reuse the symbol $\perp\!\!\!\perp$ to refer to d-separation (in $\mathcal{G}$) rather than conditional independence (in $P(x_V)$). This slight abuse of notation is justified due to the following result. If $P(x_V)$ is Markov relative to $\mathcal{G}$, then the following implication holds:

$$(A \perp\!\!\!\perp B \mid C)_{[\mathcal{G}]} \Rightarrow (X_A \perp\!\!\!\perp X_B \mid X_C)_{[P(x_V)]} \,. \tag{4}$$

Equation (4) defines the so-called *global Markov property.* A well-known result in DAG models (Lauritzen, 1996) is that these three separate definitions which link a distribution and a DAG give the same model.

Theorem 2 *Given $\mathcal{G}$ and distribution $P(x_V)$ the following are equivalent: $P(x_V)$ factorizes as in (2); the constraints (3) hold for every $v \in V$; (4) holds for all disjoint sets A, B, C where C may be empty.*

Since DAGs contain directed edges, which are suggestive of causal relationships, one might think that therefore statistical DAG models are causal. In fact, no mention of causality has been made in any of these definitions, and therefore these models are purely *statistical* objects; in other words, sets of distributions.

2.2.2 Markov Equivalence in DAG Models

Two distinct DAGs may define the same statistical DAG model, in which case they are called *Markov equivalent.* For example, for every DAG in Fig. 2, the set of Markov distributions is:

$$\left\{ P(x_1, x_2, x_3, x_4) \mid |(X_2 \perp\!\!\!\perp X_3 \mid X_1)_{[p]}, (X_4 \perp\!\!\!\perp X_1 \mid X_2, X_3)_{[p]} \right\}.$$

A result derived in Verma and Pearl (1990) gives a characterization of Markov equivalence for DAGs in terms of simple graphical structures:

We say that vertices a, b in a DAG $\mathcal{G}$ form an *adjacency* if either $a \to b$ or $b \to a$ exists in $\mathcal{G}$. We say that vertices a, b, c in a DAG $\mathcal{G}$ form an *unshielded collider* (or a *v-structure*) if edges $a \to b$ and $c \to b$ exist in $\mathcal{G}$, but a and c are not adjacent in $\mathcal{G}$.

Theorem 3 *Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if they have the same adjacencies, and unshielded colliders.*

For example, all DAGs in Fig. 2 share the adjacencies $(1, 2)$, $(1, 3)$, $(3, 4)$, and $(2, 4)$, and the unshielded collider $(2, 4, 3)$.

2.2.3 Faithfulness and Structure Learning in DAG Models

Markov equivalence places fundamental limits on our ability to solve the *structure learning problem* in DAG models. The structure learning problem assumes a distribution $P(x_V)$ is in the model of some unknown true DAG $\mathcal{G}^*$. The goal is to infer as many features of $\mathcal{G}^*$ as possible under the assumption that the converse of (4) holds, that is

$$(X_A \perp\!\!\!\perp X_B \mid X_C)_{[P(x_V)]} \Rightarrow (A \perp\!\!\!\perp B \mid C)_{[\mathcal{G}]} . \tag{5}$$

This is known as the *faithfulness* assumption, and it allows us to infer graphical features based on conditional independences. The faithfulness assumption is not uncontroversial[3]. If the underlying DAG is further assumed to be *causal* (see Section 2.3), solving the structure learning problem entails learning cause-effect relationships from observational data.

Structure learning problems in DAGs are solved by two types of approach. The constraint-based approach rules out DAGs inconsistent with a set of constraints present in $P(x_V)$ based on the faithfulness assumption; the score-based approach assigns a likelihood-based score to each model in a set, and picks the model with the highest score. The standard constraint-based algorithm for DAGs is the PC algorithm (Spirtes et al., 1993), while the GES algorithm (Chickering, 2002) is a model scoring algorithm which explicitly takes Markov equivalence into account, and has been proven to be asymptotically consistent.

Note that, in principle, constraint-based structure learning algorithms are valid for any model defined by constraints for which hypothesis tests exist (provided we have a complete list of the constraints implied by each model). In contrast, scoring-based structure learning algorithms are restricted to models in which the dimension is small enough for the likelihood function to exist with high probability.

Since DAG models are defined via conditional independence constraints, any structure learning method that makes no additional assumptions, including the PC and the GES algorithms, cannot distinguish between distinct, but Markov equivalent DAGs[4]. For instance, if the PC algorithm decided via a series of hypothesis tests that both $(X_2 \perp\!\!\!\perp X_3 \mid X_1)$ and $(X_4 \perp\!\!\!\perp X_1 \mid X_2, X_3)$ held in the data, and no other independences held, it would conclude that one of the DAGs in Fig. 2 is true, but would be unable to decide which.

[3] See Robins and Wasserman (1999) and Robins et al. (2003) for a fuller discussion of related issues.

[4] Methods that employ additional parametric assumptions may get around this limitation, see for example Shimizu et al. (2006).

2.2.4 Latent Variable DAG Models

The standard definition of DAG models given in Theorem 2 assumes that all variables in X_V are observed. If this assumption is relaxed, the result is a set of distributions which are marginals of distributions that are Markov relative to some DAG. For example, if we assume a DAG model corresponding to the graph $\mathcal{G}$ in Fig. 1(a), where only X_1, X_2, X_3, and X_4 are observed, while X_5 is hidden, we obtain the *marginal statistical DAG model*:

$$\left\{ P(x_1, x_2, x_3, x_4) \mid P(x_1, x_2, x_3, x_4, x_5) \in \mathcal{P}^d(\mathcal{G}) \right\}. \tag{6}$$

The benefit of such models is that they can leverage the machinery developed for fully observable DAG models. This powerful advantage resulted in a vast literature on latent variable graphical models (Friedman, 1997; Roweis and Ghahramani, 1999; Beal and Ghahramani, 2004; Koller and Friedman, 2009).

However, in the context of structure learning, there are multiple disadvantages. Selecting among hidden variable DAG models entails selecting elements from an infinite set. For example, if the true model is given in Fig. 1(a), it is difficult to rule out an infinite number of alternatives, such as Fig. 1(b). Even if a single model is chosen based on a combination of prior knowledge and simplicity arguments such as Occam's razor, the hidden variables in such a model must be parameterized in some way, and an incorrect parameterization may lead to misspecification bias[5].

In addition, evaluating the likelihood of the marginal in the hidden variable DAG model is computationally intractable. Finally, while DAG models form *curved exponential families* (Geiger and Meek, 1998), marginals of DAG models are known not to do so in general. This means that DAG models correspond to a smooth submanifold of a saturated model, while a hidden variable DAG model may be a subset that contains singularities. This complicates parameter fitting procedures, choosing priors for Bayesian reasoning (Settimi and Smith, 1998), and other common statistical tasks even in fairly simple cases (Drton, 2009).

2.3 Causal DAG Models

A way to formalize the natural causal interpretation of DAGs (under which an arrow $a \rightarrow b$ means that X_a is a "direct cause" of X_b) is via the *intervention operation*, which assigns to a random variable X_a a value x_a regardless of the natural behavior of X_a. This operation is denoted as $g = (x_a)$ by Robins (1986), and $\text{do}(x_a)$ by Pearl (2000)[6]. Interventions represent hypothetical idealized experiments, where a popu-

[5] For a simple example of this issue, consider how we might select the state space of discrete hidden variable DAG models. If we do not wish to impose equality constraints on such a model beyond the equality constraints implied by the structure of the DAG, we may in general have to grow the state space of hidden variables as a function of the size of the DAG.

[6] In the current literature the $\text{do}(\cdot)$ operator is used more widely than the g operator and thus we shall use the former in the sequel.

lation in which X_a would vary naturally in the absence of an intervention is instead assigned a fixed treatment value x_a for X_a. An alternative notation, first introduced by Neyman (1923), views a random variable X_y under a hypothetical intervention $\text{do}(x_a)$ as a *potential outcome*, written in our notation as $X_y(x_a)$[7].

There are a number of ways to define graphical causal models using either the do(.) or potential outcome notation that differ in the level of ontological and epistemic commitments made. A detailed discussion is beyond the scope of this manuscript (though see (Robins and Richardson, 2010) for more details). For our purposes it will be sufficient to restrict attention to a particular property that virtually all interventionist graphical causal models agree on, which equates interventional distributions over variables in X_V with functionals of $P(x_V)$ via the *g-formula* (Robins, 1986), *manipulated distribution* (Spirtes et al., 1993), or *truncated factorization* (Pearl, 2000). Specifically, if $P(x_V)$ is Markov relative to a DAG $\mathcal{G}$, for any subset $A \subseteq V$ such that the intervention $\text{do}(x_A)$ is meaningful, we have

$$P(x_{V \setminus A} \mid \text{do}(x_A)) = \prod_{i \notin A} P(x_i \mid x_{\text{pa}_{\mathcal{G}}(i)}) \tag{7}$$

where $x_{V \setminus A}$, x_A, x_i, and $x_{\text{pa}_{\mathcal{G}}(i)}$ are consistent (in that the same values are assigned to the same variables). Note that (7) is obtained from (2) by deleting all terms $P(x_j \mid x_{\text{pa}_{\mathcal{G}}(j)})$ where $j \in A$, or, equivalently, by dividing the joint distribution $P(x_V)$ by conditional densities $P(x_j \mid x_{\text{pa}_{\mathcal{G}}(j)})$ for $j \in A$.

If we assume that all variables can be intervened on, that is $A = V$, (7) implies a causal interpretation of the *absence* of directed edges. Specifically, the absence of an edge $a \to b$ means that if we were to intervene to set the value of X_a, *while intervening to set the values of all other parents of* b, then the distribution of X_b will not depend on the value assigned to X_a. In other words, for any pair of assignments x_a, x'_a to X_a, and all assignments $x_{\text{pa}_{\mathcal{G}}(b)}$,

$$P(x_b \mid \text{do}(x_{\text{pa}_{\mathcal{G}}(b)}, x_a)) = P(x_b \mid \text{do}(x_{\text{pa}_{\mathcal{G}}(b)}, x'_a)) \tag{8}$$

or that X_a is not a "direct cause" of X_b[8].

Note that any causal model that implies (7) already assumes the Markov factorization under no interventions ($A = \emptyset$), and as such implies that $P(x_V)$ is in the (statistical) model associated with $\mathcal{G}$[9]. This means that the difficulties with latent variable

[7] In fact Neyman (1923) and many subsequent papers (Rubin, 1974) on potential outcomes had no need to distinguish random variables and vertices, and considered only a single treatment variable. For this reason the potential outcome literature used e.g. $Y(1)$ where Y is the outcome, and 1 is the value assumed by the treatment variable. We write the same potential outcome as $X_y(x_a)$, in order to remove notational ambiguities which may arise with the simpler notation in our case.

[8] Note that in general we may have cases where $X_b(x_a)$ depends on x_a for each unit, but this dependence is masked on average in the sense that (8) holds. The definition given here corresponds to the absence of a direct effect on the population level.

[9] The opposite is not true! To see this, suppose the causal DAG model associated with $\mathcal{G}$ is true. In this case, for all $A \subseteq V$, $P(x_{V \setminus A} \mid \text{do}(x_A))$ obeys (7) with respect to $\mathcal{G}$. In general there may exist

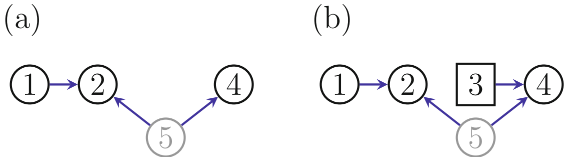

Figure 3: Two graphical representations of the truncated factorization of an interventional distribution $P(x_1, x_2, x_4, x_5 \mid \mathrm{do}(x_3))$ corresponding to a causal model associated with the graph in Fig. 1. (a) The node corresponding to X_3 is deleted from the DAG. (b) The node corresponding to X_3 is kept but is denoted by a square to denote the corresponding variable is now a fixed constant. In both cases all arrows pointing into an intervened-on node are removed.

statistical DAGs (to which we alluded earlier) translate directly to difficulties with latent variable graphical models under a causal interpretation, such as semi-Markovian models (Pearl, 2000).

This also implies that Markov equivalence places limits on the ability of structure learning algorithms to infer cause effect relationships from the observed distribution $P(x_V)$. For example, if we assume the true DAG is causal, and moreover is given by Fig. 2(a)[10], then without additional parametric assumptions no structure learning algorithm will be able to determine whether X_1 is a direct cause of X_2 or vice versa. This is because a causal DAG in Fig. 2(a), where the former is true, and a causal DAG in Fig. 2(c), where the latter is true, induce equivalent statistical models.

2.3.1 Graphical Representations of Independence in Interventional Distributions

If $P(x_V)$ is Markov relative to a DAG $\mathcal{G}$, and (7) holds for $A \subseteq V$, this immediately implies that $P(x_{V \setminus A} \mid \mathrm{do}(x_A))$ is Markov relative to a DAG obtained from $\mathcal{G}$ by removing all vertices in A and all edges adjacent to such nodes. For the DAG $\mathcal{G}$ in Fig. 1(a), and $A = \{3\}$, such a graph is shown in Fig. 3(a).

Conditional independence relationships in $P(x_{V \setminus A} \mid \mathrm{do}(x_A))$ may be read off this graph using standard d-separation. New d-separation statements may be introduced by the removal of edges adjacent to A. For instance, in Fig. 3(a), 1 is d-separated from 4, while the same is not true in Fig. 1(a).

One difficulty with simply removing intervened-on vertices is that the resulting graph no longer holds any record of the intervention performed. That is, it is not immediately obvious if Fig. 3(a) refers to $P(x_1, x_2, x_4, x_5 \mid \mathrm{do}(x_3))$ or some non-intervened density $P(x_1, x_2, x_4, x_5)$. In particular, nowhere in the graph is the dependence of (some parts of) the distribution on the value assigned to X_3 displayed.

An alternative representation that addresses this deficiency is the so-called "manipulated graph" (Spirtes et al., 1993), also known as the "mutilated graph" (Pearl,

another DAG $\mathcal{G}'$ Markov equivalent to $\mathcal{G}$. In this case $P(x_V)$ will obey (2) with respect to $\mathcal{G}'$, and hence the statistical model associated with $\mathcal{G}'$ will be true, but (7) will not hold for all $A \subseteq V$, so the causal model associated with $\mathcal{G}'$ is false.

[10] Here we also suppose that no subgraph of Fig. 2(a) is true, so that, in particular, X_1 is a direct cause of X_2.

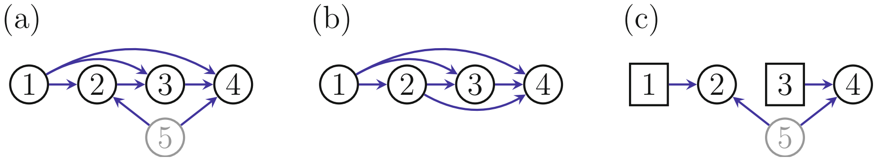

Figure 4: (a) A DAG representing a two stage longitudinal study with treatments X_1, X_3 and outcome X_4; the intermediate variable X_2 is observed; X_5 represents underlying health status, and is not observed. X_1 is randomized, while X_3 is randomized conditional on X_1 and X_2. (b) A DAG representing a similar causal model where there is no unobserved common cause of X_2 and X_4, but X_2 instead has a direct effect on X_4. (c) A "manipulated graph" representing the interventional distribution $P(x_4, x_2 \mid \mathrm{do}(x_1, x_3))$ obtained from $P(x_1, x_2, x_3, x_4)$ corresponding to the study in (a) under the null hypothesis of no direct effect of X_1 on X_4 which is represented by the absence of a directed arrow from 1 to 4.

2000), which keeps intervened-on nodes, but removes all arrows pointing to such nodes. An additional refinement is to mark intervened-on nodes in some way to signify that they are no longer random variables, but constants. In this performed we will draw intervened-on nodes as squares, and ordinary random variable nodes as circles. Fig. 3(b) shows the resulting graph, which represents $P(x_1, x_2, x_4, x_5 \mid \mathrm{do}(x_3))$ obtained after $\mathrm{do}(x_3)$ was applied to a distribution compatible with Fig. 1(a).

2.3.2 A Short History of Generalized Independence Constraints

The generalized independence constraint (1) first arose in the analysis of direct effects in longitudinal studies (Robins, 1986, §§8 & 9). The graph in Fig. 4(a) represents a simple example of a longitudinal study in which such constraints arise. Fig. 4 represents a study in which an HIV drug combination, referred to as HAART, is administered to HIV-positive patients twice (X_1 and X_3); X_2 is the immune status of the patient after the first treatment; X_4 represents the final endpoint of the study, survival; lastly X_5 is the patient's underlying health status, which is unobserved. The initial drug assignment (X_1) is randomized, which is represented in the graph by a lack of incoming arrows to 1. The second instance of the drug treatment (X_3) is given according to the policy followed by the patient's doctor, which depends only on whether the first treatment was administered (X_1) and the patient's immune status (X_2), which in turn is influenced by X_1. In general, we expect the outcome (X_4) to be influenced by both treatments. We also expect the patient's health status X_5 to influence both their immune status (X_2) and survival (X_4).

Since X_5 is not observed, the causal model is a hidden variable DAG. Assume we are interested in the causal effect of both treatments on the outcome, in other words we are interested in $P(x_4 \mid \mathrm{do}(x_1, x_3))$ (or some function thereof). Had we been interested in this effect in a similar causal model represented by a DAG with no hidden variables shown in Fig. 4(b), then (7) implies that this effect would be given by

$$P(x_4 \mid \mathrm{do}(x_1, x_3)) = \sum_{x_2} P(x_4 \mid x_1, x_2, x_3) P(x_2 \mid x_1). \tag{9}$$

Robins (1986) noticed that the same formula can be used to obtain this effect from $P(x_1, x_2, x_3, x_4)$ in Fig. 4(a). From this, it is easy to see that under the null hypothesis of no direct effect of X_1 on X_4 (where Fig. 4(a) reduces to Fig. 1(a)), the constraint (1) holds in $P(x_1, x_2, x_3, x_4)$. If we restrict attention to d-separation statements involving only vertices $1, 2, 3$ and 4, we note that no such statements hold in the graph, and thus no conditional independences hold in the marginal $P(x_1, x_2, x_3, x_4)$ obtained from a distribution factorizing according to Fig. 1(a). What Robins (1986) thus demonstrated is that in some natural marginal distributions there exist equality constraints that *are not, nor are implied by,* conditional independence constraints on variables in that marginal!

Robins (1986) did not provide a DAG-based view of this constraint. This was provided later by Spirtes et al. (1993), and Robins (1999). If $P(x_4, x_2 \mid \text{do}(x_1, x_3)) = P(x_4 \mid x_1, x_2, x_3)P(x_2 \mid x_1)$ holds in distributions compatible with Fig. 1(a) when X_5 is not observed, and we assume no direct effect of X_1 and X_3 on X_4, then Fig. 4(c) is a manipulated graph representation of the corresponding interventional distribution. In this graph vertices 2 and 4 are random and vertices 1 and 3 are fixed, and the graph represents a set of distributions over X_2, X_4, indexed by values x_1 and x_3, of the form:

$$q(x_2, x_4 \mid x_1, x_3) = P(x_4 \mid x_1, x_2, x_3)P(x_2 \mid x_1).$$

However, this distribution is not "conditional" in the sense that it was not obtained from $P(x_1, x_2, x_3, x_4)$ by conditioning on X_1 and X_3. In subsequent sections we will call such distributions "kernels," following Lauritzen (1996).

The null hypothesis of no direct effect of X_1 on X_4 is expressed in this kernel by the statement that $\sum_{x_2} q(x_2, x_4 \mid x_1, x_3)$ is not a function of x_1. The corresponding graphical statement is that 1 is d-separated from 4 in Fig. 4(c); see (Robins, 1999). Note that one of the vertices in this statement is fixed and one is random[11].

Given the existence of such constraints in certain cases, the natural question to consider is under what conditions such constraints arise in general. Systematic approaches for obtaining such constraints in hidden variable causal DAGs were given by Robins (1999) and Tian and Pearl (2002). An algorithm for listing constraints with a causal interpretation among variables that were not fixed was given by Shpitser and Pearl (2008). Shpitser et al. (2009) used (some of) these constraints to derive a general method for testing for the presence of edges in causal graphs.

For the purposes of performing model search whilst avoiding the difficulties associated with latent variables we would like to construct models that are *defined* via distributions for the observed variables that obey independence in kernels derived from these distributions. A series of papers pursuing this line of research (Richardson et al., 2012; Shpitser et al., 2013) culminated in the development of *nested Markov models.* The remainder of this paper will illustrate this development via examples,

[11] More precisely, Robins (1999) interpreted the parentless node 3 in Fig. 4(c) as random rather than fixed; the manipulated graph could thereby be identified with a clinical trial in which variable 3 had been randomly assigned.

and compare it with the 'ordinary' approach based on using *only* conditional independences.

3. Overview of ADMGs, Ordinary and Nested Markov Models.

In general a DAG possibly containing latent variables such as Fig. 4(a) implies conditional independence restrictions on the distribution of the observables, but as we saw in the last section may also imply constraints such as (1).

As noted earlier, in general, the set of observed distributions implied by a DAG containing latent variables is complicated: the set is defined by equality constraints, which include independence restrictions and generalized independences such as (1), but also by inequality constraints.

We consider two natural larger sets of distributions, defined solely by equality constraints, that contain the set of distributions over the observed variables implied by the latent variable DAG model:

- The *ordinary Markov model* consists of the set of distributions over the observed variables that obey the conditional independence relations implied by the original DAG with latents;
- The *nested Markov model* is the set of distributions obeying both the conditional independence restrictions and generalized independence restrictions such as (1) that are implied by the DAG with latents.

By definition, we will have the following relations between these models:

$$\text{latent variable DAG} \subseteq \text{nested Markov model} \subseteq \text{ordinary Markov model}.$$

Note that these inclusion relations are in general strict. In particular this means that there may be distributions in the nested Markov model associated with a DAG $\mathcal{D}$ with latents, that could not be obtained as the observed marginal of any distribution that is Markov with respect to $\mathcal{D}$. For example, there may be distributions in the nested Markov model that do not obey inequality restrictions implied by $\mathcal{D}$.

This also means that there are distributions in the nested Markov model for which there is no causal interpretation in terms of the original DAG $\mathcal{D}$. Consequently, although it is true that if we start with a causal DAG such as the one in Fig. 4(a), then the generalized conditional independence restrictions (1) defining the nested Markov model have a causal interpretation via (9), in general there are distributions in the nested Markov model that obey (1) yet for which this constraint cannot be interpreted causally since the joint distribution $P(x_1, x_2, x_3, x_4)$ is not the marginal of any DAG.

It is for this reason that in the following development we are careful to *define* the nested Markov model in terms of constraints on (functionals of) the joint distribution of the observed variables such as (9) and *not* directly in terms of intervention distributions, such as $P(x_4 \mid \text{do}(x_1, x_3))$: the assumption that such intervention distributions exist may not make any sense if our observed distribution does not obey inequality restrictions, hence there is no DAG to which we could apply the formula (7).

Since both the nested Markov and ordinary Markov models are defined by constraints implied by DAGs with latent variables, we could choose simply to work with DAGs with latent variables. However, this suffers from the disadvantages that there are infinitely many such DAGs, and many DAGs will give rise to exactly the same latent projections[12)].

Instead we work with acyclic directed mixed graphs (ADMGs) which contain two types of edges, directed ($\rightarrow$) and bidirected ($\leftrightarrow$). Given a DAG with latents the ADMG is derived via a simple "latent projection" operation. The ADMG encodes the constraints defining the ordinary Markov model via m-separation, which is the natural extension of d-separation to mixed graphs (Richardson, 2003). The ADMG also encodes the constraints defining the nested Markov model via m-separation in conjunction with a simple "fixing" operation[13)].

If one is interested *solely* in the ordinary Markov model, not the nested Markov model, then Richardson and Spirtes (2002) showed that given a DAG with latents it is possible to construct a special kind of ADMG, called a maximal ancestral graph (MAG) that encodes (via m-separation) the conditional independence restrictions implied by the original DAG (via d-separation). Furthermore, in the Gaussian case, the MAG leads directly to a simple parametrization of the model by viewing the graph as a path diagram. MAGs are also useful for characterizing when two DAGs with latents will imply the same ordinary Markov model (Ali et al., 2009).

3.1 Acyclic Directed Mixed Graphs (ADMGs)

A *directed mixed graph* $\mathcal{G}(V, E)$ is a graph with a set of vertices V and a set of edges E which may contain directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges. An *acyclic* directed mixed graph (ADMG) is a mixed graph containing no directed cycles. A vertex b is a *spouse* of a in an ADMG $\mathcal{G}$, if an edge $a \leftrightarrow b$ exists in $\mathcal{G}$. The set of spouses of a in $\mathcal{G}$ is denoted by $\mathrm{sp}_{\mathcal{G}}(a)$. Note that there may be two edges between a given pair of variables in an ADMG: $a \rightarrow b \leftrightarrow a$.

3.1.1 The m-separation criterion

We introduce the natural extension of d-separation to mixed graphs. A non-endpoint vertex z on a path is a *collider on the path* in an ADMG if the edges preceding and succeeding z on the path have an arrowhead at z, i.e. $\rightarrow z \leftarrow$, $\leftrightarrow z \leftrightarrow$, $\leftrightarrow z \leftarrow$ or $\rightarrow z \leftrightarrow$. A non-endpoint vertex z on a path which is not a collider is a *non-collider on the path*, i.e. $\leftarrow z \rightarrow$, $\leftarrow z \leftarrow$, $\rightarrow z \rightarrow$, $\leftrightarrow z \rightarrow$ or $\leftarrow z \leftrightarrow$. A path between vertices a and b in a mixed graph $\mathcal{G}$ is said to be *m-connecting given a set* C if every non-collider on the path is not in C, and every collider on the path is an ancestor of C in $\mathcal{G}$. If there is no path m-connecting a and b given C, then a and b

[12)] Or exactly the same maximal ancestral graphs, a special type of ADMG defined later.

[13)] For a distribution that is in the original DAG model with latent variables this operation has an obvious causal interpretation, but as noted above, this interpretation is not in general valid for all distributions in the nested model.

are said to be *m-separated* given C. Sets A and B are said to be *m-separated* given C, if every pair $a \in A$ and $b \in B$ are m-separated given C. Note that if $\mathcal{G}$ is a DAG then the above definition is identical to Pearl's d-separation criterion.

3.1.2 Latent Projections

Given a DAG with latent variables we associate with it the following ADMG, originally defined by Verma and Pearl (1990):

Definition 4 (latent projection) *Let $\mathcal{G}$ be a DAG with vertex set $V \dot\cup L$ where the vertices in V are observed, while those in L are latent. The* latent projection $\mathcal{G}(V)$ *is a directed mixed graph with vertex set V, where for every pair of distinct vertices $v_i, v_j \in V$:*

(i) *$\mathcal{G}(V)$ contains $v_i \rightarrow v_j$ iff there is a directed path $v_i \rightarrow \cdots \rightarrow v_j$ on which every non-endpoint vertex is in L.*
(ii) *$\mathcal{G}(V)$ contains $v_i \leftrightarrow v_j$ iff there exists a path of the form $v_i \leftarrow \cdots \rightarrow v_j$, on which every non-endpoint vertex is a non-collider and in L.*

Note that a single edge also counts as a path so that any edge between vertices in V that is present in $\mathcal{G}$ will also be present in $\mathcal{G}(V)$. A latent projection of a DAG $\mathcal{G}(V \dot\cup L)$ over $V \dot\cup L$ onto a set V is always an ADMG $\mathcal{G}(V)$.

The ADMG in Fig. 5(e) is a latent projection of the DAG in Fig. 5(a); the ADMG in Fig. 5(f) is a latent projection of both the DAG in Fig. 5(b) and in Fig. 5(c); lastly the ADMG in Fig. 5(g), which is also a DAG, is a latent projection of the DAG in Fig. 5(d).

3.1.3 Ordinary Markov Models of ADMGs

Given disjoint subsets $A, B, C \subseteq V$, if A is d-separated from B given C in a DAG $\mathcal{G}(V \dot\cup L)$ then A is m-separated from B given C in the latent projection $\mathcal{G}(V)$. Thus $\mathcal{G}(V)$ encodes all conditional independence constraints that hold in the marginal $P(x_V) = \sum_L P(x_{V \dot\cup L})$, where $P(x_{V \dot\cup L})$ is Markov relative to the DAG $\mathcal{G}(V \dot\cup L)$.

This motivates the definition of the *ordinary Markov model* of an ADMG $\mathcal{G}(V)$.

Definition 5 *A distribution $P(x_V)$ is said to be in the* ordinary Markov model *of an ADMG $\mathcal{G}$ if for every disjoint $A, B, C \subseteq V$, where C may be empty, if A is m-separated from B given C in $\mathcal{G}$, then $(X_A \perp\!\!\!\perp X_B \mid X_C)$ holds in $P(x_V)$.*

Richardson (2003) gave an (ordered) local Markov property that may be used to define this model, as well as a global property based on transforming to an undirected graph (analogous to "moralization" for DAGs (Lauritzen, 1996)).

3.1.4 Maximal Ancestral Graphs

As mentioned above, Richardson and Spirtes (2002) show that if one is interested in the ordinary Markov model associated with $\mathcal{G}$, a DAG with latents, then there is a particular ADMG, called the maximal ancestral graph (MAG) associated with $\mathcal{G}$,

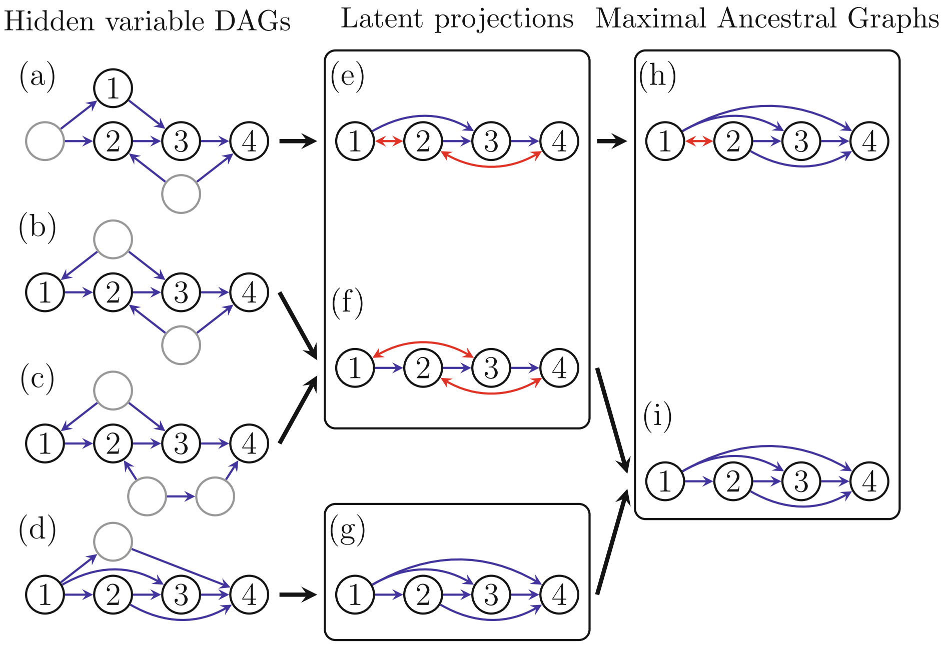

Figure 5: (a), (b), (c), (d) Hidden variable DAGs. (e) The latent projection of (a). (f) The latent projection of both (b) and (c). (g) The latent projection of (d). (h) The maximal ancestral graph resulting from (a). (i) The maximal ancestral graph resulting from (b), (c), (d). As indicated by the rectangles the nested Markov models associated with the latent DAGs form two nested Markov equivalence classes: $\{(e),(f)\}$ and $\{(g)\}$, while the ordinary Markov models are all equivalent (and saturated).

that it is sometimes useful to construct[14].

Given a DAG $\mathcal{G}$ where C may be empty, then the MAG $\mathcal{G}^m(V)$ may be constructed as follows:

- Every pair of vertices $a, b \in V$ in $\mathcal{G}$ that are connected by an *inducing path* become adjacent in $\mathcal{G}^m(V)$, where a path between a and b is *inducing* if every collider on the path is in $\text{an}_{\mathcal{G}}(\{a,b\})$, and every non-collider on the path is in L[15]. Note that a single edge is trivially an inducing path.
- An edge connecting a and b in $\mathcal{G}^m(V)$ is oriented as follows: if $a \in \text{an}_{\mathcal{G}}(b)$, then the edge is $a \to b$; If $b \in \text{an}_{\mathcal{G}(V)}(a)$, then $b \to a$; If neither is the case, then $a \leftrightarrow b$.

Note that if we have already constructed the latent projection $\mathcal{G}(V)$ then the MAG $\mathcal{G}^m(V)$ may be constructed by applying this procedure to $\mathcal{G}(V)$, rather than to the original DAG $\mathcal{G}(V \dot{\cup} L)$.

The resulting ADMG $\mathcal{G}^m(V)$ is called *ancestral* because if there is an edge between

[14] MAGs as defined (Richardson and Spirtes, 2002) in their full generality are defined with respect to a graph containing three kinds of edges: directed, bidirected, and undirected. The third kind of edge is associated with conditional distributions meant to model settings with selection bias, and will be ignored in this paper.

[15] This implies that there is no subset C of $V \setminus \{a,b\}$ such that a and b are d-separated given C in $\mathcal{G}$.

a and b in $\mathcal{G}^m(V)$ then a is an ancestor of b in $\mathcal{G}$ if and only if $a \rightarrow b$ in $\mathcal{G}^m(V)$[16]. $\mathcal{G}^m(V)$ is also called *maximal* because for any pair of non-adjacent variables a and b there is some set C such that a and b are m-separated given C. Consequently, no more pairs of vertices may be made adjacent by adding edges to $\mathcal{G}^m(V)$ without changing the set of m-separation relations that are encoded.

Continuing the discussion of the example in Fig. 5, the ADMG in Fig. 5(h) is the MAG corresponding to the DAG in Fig. 5(a); the ADMG in Fig. 5(i) is the MAG corresponding to the DAGs in 5(b), (c), (d). Note that in Fig. 5(i), 1 and 4 are adjacent even though they were not adjacent in Fig. 5(f). This reflects the fact that even though 1 and 4 are not adjacent, there is no subset of the other observed variables ($\{2, 3\}$) that d-separates 1 and 4 in the DAGs in Fig. 5(b), (c); this is a consequence of the fact that $1 \rightarrow 2 \leftarrow \cdots 4$ forms an inducing path in both of these DAGs. Both MAGs in Fig. 5(h) and (i) imply no m-separation statements, and thus both define a saturated ordinary Markov model over 4 variables.

The following generalizes the notion of equivalence for DAGs:

Definition 6 (ordinary Markov equivalence) *Two MAGs $\mathcal{G}_1^m$ and $\mathcal{G}_2^m$ are said to be* (ordinary) Markov equivalent *if their associated ordinary Markov models are the same, in other words if $\mathcal{G}_1^m$ and $\mathcal{G}_2^m$ encode the same m-separation relations.*

There is a characterization of (ordinary) Markov equivalence for MAGs that generalizes Theorem 3. It states that two MAGs $\mathcal{G}_1^m$, $\mathcal{G}_2^m$ are (ordinary) Markov equivalent if and only if they share adjacencies, and the same so-called *colliders with order*, which include unshielded colliders, and certain other colliders that depend on the arrangement of bidirected arcs (the full definition is beyond the scope of this paper; see (Ali et al., 2009)).

The latent projection $\mathcal{G}(V)$ and the MAG $\mathcal{G}^m(V)$ obtained from a DAG $\mathcal{G}(V \dot{\cup} L)$ always correspond to the same ordinary Markov model. It follows that if two DAGs with latents, $\mathcal{G}_1(V \dot{\cup} L_1)$, $\mathcal{G}_2(V \dot{\cup} L_2)$ lead to MAGs $\mathcal{G}_1^m(V)$ and $\mathcal{G}_2^m(V)$ that are ordinary Markov equivalent, then the ordinary Markov models associated with the latent projections $\mathcal{G}_1(V)$, $\mathcal{G}_2(V)$ will likewise be the same. Consequently, the class of ordinary Markov models represented by some ADMG, and those represented by some MAG are the same. Hence for the purpose of model search it is sufficient to consider MAGs.

Both constraint-based (Spirtes et al., 1993; Colombo et al., 2012; Claassen et al., 2013) and score-based (Evans and Richardson, 2010) structure learning algorithms exist for MAGs. Under the assumption of faithfulness these algorithms will recover, in sufficiently large samples, the (ordinary) Markov equivalence class of MAGs corresponding to the ordinary Markov model for the hidden variable DAG that generated the data. Furthermore, if there is an edge $a \rightarrow b$ in every MAG in the ordinary Markov equivalence class, then a is an ancestor of b in the hidden variable DAG that generated the data.

[16] Here, as stated earlier, we are ignoring selection bias, and hence restricting to ancestral graphs without undirected edges.

3.2 Conditional ADMGs and Kernels

To give a general treatment of constraint (1), and nested Markov models that are defined via such constraints, we viewed the functional $\sum_{x_2} P(x_4 \mid x_1, x_2, x_3)P(x_2 \mid x_1)$ as a "kernel" $q(x_4 \mid x_1, x_3)$, that is a mapping from values x_1, x_3 onto distributions over X_4, and the constraint itself as an independence statement in this kernel. We will display constraints in kernels via a modification of an ADMG which allows for fixed nodes, called a *conditional* ADMG (CADMG). A CADMG $\mathcal{G}(V, W, E)$ is an ADMG with two disjoint sets of vertices V and W, subject to the restriction that for all $w \in W$, $\mathrm{pa}_{\mathcal{G}}(w) = \emptyset = \mathrm{sp}_{\mathcal{G}}(w)$. The rationale for excluding edges between vertices in W or with arrowheads in W is that the CADMG represents the structure of a kernel $q_V(x_V \mid x_W)$; the vertices in W merely index distributions over V.

Following Lauritzen (1996, p.46), we define a *kernel* to be a non-negative function $q_V(x_V \mid x_W)$ satisfying:

$$\sum_{x_V \in \mathfrak{X}_V} q_V(x_V \mid x_W) = 1 \qquad \text{for all } x_W \in \mathfrak{X}_W. \tag{10}$$

where $\mathfrak{X}_W$ is the probability space corresponding to X_W.

We use the term "kernel" and write $q_V(\cdot|\cdot)$ (rather than $P(\cdot|\cdot)$) to emphasize that these functions, though they satisfy (10) and thus most properties of conditional densities, will not, in general, be formed via the usual operation of conditioning on the event $X_W = x_W$. To conform with standard notation for densities, we define for every $A \subseteq V$,

$$q_V(x_A|x_W) \equiv \sum_{V \setminus A} q_V(x_V|x_W), \quad \text{and} \quad q_V(x_{V\setminus A}|x_{W\cup A}) \equiv \frac{q_V(x_V|x_W)}{q_V(x_A|x_W)}.$$

3.2.1 Independence in Kernels

As discussed in Section 2.3.2, generalized independence constraints are really independences in kernels, and may involve both random and fixed variables. Here we give a formal definition of this kind of independence.

Definition 7 *For disjoint subsets $A, B, C \subseteq V \cup W$, we define X_A to be conditionally independent of X_B given X_C under the kernel q_V, written:*

$$X_A \perp\!\!\!\perp X_B \mid X_C \quad [q_V(X_V|X_W)]$$

if **either***:*

(a) *$A \cap W = \emptyset$ and $q_V(x_A \mid x_B, x_C, x_{W\setminus(B\cup C)})$ is a function only of x_A and x_C (whenever this kernel is defined),* **or**

(b) *$B \cap W = \emptyset$ and $q_V(x_B \mid x_A, x_C, x_{W\setminus(A\cup C)})$ is a function only of x_B and x_C (whenever this kernel is defined).*

Note that it follows immediately from this definition that $(X_A \perp\!\!\!\perp X_B \mid X_C)$ if and only if either $(X_A \perp\!\!\!\perp X_{B\cup(W\setminus C)} \mid X_C)$ or $(X_B \perp\!\!\!\perp X_{A\cup(W\setminus C)} \mid X_C)$.

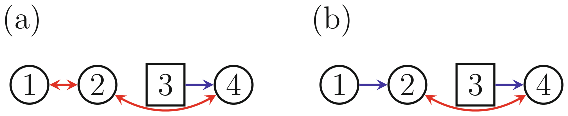

Figure 6: (a), (b) CADMGs obtained by fixing 3 in Fig. 5(e) and (f) respectively. Note that 1 is m-separated from 4 (given 3).

3.3 Nested Markov Models

The constraints defining a nested Markov model may be derived from the latent projection (ADMG) corresponding to the original DAG with latents.

3.3.1 The Global Nested Markov Property

There are a number of alternative definitions of the nested Markov model. One such definition, the nested global Markov property, is described in the Appendix. It links a marginal distribution $P(x_V)$ with an ADMG $\mathcal{G}$ by associating m-separation statements in CADMGs (obtained from $\mathcal{G}$ by sequences of fixing operations) with kernel objects (obtained from $P(x_V)$ by sequences of divisions by certain kernels). Examples of this link between separation statements in CADMGs and independence statements in kernels are given below.

3.3.2 Relationship Between Ordinary and Nested Markov Models of Hidden Variable DAGs

Both ordinary and nested Markov models imply conditional independence constraints via m-separation in the ADMG. For example, no m-separation statements hold in the ADMGs in Fig. 5(e) and (f), and thus the nested Markov model entails no conditional independences on the observed margin.

The nested model also implies additional generalized independence constraints by means of additional m-separation statements in certain CADMGs derived from the original ADMG. For example, the CADMGs in Fig. 6(a) and (b) are obtained from Fig. 5(e) and (f) respectively by "fixing" the vertex 3. This is represented graphically by removing all incoming arrows to 3 (including bidirected edges) and rendering 3 as a square[17]. In these CADMGs, 1 is m-separated from 4 given 3. This m-separation corresponds to the independence $(X_1 \perp\!\!\!\perp X_4 \mid X_3)$ in the kernel obtained by dividing the marginal distribution $P(x_1, x_2, x_3, x_4)$ by $P(x_3 \,|\, x_2, x_1)$, to obtain $q_{1,2,4}(x_1, x_2, x_4 \mid x_3) \equiv P(x_4 \,|\, x_3, x_2, x_1)P(x_2, x_1)$. In other words, these nested models imply (1).

Since nested Markov models are defined via both conditional independence constraints and generalized constraints such as (1), we can use the presence or absence of such constraints to distinguish between graphs that correspond to the same ordinary Markov model. In fact, (1) holds in hidden variable DAGs corresponding to Fig. 5(e)

[17] However, in general, variables may be fixed in this way only if they have no descendant that can also be reached by a path containing only bi-directed edges; see the definition of 'fixable' in the Appendix.

and (f), but not in a hidden variable DAGs corresponding to Fig. 5 (g).

Thus, the latent DAGs in Fig. 5 that all give rise to the same ordinary Markov model can be partitioned into multiple equivalence classes since they imply different nested Markov models. Specifically, the DAGs in Fig. 5(a), (b), (c) give rise to the latent projections in Fig. 5(e) and (f) both of which imply (1). In contrast the DAG Fig. 5(d) results in the latent projection shown in Fig. 5(g) that does not imply (1).

A major motivation for the development of nested Markov models is that constraints such as (1) are more informative than simple conditional independence constraints for learning structural features of a hidden variable DAG.

For example, the DAG in Fig. 5(a) does not imply any independence relations over the observed variables, and hence gives rise to an ordinary Markov model that is saturated. This (empty) set of conditional independence relations is encoded by both the latent projection Fig. 5(e) for this DAG and the MAG, shown in Fig. 5(h). However, the DAG in Fig. 5(a) implies (1), which is encoded in a CADMG shown in Fig. 6 derived from the latent projection in Fig. 5(e).

Similarly, like (a), none of the DAGs in Fig. 5(b), (c) and (d) imply any conditional independences hence they all give rise to the same ordinary Markov model, which is saturated. However, also like (a), (b) and (c) both imply (1), and lead to the same latent projection (f) that is associated with the same nested Markov model as (e).

In contrast, the DAG in Fig. 5(d) does not imply (1). The corresponding latent projection shown in (g) gives rise to a saturated nested Markov model. Consequently the ADMG in Fig. 5(g) is associated with a different nested Markov model from the ADMGs in (e) or (f). Notwithstanding this, as noted already, all of the DAGs in Fig. 5(a), (b), (c) and (d) are associated with the same (saturated) ordinary Markov model.

3.3.3 Connections with Structural Nested Models

As discussed in Section 2.3.2, Robins (1986) motivated the general problem of testing generalized independence constraints as a way of establishing the lack of direct causal effects in longitudinal trials. Robins (1999) viewed these constraints graphically, as corresponding to missing edges in causal DAGs (and thus as d-separation statements in a world in which interventions on certain variables had been randomly assigned as in a clinical trial), and gave a general procedure for testing for the presence of certain types of generalized independence constraints that can be viewed as an independence of the outcome variable and some randomized longitudinal treatment variable, given that some other treatment variables were also randomized. This was done by reparametrizing a hidden variable causal DAG model using direct-effect structural nested models.

The procedure is quite general, and gives correct tests in a wide variety of graphs, in particular it gives the correct test for (1) in the graph in Fig. 1(a), and the constraint in Fig. 7(c). However, the procedure only considers tests in settings where direct effects can identified by means of the g-computation algorithm (Robins, 1986), which excludes some situations where generalized constraints are present, and indeed are en-

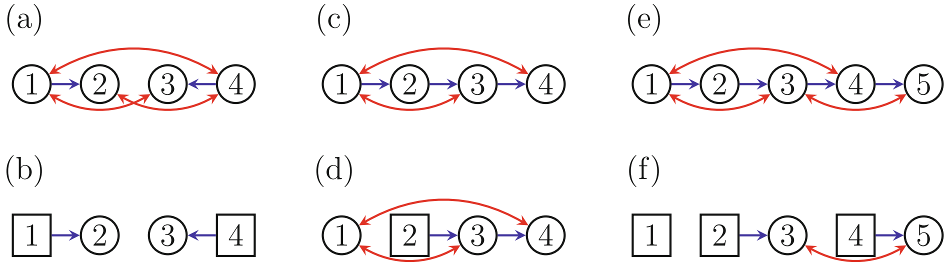

Figure 7: Given any causal DAG giving rise to the latent projection (a), the dormant independence $(X_3 \perp\!\!\!\perp X_2 \mid \text{do}(x_1, x_4))$ holds. However, there is no constraint implied by the nested Markov property; (b) A CADMG corresponding to $P(x_2, x_3 \mid \text{do}(x_1, x_4))$ given a causal DAG whose latent projection is (a). (c) An ADMG where the independence constraint $(X_4 \perp\!\!\!\perp X_2 \mid X_3)$ in $q_{1,3,4}(x_1, x_3, x_4 \mid x_2)$ is implied by the nested global Markov property, but does not correspond to a dormant independence because X_2 must be fixed in order to obtain this independence; (d) A CADMG corresponding to the kernel $q_{1,3,4}(x_1, x_3, x_4 \mid x_2)$ obtained by fixing x_2 under the nested Markov model associated with (c). (e) An ADMG corresponding to a nested Markov model where a generalized independence between X_2 and X_5 in a kernel $q_{3,5}(x_3, x_5 \mid x_4, x_2)$ is encoded by (f). (f) A CADMG obtained from (e) which implies $X_2 \perp\!\!\!\perp X_5 \mid X_4$.

coded by the nested global Markov property. For instance, in the graph in Fig. 7(e), there is a constraint that can be viewed causally as stating that $P(x_5 \mid \text{do}(x_2, x_4))$ is not a function of x_2, and graphically encoded by m-separation of 2 and 5 given 4 in Fig. 7(f), the CADMG resulting from fixing 2, 1 and 4. The corresponding generalized constraint on the observed margin is

$$\frac{\partial}{\partial x_2} \frac{\left(\sum_{x_1} P(x_5, x_4, x_3 \mid x_2, x_1) P(x_1)\right) \cdot \left(\sum_{x_1} P(x_3 \mid x_2, x_1) P(x_1)\right)}{\sum_{x_1} P(x_4, x_3 \mid x_2, x_1) P(x_1)} = 0$$

which is equivalent to $X_2 \perp\!\!\!\perp X_5 \mid X_4$ in the kernel $q_{3,5}(x_5, x_3 \mid x_2, x_4)$; this kernel is obtained from $P(x_1, x_2, x_3, x_4, x_5)$ by first dividing by $P(x_2 \mid x_1)$, then marginalizing X_1, to obtain $q_{3,4,5}(x_3, x_4, x_5 \mid x_2)$, and then dividing $q_{3,4,5}(x_3, x_4, x_5 \mid x_2)$ by $q_{3,4,5}(x_4 \mid x_3, x_2)$ to obtain $q_{3,5}(x_5, x_3 \mid x_2, x_4)$[18].

The reason Robins (1999) excludes this case is because the g-computation algorithm sequentially applies the g-formula to a fixed ordering of the variables, whereas in this example the correct strategy for identifying the causal effect entails applying the g-formula to the past treatment X_2 and then marginalizing over X_1 before applying the g-formula to X_4. The marginalization of the earlier variable X_1 is equivalent to a reordering the variables before applying the g-formula to X_4.

3.3.4 Relationship to the Null Paradox in Causal Inference

Consider again the longitudinal study in Section 2.3.2, represented by the graph in

[18] The name *nested Markov models* is partly a reference to the nature of the Markov properties and factorizations of these models, which entail independences in kernels which "nest" within each other like Matryoshka dolls (for example the kernel $q_{3,5}(x_5, x_3 \mid x_2, x_4)$ "nests" inside $q_{3,4,5}(x_3, x_4, x_5 \mid x_2)$), and partly an allusion to structural *nested* models (Robins and Wasserman, 1997; Robins, 1997, 1999) which were the first general model that dealt with generalized constraints such as (1).

Fig. 4(a), where we wish to test for the absence of a direct effect of 1 on 4, that is, we wish to test whether $P(x_4 \mid \text{do}(x_1, x_3))$ is a function of x_1.

Since $P(x_4 \mid \text{do}(x_1, x_3)) = \sum_{x_2} P(x_4 \mid x_3, x_2, x_1)P(x_2 \mid x_1)$, a natural approach here is to model the conditional densities $P(x_4 \mid x_3, x_2, x_1)$ and $P(x_2 \mid x_1)$ in some way, for example by using regression models, and then construct a hypothesis test to check whether $\sum_{x_2} P(x_4 \mid x_3, x_2, x_1)P(x_2 \mid x_1)$ is a function of x_1.

A startling fact is that assuming the null hypothesis (no direct effect of 1 on 4) is true, very natural models are guaranteed to be misspecified, and hypothesis tests based on these models will reject the null with probability converging to 1. In particular, this occurs if X_2 is discrete, X_1 is either continuous or categorical with a large number of levels, $P(x_2 \mid x_1)$ is modeled using logistic regression, X_4 is continuous, and $P(x_4 \mid x_3, x_2, x_1)$ is represented using a linear Gaussian regression model (Robins and Wasserman, 1997). This is known as the "null paradox"

This paradox arises because the null hypothesis in question corresponds to a generalized independence constraint on the observed marginal. These constraints are naturally represented in terms of kernel objects, rather than conditional distribution objects.

There exist modeling approaches which avoid the null paradox for specific subsets of the entire set of causal effects defined by the causal model. Such approaches include structural nested models (Robins, 1989, 1992; Robins and Wasserman, 1997), direct effect structural nested models (Robins, 1999, 1997, Appendix C) and marginal structural models (Robins, 2000). However, these approaches were not designed to avoid the null paradox for *all* causal effects defined by the model. Since nested Markov models parameterize the entire marginal density using kernel objects directly, hypothesis tests using these models avoid the null paradox by construction. The disadvantage of nested Markov models is that currently only parameterizations for discrete state spaces are known, whereas structural nested models can accommodate continuous state spaces.

3.3.5 Connections with "Dormant Independences"

Given an ADMG $\mathcal{G}(V)$ representing a latent projection of a causal DAG $\mathcal{G}(V \dot{\cup} L)$ onto V, and a corresponding marginal distribution $P(x_V)$, an algorithm given by Shpitser and Pearl (2008) gave a way to determine for disjoint subsets $A, B \subseteq V$, if there existed sets W, Z such that A was independent from B given Z in the interventional distribution $P(x_A, x_B \mid x_Z, \text{do}(x_W))$, and such that this interventional distribution is *identifiable* from $P(x_V)$ assuming $P(x_V)$ is a marginal of $P(x_{V \dot{\cup} L})$ which is Markov relative to $\mathcal{G}(V \dot{\cup} L)$. Such independences were called *identifiable dormant independence constraints.*

An interventional distribution is identifiable if, assuming $P(x_{V \dot{\cup} L})$ is in the DAG causal model for $\mathcal{G}(V \dot{\cup} L)$, it is possible to exploit (7), and independences embedded in $\mathcal{G}(V)$ to express the interventional distribution as a functional of $P(x_V)$. A complete algorithm for identification of conditional interventional distributions was given by Shpitser and Pearl (2006).

The requirement that the interventional distributions in which the dormant independence held be identifiable was a necessary step to re-express dormant independences as constraints on an observable distribution, as is the case with (1). While many dormant independences correspond to constraints that hold in the nested global Markov property, there are some dormant independences that do not correspond to a constraint on the observable distribution. Such constraints are also not implied by the nested global Markov property. In addition, some constraints implied by the nested global Markov property do not correspond to dormant independences as defined (currently).

As an example of the former, consider a causal DAG whose latent projection is the ADMG shown in Fig. 7(a). Under any such DAG, the distribution $P(x_2, x_3 \mid \mathrm{do}(x_1, x_4))$ is identifiable as $q_{2,3}(x_2, x_3 \mid x_1, x_4) = P(x_2 \mid x_1) \cdot P(x_3 \mid x_4)$, which means that the constraint that X_3 is independent of X_2 given that we perform $\mathrm{do}(x_1, x_4)$ (intervene on X_1 and X_4 to force their values to x_1 and x_4), is an identifiable dormant independence constraint, displayed by d-separation in the CADMG in Fig. 7(b). However, there is no direct analogue of this constraint in the nested global Markov property.

This reflects the fact that, because the nested Markov model is designed to be a tool for causal discovery, it only imposes constraints on the distribution of the observed variables. The identifiable dormant independence constraint in Fig. 7(b) does not in fact constrain the observable margin in any way, since it holds in the kernel $q_{2,3}(x_2, x_3 \mid x_1, x_4)$ by construction!

At the same time, identifiable dormant independences were only defined to hold between sets of random variables, and some constraints that define nested Markov models hold between random and fixed variables. For example, any distribution $P(x_1, x_2, x_3, x_4)$ that obeys the nested global Markov property for the graph in Fig. 7(c) will imply the constraint $(X_4 \perp\!\!\!\perp X_2 \mid X_3)$ in $q_{1,3,4}(x_1, x_3, x_4 \mid x_2) = \frac{P(x_1,x_2,x_3,x_4)}{P(x_2 \mid x_1)}$, which follows from m-separation in the CADMG in Fig. 7(d), obtained by fixing 2. However, because this independence does not correspond to an independence between two random variables, there is no corresponding identifiable dormant independence. While it is possible to consider an additional class of causal constraints of the form "$P(x_4 \mid x_3, \mathrm{do}(x_2))$ is not a function of x_2, and $P(x_4 \mid x_2, \mathrm{do}(x_2))$ is identifiable from $P(x_1, x_2, x_3, x_4)$ given Fig. 7(c)," this was not done in (Shpitser and Pearl, 2008). The graphical approach of Robins (1999) in which the intervention variables are made parentless but left random naturally encodes such constraints.

Finally, dormant independences hold in *interventional* distributions derived from hidden variable causal DAG models. In contrast, as noted in Section 3.3.2 some distributions in the nested Markov model associated with an ADMG $\mathcal{G}$ could not have arisen from any DAG (whose latent projection is $\mathcal{G}$). As a consequence, there are distributions in the nested Markov model for which the (generalized) independence constraints that are implied to hold in certain kernels do not have a causal interpretation in terms of the latent projection of a DAG.

3.3.6 Connections with Tian's Constraint-finding Algorithm

As mentioned earlier, an algorithm for finding constraints similar to (1) was given by Tian and Pearl (2002). This algorithm takes as input a hidden variable DAG $\mathcal{G}(V \dot\cup L)$ where vertices V correspond to observed variables, and vertices in L to hidden variables, and a marginal distribution $P(x_V) = \sum_{x_L} P(x_{V \dot\cup L})$ such that $P(x_{V \dot\cup L})$ is Markov relative to $\mathcal{G}(V \dot\cup L)$.

The algorithm is phrased in terms of "q-factors" over sets of random variables X_R, written $Q[R]$, which correspond to kernels $q_R(x_R \mid x_{V \setminus R})$ (except with fixed variables suppressed from the notation), and subgraphs of $\mathcal{G}(V \dot\cup L)$ which correspond to CADMGs (except with fixed nodes suppressed).

The set of distributions $P(x_V)$ obeying the constraints resulting from an invocation of Tian's algorithm given a DAG $\mathcal{G}(V \dot\cup L)$ as input is conjectured to be equal to the nested Markov model on the latent projection $\mathcal{G}(V)$.

4. Parameterization and Equivalence in Ordinary and Nested Markov Models

DAG models can be parameterized by associating a set of parameters with each Markov factor $P(x_v \mid x_{\mathrm{pa}_{\mathcal{G}}(v)})$ independently. For example, a distribution $P(x_1, x_2, x_3, x_3)$ over binary variables in the model of the DAG in Fig. 2(a) can be parameterized by:

$$P(0_1), P(0_2 \mid x_1), P(0_3 \mid x_1), P(0_4 \mid x_2, x_3)^{19)} \tag{11}$$

which gives a total of $1+2+2+4 = 9$ parameters. These parameters are all conditional probabilities.

Similar parameterizations exist for discrete ordinary and nested Markov models. These parameterizations are complicated by the fact that, in general, there may be multiple parameters that are probabilities of events that include the same random variable. Due to this overlap, both ordinary and nested Markov parameters are in general variation dependent. That is, fixing the value of a particular parameter may constrain the values of others. In addition, while all ordinary Markov parameters are conditional probabilities, nested Markov model parameters are conditional probabilities in kernels.

4.1 The Möbius Parameterization of Binary Ordinary Markov Models

We illustrate the parameterization of binary ordinary Markov models with an example shown in Fig. 8. This ADMG represents the set of densities $P(x_1, x_2, x_3, x_4)$ where conditional independences $(X_1 \perp\!\!\!\perp X_4)$, $(X_2 \perp\!\!\!\perp X_4 \mid X_1)$, and $(X_3 \perp\!\!\!\perp X_1 \mid X_4)$ hold. Note that no DAG model on 4 variables exists that represents exactly this set of independences.

[19)] As a shorthand, we denote by 0_i the event "X_i attains value 0." 1_i is defined similarly.

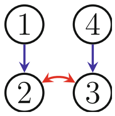

Figure 8: An ADMG for which the nested and ordinary Markov models coincide.

If the vertices in this graph represent binary random variables, then the ordinary Markov model is parameterized as follows:

$$P(0_1), P(0_4), P(0_2 \mid x_1), P(0_3 \mid x_4), P(0_2, 0_3 \mid x_1, x_4) \tag{12}$$

which gives a total of $1+1+2+2+4 = 10$ parameters. Like DAG parameters, these are conditional probabilities. Unlike binary DAG parameters, where each variable X_i corresponds to precisely one family of parameters $P(0_i \mid x_{\mathrm{pa}_{\mathcal{G}}(i)})$ of size $2^{|\mathrm{pa}_{\mathcal{G}}(i)|}$, here a single variable may be present in multiple families. For example, X_2 is present in both the two parameter family $P(0_2 \mid x_1)$, and the four parameter family $P(0_2, 0_3 \mid x_1, x_4)$. Note also that some parameters are joint (conditional) probabilities over more than one variable, whereas DAG parameters are always (conditional) probabilities over one variable.

The mapping from parameters to joint probabilities is more complicated for ordinary Markov models associated with ADMGs containing bi-directed edges than it is for DAGs. For example, the joint probability $P(0_1, 0_2, 1_3, 1_4)$ can be obtained from DAG parameters in (11) by a simple formula

$$P(0_1) \cdot P(0_2 \mid 0_1) \cdot (1 - P(0_3 \mid 0_1)) \cdot (1 - P(0_4 \mid 0_2, 1_3)).$$

The corresponding formula for obtaining $P(0_1, 0_2, 1_3, 1_4)$ from the ordinary Markov parameters in (12) is a generalization of the Möbius inversion formula, which is a kind of alternating sum:

$$\begin{aligned} P(0_1) \cdot P(0_2 \mid 0_1) - P(0_2, 0_3 \mid 0_1, 1_4) \cdot P(0_1) - P(0_2 \mid 0_1) \cdot P(0_1) \cdot P(0_4) + \\ P(0_2, 0_3 \mid 0_1, 1_4) \cdot P(0_1) \cdot P(0_4). \end{aligned}$$

The general Möbius inversion formula for ordinary Markov parameters is given in the Appendix; see Evans and Richardson (2013b,a) for details and proofs.

As another example, a parametrization of the ordinary Markov model corresponding to the ADMGs in Fig. 5(e) and (h) is given by:

$$P(0_1), P(0_2), P(0_1, 0_2), P(0_3 \mid x_1, x_2), P(0_4 \mid x_1, x_2, x_4) \tag{13}$$

which gives a total of $1+1+1+4+8 = 15$ parameters. Since the model is binary, this parameterization represents a saturated model with $2^4 - 1$ parameters. This is as we would expect, since no m-separation relations hold in the ADMGs in Fig. 5(e) and (h).

4.2 The Möbius Parameterization of Binary Nested Markov Models

The parameterization of binary nested Markov models builds on the parametrization of the ordinary Markov model in many respects. However, as mentioned earlier, binary nested Markov parameters, unlike ordinary Markov parameters, are not in general conditional probabilities – they are probabilities in kernels derived from marginal distributions by means of an iterated application of the g-formula (7). In general the nested Markov model associated with an ADMG $\mathcal{G}$ may have fewer parameters than the ordinary Markov model associated with $\mathcal{G}$. This is as we would expect since the nested model imposes additional constraints.

We illustrate these differences with an example. Consider the hidden variable DAG shown in Fig. 5(a). As stated earlier, ordinary Markov models are defined by d-separation relations among observed variables, which are encoded via m-separation in the associated latent projection (or alternatively in the MAG $\mathcal{G}^m$). For the DAG with hidden variables in Fig. 5(a), the latent projection is given in Fig. 5(e) and the MAG in Fig. 5(h); as these show, there are no m-separation relations implied. Consequently the ordinary Markov model is saturated and parametrized via the 15 parameters given in (13).

In contrast, the nested Markov model associated with the hidden variable DAG in Fig. 5(a), is defined by the global nested Markov property applied to the latent projection, given in Fig. 5(e). This property implies the constraint (1) via m-separation of 1 and 4 in the CADMG in Fig. 6(a), obtained by fixing 3 in Fig. 5(e). A parametrization of the nested model is given by:

$$\begin{aligned}&P(0_1), P(0_2), P(0_1, 0_2), P(0_3 \mid x_1, x_2),\\ &q_{1,2,4}(0_1, 0_2, 0_4 \mid x_3), q_{2,4}(0_2, 0_4 \mid x_3), q_4(0_4 \mid x_3)\end{aligned} \tag{14}$$

which gives a total of $1 + 1 + 1 + 4 + 2 + 2 + 2 = 13$ parameters.

While the first four parameter families are identical to parameter families in (13), the remainder are different and involve kernels obtained from $P(x_1, x_2, x_3, x_4)$ by applications of the g-formula. In particular[20],

$$\begin{aligned}q_{1,2,4}(0_1, 0_2, 0_4 \mid x_3) &= \frac{P(0_1, 0_2, x_3, 0_4)}{P(x_3 \mid 0_2, 0_1)} = P(0_4 \mid x_3, 0_2, 0_1)P(0_2, 0_1),\\ q_{2,4}(0_2, 0_4 \mid x_3) &= \sum_{x_1} \frac{P(x_1, 0_2, x_3, 0_4)}{P(x_3 \mid 0_2, x_1)} = \sum_{x_1} P(0_4 \mid x_3, 0_2, x_1)P(0_2, x_1),\\ q_4(0_4 \mid x_3) &= \sum_{x_1, x_2} \frac{P(x_1, x_2, x_3, 0_4)}{P(x_3 \mid x_2, x_1)} = \sum_{x_1, x_2} P(0_4 \mid x_3, x_2, x_1)P(x_2, x_1).\end{aligned}$$

[20] Note that we can view marginalizations as applications of the g-formula as well. For example, $q_{2,4}(0_2, 0_4 \mid x_3) = \frac{q_{1,2,4}(0_1, 0_2, 0_4 \mid x_3)}{q_{1,2,4}(0_1 \mid 0_2, 0_4, x_3)}$, where $q_{1,2,4}(0_1 \mid 0_2, 0_4, x_3)$ is obtained by standard conditioning: $q_{1,2,4}(0_1 \mid 0_2, 0_4, x_3) = \frac{q_{1,2,4}(0_1, 0_2, 0_4 \mid x_3)}{q_{1,2,4}(0_2, 0_4 \mid x_3)}$. Full details are beyond the scope of this manuscript, but can be found for instance in (Shpitser et al., 2013).

Given a marginal distribution obtained from a distribution Markov relative to the hidden variable DAG in Fig. 5(a), and under the assumption that this DAG is causal (in the sense of Section 2.3) then the kernel parameters can be interpreted as interventional probabilities, as follows:

$$\begin{aligned} q_{1,2,4}(0_1, 0_2, 0_4 \mid x_3) &= P(0_1, 0_2, 0_4 \mid \mathrm{do}(x_3)), \\ q_{2,4}(0_2, 0_4 \mid x_3) &= P(0_2, 0_4 \mid \mathrm{do}(x_3)), \\ q_4(0_4 \mid x_3) &= P(0_4 \mid \mathrm{do}(x_3)). \end{aligned}$$

However, as noted before, in general a distribution $P(x_1, x_2, x_3, x_4)$ may be in the nested Markov model (encoded by the ADMG in Fig. 5(e)) and yet not be the marginal of any joint distribution corresponding to Fig. 5(a), since the latter implies inequality restrictions that may not be satisfied by $P(x_1, x_2, x_3, x_4)$.

As an additional example, consider the hidden variable DAG in Fig. 5(b). The corresponding ordinary Markov model is saturated since there are no m-separation relations holding in Fig. 5(f) (likewise for the MAG in Fig. 5(i)). In this case the ordinary Markov model parameterization associated with the ADMGs in Fig. 5(f), (i) is a DAG parameterization:

$$P(0_1), P(0_2 \mid x_1), P(0_3 \mid x_2, x_1), P(0_4 \mid x_3, x_2, x_1)$$

for a total of 1+2+4+8 = 15 parameters, giving a saturated model. The DAG encodes no d-separation statements, and thus no conditional independence constraints.

The nested Markov model associated with the ADMG in Fig. 5(f) has the following parameterization:

$$P(0_1), P(0_2 \mid x_1), q_{1,3}(0_1, 0_3 \mid x_2), q_3(0_3 \mid x_2), q_{2,4}(0_2, 0_4 \mid x_1, x_3), q_4(0_4 \mid x_3) \qquad (15)$$

for a total of $1 + 2 + 2 + 2 + 4 + 2 = 13$ parameters. We have

$$\begin{aligned} q_{1,3}(0_1, 0_3 \mid x_2) &= P(0_3 \mid x_2, 0_1)P(0_1), \\ q_3(0_3 \mid x_2) &= \sum_{x_1} P(0_3 \mid x_2, x_1)P(x_1), \\ q_{2,4}(0_2, 0_4 \mid x_1, x_3) &= P(0_4 \mid x_3, 0_2, x_1)P(0_2 \mid x_1), \\ q_4(0_4 \mid x_3) &= \sum_{x_2} P(0_4 \mid x_3, 0_2, x_1)P(0_2 \mid x_1). \end{aligned}$$

The nested Markov model encodes m-separation between 1 and 4 in the CADMG in Fig. 6(b) corresponding to (1). Note that while the nested Markov models corresponding to Fig. 5(e) and (f) are defined by the same constraint, they have rather different parameterizations (though of the same dimension). The same situation arises with equivalent DAGs and ordinary Markov models.

As is the case with the parameters for ordinary Markov models, the mapping from parameters to joint probabilities is given by a generalization of the Möbius inversion formula. In particular, given the parameterization in (15), $P(0_1, 0_2, 1_3, 1_4)$ is given by

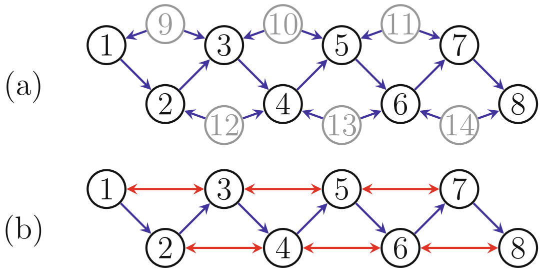

Figure 9: (a) A hidden variable DAG with binary observed nodes, and ternary hidden nodes. (b) The latent projection of this DAG.

$$P(0_1) \cdot P(0_2 \mid 0_1) - q_{1,3}(0_1, 0_3 \mid 0_2) \cdot P(0_2 \mid 0_1) - q_{2,4}(0_2, 0_4 \mid 0_1, 1_3) \cdot P(0_1) + \\ q_{2,4}(0_2, 0_4 \mid 0_1, 1_3) \cdot P(0_1, 0_3 \mid 0_2).$$

More details on this mapping can be found in Shpitser et al. (2012), and in the Appendix.

4.3 Sparse Modeling Using Nested Markov Models

One of the advantages of DAG models is that they are able to exploit conditional independence constraints to *concisely* represent large, multivariate joint densities as a product of Markov factors $P(x_v \mid x_{\mathrm{pa}_{\mathcal{G}}(v)})$. This has both statistical and computational advantages: Statistically, a model with fewer parameters is easier to fit. At the same time, computing marginal and conditional densities from the joint density becomes tractable for DAGs where the number of parents of all nodes is low. Well known algorithms that exploit this include belief propagation (Pearl, 1988), and variable elimination (Zhang and Poole, 1994).

Sparse modeling becomes more complex in the presence of hidden variables. Consider a DAG with many variables, where some are hidden, for instance the graph in Fig. 9(a), where nodes 9 through 14 are hidden. A simple and popular approach here is to posit a DAG model with hidden variables. This approach certainly can provide a sparse model, but has a number of disadvantages discussed earlier.

An ordinary Markov model avoids these difficulties, but since it only exploits conditional independence constraints on the observed marginal, and does not make assumptions about the dimension of the hidden variables, it often requires a large number of parameters. For example, the ordinary Markov model implied by the hidden variable DAG in Fig. 9(a) is given by the m-separation relations encoded in the ADMG in Fig. 9(b) (of which there are none), and contains $2^8 - 1 = 255$ parameters if variables X_1 through X_8 are binary.

By contrast, the nested Markov model for the same ADMG only contains 89 parameters. This is because the nested Markov model is able to exploit generalized independence constraints, such as the one that states that the kernel $q_4(x_4 \mid x_1, x_3) = \sum_{x_2} P(x_4 \mid x_3, x_2, x_1) P(x_2 \mid x_1)$ is not a function of x_1. Shpitser et al. (2011) have shown how to exploit generalized independence constraints to give a nested Markov version of variable elimination.

The ADMG in Fig. 9(b) can be viewed as a pair of bidirected chains of length 4, connected by a single directed path. A binary ordinary Markov parameterization of a pair of bidirected chains of length k connected in this way will always contain $2^{2k} - 1$ parameters, since such a graph does not encode any m-separation statements. The number of parameters of a binary nested Markov model of the same graph is $7 \cdot (2^k - 1) - 4k$, which is a dramatic improvement.

An alternative parameterization for discrete ordinary Markov models (Evans, 2011; Evans and Richardson, 2013a), and nested Markov models (Shpitser et al., 2013) replaces Möbius parameters by log-linear parameters that generalize log odds ratios of conditional and kernel probabilities, respectively. For example, the two parameter family $q_{1,3}(0_1, 0_3 \mid x_2) = P(0_3 \mid x_2, 0_1) \cdot P(0_1)$ of the nested model corresponding to Fig. 7 (c) is replaced by the following two parameters:

$$\lambda_{1,3}^{1,2,3} = \frac{1}{8} \log \frac{q_{1,3}(0_1, 0_3 \mid 0_2) \cdot q_{1,3}(0_1, 0_3 \mid 1_2) \cdot q_{1,3}(1_1, 1_3 \mid 0_2) \cdot q_{1,3}(1_1, 1_3 \mid 1_2)}{q_{1,3}(0_1, 1_3 \mid 0_2) \cdot q_{1,3}(0_1, 1_3 \mid 1_2) \cdot q_{1,3}(1_1, 0_3 \mid 0_2) \cdot q_{1,3}(1_1, 0_3 \mid 1_2)}$$

$$\lambda_{1,2,3}^{1,2,3} = \frac{1}{8} \log \frac{q_{1,3}(0_1, 0_3 \mid 1_2) \cdot q_{1,3}(0_1, 1_3 \mid 0_2) \cdot q_{1,3}(1_1, 0_3 \mid 0_2) \cdot q_{1,3}(1_1, 1_3 \mid 1_2)}{q_{1,3}(0_1, 1_3 \mid 1_2) \cdot q_{1,3}(1_1, 1_3 \mid 0_2) \cdot q_{1,3}(1_1, 0_3 \mid 1_2) \cdot q_{1,3}(0_1, 0_3 \mid 0_2)}$$

The advantage of these parameters is that they can be viewed as interactions within a kernel. For example, the parameter $\lambda_{1,3}^{1,2,3}$ is a two-way interaction and the parameter $\lambda_{1,2,3}^{1,2,3}$ is a three-way interaction within $q_{1,3}(x_1, x_3 \mid x_2)$. Experiments reported by Shpitser et al. (2013) indicate that, at moderate sample sizes the presence of high order interaction parameters does not make a significant difference. Specifically, under the reported experimental setup, when simulating from Fig. 9(a) removing 6-way and higher interactions from the nested Markov model of the ADMG in Fig. 9(b) removed 18 parameters from the model (giving the remaining total of 71 parameters), without a significant effect on the fit.

The log-linear parameterizations reported in Evans (2011) and Shpitser et al. (2013) can be viewed as ADMG analogues of sparse log-linear parameterizations in undirected graphical models, of which Boltzmann machines (Ackley et al., 1985) form a well-known special case. More general marginal log-linear parameterizations were introduced by Bergsma and Rudas (2002). Since ordinary and nested Markov models are associated with ADMGs, they can, unlike undirected graphical models, concisely represent causal hypotheses in the presence of hidden variables.

4.4 Equivalence in Nested Markov Models

As we saw in Section 2.2.2, there is a simple characterization of Markov equivalence in DAG models. A more complex characterization exists for ordinary Markov models (Ali et al., 2009). While no such characterization yet exists for nested Markov models,

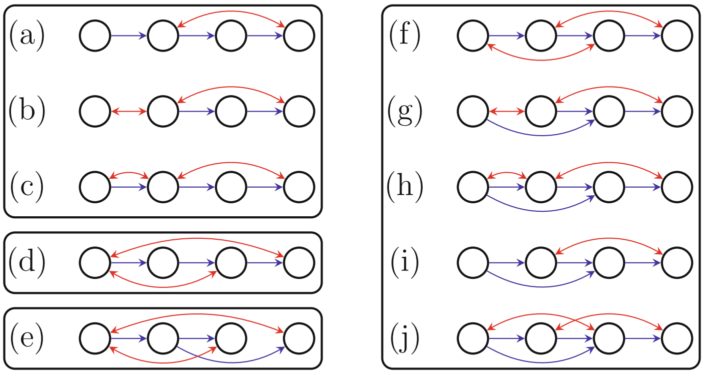

Figure 10: The conjectured equivalence classes among graph patterns (with vertex labeling suppressed) of 4 node ADMGs corresponding to nested Markov models that are strict submodels of the ordinary Markov models given by the same ADMG.

we report here on numerical experiments that reveal the equivalence classes for nested Markov models over 4 node graphs.

We say two ADMGs are *nested Markov equivalent* if they define the same nested Markov model. In particular, they must agree on all conditional independence constraints and in addition agree on all generalized independence constraints such as (1). It immediately follows that a nested Markov equivalence class of ADMGs is always contained within an ordinary Markov equivalence class of ADMGs.

On four variables there are precisely 543 DAGs (sequence A003024 in OEIS) and $2^{\binom{4}{2}} = 2^6$ bi-directed graphs. Thus there are $543 \cdot 2^6 = 34752$ ADMGs with 4 vertices. For many of these ADMGs the nested and ordinary Markov models are the same. The set of (unlabeled) ADMG "patterns" for which the nested model is strictly smaller is shown in Fig. 10. The set of labeled ADMGs can be obtained from these patterns by assigning labels $1, 2, 3, 4$ to unlabeled vertices in all possible ways. For instance, there are $4! = 24$ possible vertex labelings of Fig. 10(a), but, owing to symmetry, only $\binom{4}{2} \cdot 2! = 12$ vertex labelings of Fig. 10(e) lead to different ADMGs. In total, there are $24 \cdot 9 + 12 \cdot 1 = 228$ such ADMGs arranged in what are conjectured to be 84 equivalence classes, shown schematically by boxes around the ADMG patterns in Fig. 10.

This conjecture is supported by experiments in which we fitted the nested models associated with these ADMGs to simulated datasets. We observed that distinct ADMGs within the same (conjectured) equivalence class always gave the same likelihood values, while no two models in different classes ever yielded the same value.

We now give a constraint-based view of these equivalence classes. If we label the vertices in the graphs in Fig. 10(a), (b), (c), sequentially (from left to right) by $1, 2, 3, 4$, the nested property consists of two constraints: $(X_1 \perp\!\!\!\perp X_3 \mid X_2)$, which is a regular conditional independence, and $(X_4 \perp\!\!\!\perp X_1 \mid X_3)$ in $\frac{P(x_1,x_2,x_3,x_4)}{P(x_3|x_2,x_1)}$, which is not. The graph in Fig. 10(d) under similar labeling gives a model defined by the constraint $(X_2 \perp\!\!\!\perp X_4 \mid X_3)$ in $\frac{P(x_1,x_2,x_3,x_4)}{P(x_2|x_1)}$. The graph in Fig. 10(e) gives a model defined by the

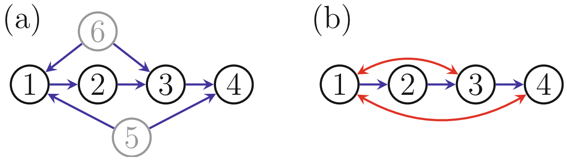

Figure 11: (a) A hidden variable causal DAG. (b) An ADMG representation of the DAG in (a) based on structure learning from the marginal $P(x_1, x_2, x_3, x_4)$ based on the nested Markov model parameterizations.

constraint $(X_3 \perp\!\!\!\perp X_4 \mid X_2)$ in $\frac{P(x_1,x_2,x_3,x_4)}{P(x_2|x_1)}$. The graphs in Fig. 10(f), (g), (h), (i) and (j) give a model under the constraint $(X_4 \perp\!\!\!\perp X_1 \mid X_3)$ in $\frac{P(x_1,x_2,x_3,x_4)}{P(x_3|x_2,x_1)}$. Note that this constraint is precisely (1).

If we assume that the truth is a hidden variable causal DAG shown in Fig. 11(a), and assume that the observable distribution is faithful[21) any structure learning method based on conditional independences among observable variables, such as the FCI algorithm, or a scoring algorithm applied to the likelihood function of a DAG or ordinary Markov model involving variables X_1, X_2, X_3, X_4, would, asymptotically, find no conditional independences, and hence return the ordinary Markov equivalence class containing all complete graphs, and any other ADMGs entailing no conditional independences over $P(x_1, x_2, x_3, x_4)$.

By contrast, a search and score method based on nested Markov models would, asymptotically, return the ADMG in Fig. 11(b)[22). Given the assumption that the underlying true DAG is causal, a structure learning procedure based on nested Markov models has recovered essentially all causal structure from a single independence

$$(X_2 \perp\!\!\!\perp X_4 \mid X_3) \text{ in } \frac{P(x_1, x_2, x_3, x_4)}{P(x_2 \mid x_1)}. \tag{16}$$

Thus, structure learning based on nested Markov models is able to reconstruct the exact causal pathway among $1, 2, 3, 4$ present in the true DAG, and the presence of unobserved confounders (hidden common causes) between 1 and 3 and between 1 and 4. Preliminary results along these lines based on a nested Markov parameterization reported in Section 4.2 can be found in (Shpitser et al., 2012).

5. Discussion and Future Work

We have shown how generalized equality constraints first considered in the context of direct effect models in longitudinal studies have given rise to a new kind of graphical model called the *nested Markov model.*

We have illustrated a Möbius parameterization of discrete nested Markov models by means of simple examples, and compared it to the earlier ordinary Markov

[21) That is, we assume that the only independence or generalized independence constraints that hold are those implied by the nested Markov property.

[22) Here we assume that the nested Markov equivalence classes are as conjectured in Fig. 10.

parametrization of Evans and Richardson (2013a) on which it is based. We have shown how, assuming faithfulness, a structure learning algorithm can distinguish between DAGs with latents that cannot be discriminated solely on the basis of conditional independence. In particular, we have shown a case in Section 4.4 where, under appropriate assumptions, we conjecture that a single generalized independence constraint is sufficient to conclude that a unique ADMG is the latent projection of the true hidden variable DAG. We have also shown that under the (additional) assumption of an underlying causal DAG the parameters of discrete nested Markov models have an intuitive causal interpretation as probabilities corresponding to certain identifiable interventional distributions.

A number of open problems remain: There is as yet no graphical characterization of equivalence in nested Markov models. Structure learning with nested Markov models from data with small and moderate sample sizes appears to be challenging. An alternative log-linear parameterization for nested Markov models well suited for *sparsity methods* was recently developed to address this (Shpitser et al., 2013).

There is no analogue of computationally efficient methods of obtaining arbitrary marginals in nested Markov models as there is in DAGs and undirected graphical models. Preliminary work in this area gave a nested Markov analogue of variable elimination (Shpitser et al., 2011).

Finally, nested Markov models are models defined only in terms of equality constraints. It is known, however, that hidden variables may induce inequality constraints on the observed margin, such as the instrumental inequality (Pearl, 1995). Graphically characterizing these inequalities, and extending nested Markov models to take advantage of them is a challenging avenue of future research, though see (Evans, 2012) for some preliminary results.

Appendix

Here we give the formal definition of the global nested Markov property, the ordinary and nested Möbius parameters for ADMGs (restricted to binary state-spaces for simplicity), and the corresponding inversion formulas. The inversion formulas rely on rather lengthy definitions of vertex set partitions, which we do not include for space reasons, and for which we give references.

The Global Nested Markov Property

For a CADMG $\mathcal{G}(V, W)$, we call a set $D \subseteq V$ *bidirected-connected* if for any $a, b \in D$, there exists a path $a \leftrightarrow \cdots \leftrightarrow b$ in $\mathcal{G}$. $D \subseteq V$ is a *district* in $\mathcal{G}$ if it is an inclusion maximal bidirected connected set in $\mathcal{G}$. We will denote the set of districts of $\mathcal{G}$ by $\mathcal{D}(\mathcal{G})$. $\mathcal{D}(\mathcal{G})$ partitions V in $\mathcal{G}$. If a set S lies in a single district D of $\mathcal{G}$, we will denote D by $\text{dis}_{\mathcal{G}}(S)$.

Definition 8 *Given a CADMG $\mathcal{G}(V, W)$ the set of* fixable vertices*,*

$$\mathbb{F}(\mathcal{G}) \equiv \{v \mid v \in V, \mathrm{dis}_{\mathcal{G}}(v) \cap \mathrm{de}_{\mathcal{G}}(v) = \{v\}\}\,.$$

In words, a vertex v is fixable in $\mathcal{G}$ if there is no vertex v^* that is both a descendant of v and in the same district as v in $\mathcal{G}$.

Definition 9 *Given a CADMG $\mathcal{G}(V, W, E)$, and a kernel $q_V(X_V \mid X_W)$, for every $r \in \mathbb{F}(\mathcal{G})$ we associate a* fixing transformation *ϕ_r on the pair $(\mathcal{G}, q_V(X_V \mid X_W))$ defined as follows:*

$$\phi_r(\mathcal{G}) \equiv \mathcal{G}^*(V \setminus \{r\}, W \cup \{r\}, E_r),$$

where E_r is the subset of edges in E that do not have arrowheads into r, and

$$\phi_r(q_V(x_V \mid x_W); \mathcal{G}) \equiv \frac{q_V(x_V \mid x_W)}{q_V(x_r \mid x_{\mathrm{mb}_{\mathcal{G}}(r, \mathrm{an}_{\mathcal{G}}(\mathrm{dis}_{\mathcal{G}}(r)))})}. \tag{17}$$

We use $\circ$ to indicate composition of operations in the natural way, so that:

$$\begin{aligned} \phi_r \circ \phi_s(\mathcal{G}) &\equiv \phi_r(\phi_s(\mathcal{G})), \\ \phi_r \circ \phi_s(q_V(X_V \mid X_W); \mathcal{G}) &\equiv \phi_r\left(\phi_s\left(q_V(X_V \mid X_W); \mathcal{G}\right); \phi_s(\mathcal{G})\right). \end{aligned}$$

Definition 10 *Given a CADMG $\mathcal{G}(V, W, E)$, we define $\mathcal{G}^{|W}$ to be a mixed graph with vertex set $V^* = V \cup W$, and edge set*

$$E^* \equiv E \cup \{w \leftrightarrow w' \mid w, w' \in W\}.$$

In words, the graph $\mathcal{G}^{|W}$ is formed from $\mathcal{G}$ by adding bi-directed edges between all pairs of vertices $w, w' \in W$, and then eliminating the distinction between vertices in V and W.

A kernel $q_V(X_V|X_W)$ satisfies the *global Markov property* for a CADMG $\mathcal{G}(V, W, E)$ if for arbitrary disjoint sets A, B, C, (C may be empty)

$$\text{if } A \text{ is m-separated from } B \text{ given } C \text{ in } \mathcal{G}^{|W} \;\Rightarrow\; X_A \perp\!\!\!\perp X_B \mid X_C \; [q_V(X_V|X_W)]$$

Definition 11 *A CADMG $\mathcal{G}(V, W)$ is* reachable *from ADMG $\mathcal{G}^*(V \cup W)$ if there is an ordering of the vertices in $W = \langle w_1, \dots, w_k \rangle$, such that for $j = 1, \dots, k$,*

$$\begin{aligned} &w_1 \in \mathbb{F}(\mathcal{G}^*) \textit{ and for } j = 2, \dots, k, \\ &\quad w_j \in \mathbb{F}(\phi_{w_{j-1}} \circ \cdots \circ \phi_{w_1}(\mathcal{G}^*)). \end{aligned}$$

Such an ordering is called a valid fixing sequence.

Let $\mathbb{G}(\mathcal{G}) \equiv \{(\mathcal{G}^*, \mathbf{w}^*) \mid \mathcal{G}^* = \phi_{\mathbf{w}^*}(\mathcal{G})\}$. In words, $\mathbb{G}(\mathcal{G})$ is the set of valid fixing sequences and the CADMGs that they reach.

Definition 12 *We say that a distribution $P(x_V)$ obeys the* global nested Markov property *for $\mathcal{G}(V)$ if for all $(\mathcal{G}^*, \mathbf{w}^*) \in \mathbb{G}(\mathcal{G})$, $\phi_{\mathbf{w}^*}(P(x_V); \mathcal{G})$ obeys the global Markov* *property for* $\phi_{\mathbf{w}^*}(\mathcal{G}) \equiv \mathcal{G}^*$*.*

Parameterization of Binary Ordinary Markov Models of ADMGs

Definition 13 *A set of vertices* H *in an ADMG* $\mathcal{G}$ *is called a* head *if* $H = \{v \in H \mid \mathrm{de}_{\mathcal{G}}(v) \cap H = \{v\}\}$*, and* H *is contained within a single district in* $\mathcal{G}(\mathrm{an}_{\mathcal{G}}(H))$ *(the latent projection of* $\mathcal{G}$ *onto* $\mathrm{an}_{\mathcal{G}}(H)$*).*

Definition 14 *The* tail *associated with a head* H *in an ADMG* $\mathcal{G}$ *is given by:* $\mathrm{tail}(H) \equiv (\mathrm{dis}_{\mathcal{G}(\mathrm{an}(H))}(H) \setminus H) \cup \mathrm{pa}(\mathrm{dis}_{\mathcal{G}(\mathrm{an}(H))}(H))$*.*

We denote the set of all heads of $\mathcal{G}$ as $\mathcal{H}(\mathcal{G})$.

Definition 15 *The* ordinary Möbius parameters *associated with an ADMG* $\mathcal{G}$ *are a set of probabilities:* $\mathfrak{P}_{\mathcal{G}} \equiv \big\{P(X_H = \mathbf{0} \mid x_{\mathrm{tail}(H)}) \mid H \in \mathcal{H}(\mathcal{G})\big\}$*.*

Definition 16 *Let* $\nu : V \to \{0,1\}^{|V|}$ *be an assignment of values to the variables indexed by* V*. Define* $\nu(T)$ *to be the values assigned to variables indexed by a subset* $T \subseteq V$*. Let* $\nu^{-1}(0) = \{v \mid v \in V, \nu(v) = 0\}$*.*

A distribution $P(x_V)$ *is said to be* parameterized by *the set of ordinary Möbius parameters* $\mathfrak{P}_{\mathcal{G}}$*, for an ADMG* $\mathcal{G}$ *if, for all* ν*::*

$$P(X_V = \nu(V)) = \sum_{B \,:\, \nu^{-1}(0) \subseteq B \subseteq V} (-1)^{\left|B \setminus \nu^{-1}(0)\right|} \times \prod_{H \in [B]_{\mathcal{G}}} P(X_H = \boldsymbol{0} \mid X_{\mathrm{tail}(H)} = \nu(\mathrm{tail}(H)))$$

where the empty product is defined to be 1*, and* $[B]_{\mathcal{G}}$ *is a partition of* B *into "heads" given in (Evans and Richardson, 2013b).*

Parameterization of Binary Nested Markov Models of ADMGs

Definition 17 *A set of vertices* S *is* intrinsic *in* $\mathcal{G}$ *if it is a district in a reachable subgraph of* $\mathcal{G}$*. The set of intrinsic sets in an ADMG* $\mathcal{G}$ *is denoted by* $\mathcal{I}(\mathcal{G})$*.*

Definition 18 *For an intrinsic set* $S \in \mathcal{I}(\mathcal{G})$ *of a CADMG* $\mathcal{G}$*, define the recursive head (rh) as:* $\mathrm{rh}(S) \equiv \{x \in S \mid \mathrm{ch}_{\mathcal{G}}(x) \cap S = \emptyset\}$*.*

Definition 19 *The (recursive)* tail *associated with a recursive head* H *of an intrinsic set* S *in a CADMG* $\mathcal{G}$ *is given by:* $\mathrm{rt}(H) \equiv (S \setminus H) \cup \mathrm{pa}_{\mathcal{G}}(S)$*.*

Definition 20 *The* nested Möbius parameters *associated with an ADMG* $\mathcal{G}$ *are a set of functions:* $\mathfrak{Q}_{\mathcal{G}} \equiv \big\{q_S(X_H = \mathbf{0} \mid x_{\mathrm{rt}(H)})$ *for* $H = \mathrm{rh}(S), S \in \mathcal{I}(\mathcal{G})\big\}$*.*

Intuitively, a parameter $q_S(X_H = \mathbf{0} | x_{\mathrm{rt}(H)})$ is the probability that the variable set X_H assumes values $\mathbf{0}$ in a kernel obtained from $P(x_V)$ by fixing $X_{V \setminus S}$, and conditioning on $X_{\mathrm{rt}(H) \setminus (V \setminus S)}$.

Definition 21 *Let* $\nu : V \to \{0,1\}^{|V|}$ *be an assignment of values to the variables indexed by* V*. Define* $\nu(T)$ *to be the values assigned to variables indexed by a subset* $T \subseteq V$*. Let* $\nu^{-1}(0) = \{v \mid v \in V, \nu(v) = 0\}$*.*

A distribution $P(x_V)$ *is said to be* parameterized by *the set of nested Möbius parameters* $\mathfrak{Q}_\mathcal{G}$ *for an ADMG* $\mathcal{G}$ *if, for all* ν*:*

$$P(X_V = \nu(V)) = \sum_{B\,:\,\nu^{-1}(0)\subseteq B\subseteq V} (-1)^{\left|B\setminus\nu^{-1}(0)\right|} \times \prod_{H\in[\![B]\!]_\mathcal{G}} q_S(X_H = \boldsymbol{0} \mid X_{\mathrm{rt}(H)} = \nu(\mathrm{rt}(H)))$$

where the empty product is defined to be 1*, and* $[\![B]\!]_\mathcal{G}$ *is a partition of* B *into "recursive heads" given in (Shpitser et al., 2011).*

References

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169.

Ali, A., Richardson, T. S., and Spirtes, P. (2009). Markov equivalence for ancestral graphs. *Annals of Statistics*, 37:2808–2837.

Beal, M. J. and Ghahramani, Z. (2004). Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 1(1):1–44.

Bergsma, W. P. and Rudas, T. (2002). Marginal models for categorical data. *Annals of Statistics*, 30(1):140–159.

Chickering, D. M. (2002). Optimal structure identifiation with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Claassen, T., Mooij, J. M. and Heskes, T. (2013) Learning Sparse Causal Models is not NP-hard, Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI-13), abs/1309.6824

Colombo, D., Maathuis, M. H., Kalisch, M. and Richardson T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. Annals of Statistics, 40: pp.294–321.

Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

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

Drton, M. and Plummer, M. (2013). A Bayesian information criterion for singular models. *ArXiv e-prints.*

Drton, M., Sturmfels, B., and Sullivant, S. (2009). *Lectures on Algebraic Statistics*, volume 40. Birkhäuser, Basel.

Evans, R. J. (2011). *Parameterizations of Discrete Graphical Models.* PhD thesis, Department of Statistics, University of Washington.

Evans, R. J. (2012). Graphical methods for inequality constraints in marginalized DAGs. In *22nd Workshop on Machine Learning and Signal Processing.*

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

Evans, R. J. and Richardson, T. S. (2013a). Marginal log-linear parameters for graphical Markov models. *Journal of the Royal Statistical Society: Series B*, 75(4):743–768.

Evans, R. J. and Richardson, T. S. (2013b). Markovian acyclic directed mixed graphs for discrete data. Annals of Statistics, accepted for publication. abs/1301.6624.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the 14th International Conference on Machine Learning*, pages 125–133. Morgan Kaufmann.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3/4):601–620.

Geiger, D. and Meek, C. (1998). Graphical models and exponential families. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 156–165.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT Press.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford, U.K.: Clarendon.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle. Excerpts reprinted (1990) in English. *Statistical Science*, 5:463–472.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan and Kaufmann, San Mateo.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 435–443, San Francisco, CA. Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

Pourret, O., Naïm, P., and Marcot, B. (2008). *Bayesian Networks: A Practical Guide to Applications.* Wiley.

Richardson, T. S. (2003). Markov properties for acyclic directed mixed graphs. *The Scandinavian Journal of Statistics*, 30(1):145–157.

Richardson, T. S., Robins, J. M., and Shpitser, I. (2012). Nested Markov properties for acyclic directed mixed graphs. Presented at the *28th Conference on Uncertainty in Artificial Intelligence (UAI-12).*

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

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

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. In Sechrest, L., Freeman, H., and Mulley, A., editors, *Health Service Research Methodology: A Focus on AIDS*, pages 113–159. NCHSR, U.S. Public Health Service.

Robins, J. M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79:321–334.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In Berkane, M., editor, *Latent variable modelling and applications to causality*, number 120 in Lecture notes in statistics, pages 69–117. Springer-Verlag, New York.

Robins, J. M. (1999). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In Glymour, C. and Cooper, G., editors, *Computation, Causation, and Discovery*, pages 349 – 405. Menlo Park, CA, CAmbridge, MA: AAAI Press/The MIT Press.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In Halloran, M. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, volume 116 of *The IMA Volumes in Mathematics and its Applications*, pages 95–133. Springer New York.

Robins, J. M. and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. In P. Shrout, K. Keyes, and K. Ornstein, editors, *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures.* Chapter 6, pages 1–52, Oxford University Press.

Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, 90(3):491 – 515.

Robins, J. M. and Wasserman, L. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, pages 309–420. Morgan Kaufmann.

Robins, J. M. and Wasserman, L. (1999). On the impossibility of inferring causation from association without background knowledge. In *Computation, Causation, and Discovery*, pages 305–321. Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345.

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

Rusakov, D. and Geiger, D. (2005). Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, 6:1–35.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Settimi, R. and Smith, J. Q. (1998). On the geometry of Bayesian graphical models with hidden variables. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 472–479.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. J. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030.

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 29th Conference on Uncertainty in Artificial Intelligence (UAI-13)*. AUAI Press.

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, (UAI-2006). AUAI Press.

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

Shpitser, I., Richardson, T. S., and Robins, J. M. (2009). Testing edges by truncations. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, volume 21, pages 1957–1963.

Shpitser, I., Richardson, T. S., and Robins, J. M. (2011). An efficient algorithm for computing interventional distributions in latent variable causal models. In *27th Conference on Uncertainty in Artificial Intelligence (UAI-11)*. AUAI Press.

Shpitser, I., Richardson, T. S., Robins, J. M., and Evans, R. J. (2012). Parameter and structure learning in nested Markov models. In *Proceedings of the Causal Structure Learning Workshop of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-12)*.

Shwe, M. A., Middleton, B., Heckerman, D. E., Henrion, M., Horvitz, E. J., Lehmann, H. P., and Cooper, G. F. (1991). Probabilistic diagnosis using a reformulation of the INTERN-IST-1/QMR knowledge base. *Methods of Information in Medicine*, 30(4):241–267.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search.* Springer Verlag, New York.

Tian, J. and Pearl, J. (2002). On the testable implications of causal models with hidden variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02)*, pages 519–527. AUAI Press.

Verma, T. S. and Pearl, J. (1990). Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles.

Zhang, N. L. and Poole, D. (1994). A simple approach to Bayesian network computations. In *Tenth Canadian Conference on AI*, pages 171–178.

(Received August 22 2013, Revised January 14 2014)