Nested Markov Properties for Acyclic Directed Mixed Graphs

Thomas S. Richardson

University of Washington, USA

Robin J. Evans

University of Oxford, UK

James M. Robins

Harvard School of Public Health, USA

Ilya Shpitser

Johns Hopkins University, USA

January 26, 2017

Abstract

Directed acyclic graph (DAG) models may be characterized in at least four different ways: via a factorization, the d-separation criterion, the moralization criterion, and the local Markov property. As pointed out by Robins (1986, 1999), Verma and Pearl (1990), and Tian and Pearl (2002b), marginals of DAG models also imply equality constraints that are not conditional independences. The well-known 'Verma constraint' is an example. Constraints of this type were used for testing edges (Shpitser et al., 2009), and an efficient marginalization scheme via variable elimination (Shpitser et al., 2011).

We show that equality constraints like the 'Verma constraint' can be viewed as conditional independences in kernel objects obtained from joint distributions via a fixing operation that generalizes conditioning and marginalization. We use these constraints to define, via Markov properties and a factorization, a graphical model associated with acyclic di-

1

rected mixed graphs (ADMGs). We show that marginal distributions of DAG models lie in this model, prove that a characterization of these constraints given in (Tian and Pearl, 2002b) gives an alternative definition of the model, and finally show that the fixing operation we used to define the model can be used to give a particularly simple characterization of identifiable causal effects in hidden variable graphical causal models.

1   Introduction

Graphical models provide a principled way to take advantage of independence constraints for probabilistic modeling, learning and inference, 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 vertex in the graph corresponds to a random variable. The popularity of DAG models, also known as Bayesian networks, stems from their well understood theory and from the fact that they admit an intuitive causal interpretation (under the assumption that there are no unmeasured common causes; see (Spirtes et al., 1993)). An arrow from a variable A to a variable B in a DAG model can be interpreted, in a way that can be made precise, to mean that A is a "direct cause" of B.

Starting from a causally interpreted DAG, the consequences of intervention in the system under study can be understood by modifying the graph via removing certain edges, and modifying the corresponding joint probability distribution via re-weighting (Pearl, 2000, Spirtes et al., 1992, Strotz and Wold, 1960). For example, the DAG in Figure 1(i) represents distributions that factorize as

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

If the model is interpreted causally, an experiment to externally set the value of X_3 will break the dependence of X_3 on X_1 and X_2; however, the dependence of X_4 upon X_3 will be preserved. This is represented graphically by severing incoming edges to X_3 (an operation some authors call 'mutilation' and we later

2

Figure 1: (i) A DAG on five variables, and (ii) a DAG representing the model after an experiment to externally fix X_3.

call 'fixing'), and probabilistically by removing the factor $p(x_3 \mid x_1, x_2)$ from the factorization of the joint distribution:

$$p^*(x_0, x_1, x_2, x_4 \mid x_3) = p(x_0)\, p(x_1)\, p(x_2 \mid x_0, x_1)\, p(x_4 \mid x_0, x_3). \qquad (1)$$

The functional in (1) is sometimes called the g-formula (Robins, 1986), the manipulated distribution (Spirtes et al., 1993), or the truncated factorization (Pearl, 2000). The distribution $p^*(x_0, x_1, x_2, x_4 \mid x_3)$ is commonly denoted by $p(x_0, x_1, x_2, x_4 \mid \mathrm{do}(x_3))$. In a causal model given by a DAG where all variables are observed, any interventional probability distribution can be identified by this method.

Often not all unmeasured common causes are measured, or there is no way to know *a priori* whether this is the case. This motivates the study of DAG models containing latent variables. Existing theoretical machinery based on DAGs can be applied to such settings, simply by treating the unobserved variables as missing data. However, this creates a number of problems that are particularly severe in the context where the structure of the underlying DAG model with latents is unknown. First, there are, in general, an infinite number of DAG models with latent variables that imply the (independence) constraints holding in the observed distribution. Second, assumptions concerning the state space or distribution of latent variables may have a profound effect on the model. This is problematic given that prior knowledge about latent variables is often scarce.

An alternative approach considers a supermodel defined by taking a subset of the constraints implied by a DAG model with latent variables on the observed

marginal distribution. More specifically, we consider models defined by equality constraints that are implied by the factorization of the DAG with latents, but do not depend on assumptions regarding the state-space or distribution of the latent variables. Models defined by these constraints are naturally represented by a mixed graph, i.e. a graph containing directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges, obtained from the DAG via a *latent projection* operation (Verma and Pearl, 1990); see the graph in Figure 3(i) for the latent projection of the DAG in Figure 1(i).

Previous work (Ali et al., 2009, Evans and Richardson, 2010, Richardson and Spirtes, 2002) has considered models defined by the conditional independence constraints implied by latent variable models on the observed margin. Parameterizations, fitting and search algorithms exist for these models under multivariate Gaussian and discrete distributions over the observed variables. It is well-known, however, that DAG models with latent variables imply non-parametric constraints which are not conditional independence constraints. For example, consider the DAG shown in Figure 1(i), and take the vertex X_0 as hidden. This DAG implies no conditional independence restrictions on the observed margin $p(x_1, x_2, x_3, x_4)$. This is because all vertex sets which d-separate pairs of observed variables — i.e. the pairs (x_2, x_4) and (x_1, x_4) — include the unobserved variable x_0. However, it may be shown that the $p(x_1, x_2, x_3, x_4)$ margin of any distribution $p(x_0, x_1, x_2, x_3, x_4)$ which factorizes according to the DAG in Figure 1(i), obeys the constraint that:

$$\sum_{x_2} p(x_4 \mid x_1, x_2, x_3) \cdot p(x_2 \mid x_1) \text{ is a function of only } x_3 \text{ and } x_4, \qquad (2)$$

see (Robins, 1986, Verma and Pearl, 1990). In (Robins, 1999) it is shown that this constraint is equivalent to the requirement that X_4 is independent of X_1 in the distribution obtained from $p(x_1, x_2, x_3, x_4)$ after dividing by the conditional $p(x_3 \mid x_2, x_1)$. Note that this is the same manipulation performed in (1), but the operation, which we later call 'fixing', is purely probabilistic and can be performed *without* requiring that the model has any causal interpretation.

If we interpret the original DAG as causal, then the constraint (2) is an (iden-

tifiable) dormant independence constraint (Shpitser and Pearl, 2008), denoted by $X_1 \perp\!\!\!\perp X_4 \mid \mathrm{do}(x_3)$ in (Pearl, 2000); see (Shpitser et al., 2014).

Since, as we have seen, the DAG in Figure 1(i) implies no conditional independence restrictions on the joint $p(x_1, x_2, x_3, x_4)$, the set of distributions obeying these independence relations is (trivially) saturated. Consequently, a structure learning algorithm such as FCI (Spirtes et al., 1993) that learns a Markov equivalence class of DAGs with latent variables, under the assumption of faithfulness, will return a (maximally uninformative) unoriented complete graph. (The assumption of 'faithfulness' is that if $A \perp\!\!\!\perp B \mid C$ in the observed distribution then A is d-separated from B given C in the underlying DAG with latent variables.)

Indeed, as originally pointed out by Robins (1999), if we assume a generalization of faithfulness, it is possible to use constraints such as (2) to distinguish between models. Shpitser et al. (2009) used pairwise constraints of this form to test for the presence of certain directed edges (in the context of a specific graph). Further, Tian and Pearl (2002b) presented a general algorithm for finding non-parametric constraints from DAGs with latent variables.

In this paper we introduce a (statistical) model called the *nested Markov model*, defined by these non-parametric constraints, and associated with a mixed graph called an acyclic directed mixed graph (ADMG). We give equivalent characterizations of the model in terms of global and ordered local Markov properties, and a factorization. We also show that the model can be defined by a set of constraints obtained from an algorithm by Tian and Pearl (2002b). We show that any saturated model, as well as any distribution that is a marginal of a DAG model distribution, is in the naturally associated nested Markov model. Finally, we show that our results give a particularly simple characterization of identifiable causal effects in hidden variable causal DAG models.

Building on the theory in this paper, and a parametrization of the nested model for categorical variables given in (Evans and Richardson, 2015), it is possible to determine nested Markov equivalence classes of ADMGs by evaluating likelihoods (on generic data). Evans (2015) shows that for categorical variables,

the algorithm of Tian and Pearl (2002b) is 'complete' for finding equality constraints, and that therefore the nested Markov model is, in a well-defined sense, the closest approximation to the margin of a DAGs model without considering inequality constraints. The review paper (Shpitser et al., 2014) includes all nested Markov equivalence classes over four variables. A general theory of nested Markov equivalence and search for ADMGs remains a subject for future work.

1.1 Overview of Nested Markov Models

We now outline our strategy for defining the nested Markov model in terms of ordinary conditional independence models by analogy to a way of defining DAG models in terms of undirected graphical models. We give a specific example of a nested Markov model, outlining the key concepts, while providing references to the formal definitions within the paper.

A nested Markov model is represented by an acyclic directed mixed graph (ADMG). ADMGs are naturally derived from DAGs with latent variables via an operation called *latent projection*. Intuitively, the ADMG does not contain latent variables, but indicates the presence of such variables by the inclusion of bidirected ($\leftrightarrow$) edges. Earlier work (Richardson, 2003, Richardson and Spirtes, 2002) established Markov properties for *independence* models defined by ADMGs. Such an independence model is a supermodel of the nested Markov model represented by the same ADMG, as it is defined by fewer constraints. The global Markov property for these independence models simply corresponds to the natural extension of d-separation (Pearl, 1988) to ADMGs, whereby we allow 'colliders' to involve bidirected edges, sometimes called m-separation. Latent projection is defined in Section 2.3; ADMGs and m-separation in sections 2.1 and 2.2 respectively.

We also consider conditional ADMGs (CADMGs) where certain vertices are fixed constants rather than random variables. Such vertices are treated similarly to the so called "strategy nodes" in influence diagrams (Dawid, 2002).

6

The Markov property for CADMGs is a simple extension of m-separation that takes into account fixed nodes. CADMGs are defined formally in Section 2.4; the corresponding global Markov property for a CADMG is given in section 2.8.1. Note that an ADMG is a CADMG with no fixed vertices.

As mentioned, CADMGs and their associated Markov models characterize the nested Markov model in much the same way that undirected graphs and their associated models can be used to describe a DAG model. We first briefly review the characterization of DAGs via undirected models.

The global Markov property for DAGs may be obtained from the (union of) the Markov properties associated with undirected graphs derived from the DAG by the moralization operation (Lauritzen, 1996); the resulting property is equivalent to d-separation (Pearl, 1988). More precisely, the DAG Markov property corresponds to (the union of) the Markov properties or factorizations associated with undirected graphs representing 'ancestral' margins; likewise the set of distributions corresponding to the DAG is the intersection of the sets of distributions obeying the factorization properties associated with these undirected graphs; again, this is equivalent to the characterization that the joint distribution factor into the product of each variable given its parents in the graph.

As an example, consider the DAG in Fig. 2 (i). Undirected graphs associated with some ancestral margins, and their factorizations, are shown in Fig. 2 (ii), (iii), and (iv).

Likewise, the set of distributions in the nested Markov model associated with an ADMG corresponds to the intersection of the sets of distributions obeying factorization properties encoded by specific CADMGs obtained from the original ADMG. However, whereas the undirected graphs corresponding to a DAG may be seen as representing specific (ancestral) *margins*, the CADMGs obtained from an ADMG represent 'kernel' distributions obtained by sequentially applying a new 'fixing' operation on distributions, one that generalizes conditioning and marginalizing. This fixing operation has a natural causal interpretation, as do the kernels that form CADMG factorizations of the nested Markov model.

Specifically, in the context of a latent variable causal model whose projection is a given ADMG, kernels can be viewed as (identified) interventional distributions. Not all variables are fixable. From a causal perspective this is natural since in the presence of latent variables, not all interventional distributions are identifiable. The fixing operation and the set of fixable vertices are defined in section 2.11.

As a specific example, consider the graph shown in Fig. 3(i). In this ADMG, the vertex X_3 may be fixed to give the CADMG shown in Fig. 3(ii), where the corresponding distribution and factorization are also shown. Here

$$q_1(x_1) = p(x_1), \qquad q_{24}(x_2, x_4 \mid x_1, x_3) = \frac{p(x_1, x_2, x_3, x_4)}{p(x_1)p(x_3 \mid x_2, x_1)}. \qquad (3)$$

Note that, although the original graph implied no conditional independences, the graph in Figure 3(ii) implies the independence $X_1 \perp\!\!\!\perp X_4 \mid X_3$ via m-separation.

Whereas the undirected graphs associated with a DAG correspond to distributions obtained by *specific* marginalizations (namely those that remove vertices that have no children), CADMGs correspond to certain *specific* ordered sequences of fixing operations. Not all such sequences are allowed: in some cases a vertex may be fixable only after another vertex has already been fixed. The global nested Markov property corresponds to the (union of the) m-separation relations encoded in the CADMGs derived via allowed sequences of these fixing operations, which we call 'valid'. Section 2.13 defines valid fixing sequences. These fixing sequences are closely related to a particular identification strategy for interventional distributions consisting of recursively applying the *g-formula* (Robins, 1986) to an already identified intervention distribution to obtain the result of further interventions; this connection is explored further in section 4.2.

Returning to the example, given the CADMG in Fig. 3(ii), the vertex X_1 may be fixed to give the CADMG in Fig. 3(iii). Further, given the CADMG in Fig. 3(iii), we may fix X_2. The kernel in this graph is

$$q_4(x_4 \mid x_3) = \sum_{x_2} p(x_4 \mid x_3, x_2, x_1)p(x_2 \mid x_1). \qquad (4)$$

The quantity on the RHS of (4) is a function only of x_3 and x_4, and not x_1. This

(i) $x_1 \rightarrow x_2 \quad x_3 \rightarrow x_4 \qquad p(x_1, x_2, x_3, x_4)$

(ii) $x_1 - x_2 - x_3 - x_4 \qquad p(x_1, x_2, x_3, x_4) = \varphi_{12}(x_1, x_2)\varphi_{234}(x_2, x_3, x_4)$

(iii) $x_1 - x_2 \quad x_3 \quad \boxed{x_4} \qquad \sum_{x_4} p(x_1, x_2, x_3, x_4) = \varphi_{12}(x_1, x_2)\varphi_3(x_3)$

(iv) $x_1 \quad \boxed{x_2} \quad x_3 \quad \boxed{x_4} \qquad \sum_{x_2, x_4} p(x_1, x_2, x_3, x_4) = \varphi_1(x_1)\varphi_3(x_3)$

Figure 2: Reduction of a DAG model to a set of undirected models via marginalization and moralization: (i) The original DAG $\mathcal{G}(\{x_1, x_2, x_3, x_4\})$. Undirected graphs representing the factorization of different ancestral margins: (ii) $p(x_1, x_2, x_3, x_4)$; (iii) $p(x_1, x_2, x_3)$; (iv) $p(x_1, x_3)$. Note that we have also included the marginalized variables on the graph in square nodes. The DAG model may be characterized by (the union of) the conditional independence properties implied by the undirected graphs for all ancestral margins.

(i) $x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4$ $p(x_1, x_2, x_3, x_4)$

(ii) $x_1 \rightarrow x_2 \quad \boxed{x_3} \rightarrow x_4$ $\dfrac{p(x_1, x_2, x_3, x_4)}{p(x_3 \mid x_1, x_2)} = q_1(x_1) q_{24}(x_2, x_4 \mid x_1, x_3)$

(iii) $\boxed{x_1} \rightarrow x_2 \quad \boxed{x_3} \rightarrow x_4$ $\dfrac{p(x_1, x_2, x_3, x_4)}{p(x_1) p(x_3 \mid x_1, x_2)} = q_{24}(x_2, x_4 \mid x_1, x_3)$

(iv) $\boxed{x_1} \quad \boxed{x_2} \quad \boxed{x_3} \rightarrow x_4$ $\displaystyle\sum_{x_2} \frac{p(x_1, x_2, x_3, x_4)}{p(x_1) p(x_3 \mid x_1, x_2)} = \sum_{x_2} p(x_2 \mid x_1) p(x_4 \mid x_1, x_2, x_3)$

$$= q_4(x_4 \mid x_3)$$

Figure 3: Reduction of the nested Markov model for an ADMG to a set of ordinary Markov models associated with CADMGs: (i) The ADMG $\mathcal{G}(\{x_1, x_2, x_3, x_4\})$, which is the latent projection of the graph $\mathcal{G}$ from Figure 1(i). CADMGs, representing the Markov structure of derived distributions, resulting from sequences of fixing operations in $\mathcal{G}(\{x_1, x_2, x_3, x_4\})$: (ii) $\langle X_3 \rangle$; (iii) $\langle X_3, X_1 \rangle$ or alternatively $\langle X_1, X_3 \rangle$; (iv) any of the sequences $\langle X_1, X_3, X_2 \rangle$, $\langle X_3, X_1, X_2 \rangle$, $\langle X_3, X_2, X_1 \rangle$; it is not valid to fix X_2 before X_3. The nested Markov model may be defined via the conditional independence properties for all CADMGs and associated distributions obtained (via valid fixing sequences) from the original ADMG and distribution. See also Figure 2 and text.

is precisely the constraint (2) given by the original latent variable DAG model Fig. 1(i). This constraint characterizes the nested Markov model corresponding to the ADMG in Fig. 3(i).

Returning to the original ADMG in Fig. 3(i), we could have fixed X_1 in this graph, also X_4. Had we chosen to fix X_1, we could then subsequently have fixed X_3, and would have arrived at the same CADMG and distribution as shown in Fig. 3(iii). Thus in this case the operations of fixing X_3 and X_1 commute. However, not all such operations commute: X_2 may only be fixed after X_3.

Like DAGs, the nested Markov model may be characterized via a factorization property, as well as by local and global Markov properties described in section 3. In each case these properties are defined via the set of CADMGs that are 'reached' via valid fixing sequences.

The rest of the paper is organized as follows. In section 3.5 we show that the nested Markov model corresponding to a complete ADMG gives the saturated model. In section 4.1, we prove that any marginal distribution in a hidden variable DAG model lies in the appropriate nested Markov model. A simple characterization of identifiable causal effects in hidden variable causal DAG models, based on the fixing operation, is given in section 4.2. Finally, a proof that the nested Markov model may be defined by the set of constraints found by the algorithm in (Tian and Pearl, 2002b) is given in section 4.4.

2 DAGs with Latents and Acyclic Directed Mixed Graphs

In this section we first briefly review DAG models and then introduce mixed graphs. We associate mixed graphs with DAG models containing latent variables via the operation of 'projection'.

A directed acyclic graph (DAG) $\mathcal{G}(V, E)$ is a graph containing directed edges ($\rightarrow$) subject to the restriction that there are no directed cycles $v \rightarrow \cdots \rightarrow v$. We define the *parents* of v to be $\mathrm{pa}_{\mathcal{G}}(v) \equiv \{x \mid x \rightarrow v\}$.

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

$$p(x_V) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}_\mathcal{G}(v)}).$$ (5)

We denote the set of distributions that are Markov relative to a DAG $\mathcal{G}$ by $\mathcal{P}^d(\mathcal{G})$.

2.1 Acyclic Directed Mixed Graphs (ADMGs)

Our motivation for introducing mixed graphs is two-fold. First, by removing latent variables and replacing them with bidirected edges we simplify the representation. For example, to perform a search, instead of considering a potentially infinite class of DAGs with arbitrarily many latent variables, we need only consider a finite set of mixed graphs. Second, although the statistical models that we associate with mixed graphs capture many of the constraints implied by latent variable models, the resulting model will still, in general, be a superset of the set of distributions over the observables that are implied by the original DAG with latents. The use of a mixed graph to represent our model serves to emphasize that in spite of this connection, the set of distributions we are constructing is nonetheless *not* a latent variable model.

A *directed mixed graph* $\mathcal{G}(V, E)$ is a graph with a set of vertices V, and a set of edges E which are each either directed ($\rightarrow$) or bidirected ($\leftrightarrow$). A *path* in $\mathcal{G}$ is a sequence of distinct, adjacent edges, of any type or orientation, between distinct vertices. The first and last vertices on the path are the *endpoints*. It is necessary to specify a path as a sequence of edges rather than vertices because it is possible that there is both a directed and a bidirected edge between the same pair of vertices. A path of the form $a \rightarrow \cdots \rightarrow b$ is a *directed path* from a to b; similarly, a path of the form $a \leftrightarrow \cdots \leftrightarrow b$ is a *bidirected* path between a and b.

A *directed cycle* is a path of the form $v \rightarrow \cdots \rightarrow w$ along with an edge $w \rightarrow v$. An *acyclic* directed mixed graph (ADMG) is a mixed graph containing no directed cycles. For any $T \subset V$, the induced subgraph $\mathcal{G}_T$ of $\mathcal{G}$ contains the

12

vertex set T, and the subset of edges in E that have both endpoints in T.

Let a, b and c be vertices in a mixed graph $\mathcal{G}$. If $b \to a$ then we say that b is a *parent* of a, and a is a *child* of b. A vertex a is said to be an *ancestor* of a vertex d if *either* there is a directed path $a \to \cdots \to d$ from a to d, *or $a = d$*; similarly d is said to be a *descendant* of a. If this is not the case we say that d is a *non-descendant* of a. The set of parents, children, ancestors, descendants, and non-descendants of a in $\mathcal{G}$ are written $\mathrm{pa}_{\mathcal{G}}(a)$, $\mathrm{ch}_{\mathcal{G}}(a)$, $\mathrm{an}_{\mathcal{G}}(a)$, $\mathrm{de}_{\mathcal{G}}(a)$, and $\mathrm{nd}_{\mathcal{G}}(a)$ respectively. 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 $a \notin \mathrm{de}_{\mathcal{G}}(b)$; note that this definition is the same as that for a DAG. We define the set $\mathrm{pre}_{\mathcal{G}, \prec}(b) \equiv \{a \mid a \prec b\}$. We apply these definitions disjunctively to sets, e.g. $\mathrm{an}_{\mathcal{G}}(A) = \bigcup_{a \in A} \mathrm{an}_{\mathcal{G}}(a)$. A set of vertices A in $\mathcal{G}$ is called *ancestral* if $a \in A \Rightarrow \mathrm{an}_{\mathcal{G}}(a) \subseteq A$.

2.2 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* if the edges preceding and succeeding z on the path both have an arrowhead at z, i.e. $\to z \leftarrow$, $\leftrightarrow z \leftrightarrow$, $\leftrightarrow z \leftarrow$, $\to z \leftrightarrow$. A non-endpoint vertex z on a path which is not a collider is a *non-collider on the path*, i.e. $\leftarrow z \to$, $\leftarrow z \leftarrow$, $\to z \to$, $\leftrightarrow z \to$, $\leftarrow z \leftrightarrow$. A path between vertices a and b in a mixed graph $\mathcal{G}$ is said to be *m-connecting given a set C 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 m-connecting a and b given C, then a and b are said to be *m-separated* given C. Sets A and B are said to be *m-separated* given C, if for all a, b, with $a \in A$ and $b \in B$, a and 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; see (Pearl, 1988).

2.3 Latent Projections

Given a DAG with latent variables we associate a mixed graph via the following operation; see (Pearl and Verma, 1991).

Definition 2 (latent projection). *Let $\mathcal{G}$ be an ADMG with vertex set $V \mathbin{\dot{\cup}} L$ where the vertices in V are observed, those in L are latent and $\dot{\cup}$ indicates a disjoint union. The* latent projection $\mathcal{G}(V)$ *is a directed mixed graph with vertex set V, where for every pair of distinct vertices $v, w \in V$:*

(i) *$\mathcal{G}(V)$ contains an edge $v \to w$ if there is a directed path $v \to \cdots \to w$ on which every non-endpoint vertex is in L.*

(ii) *$\mathcal{G}(V)$ contains an edge $v \leftrightarrow w$ if there exists a path between v and w such that the non-endpoints are all non-colliders in L, and such that the edge adjacent to v and the edge adjacent to w both have arrowheads at those vertices. For example, $v \leftrightarrow \cdots \to w$.*

Generalizations of this construction are considered by Wermuth (2011) and Koster (2002) in the context of marginalizing *and* conditioning. As an example, the mixed graph in Figure 3(i) is the latent projection of the DAG shown in Figure 1(i).

Proposition 3. *If $\mathcal{G}$ is a DAG with vertex set $V \mathbin{\dot{\cup}} L$ then $\mathcal{G}(V)$ is an ADMG.*

The latent projection $\mathcal{G}(V)$ represents the set of d-separation relations holding among the variables in V in $\mathcal{G}$:

Proposition 4. *Let $\mathcal{G}$ be a DAG with vertex set $V \mathbin{\dot{\cup}} L$. For disjoint subsets $A, B, C \subseteq V$, (C may be empty), A is d-separated from B given C in $\mathcal{G}$ if and only if A is m-separated from B given C in $\mathcal{G}(V)$.*

However, as we will see later, the latent projection $\mathcal{G}(V)$ captures much more than simply the d-separation relations holding in V. As suggested by Figures 1(i) and 3(i), $\mathcal{G}(V)$ also represents constraints such as (2), and further all those found by the algorithm in (Tian and Pearl, 2002b). However, Evans (2016) shows that some inequality constraints on DAGs with latent variables are not captured by ADMGs.

Figure 4: (i) A conditional mixed graph $\mathcal{G}(V = \{x_2, x_4\}, W = \{x_1, x_3\}, E)$ describing the structure of a kernel $q_{24}(x_2, x_4 \mid x_1, x_3)$. (ii) The corresponding graph $\mathcal{G}^{|W}$ from which the conditional Markov property given by $\mathcal{G}$ may be obtained by applying m-separation.

2.4 Kernels

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

In order to introduce the nested Markov property we introduce a type of bipartite mixed graph that we term a *conditional* ADMG. Whereas an ADMG with vertex set V represents a joint density $p(x_V)$, a conditional ADMG is a graph with two disjoint sets of vertices, V and W that is used to represent the Markov structure of a 'kernel' $q_V(x_V|x_W)$. Following (Lauritzen, 1996, p.46), we define a *kernel* to be a non-negative function $q_V(x_V|x_W)$ satisfying:

$$\sum_{x_V \in \mathfrak{X}_V} q_V(x_V \mid x_W) = 1 \qquad \text{for all } x_W \in \mathfrak{X}_W. \tag{6}$$

We use the term 'kernel' and write $q_V(\cdot|\cdot)$ (rather than $p(\cdot|\cdot)$) to emphasize that these functions, though they satisfy (6) 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 $A \subseteq V$:

$$q_V(x_A|x_W) \equiv \sum_{x_{V\setminus A} \in \mathfrak{X}_{V\setminus A}} q_V(x_V|x_W); \tag{7}$$

$$q_V(x_{V\setminus A}|x_{W\cup A}) \equiv \frac{q_V(x_V|x_W)}{q_V(x_A|x_W)}. \tag{8}$$

For disjoint $V_1 \dot{\cup} V_2 = V$ and $W_1 \dot{\cup} W_2 = W$, we will sometimes write $q_V(x_{V_1}, x_{V_2} \mid x_{W_1}, x_{W_2})$ to mean $q_V(x_{V_1 \cup V_2} \mid x_{W_1 \cup W_2})$.

2.4.1 Conditional ADMGs

A *conditional* acyclic directed mixed graph (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 $w \in W$, $\text{pa}_{\mathcal{G}}(w) = \emptyset$ and there are no bidirected edges involving w. Equivalently, in a CADMG $\mathcal{G}(V, W, E)$ the induced subgraph $\mathcal{G}_W$ has no edges, and all edges from $w \in W$ to $v \in V$ take the form $w \to v$. The rationale for excluding edges between vertices in W or with arrowheads in W is that the CADMG represents the structure of a kernel; the vertices in W merely index distributions over V. (We note that CADMGs represent kernels that are not, in general, formed by standard conditioning from the original observed distribution.)

In a CADMG $\mathcal{G}(V, W, E)$ we will refer to the sets V and W as the *random* and *fixed* nodes respectively. We also introduce operators $\mathbb{V}(\mathcal{G})$ and $\mathbb{W}(\mathcal{G})$ that return, respectively, the sets of random and fixed nodes associated with a CADMG $\mathcal{G}$. We will use circular nodes to indicate the random vertices (in $\mathbb{V}(\mathcal{G})$), and square nodes to indicate the fixed vertices (in $\mathbb{W}(\mathcal{G})$). See, for instance, the CADMGs in Figures 1(ii), 3 and 4(i).

When the edge set or vertex sets are clear from context we will abbreviate $\mathcal{G}(V, W, E)$ as $\mathcal{G}(V, W)$ or $\mathcal{G}$.

2.5 Induced Subgraphs and Districts

An ADMG $\mathcal{G}(V, E)$ may be seen as a CADMG in which $W = \emptyset$. In this manner, though we will state subsequent definitions for CADMGs, they will also apply to ADMGs.

The *induced subgraph of a CADMG* $\mathcal{G}(V, W, E)$ *on a set* A, denoted $\mathcal{G}_A$ consists of $\mathcal{G}(V \cap A, W \cap A, E_A)$, where E_A is the set of edges in $\mathcal{G}$ with both endpoints in A. Note that in forming $\mathcal{G}_A$, the status of the vertices in A with regard to whether they are in V or W is preserved.

Proposition 5. *Given an ADMG* $\mathcal{G}$ *and an ancestral set* A, $\mathcal{G}_A = \mathcal{G}(A)$.

Thus the induced subgraph on an ancestral set A is the same as the latent projection onto A.

Definition 6. *A set of vertices* C *is called* bidirected-connected *if for every pair of vertices* $c, d \in C$ *there is a bidirected path between* c *and* d *with every node on the path in* C. *A maximal bidirected-connected set of random vertices is referred to as a* district. *Let*

$$\mathcal{D}(\mathcal{G}) \equiv \{D \mid D \text{ is a district in } \mathcal{G}\}$$

be the set of districts in $\mathcal{G}$. *For* $v \in \mathbb{V}(\mathcal{G})$, *let* $dis_{\mathcal{G}}(v)$ *be the district containing* v *in* $\mathcal{G}$. *We write* $dis_A(v)$ *as a shorthand for* $dis_{\mathcal{G}_A}(v)$, *the district of* v *in the induced subgraph* $\mathcal{G}_A$.

Tian and Pearl (2002b) refer to districts in ADMGs as 'c-components'. Note that, by definition, in a CADMG, $\mathcal{D}(\mathcal{G})$ forms a partition of the set of random vertices $\mathbb{V}(\mathcal{G})$. By definition, nodes in $\mathbb{W}(\mathcal{G})$ are not included in districts. In a DAG $\mathcal{G}(V, E)$, $\mathcal{D}(\mathcal{G}) = \{\{v\} \mid v \in V\}$, the set of singleton subsets of V.

2.6 Independence in Kernels

We extend the notion of conditional independence to kernels over $\mathfrak{X}_V$ indexed by $\mathfrak{X}_W$. A rigorous treatment of conditional independence in settings where not all variables are random was given in (Constantinou, 2013). With slight abuse of notation we define

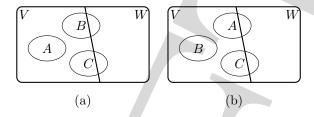$$q_V \equiv \{q_V(x_V \mid x_W), x_W \in \mathfrak{X}_W\}.$$

Figure 5: Illustration of cases in Definition 7:

(a) $A \cap W = \emptyset$; (b) $B \cap W = \emptyset$.

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 kernel q_V, written:*

$$X_A \perp\!\!\!\perp X_B \mid X_C \quad [q_V]$$

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).*

See Figure 5 for an illustration of cases. The condition that the density should exist simply addresses the situation where the conditioning event has zero probability. Note that the kernels appearing in (a) and (b) specify values for all of the variables X_W, and are defined via conditioning in the kernel over X_V that is then specified. For example in (a),

$$q_V(x_A \mid x_B, x_C, x_{W \setminus (B \cup C)}) = q_V(x_{V \cap (A \cup B \cup C)} \mid x_W) / q_V(x_{V \cap (B \cup C)} \mid x_W).$$

Since, if (a) holds, the value of the density does not depend either on the values assigned to X_B or to $X_{W \setminus (B \cup C)}$ it would be natural to express this constraint as:

$$q_V(x_A \mid x_B, x_C, x_{W \setminus (B \cup C)}) = q_V(x_A \mid x_C).$$

18

However, in general, for $R, S \subseteq V$ and $T \subseteq W$, the density $p(x_R \mid x_{S \cup T})$ may not be defined, since in the absence of a distribution over X_W, $X_{W \setminus T}$ cannot be integrated out. Thus we opt for the formulation above. Similar comments apply to (b), interchanging A and B.

Proposition 8. *In a set of kernels $q_V(x_V \mid x_W)$, $X_A \perp\!\!\!\perp X_B \mid X_C$ if and only if either $X_A \perp\!\!\!\perp X_{B \cup (W \setminus C)} \mid X_C$ or $X_B \perp\!\!\!\perp X_{A \cup (W \setminus C)} \mid X_C$.*

Proof. This follows directly from Definition 7. $\qquad\square$

2.6.1 Semi-Graphoid Axioms In Kernels

Classical conditional independence constraints can logically imply others. Though no finite axiomatization of conditional independence is possible (Studený, 1992), deductive derivations of conditional independence constraints in many graphical models can be restricted, without loss of generality, to the semi-graphoid axioms (Dawid, 1979) of symmetry and the 'chain rule', which we reproduce here:

$$(X_A \perp\!\!\!\perp X_B \mid X_C) \Leftrightarrow (X_B \perp\!\!\!\perp X_A \mid X_C),$$

$$(X_A \perp\!\!\!\perp X_B \mid X_{C \cup D}) \wedge (X_A \perp\!\!\!\perp X_D \mid X_C) \Leftrightarrow (X_A \perp\!\!\!\perp X_{B \cup D} \mid X_C).$$

(The chain rule axiom is sometimes written as the three separate axioms of contraction, decomposition and weak union.) We now show that, unsurprisingly, conditional independence constraints defined for kernels also obey these axioms. An additional set of axioms called *separoids* has been shown to apply to versions of conditional independence involving non-stochastic variables like X_W in (Constantinou, 2013).

Proposition 9. *The semi-graphoid axioms are sound for kernel independence.*

Proof. Symmetry follows directly from Definition 7.

Let $q_V(x_V \mid x_W)$ be a kernel for which $(X_A \perp\!\!\!\perp X_{B \cup D} \mid X_C)$ holds. Assume condition (a) for this independence holds, that is $A \cap W = \emptyset$, and assume $q_V(x_A \mid x_B, x_C, x_D, x_{W \setminus (B \cup C \cup D)})$ is only a function of x_A and x_C. Then it

19

immediately follows that condition (a) for $(X_A \perp\!\!\!\perp X_B \mid X_{C \cup D})$ also holds. To see that $(X_A \perp\!\!\!\perp X_D \mid X_C)$ holds, consider the following derivation.

$$q_V(x_A \mid x_C, x_D, x_{W \setminus (C \cup D)})$$

$$= \frac{\displaystyle\sum_{x_{B \cap V}} q_V(x_A, x_{B \cap V}, x_{C \cap V}, x_{D \cap V} \mid x_W)}{\displaystyle\sum_{x_{B \cap V}} q_V(x_{B \cap V}, x_{C \cap V}, x_{D \cap V} \mid x_W)}$$

$$= \frac{\left(\displaystyle\sum_{x_{B \cap V}} q_V(x_A \mid x_B, x_C, x_D, x_{W \setminus (B \cup C \cup D)}) \cdot q_V(x_{B \cap V}, x_{C \cap V}, x_{D \cap V} \mid x_W) \right)}{\displaystyle\sum_{x_{B \cap V}} q_V(x_{B \cap V}, x_{C \cap V}, x_{D \cap V} \mid x_W)}$$

$$= \frac{\left(q_V(x_A \mid x_B, x_C, x_D, x_{W \setminus (B \cup C \cup D)}) \cdot \displaystyle\sum_{x_{B \cap V}} q_V(x_{B \cap V}, x_{C \cap V}, x_{D \cap V} \mid x_W) \right)}{\displaystyle\sum_{x_{B \cap V}} q_V(x_{B \cap V}, x_{C \cap V}, x_{D \cap V} \mid x_W)}$$

$$= q_V(x_A \mid x_B, x_C, x_D, x_{W \setminus (B \cup C \cup D)}).$$

Here the first equality follows by (7)&(8), the second by the chain rule of probability, which applies to kernels also by (7)&(8), the third since we established above that $(X_A \perp\!\!\!\perp X_B \mid X_{C \cup D})$ holds in $q_V(x_V \mid x_W)$, and the last by cancellation. Since $(X_A \perp\!\!\!\perp X_{B \cup D} \mid X_C)$, the final term does not depend upon x_B or x_D, so the independence $(X_A \perp\!\!\!\perp X_D \mid X_C)$ follows.

Now assume $(X_A \perp\!\!\!\perp X_{B \cup D} \mid X_C)$ holds due to condition (b), that is $q_V(x_{B \cap V}, x_{D \cap V} \mid x_A, x_C, x_{W \setminus (A \cup C)})$ is only a function of x_B, x_C and x_D. Then $q_V(x_{B \cap V} \mid x_A, x_C, x_D, x_{W \setminus (A \cup C \cup D)})$ is also only a function of x_B, x_C and x_D, which in turn implies $(X_A \perp\!\!\!\perp X_B \mid X_{C \cup D})$. To see that $(X_A \perp\!\!\!\perp X_D \mid X_C)$ holds, repeat the above argument, but swap x_A and x_D, and use the fact that $(X_A \perp\!\!\!\perp X_{B \cup D} \mid X_C)$ implies $(X_A \perp\!\!\!\perp X_D \mid X_{B \cup C})$ under either (a) or (b). This was already shown above.

To show the converse, let $q_V(x_V \mid x_W)$ be a kernel in which $(X_A \perp\!\!\!\perp X_B \mid X_{C \cup D})$ and $(X_A \perp\!\!\!\perp X_D \mid X_C)$ hold. If this is due to condition (a) (where $A \cap W = \emptyset$), then $(X_A \perp\!\!\!\perp X_{B \cup C} \mid X_D)$ follows by above derivation, and the two assumed independences. If this is due to condition (b) (where $(B \cup D) \cap W = \emptyset$), then $(X_A \perp\!\!\!\perp X_{B \cup C} \mid X_D)$ follows by the above derivation where x_A and x_D are
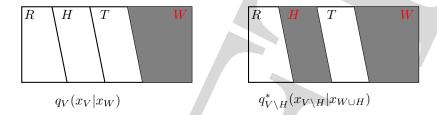
20

Figure 6: Structure of sets for invariance properties considered in section 2.7; $V = R \dot\cup H \dot\cup T$; shaded sets are fixed.

swapped, and the two assumed independences. □

2.7 Constructing kernels

We will typically construct new kernels via the operation of dividing either a distribution $p(x_V)$ by $p(x_H \mid x_T)$ or an existing kernel $q_V(x_V \mid x_W)$ by $q_V(x_H \mid x_{T \cup W})$, where $H \dot\cup T \subseteq V$. For the results in the remainder of this section, we will consider a kernel $q_V(x_V \mid x_W)$ where $V = R \dot\cup H \dot\cup T$, and a new kernel

$$q^*_{V \setminus H}(x_{V \setminus H} \mid x_H, x_W) = q^*_{V \setminus H}(x_R, x_T \mid x_H, x_W) \equiv \frac{q_V(x_R, x_H, x_T \mid x_W)}{q_V(x_H \mid x_T, x_W)}. \quad (9)$$

See Figure 6 for an illustration.

Lemma 10.

$$q^*_{V \setminus H}(x_R, x_T \mid x_H, x_W) = q_V(x_R \mid x_H, x_T, x_W) q_V(x_T \mid x_W), \quad (10)$$

and hence

$$q^*_{V \setminus H}(x_R \mid x_H, x_T, x_W) = q_V(x_R \mid x_H, x_T, x_W); \quad (11)$$

$$q^*_{V \setminus H}(x_T \mid x_W) = q^*_{V \setminus H}(x_T \mid x_H, x_W) = q_V(x_T \mid x_W). \quad (12)$$

Proof. By the chain rule of probability:

$$q_V(x_R, x_H, x_T \mid x_W) = q_V(x_R \mid x_H, x_T, x_W) q_V(x_H \mid x_T, x_W) q_V(x_T \mid x_W).$$

21

Hence (10) follows directly from (9). Since

$$q^*_{V \setminus H}(x_R, x_T \mid x_H, x_W) = q^*_{V \setminus H}(x_R \mid x_H, x_T, x_W) q^*_{V \setminus H}(x_T \mid x_H, x_W), \quad (13)$$

the second equality in (12) follows by summing the right-hand sides of (10) and (13) over x_R. The first then follows directly since $q_V(x_T \mid x_W)$ is not a function of x_H. Finally (11) follows from (12) by canceling $q^*_{V \setminus H}(x_T \mid x_H, x_W)$ and $q_V(x_T \mid x_W)$ from the right-hand sides of (10) and (13). $\qquad \square$

2.7.1 New independences resulting from kernel construction

Lemma 11. $(X_H \perp\!\!\!\perp X_T \mid X_W)$ in $q^*_{V \setminus H}(x_{V \setminus H} \mid x_H, x_W)$.

Proof. By definition and the chain rule of probability, $q^*_{V \setminus H}(x_{V \setminus H} \mid x_H, x_W) = q_V(x_{V \setminus (H \cup T)} \mid x_H, x_T, x_W) \cdot q_V(x_T \mid x_W)$. By definition of marginalization in kernels,

$$
\begin{aligned}
q^*_{V \setminus H}(x_T \mid x_H, x_W) &= \sum_{x_{V \setminus (H \cup T)}} q^*_{V \setminus H}(x_{V \setminus H} \mid x_H, x_W) \\
&= \sum_{x_{V \setminus (H \cup T)}} q_V(x_{V \setminus (H \cup T)} \mid x_H, x_T, x_W) \cdot q_V(x_T \mid x_W) \\
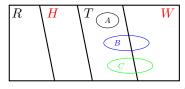&= q_V(x_T \mid x_W).
\end{aligned}
$$

But this kernel is only a function of x_T and x_W by construction, which implies our conclusion. $\qquad \square$
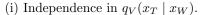
Corollary 12. *If* $X_V \perp\!\!\!\perp X_{W \setminus W_1} \mid X_{W_1}$ *in* q_V, *then* $X_{H \cup (W \setminus W_1)} \perp\!\!\!\perp X_T \mid X_{W_1}$ *in* q^*_V.
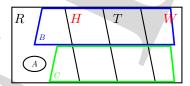
Proof. As above, but q_V is now only a function of x_T and x_{W_1} by construction, which implies our conclusion. $\qquad \square$

2.7.2 Preservation of existing independences in a kernel

We now state two important properties that capture the way conditional independence and the fixing operation interact. These properties will let us transfer conditional independence statements from one kernel to another. We state these

(i) Independence in $q_V(x_T \mid x_W)$.   (ii) Independence in $q_V(x_R \mid x_H, x_T, x_W)$.

Figure 7: Preservation of independence.

results as theorems about probability distributions, but they can also be viewed abstractly as properties of irrelevance given fixing and conditioning, much in the same way as the graphoid axioms (Dawid, 1979) can be viewed either as results about probability, or as axioms characterizing "irrelevance."

Proposition 13 (ordering)**.** *Given disjoint sets A, B, C, where C may be empty, if $A \subseteq T$ and $B, C \subseteq T \cup W$, then*

$$X_A \perp\!\!\!\perp X_B \mid X_C \ [q_V] \quad \Leftrightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_C \ [q^*_{V \setminus H}].$$

This result follows directly from (12). In words, it states that, given an ordering in which W precedes V, fixing variables in H preserves conditional independence statements among variables that precede H in the ordering; see Fig. 7(i). If we interpret fixing causally as an intervention operation, then the result states that interventions in the future cannot causally affect the past; that is, 'retrocausality' is forbidden!

Proposition 14 (modularity)**.** *Given disjoint sets A, B, C, where C may be empty, if $A \subseteq R$ and $(B \cup C) \supseteq H \cup T$, then*

$$X_A \perp\!\!\!\perp X_B \mid X_C \ [q_V] \quad \Leftrightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_C \ [q^*_{V \setminus H}].$$

This result, which follows directly from (11), is illustrated in Fig. 7(ii). In words, it says that if we express a probability distribution as a set of factors

23

via the chain rule of probability, and a conditional independence statement can be stated exclusively in terms of variables in one of the factors, then fixing a variable such that another factor is dropped does not affect this conditional independence statement. In other words, "factors are modular." If we interpret fixing causally, then this result can be seen as stating that 'local' causal systems stay invariant after interventions.

2.8 Markov Properties for CADMGs

As described earlier, a CADMG $\mathcal{G}(V, W)$ represents the structure of a kernel $q_V(x_V \mid x_W)$. We now introduce a number of Markov properties, whose equivalence we will prove in section 2.10.

2.8.1 The CADMG global Markov property

The global Markov property for CADMGs may be derived from m-separation via the following simple construction:

Definition 15. *Given a CADMG $\mathcal{G}(V, W)$, 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 bidirected edges between all pairs of vertices $w, w' \in W$, and then eliminating the distinction between vertices in V and W. See Figure 4(ii) for an example.

Definition 16. *A kernel q_V satisfies the global Markov property for a CADMG $\mathcal{G}(V, W)$ if for arbitrary disjoint sets A, B, C, (C may be empty)*

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

We denote the set of such kernels by $\mathcal{P}^c_m(\mathcal{G})$.

24

2.8.2   The CADMG local Markov property

The local Markov property for a DAG states that each vertex v is independent of vertices prior to v under a topological ordering conditional on the parents of v. In the context of CADMG, $\mathcal{G}(V, W)$, the Markov blanket plays the same role as the set of parents.

If $t \in V$ then the *Markov blanket* of t in $\mathcal{G}$ is defined as:

$$\mathrm{mb}_{\mathcal{G}}(t) \equiv \mathrm{pa}_{\mathcal{G}}\Big(\mathrm{dis}_{\mathcal{G}}(t)\Big) \cup \Big(\mathrm{dis}_{\mathcal{G}}(t) \setminus \{t\}\Big). \tag{14}$$

Given a vertex $t \in V$ such that $\mathrm{ch}_{\mathcal{G}}(t) = \emptyset$, a kernel q_V obeys the *local Markov property for $\mathcal{G}$ at t* if

$$X_t \perp\!\!\!\perp X_{(V \cup W) \setminus (\mathrm{mb}_{\mathcal{G}}(t) \cup \{t\})} \mid X_{\mathrm{mb}_{\mathcal{G}}(t)} \ [q_V]. \tag{15}$$

If $\prec$ is a topological total ordering on the vertices in $\mathcal{G}$, then for a subset A define $\max_{\prec}(A)$ to be the $\prec$-greatest vertex in A.

We define the set of kernels obeying the *ordered local Markov property for the CADMG $\mathcal{G}(V, W)$ under the ordering $\prec$* as follows:

$$\mathcal{P}_l^c(\mathcal{G}, \prec) \equiv \{q_V(x_V \mid x_W) \mid \text{for every ancestral set } A, \text{with } \max_{\prec}(A) \in V,$$
$$q_V(x_{V \cap A} \mid x_W) \text{ obeys the local Markov property for} \tag{16}$$
$$\mathcal{G}(V \cap A, W) \text{ at } \max_{\prec}(A)\}.$$

In what follows we will make use of the following extension: for an ancestral set A in a CADMG $\mathcal{G}$ and a vertex $t \in V \cap A$ such that $\mathrm{ch}_{\mathcal{G}}(t) \cap A = \emptyset$, let

$$\mathrm{mb}_{\mathcal{G}}(t, A) \equiv \mathrm{mb}_{\mathcal{G}_A}(t). \tag{17}$$

Proposition 17. *Given a CADMG $\mathcal{G}$, an ancestral set A, and a random vertex $t \in A$ such that $\mathrm{ch}_{\mathcal{G}}(t) \cap A = \emptyset$,*

(i) $\mathrm{mb}_{\mathcal{G}}(t, A) \subseteq D \cup \mathrm{pa}_{\mathcal{G}}(D)$, where $t \in D \in \mathcal{D}(\mathcal{G})$;

(ii) if A^* is an ancestral set and $t \in A^* \subseteq A$, then $\mathrm{mb}_{\mathcal{G}}(t, A^*) \subseteq \mathrm{mb}_{\mathcal{G}}(t, A)$.

2.8.3 The CADMG augmented Markov property

The following is the analog of moralization in DAGs for CADMGs. For a CADMG $\mathcal{G}(V,W)$, the *augmented graph derived from* $\mathcal{G}$, denoted $(\mathcal{G})^a$, is an undirected graph with the same vertex set as $\mathcal{G}$ such that $c - d$ in $(\mathcal{G})^a$ if and only if c and d are connected by a path containing only colliders in $\mathcal{G}^{|W}$. For any three disjoint sets A, B, C in an undirected graph $\mathcal{G}$, we say that A *is separated from B given C* if every undirected path from a vertex $a \in A$ to a vertex $b \in B$ contains at least one vertex in C. A kernel q_V obeys the *augmented Markov property for a CADMG $\mathcal{G}(V,W)$* if for arbitrary disjoint sets A, B, C (C may be empty), whenever A is separated from B given C in $(\mathcal{G}_{\mathrm{an}(A \cup B \cup C)})^a$ it follows that $X_A \perp\!\!\!\perp X_B \mid X_C$ in q_V. We denote the set of such kernels by $\mathcal{P}_a^c(\mathcal{G})$.

2.9 Tian Factorization for CADMGs

The joint distribution under a DAG model may be factorized into univariate densities. In DAG models, these factors take the form $p(x_a | x_{\mathrm{pa}_{\mathcal{G}}(a)})$. This factor is a conditional distribution for a singleton variable X_a, given the set of variables corresponding to parents of a in the graph. The factorization property may be generalized to CADMGs by requiring factorization of q_V into kernels over districts.

We define the set of kernels that *Tian factorize* with respect to a CADMG:

$$\mathcal{P}_f^c(\mathcal{G}) \equiv \Big\{ q_V(x_V \mid x_W) \mid \text{for every ancestral set } A, \text{ there exist kernels } f_D^A(\cdot | \cdot)$$
$$\text{s.t. } q_V(x_{V \cap A} \mid x_W) = \prod_{D \in \mathcal{D}(\mathcal{G}_A)} f_D^A(x_D \mid x_{\mathrm{pa}_{\mathcal{G}}(D) \setminus D}) \Big\}. \quad (18)$$

In the next Lemma we show that the terms $f_D^A(\cdot \mid \cdot)$ arising in (18) are equal to products of univariate conditional densities, i.e. instances of the g-formula of Robins (1986), with conditioning sets determined by the Markov blankets, and thus do not depend on A other than through D.

Lemma 18. *For every topological ordering $\prec$ on the vertices in a CADMG $\mathcal{G}$, if $q_V(x_V \mid x_W) \in \mathcal{P}_f^c(\mathcal{G})$ then for every ancestral set A, and every $D \in \mathcal{D}(\mathcal{G}_A)$,*

we have:

$$f_D^A(x_D \mid x_{\mathrm{pa}(D)\setminus D}) = \prod_{d \in D} q_V(x_d \mid x_{T_{\prec}^{(d,D)}}), \tag{19}$$

where $T_{\prec}^{(d,D)} \equiv \mathrm{mb}_\mathcal{G}(d, \ \mathrm{an}_\mathcal{G}(D) \cap \mathrm{pre}_{\mathcal{G},\prec}(d))$, *so that*

$$q_V(x_d \mid x_{T_{\prec}^{(d,D)}}) = q_V(x_d \mid x_{A \cap \mathrm{pre}_{\mathcal{G},\prec}(d)}, x_W). \tag{20}$$

Conversely if, under some topological ordering $\prec$, (20) *holds for all ancestral sets* A *and all* $d \in A$ *then* $q_V(x_V \mid x_W) \in \mathcal{P}_f^c(\mathcal{G})$.

Note that by Proposition 17, $\mathrm{mb}_\mathcal{G}(d, \mathrm{an}_\mathcal{G}(D) \cap \mathrm{pre}_{\mathcal{G},\prec}(d)) \subseteq D \cup \mathrm{pa}_\mathcal{G}(D)$. Lemma 18 has the following important consequence:

$$\mathcal{P}_f^c(\mathcal{G}) = \{q_V(x_V \mid x_W) \mid \text{for every ancestral set } A,$$
$$q_V(x_{V \cap A} \mid x_W) = \prod_{D \in \mathcal{D}(\mathcal{G}_A)} q_D(x_D \mid x_{\mathrm{pa}(D)\setminus D}) \ \} \tag{21}$$

where $q_D(x_D \mid x_{\mathrm{pa}(D)\setminus D})$ is defined via the right-hand side of (19) under any topological ordering. In a context where we refer to q_V and q_D where $D \in \mathcal{D}(\mathcal{G})$ it is implicit that q_D is derived from q_V in this way. We will subsequently extend this notation to include other sets.

Proof. (Cf. proof of Lemma 1 in (Tian and Pearl, 2002a)):

($\Rightarrow$) The proof is by induction on the size of the set A in (18). If $|A| = 1$, the claim is trivial. Suppose that the claim holds for sets A of size less than n. Specifically, we assume that all factors $f_D^A(\cdot|\cdot)$ occurring in (18) for sets A such that $|A| < n$, obey (19) and (20).

Now suppose that A contains n variables and that $A \subseteq \{t\} \cup \mathrm{pre}_{\mathcal{G},\prec}(t)$ for some vertex $t \in A$. Let $D^t \equiv \mathrm{dis}_{\mathcal{G}_A}(t)$ be the district containing t in $\mathcal{G}_A$, so that by hypothesis:

$$q_V(x_{A \cap V} \mid x_W) = f_{D^t}^A(x_{D^t} \mid x_{\mathrm{pa}(D^t)\setminus D^t}) \prod_{D \in \mathcal{D}(\mathcal{G}_A)\setminus\{D^t\}} f_D^A(x_D \mid x_{\mathrm{pa}(D)\setminus D}). \tag{22}$$

Since $A \setminus \{t\} \subseteq \mathrm{pre}_{\mathcal{G},\prec}(t)$, for all $D \in \mathcal{D}(\mathcal{G}_A) \setminus \{D^t\}$, $t \notin \mathrm{pa}_\mathcal{G}(D) \setminus D$. Thus

summing both sides of (22) over x_t leads to:

$$q_V(x_{(A \cap V) \setminus \{t\}} \mid x_W) = \left(\sum_{x_t} f_{D^t}^A(x_t, x_{D^t \setminus \{t\}} \mid x_{\mathrm{pa}(D^t) \setminus D^t}) \right)$$
$$\times \prod_{D \in \mathcal{D}(\mathcal{G}_A) \setminus \{D^t\}} f_D^A(x_D \mid x_{\mathrm{pa}(D) \setminus D}). \qquad (23)$$

Now since $\prec$ is a topological ordering, $A \setminus \{t\}$ is an ancestral set in $\mathcal{G}$; further every district in $\mathcal{D}(\mathcal{G}_A) \setminus \{D^t\}$ is also a district in $\mathcal{G}_{A \setminus \{t\}}$ hence, by the induction hypothesis, all of the corresponding densities $f_D^A(\cdot \mid \cdot)$ in (23) obey (19) and (20). Rearranging (22) gives:

$$f_{D^t}(x_{D^t} \mid x_{\mathrm{pa}(D^t) \setminus D^t}) = \frac{q_V(x_{A \cap V} \mid x_W)}{\prod_{D \in \mathcal{D}(\mathcal{G}_A) \setminus \{D^t\}} f_D^A(x_D \mid x_{\mathrm{pa}(D) \setminus D})}.$$

By the chain rule of probability,

$$q_V(x_{A \cap V} \mid x_W) = \prod_{a \in A \cap V} q_V(x_a \mid x_{A \cap \mathrm{pre}_{\mathcal{G}, \prec}(a)}, x_W).$$

Since for every $D \in \mathcal{D}(\mathcal{G}) \setminus \{D^t\}$, $f_D^A(\cdot \mid \cdot)$ obeys (19) and (20) so

$$f_{D^t}(x_{D^t} \mid x_{\mathrm{pa}(D^t) \setminus D^t}) = \prod_{d \in D^t} q_V(x_d \mid x_{A \cap \mathrm{pre}_{\mathcal{G}, \prec}(d)}, x_W). \qquad (24)$$

By the inductive hypothesis applied to $A \setminus \{t\}$, we have that (20) holds for all $d \in D^t \setminus \{t\} \subseteq A \setminus \{t\}$. It is thus sufficient to prove that (20) also holds for t. Rearranging (24) we obtain:

$$q_V(x_t \mid x_{A \cap \mathrm{pre}_{\mathcal{G}, \prec}(t)}, x_W) = \frac{f_{D^t}(x_{D^t} \mid x_{\mathrm{pa}(D^t) \setminus D^t})}{\prod_{d \in D^t \setminus \{t\}} q_V(x_d \mid x_{T_{\prec}^{(d,A)}})}. \qquad (25)$$

By Proposition 17, for all $d \in D^t \setminus \{t\}$ we have $T^{(d,A)} \subseteq (D^t \setminus \{t\}) \cup \mathrm{pa}_{\mathcal{G}}(D^t)$, so the RHS is a function of $x_{D^t \cup \mathrm{pa}(D^t)}$. Hence:

$$X_t \perp\!\!\!\perp X_{(V \cup W) \setminus (D^t \cup \mathrm{pa}(D^t))} \mid X_{(D^t \setminus \{t\}) \cup \mathrm{pa}(D^t)} \quad [q_V]$$

from which (20) follows.

($\Rightarrow$) Follows from construction of the kernels $f_D^A(\cdot \mid \cdot)$ via (19). $\qquad \square$

2.10 Equivalence of Factorizations and Markov Properties for CADMGs

The above definitions describe the same set of kernels due to the following result.

Theorem 19. $\mathcal{P}_f^c(\mathcal{G}) = \mathcal{P}_l^c(\mathcal{G}, \prec) = \mathcal{P}_m^c(\mathcal{G}) = \mathcal{P}_a^c(\mathcal{G})$

The proof is given in the supplementary materials, and the argument for the last two equalities follows that given in (Richardson, 2003). Given this result we use $\mathcal{P}^c(\mathcal{G})$ to denote the set of such kernels, and simply refer to a kernel $q_V \in \mathcal{P}^c(\mathcal{G})$ as being *Markov* with respect to a CADMG $\mathcal{G}$.

2.11 The fixing operation and fixable vertices

We now introduce a 'fixing' operation on an ADMG or CADMG that has the effect of transforming a random vertex into a fixed vertex, thereby changing the graph. However, we define this operation only for a subset of the vertices in the graph, which we term the set of (potentially) fixable vertices.

Definition 20. *Given a CADMG $\mathcal{G}(V,W)$ the set of* fixable vertices *is*

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

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

Proposition 21. *In a CADMG $\mathcal{G}(V,W)$, for every district $D \in \mathcal{D}(\mathcal{G})$, $D \cap \mathbb{F}(\mathcal{G}) \neq \emptyset$.*

That is, in every district there is at least one vertex that is fixable.

Proof. Let $\prec$ be any topological ordering of $\mathcal{G}$. In every district D, the $\prec$-maximal vertices in D are fixable in $\mathcal{G}$. □

Proposition 22. *If $D \in \mathcal{D}(\mathcal{G})$, $v \in D$ but $v \notin \mathbb{F}(\mathcal{G})$, then $\mathrm{de}_{\mathcal{G}}(v) \cap D \cap \mathbb{F}(\mathcal{G}) \neq \emptyset$.*

Thus, if a vertex in a district is not fixable then there is a descendant of the vertex within the district that is fixable.

29

Proof. As in the proof of Proposition 21 consider a maximal vertex in the set $\mathrm{de}_{\mathcal{G}}(v) \cap D$. $\qquad\square$

Proposition 23. *In a CADMG $\mathcal{G}(V, W)$ if $v \in V$ and $\mathrm{ch}_{\mathcal{G}}(v) = \emptyset$ then $v \in \mathbb{F}(\mathcal{G})$.*

Thus any vertex $v \in V$ in a CADMG $\mathcal{G}$ that has no children is fixable.

Proof. This follows by definition of $\mathbb{F}(\mathcal{G})$. $\qquad\square$

We recall that $\mathrm{mb}_{\mathcal{G}}(t)$, defined in (14) is the set of vertices $v \in V \cup W$ which can be reached via paths of the form:

$$t \leftarrow v, \qquad\qquad t \leftrightarrow \cdots \leftrightarrow v, \qquad\qquad t \leftrightarrow \cdots \leftrightarrow\leftarrow v.$$

If $\mathrm{ch}_{\mathcal{G}}(t) = \emptyset$ then we have (15) which is the CADMG local Markov property. However, if t is fixable then

$$X_t \perp\!\!\!\perp X_{\mathrm{nd}_{\mathcal{G}}(t) \setminus \mathrm{mb}_{\mathcal{G}}(t)} \mid X_{\mathrm{mb}_{\mathcal{G}}(t)} \quad [q_V] \tag{26}$$

follows from m-separation in $\mathcal{G}^{|W}$. This holds even if $\mathrm{ch}_{\mathcal{G}}(t) \neq \emptyset$.

Definition 24. *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

$$\phi_r(q_V(x_V \mid x_W); \mathcal{G}) \equiv \frac{q_V(x_V \mid x_W)}{q_V(x_r \mid x_{\mathrm{mb}_{\mathcal{G}}(r)})}. \tag{27}$$

Note that $\mathbb{V}(\phi_r(\mathcal{G})) = \mathbb{V}(\mathcal{G}) \setminus \{r\}$ and $\mathbb{W}(\phi_r(\mathcal{G})) = \mathbb{W}(\mathcal{G}) \cup \{r\}$, so that $\phi_r(\mathcal{G})$ is a new CADMG in which the status of r changes from random to fixed, while $\phi_r(q_V; \mathcal{G})$ forms a new kernel, as we show below in Proposition 26. Although the CADMG $\phi_r(\mathcal{G})$ is determined solely by the graph $\mathcal{G}$ given as input, the transformation on the kernel $\phi_r(q_V(x_V \mid x_W); \mathcal{G})$, is a function of both the graph and the kernel itself.

Proposition 25. *If q_V is Markov with respect to $\mathcal{G}$, and $r \in \mathbb{F}(\mathcal{G})$ then*

$$\phi_r(q_V(x_V \mid x_W); \mathcal{G}) = q_V(x_V \mid x_W)/q_V(x_r \mid x_{\mathrm{nd}_{\mathcal{G}}(r)}). \tag{28}$$

Proof. This follows from Theorem 19 and (26) with $r = t$ and $V = \mathrm{an}_{\mathcal{G}}(\mathrm{dis}_{\mathcal{G}}(r))$. $\qquad\square$

Proposition 26. *If $\mathcal{G}(V, W)$ is a CADMG and $r \in \mathbb{F}(\mathcal{G})$, then for all $x_r \in \mathfrak{X}_r$ and $x_W \in \mathfrak{X}_W$, we have:*

$$\sum_{x_{V \setminus \{r\}}} \phi_r(q_V(x_V \mid x_W); \mathcal{G}) = 1,$$

so that $\phi_r(q_V(x_V \mid x_W); \mathcal{G})$ forms a new kernel that maps values of $x_{\{r\} \cup W}$ to normalized probability distributions over $X_{V \setminus \{r\}}$.

Lemma 27. *If $r \in \mathbb{F}(\mathcal{G})$ then $\mathbb{F}(\mathcal{G}) \setminus \{r\} \subseteq \mathbb{F}(\phi_r(\mathcal{G}))$.*

This Lemma implies that any vertex s that was fixable before r was fixed is still fixable after r has been fixed (with the obvious exception of r itself). This Lemma is important because it shows that when fixing vertices, although our choices may be limited at various stages, we never have to backtrack. In other words, it is never the case that when faced with a choice between fixing r and r', by choosing r we preclude subsequently fixing r'.

Proof. This follows from the definition of $\mathbb{F}(\mathcal{G})$ and $\phi_r(\mathcal{G})$. Since the set of edges in $\phi_r(\mathcal{G})$ is a subset of the set of edges in $\mathcal{G}$, any vertex $t \in V \setminus \{r\}$ that is in $\mathbb{F}(\mathcal{G})$ is also in $\mathbb{F}(\phi_r(\mathcal{G}))$. $\qquad\square$

Proposition 28. *If $\mathcal{G}$ is a subgraph of $\mathcal{G}^*$ with the same random and fixed vertex sets then $\mathbb{F}(\mathcal{G}^*) \subseteq \mathbb{F}(\mathcal{G})$.*

Proof. If r has no descendant within the district containing it in $\mathcal{G}^*$ then this also holds in $\mathcal{G}$. $\qquad\square$

Proposition 29. *Let $\mathcal{G}$ be a CADMG, with $r \in \mathbb{F}(\mathcal{G})$. If $r \in D^r \in \mathcal{D}(\mathcal{G})$ then*

$$\mathcal{D}(\phi_r(\mathcal{G})) = (\mathcal{D}(\mathcal{G}) \setminus \{D^r\}) \,\dot\cup\, \mathcal{D}(\mathcal{G}_{D^r \setminus \{r\}}).$$

31

Thus, if $D \in \mathcal{D}(\phi_v(\mathcal{G}))$ then $D \subseteq D^*$ for some $D^* \in \mathcal{D}(\mathcal{G})$; further if $D \neq D^*$, then $r \in D^*$.

In words, the set of districts in $\phi_r(\mathcal{G})$, the graph obtained by fixing r, consist of the districts in $\mathcal{G}$ that do not contain r, together with new districts that are subsets of D^r, the district in $\mathcal{G}$ that contains r. The new districts are bidirected-connected subsets of D^r after removing r.

Proposition 30. *For $t \in \mathbb{F}(\mathcal{G})$, and $v \in V \setminus \{t\}$, $\mathrm{pa}_{\phi_t(\mathcal{G})}(v) = \mathrm{pa}_{\mathcal{G}}(v)$.*

Proof. This follows since when forming $\phi_t(\mathcal{G})$ we only remove edges that are into t. $\qquad\square$

Proposition 31. *For $t \in \mathbb{F}(\mathcal{G})$, and $v \in V$, $\mathrm{de}_{\phi_t(\mathcal{G})}(v) \subseteq \mathrm{de}_{\mathcal{G}}(v)$.*

Proof. This follows since no new directed edges are introduced in $\phi_t(\mathcal{G})$. $\qquad\square$

2.12 Fixing and factorization

Proposition 32. *Take a CADMG $\mathcal{G}(V, W, E)$ with kernel $q_V \in \mathcal{P}^c(\mathcal{G})$ with associated district factorization:*

$$q_V(x_V \mid x_W) = \prod_{D \in \mathcal{D}(\mathcal{G})} q_D(x_D \mid x_{\mathrm{pa}_{\mathcal{G}}(D) \setminus D}), \qquad (29)$$

where the kernels $q_D(x_D \mid x_{\mathrm{pa}_{\mathcal{G}}(D) \setminus D})$ are defined via the right-hand side of (19). If $r \in \mathbb{F}(\mathcal{G})$ and $D^r \in \mathcal{D}(\mathcal{G})$ is the district containing r then

$$\phi_r(q_V(x_V \mid x_W); \mathcal{G}) = q_{D^r}(x_{D^r \setminus \{r\}} \mid x_{\mathrm{pa}_{\mathcal{G}}(D^r) \setminus D^r}) \prod_{D \in \mathcal{D}(\mathcal{G}) \setminus \{D^r\}} q_D(x_D \mid x_{\mathrm{pa}_{\mathcal{G}}(D) \setminus D}). \tag{30}$$

In words, the result of a fixing operation is solely to marginalize the variable X_r from the density q_{D^r} associated with the district D^r in which the vertex r occurs, while leaving unchanged all of the other terms q_D in the factorization.

Proof.

$$\phi_r(q_V(x_V \mid x_W); \mathcal{G})$$

$$= \left(\prod_{D \in \mathcal{D}(\mathcal{G}_{V \cup W})} q_D(x_D \mid x_{\mathrm{pa}_{\mathcal{G}}(D) \setminus D}) \right) \Big/ q_V(x_r \mid x_{\mathrm{mb}_{\mathcal{G}}(r)})$$

$$= \frac{q_{D^r}(x_{D^r} \mid x_{\mathrm{pa}_{\mathcal{G}}(D^r) \setminus D^r})}{q_V(x_r \mid x_{\mathrm{mb}_{\mathcal{G}}(r)})} \prod_{D \in \mathcal{D}(\mathcal{G}) \setminus \{D^r\}} q_D(x_D \mid x_{\mathrm{pa}_{\mathcal{G}}(D) \setminus D}).$$

Now consider an ordering $\prec$ on the vertices in $\mathcal{G}$ under which r is the last vertex in D, so that $D^r \setminus \{r\} \subseteq \mathrm{pre}_{\mathcal{G}, \prec}(r)$; since $r \in \mathbb{F}(\mathcal{G})$, such an ordering exists. By (19) we have that:

$$q_{D^r}(x_{D^r} \mid x_{\overline{\mathrm{pa}(D^r)} \setminus D^r}) = \prod_{d \in D^r} q_{D^r}(x_d \mid x_{T_{\prec}^{(d,A)}}), \tag{31}$$

where $T_{\prec}^{(d,A)} \equiv \mathrm{mb}\left(d, \ A \cap \mathrm{pre}_{\mathcal{G}, \prec}(d)\right) \subseteq D^r \cup \mathrm{pa}(D^r)$ by Proposition 17. Finally, $q_V(x_r \mid x_{\mathrm{mb}_{\mathcal{G}}(r)}) = q_V(x_r \mid x_{\mathrm{nd}_{\mathcal{G}}(r)}) = q_{D^r}(x_r \mid x_{T_{\prec}^{(r,A)}})$, by the local Markov property and (19). Hence these terms cancel as required. $\square$

Corollary 33. *If* $\mathcal{G}(V, W)$ *is a CADMG,* $r \in \mathbb{F}(\mathcal{G})$ *and* $\mathrm{ch}_{\mathcal{G}}(r) = \emptyset$ *then*

$$\phi_r(q_V(x_V \mid x_W); \mathcal{G}) = q_V(x_{V \setminus \{r\}} \mid x_W).$$

Thus if $\mathrm{ch}_{\mathcal{G}}(r) = \emptyset$, then ϕ_r simply marginalizes over X_r: the conditioning on X_r in $\phi_r(q_V(x_V \mid x_W))$ is vacuous in the sense that the resulting kernel does not depend on the value of X_r. Though it may appear unnatural to think of marginalization in this way, it greatly simplifies our development to only need to consider one operation that fixes vertices in a graph.

2.13 Reachable subgraphs of an ADMG

.

We use $\circ$ to indicate composition of operations in the natural way, so that if $s \in \mathbb{F}(\mathcal{G})$ and $r \in \mathbb{F}(\phi_s(\mathcal{G}))$ then

$$\phi_r \circ \phi_s(\mathcal{G}) \equiv \phi_r(\phi_s(\mathcal{G}))$$

$$\phi_r \circ \phi_s(q_V; \mathcal{G}) \equiv \phi_r\left(\phi_s\left(q_V; \mathcal{G}\right); \phi_s(\mathcal{G})\right).$$

Similarly for a sequence $\mathbf{w} = \langle w_1, \ldots, w_k \rangle$, such that $w_1 \in \mathbb{F}(\mathcal{G})$ and for $j = 2, \ldots, k$, $w_j \in \mathbb{F}(\phi_{w_{j-1}} \circ \cdots \circ \phi_{w_1}(\mathcal{G}))$, we define $\phi_{\mathbf{w}}(\mathcal{G}) \equiv \phi_{w_k} \circ \cdots \circ \phi_{w_1}(\mathcal{G})$, and similarly $\phi_{\mathbf{w}}(q_V; \mathcal{G}) \equiv \phi_{w_k} \circ \cdots \circ \phi_{w_1}(q_V; \mathcal{G})$. We refer to such an ordering $\mathbf{w}$ as a *valid fixing sequence for* W.

Definition 34. *A CADMG* $\mathcal{G}(V, W)$ *is* reachable *from ADMG* $\mathcal{G}^*(V \cup W)$ *if there exists an ordering* $\mathbf{w}$ *of the vertices in* W *such that* $\mathcal{G} = \phi_{\mathbf{w}}(\mathcal{G}^*)$.

In words, a subgraph is reachable if, under some ordering, each of the vertices in W may be fixed, first in $\mathcal{G}^*$, and then in $\phi_{w_1}(\mathcal{G}^*)$, then in $\phi_{w_2}(\phi_{w_1}(\mathcal{G}^*))$, and so on. If a CADMG $\mathcal{G}(V, W)$ is reachable from $\mathcal{G}^*(V \cup W)$, we say that *the set* V *is reachable in* $\mathcal{G}^*$.

A key result of our paper, which we will show later, is that reachable CAD-MGs and their associated kernels are invariant with respect to any valid fixing sequence. It is not hard to see that if there are two valid fixing sequences $\mathbf{w}$ and $\mathbf{u}$ for W then $\phi_{\mathbf{w}}(\mathcal{G}) = \phi_{\mathbf{u}}(\mathcal{G})$. However, it requires more work to show that $\phi_{\mathbf{w}}(q_V; \mathcal{G}) = \phi_{\mathbf{u}}(q_V; \mathcal{G})$; see Theorem 38 below.

Proposition 35. *If* $\mathcal{G}^*$ *is a CADMG that is reachable from* $\mathcal{G}$, *and* $v \in \mathbb{V}(\mathcal{G}^*)$ *then* $\mathrm{pa}_{\mathcal{G}}(v) = \mathrm{pa}_{\mathcal{G}^*}(v)$.

Thus any vertex that is not fixed in a CADMG $\mathcal{G}^*$ that is reachable from $\mathcal{G}$ has the same parents that it had in $\mathcal{G}$.

3   Nested Markov Models

In this section we define a set of recursive Markov properties and a factorization, and show their equivalence. The models which obey these properties will be called 'nested' Markov models. 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. Note that the same graph may be reached by more than one sequence.

For all the following definitions, we will fix an ADMG $\mathcal{G}(V)$, a density $p(x_V)$, and a topological ordering $\prec$ for V.
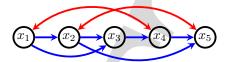
Figure 8: A graph where $\langle x_4, x_3, x_1 \rangle$ and $\langle x_3, x_4, x_1 \rangle$ are valid fixing sequences.

Definition 36. *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 $\phi_{\mathbf{w}^*}(\mathcal{G}) \equiv \mathcal{G}^*$.*

We denote the set of such distributions by $\mathcal{P}_m^n(\mathcal{G})$.

3.1 Invariance to the order of fixing in an ADMG

In this section we show that, given a distribution that obeys the nested Markov property with respect to an ADMG, any two valid fixing sequences that fix the same vertices will lead to the same reachable graph and kernel. For marginal distributions obtained from a hidden variable DAG this claim follows by results in (Tian and Pearl, 2002b). However, for distributions which obey the nested Markov property for an ADMG, but which are not derived from any hidden variable DAG, the claim is far less obvious. For instance in the ADMG in Fig. 8, the fixing sequence $\langle x_4, x_3, x_1 \rangle$, which leads to the kernel

$$q_{2,5}^1(x_2, x_5 \mid x_4, x_3, x_1) \equiv \frac{\sum_{x_3} p(x_5 \mid x_4, x_3, x_2, x_1) p(x_3, x_2, x_1)}{\sum_{x_3, x_2, x_5} p(x_5 \mid x_4, x_3, x_2, x_1) p(x_3, x_2, x_1)}$$

and the fixing sequence $\langle x_3, x_4, x_1 \rangle$ which leads to the kernel

$$q_{2,5}^2(x_2, x_5 \mid x_4, x_3, x_1) \equiv \frac{p(x_5 \mid x_4, x_3, x_2, x_1) p(x_2, x_1)}{\sum_{x_5, x_2} p(x_5 \mid x_4, x_3, x_2, x_1) p(x_2, x_1)}$$

are both valid, and these two kernels are therefore the same, in the context of our model. That this is so is not entirely obvious from inspecting these expressions. In addition, $q_{2,5}^1$ and $q_{2,5}^2$ are not functions of x_3 in our model; this is clear for $q_{2,5}^1$ since x_3 is summed out, but not so obvious for $q_{2,5}^2$.

35

Lemma 37. *Given a CADMG $\mathcal{G}(V, W)$ and a kernel q_V such that q_V is Markov w.r.t. $\mathcal{G}$, $r, s \in \mathbb{F}(\mathcal{G})$ and (a) $\phi_r(q_V; \mathcal{G})$ is Markov w.r.t. $\phi_r(\mathcal{G})$; (b) $\phi_s(q_V; \mathcal{G})$ is Markov w.r.t. $\phi_s(\mathcal{G})$ then*

$$
\begin{aligned}
\phi_r \circ \phi_s(\mathcal{G}) &= \phi_s \circ \phi_r(\mathcal{G}); \\
\phi_r \circ \phi_s(q_V; \mathcal{G}) &= \phi_s \circ \phi_r(q_V; \mathcal{G}).
\end{aligned}
$$

In words, if we have a choice to fix two vertices in $\mathcal{G}$ then the order in which we do this does not affect the resulting graph, or kernel, provided that the original kernel is Markov w.r.t. $\mathcal{G}$ and the kernels resulting from fixing r and s are, respectively, Markov w.r.t. $\phi_r(\mathcal{G})$ and $\phi_s(\mathcal{G})$.

Proof. That the resulting graphs are the same is immediate since ϕ_r removes edges into r, while ϕ_s removes edges into s.

To show that the resulting kernels are the same, we will show that if $r, s \in \mathbb{F}(\mathcal{G})$ then the product of the two divisors arising in (27) in $\phi_r(q_V(x_V \mid x_W); \mathcal{G})$ and $\phi_s((\phi_r(q_V(x_V \mid x_W); \mathcal{G})); \phi_r(\mathcal{G}))$, are the same as the product of the divisors in $\phi_s(q_V(x_V \mid x_W); \mathcal{G})$ and $\phi_r((\phi_s(q_V(x_V \mid x_W); \mathcal{G})); \phi_s(\mathcal{G}))$

Let $D^r \in \mathcal{D}(\mathcal{G})$, be the district containing r in $\mathcal{G}$. The divisor when fixing r is given by:

$$
q_V(x_r \mid x_{\mathrm{nd}_\mathcal{G}(r)}) = q_{D^r}(x_r \mid x_{\mathrm{mb}_\mathcal{G}(r)}), \tag{32}
$$

where q_{D^r} is given by (20).

Further, by (30), the resulting kernel is given by:

$$
\phi_r(q_V(x_V \mid x_W); \mathcal{G}) = q_{D^r}(x_{D^r \setminus \{r\}} \mid x_{\mathrm{pa}_\mathcal{G}(D^r) \setminus D^r}) \prod_{D \in \mathcal{D}(\mathcal{G}) \setminus \{D^r\}} q_D(x_D \mid x_{\mathrm{pa}_\mathcal{G}(D) \setminus D}). \tag{33}
$$

Here and in the remainder of the proof we use q_D, with $D \in \mathcal{D}(\mathcal{G})$, to refer to terms in the decomposition (29) associated with $\mathcal{G}$.

Set $\tilde{\mathcal{G}} = \phi_r(\mathcal{G})$. If $r, s \in \mathbb{F}(\mathcal{G})$ then either (i) $\mathrm{dis}_\mathcal{G}(r) = \mathrm{dis}_\mathcal{G}(s)$, but $r \notin \mathrm{de}_\mathcal{G}(s)$ and $s \notin \mathrm{de}_\mathcal{G}(r)$, or (ii) $\mathrm{dis}_\mathcal{G}(r) \neq \mathrm{dis}_\mathcal{G}(s)$. We now consider each case in turn:

(i) In this case, $s \in D^r$ since r and s are in the same district in $\mathcal{G}$. By

36

definition, the divisor when fixing s, having already fixed r, is given by:

$$(\phi_r(q_V))(x_s \mid x_{\mathrm{nd}_{\tilde{\mathcal{G}}}(s)}).$$

Now, $\{s\} \cup \mathrm{mb}_{\tilde{\mathcal{G}}}(s)$ is a subset of $\{s\} \cup \mathrm{mb}_{\mathcal{G}}(s) = D^r \cup \mathrm{pa}_{\mathcal{G}}(D^r)$, because fixing removes edges. Hence by Proposition 17(i)

$$\{s\} \cup \mathrm{mb}_{\tilde{\mathcal{G}}}(s) \subseteq (D^r \setminus \{r\}) \cup \mathrm{pa}_{\mathcal{G}}(D^r) = \mathrm{mb}_{\mathcal{G}}(s).$$

It follows from the Markov property for $\tilde{\mathcal{G}} = \phi_r(\mathcal{G})$, that

$$s \;\perp\!\!\!\perp\; ((D^r \setminus \{r,s\}) \cup \mathrm{pa}_{\mathcal{G}}(D^r)) \setminus \mathrm{mb}_{\tilde{\mathcal{G}}}(s) \mid \mathrm{mb}_{\tilde{\mathcal{G}}}(s) \quad [\phi_r(q_V)].$$

It then follows from (33) that:

$$(\phi_r(q_V))(x_s \mid x_{\mathrm{mb}_{\tilde{\mathcal{G}}}(s)}) = q_{D^r}(x_s \mid x_{\mathrm{mb}_{\mathcal{G}}(r)\setminus\{s\}}). \tag{34}$$

Thus the product of the two divisors (32) and (34) is: $q_{D^r}(x_{\{r,s\}} \mid x_{\mathrm{mb}_{\mathcal{G}}(r)\setminus\{s\}})$. Note that since, by hypothesis, r and s are in the same district in $\mathcal{G}$, this last expression is symmetric in r and s.

(ii) Let D^s be the district in $\mathcal{D}(\mathcal{G})$ that contains s. Since, by assumption, $D^s \neq D^r$, by Proposition 29 it follows that $s \in D^s \in \mathcal{D}(\tilde{\mathcal{G}})$. It then follows from (33) that

$$(\phi_r(q_V))(x_s \mid x_{\mathrm{mb}_{\tilde{\mathcal{G}}}(s)}) = q_{D^s}(x_s \mid x_{\mathrm{mb}_{\mathcal{G}}(s)}). \tag{35}$$

Thus the product of divisors is given by

$$q_{D^r}(x_r \mid x_{\mathrm{mb}_{\mathcal{G}}(r)}) \cdot q_{D^s}(x_s \mid x_{\mathrm{mb}_{\mathcal{G}}(s)}).$$

Hence in both cases, the product of the divisors is symmetric in r and s, and a symmetric argument shows that the same divisor is obtained when fixing s first, and r second. $\qquad \square$

Theorem 38. *Let $p(x_V)$ be a distribution that is nested Markov with respect to an ADMG $\mathcal{G}$. Let $\mathbf{u}, \mathbf{w}$ be different valid fixing sequences for the same set $W \subset V$. Then $\phi_{\mathbf{u}}(\mathcal{G}) = \phi_{\mathbf{w}}(\mathcal{G})$ and*

$$\phi_{\mathbf{u}}(p(x_V); \mathcal{G}) = \phi_{\mathbf{w}}(p(x_V); \mathcal{G}). \tag{36}$$

Due to this theorem our fixing operations $\phi_{\mathbf{w}^*}$, which were defined for a specific fixing sequence $\mathbf{w}^*$, can be defined purely in terms of the set W of nodes that were fixed; the order does not matter (provided that at least one valid fixing sequence exists).

Since we have shown all valid fixing sequences lead to the same graph and kernel, we will subscript the fixing operator ϕ by a set rather than a sequence. That is, we write $\phi_{V \setminus R}(\mathcal{G})$ and $\phi_{V \setminus R}(p(x_V); \mathcal{G})$ to mean 'fix all elements of $V \setminus R$ in $\mathcal{G}$ and $p(x_V)$'.

We will subsequently see that if we assume the existence of a latent variable DAG model (with observed variables $V \cup W$) that has latent projection $\mathcal{G}$, then if W is fixable, the kernel $\phi_W(p(x_V, x_W), \mathcal{G})$ can be interpreted as the intervention distribution $p(x_V \mid \mathrm{do}(x_W))$; see discussion following Lemma 58 below. In this context, a valid fixing sequence corresponds to a sequence of steps in the ID algorithm that identify this intervention distribution; see section 4.3. Consequently, were we to assume the existence of a DAG with latent variables, then the soundness of the ID algorithm would directly imply the equality (36). However, since we are *not* assuming such a DAG exists, $\phi_W(p(V \cup W))$ may not correspond to an intervention distribution and hence a separate proof is required; see Example 57 for an inequality constraint that is implied by the existence of a latent variable, but does not follow from the nested Markov property associated with the latent projection.

Proof. The proof is by induction on the cardinality of W. The base case is trivial. Suppose that the result holds for sets $|W^*| < |W|$. Let u_i, w_i denote the ith vertices in sequences $\mathbf{u}, \mathbf{w}$ respectively. Further, let k be smallest i such that $u_i \neq w_i$, and let $v \equiv u_k$, so that $\mathbf{u}$ and $\mathbf{w}$ agree in the first $k - 1$ fixing operations. By definition of k,

$$\phi_{\langle u_1, \ldots, u_{k-1} \rangle}(\mathcal{G}) = \phi_{\langle w_1, \ldots, w_{k-1} \rangle}(\mathcal{G}).$$

Since $\mathbf{u}, \mathbf{w}$ both contain the same vertices, there exists $l > k$ such that $w_l = v$.

Since, by hypothesis, $\mathbf{u}, \mathbf{w}$ are both valid fixing sequences, it follows that $v \in \mathbb{F}(\phi_{\langle w_1, \ldots, w_{k-1} \rangle}(\mathcal{G}))$. It further follows by Lemma 27 that

$$v \in \mathbb{F}(\phi_{\langle w_1, \ldots, w_{i-1} \rangle}(\mathcal{G})), \quad \text{for } k - 1 \leq i \leq l.$$

Then by Lemma 37, we have that:

$$
\begin{aligned}
\phi_{\langle w_1, \ldots, w_{l-1}, v = w_l \rangle}(\mathcal{G}) &= \phi_{\langle w_1, \ldots, w_{l-2}, v, w_{l-1} \rangle}(\mathcal{G}) \\
\phi_{\langle w_1, \ldots, w_{l-1}, v = w_l \rangle}(p(x_V); \mathcal{G}) &= \phi_{\langle w_1, \ldots, w_{l-2}, v, w_{l-1} \rangle}(p(x_V); \mathcal{G}).
\end{aligned}
$$

By further applications of Lemma 37, we may show that both the graphs and kernels resulting from the fixing sequences

$$\langle w_1, \ldots, w_{l-1}, v = w_l \rangle \quad \text{and} \quad \langle w_1, \ldots, w_{k-1}, v, w_k, \ldots, w_{l-1} \rangle.$$

are the same. It further follows that the whole sequence $\mathbf{w}$ leads to the same graph and kernel as $\langle w_1, \ldots, w_{k-1}, v, w_k \ldots, w_{l-1}, w_{l+1}, \ldots, w_{|W|} \rangle$. This latter sequence now agrees with $\mathbf{u}$ in the first k fixing operations. By repeating the argument we may thus show that $\mathbf{u}$ and $\mathbf{w}$ lead to the same graph and kernel. $\qquad \square$

3.2 Simplified Definitions

In light of Theorem 38 we may now restate the global nested Markov property more simply:

Definition 39. *We say that a distribution $p(x_V)$ obeys the* global nested Markov property *for $\mathcal{G}(V)$ if for all R reachable in $\mathcal{G}(V)$, $\phi_{V \setminus R}(p(x_V); \mathcal{G})$ obeys the global Markov property for $\phi_{V \setminus R}(\mathcal{G})$.*

Definition 40. *A set C 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 41. *For a set R reachable in $\mathcal{G}$, for a vertex $v \in R$, with $\mathrm{ch}_{\phi_{V \setminus R}(\mathcal{G})}(v) = \emptyset$, we define* the Markov blanket of v in R *to be:*

$$\mathrm{mb}_{\mathcal{G}}(v, R) \equiv \mathrm{mb}_{\phi_{V \setminus R}(\mathcal{G})}(v). \tag{37}$$

Since every ancestral set A is reachable in $\mathcal{G}$ this is a natural extension of our previous definition (17). We now give two alternative definitions of the nested Markov model.

3.3 Nested Factorization

Corollary 42. *$p(x_V)$ obeys the global nested Markov property with respect to $\mathcal{G}$ if and only if for every reachable R in $\mathcal{G}$,*

$$\phi_{V \setminus R}(p(x_V); \mathcal{G}) = \prod_{D \in \mathcal{D}(\phi_{V \setminus R}(\mathcal{G}))} \phi_{V \setminus D}(p(x_V); \mathcal{G}).$$

Proof. If $p(x_V)$ obeys the global nested Markov property, then by the equivalence of the CADMG factorization and CADMG global Markov properties, for every R reachable in $\mathcal{G}$,

$$q_R(x_R \mid x_{V \setminus R}) \equiv \phi_{V \setminus R}(p(x_V); \mathcal{G}) = \prod_{D \in \mathcal{D}(\phi_{V \setminus R}(\mathcal{G}))} q_D(x_D \mid x_{\mathrm{pa}(D) \setminus D}).$$

But each factor q_D is equal to $\phi_{R \setminus D}(q_R; \phi_{V \setminus R}(\mathcal{G}))$ by an inductive application of Proposition 32. By invariance of fixing and definition of q_R, we then have

$$\phi_{V \setminus R}(p(x_V); \mathcal{G}) = \prod_{D \in \mathcal{D}(\phi_{V \setminus R}(\mathcal{G}))} \phi_{V \setminus D}(p(x_V); \mathcal{G}).$$

The converse follows immediately by equivalence of the CADMG global property and CADMG factorization. $\qquad\square$

Note that this proof implies that in a nested Markov model in $\mathcal{G}$, every kernel corresponding to a reachable set is constructed by combining a subset of the kernels corresponding to $\mathcal{I}(\mathcal{G})$. We call this the nested factorization.

3.4 The Ordered Local Nested Property

Definition 43. *We say that $p(x_V)$ obeys the* ordered local nested Markov property *for $\mathcal{G}(V)$, and a topological ordering $\prec$ if for all $C \in \mathcal{I}(\mathcal{G})$, $\phi_{V \setminus C}(p(x_V); \mathcal{G})$ obeys the local Markov property for $\phi_{V \setminus C}(\mathcal{G})$ at $\max_\prec(C)$, the largest element of C according to $\prec$.*

Theorem 44. *$p(x_V)$ obeys the global nested Markov property with respect to $\mathcal{G}$ if and only if $p(x_V)$ obeys the ordered local nested Markov property for $\mathcal{G}$.*

Proof. By earlier results on CADMGs, $p(x_V)$ obeys the global nested Markov property if and only if for every reachable set R, $\phi_{V\setminus R}(p(x_V); \mathcal{G})$ obeys the local Markov property for $\phi_{V\setminus R}(\mathcal{G})$ at $\max_\prec(C)$, the largest element of R according to $\prec$. Since every element of $\mathcal{I}(\mathcal{G})$ is reachable, $p(x_V)$ obeys the ordered local nested Markov property for $\mathcal{G}(V)$.

To see the converse, fix R reachable in $\mathcal{G}$, with vertex v maximal in R according to $\prec$. Let D^v be the element in $\mathcal{D}(\phi_{V\setminus R}(\mathcal{G}))$ containing v. Then $D^v \in \mathcal{I}(\mathcal{G})$, and therefore, $v \perp\!\!\!\perp V \setminus \mathrm{mb}_\mathcal{G}(v, D^v) \mid \mathrm{mb}_\mathcal{G}(v, D^v)$ in $\phi_{V\setminus D^v}(p(x_V); \mathcal{G})$ is part of the ordered local Markov property for $\mathcal{G}$.

Then $v \perp\!\!\!\perp V \setminus \mathrm{mb}_\mathcal{G}(v, D^v) \mid \mathrm{mb}_\mathcal{G}(v, D^v)$ holds in $\phi_{V\setminus R}(p(x_V); \mathcal{G})$ by Proposition 14. Since $\mathrm{mb}_\mathcal{G}(v, D^v) = \mathrm{mb}_\mathcal{G}(v, R)$, $v \perp\!\!\!\perp V \setminus \mathrm{mb}_\mathcal{G}(v, R) \mid \mathrm{mb}_\mathcal{G}(v, R)$ holds in $\phi_{V\setminus R}(p(x_V); \mathcal{G})$. $\square$

3.5 Nested Markov models for complete graphs are saturated

It is known that any distribution is Markov relative to a complete DAG or ADMG. We now derive an equivalent result for the nested Markov case:

Theorem 45. *Let $\mathcal{G}$ be an ADMG. $\mathcal{P}^n(\mathcal{G})$ is saturated if and only if for every fixable $v \in \mathbb{F}(\mathcal{G}^*)$ in any reachable subgraph $\mathcal{G}^*$ of $\mathcal{G}$, we have*

$$\mathbb{V}(\mathcal{G}^*) \cap \mathrm{mb}_{\mathcal{G}^*}(v) = \mathbb{V}(\mathcal{G}^*) \cap \mathrm{nd}_{\mathcal{G}^*}(v). \tag{38}$$

Proof. Suppose that the condition holds for any sequence of fixings. We will show that any distribution satisfies the ordered local nested Markov property for $\mathcal{G}$. Pick a topological ordering $\prec$, and an intrinsic set C with maximal element t. Suppose we fix $V \setminus C$ to obtain $q_C(x_C \mid x_{\mathrm{pa}(C)\setminus C})$ and $\mathcal{G}^*$. The set $\mathrm{nd}_\mathcal{G}(t)$ is ancestral and contains C, so by Theorem 38 we can organize our fixing sequence to first marginalize all strict descendants of t; hence any independence involving $\mathrm{de}_\mathcal{G}(t) \setminus \{t\}$ in the kernel q_C is trivial.

41

Now, consider any other variable v which is fixed in $\mathcal{G}^*$. We claim that either v remains in the Markov blanket for t, or that it is completely marginalized from the graph, and hence any later independences involving it are trivial. Let $\tilde{\mathcal{G}}$ be the CADMG from which v is fixed, so that, in $\tilde{\mathcal{G}}$, v is fixable and t has no strict descendants. Suppose that $t \in \mathrm{mb}_{\tilde{\mathcal{G}}}(v)$; since t has no strict descendants, this means $t \in \mathrm{dis}_{\tilde{\mathcal{G}}}(v)$ and therefore

$$\mathbb{V}(\tilde{\mathcal{G}}) \cap (\mathrm{mb}_{\tilde{\mathcal{G}}}(v) \cup \{v\}) = \mathbb{V}(\tilde{\mathcal{G}}) \cap (\mathrm{mb}_{\tilde{\mathcal{G}}}(t) \cup \{t\}) = \mathbb{V}(\tilde{\mathcal{G}}) \cap (\mathrm{nd}_{\tilde{\mathcal{G}}}(t) \cup \{t\});$$

but, by assumption t has no strict descendants so v's Markov blanket includes all the random vertices in $\tilde{\mathcal{G}}$, and hence all their fixed parents. Therefore, fixing v is marginalizing and any subsequent independence statements involving it are trivial.

Otherwise v is a parent of $\mathrm{dis}_{\tilde{\mathcal{G}}}(t)$; in fact, we claim that for any reachable graph $\mathcal{G}^*$ containing $t \in \mathbb{V}(\mathcal{G}^*)$ we have $v \in \mathrm{mb}_{\mathcal{G}^*}(t)$. To see this, take the graph defined from $\tilde{\mathcal{G}}$ by fixing everything possible in t's district except for t itself (of course, we already know that $v \notin \mathrm{dis}_{\tilde{\mathcal{G}}}(t)$). By the assumed condition, v is in the Markov blanket for t, which means there is a path $v \to c \leftrightarrow \cdots \leftrightarrow t$ (where possibly $c = t$). What is more, any intrinsic set in any reachable graph that contains t must also contain all the vertices on this path (since we have fixed everything possible). Therefore this path is present in *any* reachable graph for which t is random, and so v is always in the Markov blanket of t.

The ordered local Markov property requires that $X_t \perp\!\!\!\perp X_{V \setminus (\mathrm{mb}_{\mathcal{G}^*}(t) \cup \{t\})} \mid X_{\mathrm{mb}_{\mathcal{G}^*}(t)} [q_C]$. We have established that all vertices are either in $\mathrm{mb}_{\mathcal{G}^*}(t)$, or are completely marginalized, and hence this statement is always true for any kernel q_C derived from any distribution via this sequence of fixings. Hence the ordered local nested Markov property for $\mathcal{G}$ is satisfied by any distribution.

For the converse, suppose that for some fixing sequence $w_1, \ldots, w_k = t$ the condition is not satisfied. Let v be a point on the sequence where it fails, so that in $\mathcal{G}^* \equiv \phi_{w_1,\ldots,w_{k-1}}(\mathcal{G})$ there is some random $v \in \mathrm{nd}_{\mathcal{G}^*}(t) \setminus \mathrm{mb}_{\mathcal{G}^*}(t)$. Let p be a distribution such that all variables are independent except for X_v and X_t. All the fixings to get to $q \equiv \phi_{w_1,\ldots,w_{k-1}}(p)$ are trivial because of the independences

42

and thus $p(x_t \mid x_v) = q(x_t \mid x_v)$. But now to satisfy the local Markov property for $\mathcal{G}$ it must be that $X_t \perp\!\!\!\perp X_v \mid X_{\mathrm{mb}_{\mathcal{G}^*}(t)}$, and by construction $q(x_t \mid x_{\mathrm{mb}_{\mathcal{G}^*}(t)}, x_v) = q(x_t \mid x_v)$ generally depends upon x_v. Hence p is not generally in $\mathcal{P}^n(\mathcal{G})$ and the model is not saturated. $\qquad\square$

Corollary 46. *Let $\mathcal{G}$ be a complete ADMG; then $\mathcal{P}^n(\mathcal{G})$ is saturated.*

Proof. We need to show that for any random vertex v in $\phi_W(\mathcal{G})$ the condition (38) holds.

Let $\mathcal{G}^* = \phi_W(\mathcal{G})$. Any random vertices $w \in \mathcal{G}^*$ share an edge with v. If $w \in \mathrm{nd}_{\mathcal{G}^*}(v)$ then that means either $w \leftrightarrow v$ or $w \to v$. In either case, $w \in \mathrm{mb}_{\mathcal{G}^*}(v)$. By definition, if $w \in \mathrm{de}_{\mathcal{G}^*}(v)$ then $w \notin \mathrm{mb}_{\mathcal{G}^*}(v)$; hence $\mathbb{V}(\mathcal{G}^*) \cap \mathrm{mb}_{\mathcal{G}^*}(v) = \mathbb{V}(\mathcal{G}^*) \cap \mathrm{nd}_{\mathcal{G}^*}(v)$. $\qquad\square$

4 Connections with Causal Inference

As discussed in the introduction, there is a close relationship between the fixing operation and interventions in causal inference. Graphical causal models are defined on DAGs, possibly with hidden variables. In this section we make the connection between nested Markov models, the fixing operation and interventions in graphical causal models more explicit.

4.1 Latent Variable DAG Models are in the Nested Markov Model

We first show that if $p(x_{H \cup O})$ is Markov relative to a DAG $\mathcal{G}(H \cup O)$, then $p(x_O)$ is in the nested Markov model of $\mathcal{G}(O)$.

Definition 47 (latent projection for CADMGs). *Let $\mathcal{G}(O \cup H, W)$ be a CADMG where W is a set of fixed vertices, the random vertices in O are observed, while those in H are latent. The* latent projection $\mathcal{G}(O, W)$ *is a directed mixed graph with vertex set O, where for every pair of distinct vertices $w \in O$, $v \in O \cup W$:*

(i) $\mathcal{G}(O, W)$ *contains an edge* $v \to w$ *if there is a directed path* $v \to \cdots \to w$ *on which every non-endpoint vertex is in* H.

(ii) $\mathcal{G}(O, W)$ *contains an edge* $v \leftrightarrow w$ *if there exists a path between* v *and* w *such that the non-endpoints are all non-colliders in* H, *and such that the edge adjacent to* v *and the edge adjacent to* w *both have arrowheads at those vertices. For example,* $v \leftrightarrow \cdots \to w$.

We denote the operation of creating a latent projection of a CADMG $\mathcal{G}(H \cup O, W)$ onto the subset O as σ_H. That is $\sigma_H(\mathcal{G}(H \cup O, W)) = \mathcal{G}(O, W)$.

Lemma 48. $\mathcal{G}(O, W)^{|W} = \sigma_H(\mathcal{G}(H \cup O, W)^{|W})$.

Corollary 49. *Let* $\mathcal{G}(H \cup O, W)$ *be a CADMG. The m-separations in* $\mathcal{G}(H \cup O, W)^{|W}$ *amongst vertices in* $O \cup W$ *are the same as those in* $\mathcal{G}(O, W)^{|W}$.

Proof. Both graphs are ADMGs, and the former is a latent projection of the latter by Lemma 48. The result follows by standard results on m-connection. $\square$

We will call a CADMG which does not contain bidirected arrows a conditional acyclic directed graph (CADG). It is a corollary of the definition of $\mathcal{P}_f^c$ that if $\mathcal{G}(V, W)$ is a CADG, then $q_V(x_V \mid x_W) \in \mathcal{P}_f^c(\mathcal{G})$ if

$$q_V(x_V \mid x_W) = \prod_{a \in V} q_V(x_a \mid x_{\mathrm{pa}_{\mathcal{G}}(a)})$$

Lemma 50. *Let* $\mathcal{G}$ *be a DAG with a vertex set* V. *Then every non-empty subset* S *of* V *is reachable, and if* $p(x_V) \in \mathcal{P}^d(\mathcal{G})$,

$$\phi_{V \setminus S}(p(x_V); \mathcal{G}) = q_S(x_S \mid x_{pa_{\mathcal{G}}(S) \setminus S}) = \prod_{a \in S} p(x_a \mid x_{\mathrm{pa}_{\mathcal{G}}(a)}).$$

In other words, $\phi_{V \setminus S}(p(x_V); \mathcal{G}) \in \mathcal{P}^c(\phi_{V \setminus S}(\mathcal{G}))$.

($\mathcal{P}^d(\mathcal{G})$ is defined following Definition 1.) An easy case of the main result of this section, which occurs when H is empty, is now an immediate corollary.

44

Corollary 51. *Let $\mathcal{G}$ be a DAG with a vertex set V. Then*

$$p(x_V) \in \mathcal{P}^d(\mathcal{G}) \Leftrightarrow p(x_V) \in \mathcal{P}^n(\mathcal{G}).$$

Proof: ($\Rightarrow$) follows by Lemma 50. ($\Leftarrow$) follows by Corollary 42 (with $R = V$) since $\mathcal{D}(\mathcal{G}) = \{\{v\}; v \in V\}$ and $\phi_{V \setminus \{v\}}(p(x_V); \mathcal{G}) = p(x_v \mid x_{\mathrm{pa}_\mathcal{G}(v)})$. $\qquad\square$

Lemma 52. *Let $\mathcal{G}(H \cup O, W)$ be a CADG. Assume $v \in O$ is fixable in $\mathcal{G}(O, W) = \sigma_H(\mathcal{G}(H \cup O, W))$. Then $\sigma_H(\phi_v(\mathcal{G}(H \cup O, W))) = \phi_v(\sigma_H(\mathcal{G}(H \cup O, W)))$.[1] That is, the following commutative diagram holds:*

$$
\begin{array}{ccc}
\mathcal{G}(H \cup O, W) & \xrightarrow{\quad\sigma_H\quad} & \mathcal{G}(O, W) \\[2pt]
\phi_v \downarrow & & \downarrow \phi_v \\[2pt]
\mathcal{G}^\dagger((H \cup O) \setminus \{v\}, W \cup \{v\}) & \xrightarrow{\quad\sigma_H\quad} & \mathcal{G}^\dagger(O \setminus \{v\}, W \cup \{v\})
\end{array}
$$

Proof. Both $\sigma_H(\phi_v(\mathcal{G}(H \cup O, W)))$ and $\phi_v(\sigma_H(\mathcal{G}(H \cup O, W)))$ have the same set of random vertices $O \setminus \{v\}$ and fixed vertices $W \cup \{v\}$.

Consider the set of edges E in $\sigma_H(\mathcal{G}(H \cup O, W)) = \mathcal{G}(O, W)$. The set of edges E' in $\phi_v(\sigma_H(\mathcal{G}(H \cup O, W)))$ is a subset of E containing all edges not having an arrowhead at v. Now let π be the set of paths in $\mathcal{G}(H \cup O, W)$, where both endpoints are in $O \cup W$ and all non-endpoints are non-colliders in H. These paths d-connect marginally (given $\emptyset$). Similarly, let π' be the set of paths in $\phi_v(\mathcal{G}(H \cup O, W)) = \mathcal{G}^\dagger((H \cup O) \setminus \{v\}, W \cup \{v\})$, where both endpoints are in $O \cup W$ and all non-endpoints are non-colliders in H. π' is the subset of π formed by removing paths containing an edge with an arrowhead at v (note that since $v \notin H$, v can only occur as an endpoint).

By definition of latent projections, there is a bijection that associates each edge e in E, with a subset of paths in π with the same endpoints as e, and the

---

[1] Note that v is fixable in $\mathcal{G}(H \cup O, W)$ since this graph has no bidirected edges, and thus all random vertices are fixable.

same starting and ending orientations as e. These subsets partition π. Applying ϕ_v to $\mathcal{G}(H \cup O, W)$ means that only those paths in π' are left in the resulting graph. Paths in π' are only in subsets associated with edges in E' (by the bijection). Applying σ_H to the graph then results in the edge set E'. This establishes our conclusion. $\qquad \square$

In fact, the proof of this lemma did not rely on the vertex v being fixable, only on the specific way edges are removed by ϕ. We give a more general version of this lemma, useful for deriving properties of causal models, which we discuss later in section 4.2. To do this, we extend the graphical fixing operation ϕ to apply not only to elements in $\mathbb{F}(\mathcal{G})$, but to all vertices in V. We denote this extension of ϕ by ϕ^*. In other words, ϕ_v^* is defined just as ϕ_v, but does not require that $v \in \mathbb{F}(\mathcal{G})$. As with ϕ, for any $r, s \in V$, $\phi_r^* \circ \phi_s^*(\mathcal{G}) = \phi_s^* \circ \phi_r^*(\mathcal{G})$, and so for any $Z \subseteq V$, we define $\phi_Z^*(\mathcal{G})$ as a composition of applications of ϕ^* to elements of Z in $\mathcal{G}$ under any order.

Corollary 53. *Let $\mathcal{G}(H \cup O, W)$ be a CADG. Then for any $v \in O$, $\sigma_H(\phi_v^*(\mathcal{G}(H \cup O, W))) = \phi_v^*(\sigma_H(\mathcal{G}(H \cup O, W)))$.*

Corollary 53 is essentially equivalent to Proposition 8 of Evans (2016).

Lemma 54. *Assume $q_{H \cup O}(x_{H \cup O} \mid x_W)$ is in $\mathcal{P}^c(\mathcal{G}(H \cup O, W))$ for a CADG $\mathcal{G}(H \cup O, W)$. Then*

$$q_{H \cup O}(x_O \mid x_W) = \prod_{D \in \mathcal{D}(\mathcal{G}(O, W))} \left(\sum_{x_{H_D}} \prod_{a \in D \cup H_D} q_{H \cup O}(x_a \mid x_{\mathrm{pa}_{\mathcal{G}(H \cup O, W)}(a)}) \right)$$

$$= \prod_{D \in \mathcal{D}(\mathcal{G}(O, W))} \left(\prod_{a \in D} q_{H \cup O}(x_a \mid x_{\mathrm{pre}_{\prec, \mathcal{G}(O, W)}(a)}) \right)$$

where $H_D = \mathrm{an}_{\mathcal{G}(H \cup O, W)_{D \cup H}}(D) \cap H$, and $\prec$ is any topological ordering for $\mathcal{G}(O, W)$.

Proof. This is a simple extension of the proof for $W = \emptyset$ found in (Tian and Pearl, 2002b). $\qquad \square$

46

Lemma 55. *Let $\mathcal{G}(H \cup O, W)$ be a CADG, and assume $q_{H \cup O}(x_{H \cup O} \mid x_W) \in \mathcal{P}_f^c(\mathcal{G}(H \cup O, W))$. Assume $v \in O$ is fixable in $\mathcal{G}(O, W) = \sigma_H(\mathcal{G}(H \cup O, W))$. Then*

$$\sum_{x_H} \phi_v(q_{H \cup O}(x_{H \cup O} \mid x_W); \mathcal{G}(H \cup O, W)) = \phi_v(q_{H \cup O}(x_O \mid x_W); \sigma_H(\mathcal{G}(H \cup O, W)))$$

That is, the following commutative diagram holds:

$$
\begin{array}{ccc}
q_{H \cup O}(x_{H \cup O} \mid x_W) & \xrightarrow{\quad \sum_{x_H} \quad} & q_{H \cup O}(x_O \mid x_W) \\[2pt]
\Big\downarrow{\scriptstyle \phi_v(.; \mathcal{G}(H \cup O, W))} & & \Big\downarrow{\scriptstyle \phi_v(.; \mathcal{G}(O, W))} \\[2pt]
q^{\dagger}_{(H \cup O) \setminus \{v\}}(x_{(H \cup O) \setminus \{v\}} \mid x_{W \cup \{v\}}) & \xrightarrow[\quad \sum_{x_H} \quad]{} & q^{\dagger}_{O \setminus \{v\}}(x_{(H \cup O) \setminus \{v\}} \mid x_{W \cup \{v\}})
\end{array}
$$

Robins (1986) proves a similar result that he calls the 'collapse of the g-formula'.

Proof:

$\phi_v \to \Sigma_{x_H}$ direction:

Since $q_{H \cup O}(x_{H \cup O} \mid x_W) \in \mathcal{P}_f^c(\mathcal{G}(H \cup O, W))$, we have

$$q_{H \cup O}(x_{H \cup O} \mid x_W) = \prod_{a \in H \cup O} q_{H \cup O}(x_a \mid x_{\mathrm{pa}_{\mathcal{G}}(a)}).$$

This implies by Lemma 50 that

$$\phi_v(q_{H \cup O}(x_{H \cup O} \mid x_W); \mathcal{G}(H \cup O, W)) = \prod_{a \in (H \cup O) \setminus \{v\}} q_{H \cup O}(x_a \mid x_{\mathrm{pa}_{\mathcal{G}}(a)}),$$

which implies $\phi_v(q_{H \cup O}(x_{H \cup O} \mid x_W); \mathcal{G}(H \cup O, W)) \in \mathcal{P}_f^c(\phi_v(\mathcal{G}(H \cup O, W)))$.

Then by Lemma 54,

$$\sum_{x_H} \phi_v(q_{H \cup O}(x_{H \cup O} \mid x_W); \mathcal{G}(H \cup O, W)) = \prod_{D \in \mathcal{D}(\phi_v(\mathcal{G}(O, W)))} \left(\prod_{a \in D} q_{H \cup O}(x_a \mid x_{\mathrm{pre}_{\prec, \mathcal{G}}(a)}) \right).$$

$\Sigma_{x_H} \to \phi_v$ direction:

Similarly, by Lemma 54,

$$\sum_{x_H} q_{H \cup O}(x_{H \cup O} \mid x_W) = q_{H \cup O}(x_O \mid x_W) = \prod_{D \in \mathcal{D}(\mathcal{G}(O, W))} \left(\prod_{a \in D} q_{H \cup O}(x_a \mid x_{\mathrm{pre}_{\prec, \mathcal{G}}(a)}) \right)$$

Now let D^v be the element of $\mathcal{D}(\mathcal{G}(O, W))$ such that $v \in D^v$. Then by Proposition 32,

$$\phi_v(q_{H \cup O}(x_O \mid x_W); \mathcal{G}(O, W)) = \left(\prod_{a \in D^v \setminus \{v\}} q_{H \cup O}(x_a \mid x_{\mathrm{pre}_{\prec, \mathcal{G}}(a)}) \right) \cdot$$
$$\prod_{D \in \mathcal{D}(\mathcal{G}(O, W)) \setminus \{D^v\}} \left(\prod_{a \in D} q_{H \cup O}(x_a \mid x_{\mathrm{pre}_{\prec, \mathcal{G}}(a)}) \right)$$

$\square$

We now have enough to prove the main result of this section.

Theorem 56.

$$p(x_{O \cup H}) \in \mathcal{P}^d(\mathcal{G}(O \cup H)) \Rightarrow p(x_O) \in \mathcal{P}^n(\mathcal{G}(O)).$$

Proof. Assume $p(x_{O \cup H}) \in \mathcal{P}^d(\mathcal{G}(O \cup H))$, and for a set R reachable in $\mathcal{G}(O)$ with $A \subseteq R$ and $B, C \subseteq O$ (C possibly empty), suppose that A is m-separated from B given C in $\phi_{O \setminus R}(\mathcal{G}(O, W))^{|O \setminus R}$.

By an inductive application of Lemma 52, $\phi_{O \setminus R}(\mathcal{G}(O, W))$ is a latent projection of $\phi_{(H \cup O) \setminus R}(\mathcal{G}(H \cup O, W))$. Therefore, by Corollary 49, A is m-separated from B given C in $\phi_{(H \cup O) \setminus R}(\mathcal{G}(H \cup O, W))^{|(H \cup O) \setminus R}$. Our assumption, and Corollary 51 then imply

$$(A \perp\!\!\!\perp B \mid C)_{[\phi_{(H \cup O) \setminus R}(p(x_{H \cup O}); \mathcal{G}(H \cup O))]}$$

holds. By an inductive application of Lemma 55,

$$\sum_{x_H} \phi_{(H \cup O) \setminus R}(p(x_{H \cup O}); \mathcal{G}(H \cup O)) = \phi_{O \setminus R}(p(x_O); \mathcal{G}(O))$$

and thus

$$(A \perp\!\!\!\perp B \mid C)_{[\phi_{O \setminus R}(p(x_O); \mathcal{G}(O))]}$$

also holds. Our conclusion follows. $\square$

Note that the converse of the above theorem is not true in general. There exist distributions $p(x_O) \in \mathcal{P}^n(\mathcal{G}(O))$ for which there exists no joint distribution

$p(x_{O \cup H}) \in \mathcal{P}^d(\mathcal{G}(O \cup H))$. This is because marginals of hidden variable DAGs may induce additional *inequality constraints* which are not satisfied by all elements of the associated nested Markov model. The best known of these are the *instrumental inequalities* of Pearl (1995), which were generalized by Evans (2012). However, Evans (2015) shows that the nested Markov model accounts for all *equality* constraints that arise without making assumptions about the state-space of the latent variables.

Example 57. *Consider variables $X_0, \ldots, X_4$ under a distribution which is Markov with respect to the graph in Figure 1(i). Then the marginal distribution over $X_1, \ldots, X_4$ satisfies the nested Markov property with respect to the graph in Figure 3(i). However, if the observed variables are binary (and regardless of the state-space of X_0) their marginal distribution also satisfies the following inequality constraints not implied by the nested Markov property:*

$$0 \leq \Big\{ q_{24}(0_2 \mid x_1) + q_{24}(0_4 \mid x_3) + q_{24}(0_2, 0_4 \mid 1 - x_1, 1 - x_3)$$
$$- q_{24}(0_2, 0_4 \mid x_1, x_3) - q_{24}(0_2, 0_4 \mid x_1, 1 - x_3) - q_{24}(0_2, 0_4 \mid 1 - x_1, x_3) \Big\} \leq 1$$

for each $x_1, x_3 \in \{0, 1\}$; here, for example, 0_2 is used as a shorthand for the event $\{X_2 = 0\}$. These are related to the CHSH inequalities of Clauser et al. (1969), and are sometimes referred to as Bell inequalities *after Bell (1964).*

4.2 Causal Model of a DAG

The statistical model of a DAG $\mathcal{G}$ with vertices V, described earlier, is a set of distributions $p(x_V)$ defined by the factorization (5) formulated on $\mathcal{G}$.

We define a causal model of a DAG $\mathcal{G}$ by a set of similar factorizations that yield joint distributions under an *intervention* operation, which corresponds to setting values of variables from outside the system. Interventions can be used to formalize causal effects in both observational studies and randomized controlled trials.

Specifically, for a DAG $\mathcal{G}$ with vertices V, and any $A \subseteq V$, the causal model

for $\mathcal{G}$ defines

$$p(x_{V \setminus A} \mid \operatorname{do}_{\mathcal{G}}(x_A)) = \prod_{v \in V \setminus A} p(x_v \mid x_{\operatorname{pa}_{\mathcal{G}}(v)}). \tag{39}$$

This is known as the *g-formula, truncated factorization, or manipulated distribution.* Note that since for any DAG $\mathcal{G}$, $\mathcal{P}^d(\mathcal{G}) = \mathcal{P}^n(\mathcal{G})$, we have

$$p(x_{V \setminus A} \mid \operatorname{do}_{\mathcal{G}}(x_A)) = \prod_{v \in V \setminus A} p(x_v \mid x_{\operatorname{pa}_{\mathcal{G}}(v)}) = \phi_{V \setminus A}(p(x_V); \mathcal{G}). \tag{40}$$

We will drop the $\mathcal{G}$ subscript from do(.) when the graph is obvious.

4.3 Re-Formulation of the ID Algorithm via Fixing

Identification of causal effects is a more difficult problem in causal models where some variables are unobserved. In particular, not every distribution $p(x_y \mid \operatorname{do}(x_A))$ is identified in every $\mathcal{G}(H \cup V)$. The goal is thus to find an identifying functional for $p(x_Y \mid \operatorname{do}(x_A))$ in terms of the observed marginal distribution $p(x_V)$ obtained from $p(x_{H \cup V})$ which lies in a causal model of $\mathcal{G}(H \cup V)$, *or* to show that no such functional exists.

The problem may be formalized by considering a latent projection ADMG $\mathcal{G}(V)$ in place of the original causal DAG with hidden variables, $\mathcal{G}(H \cup V)$. A well-known "folklore" result in causal inference states that any two hidden variable DAGs $\mathcal{G}^1(H^1 \cup V)$ and $\mathcal{G}^2(H^2 \cup V)$ with the same latent projection $\mathcal{G}(V)$ will share all identifying functionals, and so there is no loss of generality in reasoning on $\mathcal{G}(V)$. We give a proof of this folklore result later in this section.

A complete algorithm for this problem—the ID algorithm, stated on ADMGs—was given in (Shpitser and Pearl, 2006, Tian and Pearl, 2002b). 'Complete' means that whenever the algorithm fails to find an expression for $p(x_Y \mid do(x_A))$ in terms of $p(x_V)$ in the causal model given by $\mathcal{G}(H \cup V)$, no other algorithm is able to do so without making more assumptions. In this section we use the fixing operation to give a simple constructive characterization (an algorithm) of all causal effects identifiable by the ID algorithm, and thus of all causal effects identifiable in any hidden variable causal DAG $\mathcal{G}(H \cup V)$. We can view this

characterization as using the fixing operation to simplify the ID algorithm from its original two page formulation down to a single line.

Lemma 58. *Let $\mathcal{G}(H \cup V)$ be a hidden variable causal DAG. Then for any set S reachable from $\mathcal{G}(V)$, $p(x_S \mid \mathrm{do}_{\mathcal{G}(H \cup V)}(x_{V \setminus S}))$ is identifiable in $\mathcal{G}(H \cup V)$ from $p(x_V)$ as the kernel $\phi_{V \setminus S}(p(x_V); \mathcal{G}(V))$ evaluated at x_S and $x_{\mathrm{pa}_{\mathcal{G}(V)}(S) \setminus S}$.*

Proof. Our conclusion follows by (40) and an inductive application of Lemma 55. The fact that the kernel $\phi_{V \setminus S}(p(x_V); \mathcal{G})$ only depends on values of S and $\mathrm{pa}_{\mathcal{G}}(S) \setminus S$ follows by the global nested Markov property, and Theorem 56. $\square$

Lemma 59. *Let $\mathcal{G}(H \cup V)$ be a hidden variable causal DAG. For any $A \subseteq V$, let $\mathcal{G}_{\overline{A}}$ be an edge subgraph of $\mathcal{G}$ where all directed arrows in $\mathcal{G}$ into A are removed. For any $Y \subseteq V \setminus A$, let $A_Y = \mathrm{an}_{\mathcal{G}_{\overline{A}}}(Y) \cap A$. Then*

$$p(x_Y \mid \mathrm{do}_{\mathcal{G}}(x_A)) = p(x_Y \mid \mathrm{do}_{\mathcal{G}}(x_{A_Y}))$$

Proof. Follows by (39), and the global Markov property for CDAG models. $\square$

Theorem 60. *Let $\mathcal{G}(H \cup V)$ be a causal DAG with latent projection $\mathcal{G}(V)$. For $A \mathbin{\dot{\cup}} Y \subset V$, let $Y^* = \mathrm{an}_{\mathcal{G}(V)_{V \setminus A}}(Y)$. Then if $\mathcal{D}(\mathcal{G}(V)_{Y^*}) \subseteq \mathcal{I}(\mathcal{G}(V))$,*

$$p(x_Y \mid \mathrm{do}_{\mathcal{G}(H \cup V)}(x_A)) = \sum_{x_{Y^* \setminus Y}} \prod_{D \in \mathcal{D}(\mathcal{G}(V)_{Y^*})} \phi_{V \setminus D}(p(x_V); \mathcal{G}(V)). \qquad (41)$$

If not, there exists $D \in \mathcal{D}(\mathcal{G}(V)_{Y^*}) \setminus \mathcal{I}(\mathcal{G}(V))$ and $p(x_Y \mid \mathrm{do}_{\mathcal{G}(H \cup V)}(x_A))$ is not identifiable in $\mathcal{G}(H \cup V)$.

Note that Y^* is the set of vertices $v \in V \setminus A$ for which, for some $y \in Y$, there is a directed path $v \to \cdots \to y$, with no vertex on the path in A. Also note that since, by Theorem 56, $X_D \perp\!\!\!\perp X_{V \setminus (D \cup \mathrm{pa}_{\mathcal{G}(V)}(D))} \mid X_{\mathrm{pa}_{\mathcal{G}(V)}(D) \setminus D}$ in $\phi_{V \setminus D}(p(x_V); \mathcal{G}(V))$, it follows that $\phi_{V \setminus D}(p(x_V); \mathcal{G}(V))$ is a function solely of x_D and $x_{\mathrm{pa}_{\mathcal{G}(V)}(D) \setminus D}$. Thus the product on the RHS of (41) is a function of the 'input' variables on the LHS, x_Y, x_A, and the 'bound' variables x_{Y^*} present in the sum.

Proof. We first prove (41). Let $A^* = V \setminus Y^* \supseteq A$. By Lemma 59, $p(x_{Y^*} \mid \text{do}_{\mathcal{G}(H \cup V)}(x_A)) = p(x_{Y^*} \mid \text{do}_{\mathcal{G}(H \cup V)}(x_{A^*}))$.

Let $\mathcal{G}^*(H \cup (V \setminus A^*), A^*) = \phi_{A^*}(\mathcal{G}(H \cup V))$; note that since $\mathcal{G}(H \cup V)$ is a DAG, A^* is fixable in $\mathcal{G}(H \cup V)$, and $\mathcal{G}^*(H \cup (V \setminus A^*), A^*)$ is a CADG. By Corollary 53, $\sigma_H(\phi_{A^*}^*(\mathcal{G}(H \cup V))) = \phi_{A^*}^*(\sigma_H(\mathcal{G}(H \cup V)))$, where σ_H is the latent projection operation, that is $\sigma_H(\mathcal{G}(H \cup V)) = \mathcal{G}(V)$. Since $\mathcal{G}^*(Y^*, A^*) = \sigma_H(\phi_{A^*}^*(\mathcal{G}(H \cup V)))$, and, by definition of induced subgraphs, $\mathcal{G}(V)_{Y^*} = (\phi_{A^*}^*(\mathcal{G}(V)))_{Y^*}$, $\mathcal{G}(V)_{Y^*} = \mathcal{G}^*(Y^*, A^*)_{Y^*}$. Consequently, $\mathcal{D}(\mathcal{G}(V)_{Y^*}) = \mathcal{D}(\mathcal{G}^*(Y^*, A^*))$.

For every $D \in \mathcal{D}(\mathcal{G}^*(Y^*, A^*))$, define $H_D \equiv H \cap \text{an}_{\mathcal{G}(H \cup V)_{D \cup H}}(D)$, and let $H^* = \bigcup_{D \in \mathcal{D}(\mathcal{G}^*(Y^*, A^*))} H_D$. Thus H_D is the set of variables $h \in H$, for which there exists a vertex $d \in D$ and a directed path $h \to \cdots \to d$ in $\mathcal{G}(H \cup V)$ on which, excepting d, all vertices are in H.

It follows from the construction that:

(a) if $D, D' \in \mathcal{D}(\mathcal{G}^*(Y^*, A^*))$, and $D \neq D'$ then $H_D \cap H_{D'} = \emptyset$;

(b) for each $D \in \mathcal{D}(\mathcal{G}^*(Y^*, A^*))$ we have $\text{pa}_{\mathcal{G}(H \cup V)}(D \cup H_D) \cap H^* = H_D$;

(c) $Y^* \cup H^*$ is ancestral in $\mathcal{G}(H \cup V)$, so if $v \in Y^* \cup H^*$, $\text{pa}_{\mathcal{G}(H \cup V)}(v) \cap H \subseteq H^*$.

We now have:

$$
\begin{aligned}
&p(x_{Y^*} \mid \text{do}_{\mathcal{G}(H \cup V)}(x_{A^*})) \\
&= \sum_{x_H} \prod_{v \in H \cup Y^*} p(x_v \mid x_{\text{pa}_{\mathcal{G}(H \cup V)}(v)}) \\
&= \sum_{x_{H^*}} \prod_{v \in H^* \cup Y^*} p(x_v \mid x_{\text{pa}_{\mathcal{G}(H \cup V)}(v)}) \underbrace{\sum_{x_{H \setminus H^*}} \prod_{v \in H \setminus H^*} p(x_v \mid x_{\text{pa}_{\mathcal{G}(H \cup V)}(v)})}_{=1} \\
&= \sum_{x_{H^*}} \prod_{D \in \mathcal{D}(\mathcal{G}^*(Y^*, A^*))} \prod_{v \in D \cup H_D} p(x_v \mid x_{\text{pa}_{\mathcal{G}(H \cup V)}(v)}) \\
&= \prod_{D \in \mathcal{D}(\mathcal{G}^*(Y^*, A^*))} \left(\sum_{x_{H_D}} \prod_{v \in D \cup H_D} p(x_v \mid x_{\text{pa}_{\mathcal{G}(H \cup V)}(v)}) \right). \quad (42)
\end{aligned}
$$

Here, the first equality follows from (40), the second follows from (c), the third

52

from (a), and the fourth from (b). Now, for any given $D \in \mathcal{D}(\mathcal{G}^*(Y^*, A^*))$,

$$\sum_{x_{H_D}} \prod_{v \in D \cup H_D} p(x_v \mid x_{\mathrm{pa}_{\mathcal{G}(H \cup V)}(v)})$$

$$= \sum_{x_{H_D}} \prod_{v \in D \cup H_D} p(x_v \mid x_{\mathrm{pa}_{\mathcal{G}(H \cup V)}(v)}) \underbrace{\sum_{x_{H \setminus H_D}} \prod_{v \in H \setminus H_D} p(x_v \mid x_{\mathrm{pa}_{\mathcal{G}(H \cup V)}(v)})}_{=1}$$

$$= \sum_{x_H} \prod_{v \in D \cup H} p(x_v \mid x_{\mathrm{pa}_{\mathcal{G}(H \cup V)}(v)})$$

$$= \sum_{x_H} \phi_{V \setminus D}(p(x_{H \cup V}); \mathcal{G}(H \cup V)). \tag{43}$$

Here the second line uses that $\mathrm{pa}_{\mathcal{G}(H \cup V)}(D \cup H_D) \cap (H \setminus H_D) = \emptyset$, which follows (b), (c) and the definition of H_D. Since, by hypothesis, $D \in \mathcal{D}(\mathcal{G}(V)_{Y^*}) = \mathcal{D}(\mathcal{G}^*(Y^*, A^*)) \subseteq \mathcal{I}(\mathcal{G}(V))$, it follows from Lemma 55 that

$$\sum_{x_H} \phi_{V \setminus D}(p(x_{H \cup V}); \mathcal{G}(H \cup V)) = \phi_{V \setminus D}(p(x_V); \mathcal{G}(V)). \tag{44}$$

Hence by (42), (43) and (44),

$$p(x_{Y^*} \mid \mathrm{do}_{\mathcal{G}(H \cup V)}(x_{A^*})) = \prod_{D \in \mathcal{D}(\mathcal{G}^*(Y^*, A^*))} \phi_{V \setminus D}(p(x_V); \mathcal{G}(V)).$$

The conclusion, (41), then follows since

$$p(x_Y \mid \mathrm{do}_{\mathcal{G}(H \cup V)}(x_A)) = \sum_{x_{Y^* \setminus Y}} p(x_{Y^*} \mid \mathrm{do}_{\mathcal{G}(H \cup V)}(x_{A^*})).$$

To establish the last claim, fix $D \in \mathcal{D}(\mathcal{G}(V)_{Y^*}) \setminus \mathcal{I}(\mathcal{G}(V))$, and let D^* be the minimal intrinsic superset of D. Assume, for a contraction, that D^* does not intersect A^*. Then $D^* \subseteq Y^*$. But since D^* is intrinsic, it must be a subset of some $D' \in \mathcal{D}(\mathcal{G}(V)_{Y^*})$. But this is impossible since $D \subset D^*$, and $D \in \mathcal{D}(\mathcal{G}(V)_{Y^*})$. Thus D^* intersects A^*.

Let $R = \{v \in D^* \mid \mathrm{ch}_{\phi_{V \setminus D^*}(\mathcal{G}(V))}(v) = \emptyset\}$. If R is not a subset of D, D^* could not be the minimal intrinsic superset of D, since any element in $R \setminus D$ is fixable in $\phi_{V \setminus D^*}(\mathcal{G}(V))$. Finally, note that by construction, $D \subsetneq D^*$, $D \cap A^* = \emptyset$, and $R \subseteq Y^* \subseteq \mathrm{an}_{\mathcal{G}(V)_{\overline{A}}}(Y)$.

This implies D and D^* satisfy the definition of a *hedge* for $p(x_Y \mid \mathrm{do}_{\mathcal{G}(H \cup V)}(x_{A^*})) = p(x_Y \mid \mathrm{do}_{\mathcal{G}(H \cup V)}(x_A))$ in $\mathcal{G}(V)$ (Shpitser and Pearl, 2007). Results in the same

53

paper show that $p(x_Y \mid \mathrm{do}_{\mathcal{G}(H \cup V)}(x_A))$ is not identifiable in a canonical hidden variable causal DAG $\mathcal{G}^\dagger(H^\dagger \cup V)$, where $H^\dagger$ consists of a hidden variable for every bidirected arc in $\mathcal{G}(V)$ (see, e.g. Richardson and Spirtes (2002) §6 for a formal definition of $\mathcal{G}^\dagger(H^\dagger \cup V)$). It follows from Theorem 2 in (Evans, 2015) that the model associated with $\mathcal{G}^\dagger(H^\dagger \cup V)$ is a submodel of that associated with $\mathcal{G}(H \cup V)$. This completes the proof. $\qquad\square$

Corollary 61. *Let $\mathcal{G}^1(H^1 \cup V)$ and $\mathcal{G}^2(H^2 \cup V)$ be two causal DAGs, with the same latent projection, so $\mathcal{G}^1(V) = \mathcal{G}^2(V)$. Then for any $A \dot\cup Y \subseteq V$,*

- *$p(Y \mid \mathrm{do}_{\mathcal{G}^1}(A))$ is identifiable if and only if $p(Y \mid \mathrm{do}_{\mathcal{G}^2}(A))$ is identifiable,*

- *If $p(Y \mid \mathrm{do}_{\mathcal{G}^1}(A))$ is identifiable, $p(Y \mid \mathrm{do}_{\mathcal{G}^1}(A)) = p(Y \mid \mathrm{do}_{\mathcal{G}^2}(A))$.*

Proof. Follows directly by Theorem 60, since Y^*, $\mathcal{D}(\mathcal{G}(V)_{Y^*})$, $\mathcal{I}(\mathcal{G}(V))$ and the terms on the RHS of (41) are defined solely in terms of the latent projection. $\quad\square$

Corollary 61 shows that we may discuss identification directly on ADMG latent projections without having to specify the particular hidden variable DAG.

Example 62. *Given some hidden variable DAG $\mathcal{G}(O \cup H)$, where $O = \{x_1, \ldots, x_4\}$ with latent projection $\mathcal{G}(O)$ given by the ADMG in Fig. 3(i), consider the problem of identifying $p(x_4 \mid \mathrm{do}_{\mathcal{G}}(x_2))$. Mapping this problem to the notation of Theorem 60, we have $Y = \{X_4\}$, $A = A^* = \{X_2\}$, $Y^* = \{X_4, X_3, X_1\}$. The districts of $\mathcal{G}_{Y^*}$ are $\{X_4\}$, $\{X_3\}$, and $\{X_1\}$. In fact, these three sets are intrinsic in $\mathcal{G}$, and thus a fixing sequence exists for each corresponding kernel:*

$$\phi_{234}(p(x_1, x_2, x_3, x_4); \mathcal{G}) = \phi_{23}\left(\frac{p(x_1, x_2, x_3, x_4)}{p(x_4 \mid x_3, x_2, x_1)}; \mathcal{G}_{\{1,2,3\}} \right)$$

$$= \phi_2\left(\frac{p(x_1, x_2, x_3)}{p(x_3 \mid x_2, x_1)}; \mathcal{G}_{\{1,2\}} \right)$$

$$= \frac{p(x_1, x_2)}{p(x_2 \mid x_1)} = p(x_1),$$

$$\phi_{124}(p(x_1, x_2, x_3, x_4); \mathcal{G}) = \phi_{12}\left(\frac{p(x_1, x_2, x_3, x_4)}{p(x_4 \mid x_3, x_2, x_1)}; \mathcal{G}_{\{1,2,3\}}\right)$$

$$= \phi_1\left(\frac{p(x_3, x_2, x_1)}{p(x_2 \mid x_1)}; \phi_2(\mathcal{G}_{\{1,2,3\}})\right)$$

$$= \frac{p(x_3 \mid x_2, x_1)p(x_1)}{p(x_1)} = p(x_3 \mid x_2, x_1),$$

$$\phi_{231}(p(x_1, x_2, x_3, x_4); \mathcal{G}) = \phi_{23}\left(\frac{p(x_1, x_2, x_3, x_4)}{p(x_1)} = p(x_2, x_3, x_4 \mid x_1); \phi_1(\mathcal{G})\right)$$

$$= \phi_2\left(\frac{p(x_2, x_3, x_4 \mid x_1)}{p(x_3 \mid x_2, x_1)} \equiv q^\dagger(x_2, x_4 \mid x_1, x_3); \phi_{31}(\mathcal{G})\right)$$

$$= \frac{q^\dagger(x_2, x_4 \mid x_1, x_3)}{q^\dagger(x_2 \mid x_1, x_3, x_4)}$$

$$= \sum_{x_2} p(x_4 \mid x_3, x_2, x_1)p(x_2 \mid x_1).$$

The last step here follows because $q^\dagger(x_2, x_4 \mid x_1, x_3) = p(x_2 \mid x_1)p(x_4 \mid x_1, x_2, x_3)$, and $q^\dagger(x_2 \mid x_1, x_3, x_4) = q^\dagger(x_2, x_4 \mid x_1, x_3) / \left(\sum_{x_2} q^\dagger(x_2, x_4 \mid x_1, x_3)\right)$. Combining these kernels as in Theorem [60] yields the same identifying functional as the one obtained by the ID algorithm applied to $\mathcal{G}$:

$$p(x_4 \mid \text{do}_{\mathcal{G}}(x_2)) = \sum_{x_3, x_1} p(x_1)p(x_3 \mid x_2, x_1) \sum_{x_2'} p(x_4 \mid x_3, x_2', x_1)p(x_2' \mid x_1),$$

where we relabel x_2 as x_2' in the last kernel to avoid confusion between a free and summation quantifier captured versions of the variable x_2 in the final expression.

4.4 Connections with Tian's Constraint Algorithm

An algorithm for enumerating constraints on kernels in marginals of DAG models was given in (Tian and Pearl, 2002b). Tian's algorithm effectively implements fixing for both graphs and kernels, with three important differences from our formalization. First, unlike CADMGs, subgraphs obtained by fixing in (Tian and Pearl, 2002b) do not show fixed nodes explicitly. Second, there is no unified fixing operation on kernels, instead the algorithm in (Tian and Pearl, 2002b) alternates steps corresponding to the application of the g-formula (division by a conditional density), and steps corresponding to marginalization.

Third, the kernel objects obtained by repeated alternative application of these two steps, called "q-factors" and written as $Q[V]$ where V is the set of nodes not yet fixed, do not explicitly show the dependence on nodes already fixed.

For a given DAG $\mathcal{G}(O \cup H)$ and a density $p(x_{O \cup H})$ Markov relative to $\mathcal{G}(O \cup H)$, a subset of observable nodes O, and a topological order $\prec$ on $\mathcal{G}$, this algorithm gives a list of constraints of the form "a kernel corresponding to a q-factor $Q[C]$ obtained by some set of applications of the g-formula and marginalization on $p(x_O)$ does not functionally depend on a set X_D, for some $D \subseteq O$."

In this section we will show that the set of constraints found by Tian's algorithm implicitly define the nested Markov model. To facilitate the proof of this result we translate Tian's algorithm into the notation used in this manuscript based on CADMGs and kernels. The result is Algorithm 1 with a subroutine Algorithm 2. In subsequent discussions, we will call the algorithm in (Tian and Pearl, 2002b) "Tian's algorithm," and our translation "Algorithm 1."

Whereas Tian's algorithm takes a latent variable DAG as input, Algorithm 1 takes a latent projection $\mathcal{G}(O)$; both algorithms require a topological order $\prec$ on O. Given this input, Algorithm 1 constructs a list $\mathbf{L}$ of constraints, that is initially empty (line 2).

The first step of Tian's algorithm considers constraints for every $v \in O$ among its predecessors $T = \mathrm{pre}_{\mathcal{G}(O), \prec}(v)$. The constraint named in their step (A1) is that v is independent of variables outside its Markov blanket in $\mathcal{G}(T)$, conditional on its Markov blanket.

The second step (A2) is recursive, and considers constraints involving v in various q-factors obtained from the q-factor $Q[S] = q_S$. Algorithm 1 implements (A2) by means of a recursive subroutine Algorithm 2, called in line 9. Note that Algorithm 2 accepts a variable, a CADMG and a kernel, while step (A2) in Tian's algorithm accepts a q-factor $Q[S]$, and an associated hidden variable DAG $\mathcal{G}$ with observable nodes S. Since step (A2) considers districts and observable subsets of S closed under descendants, Lemmas 52, and Lemma 63 and 64 below

56

imply that we can dispense with any mention of hidden variable DAGs and q-factors, and instead rephrase (A2) in terms of CADMGs, kernels and the fixing operation.

Step (A2) first iterates over all observable subsets D of S closed under descendants[2], and considers a constraint on the $S \setminus D$ margin of $Q[S]$; i.e. $Q[S \setminus D] = \sum_{x_D} Q[S]$. Note that though $v \in S$, the constraint is actually stated in terms of the margin $Q[S \setminus D]$. Lemma 63 shows that this iteration corresponds to an iteration over ancestral subsets $D' = S \setminus D$ of the corresponding CADMG containing v (shown in line 3 of Algorithm 2). Tian and Pearl (2002b) describe this part of the algorithm in terms of "effective parents" in the original DAG; these are just the parents in the latent projection.

The constraint associated with D' found by step (A2) is described by noting that some parents of D may not be parents of D', and therefore the q-factor $q_{D'} = Q[D'] = \sum_{x_D} Q[S]$ is independent of these missing parents. That is to say, $q_S(x_{D'} \mid x_{\mathrm{pa}_{\mathcal{G}}(S) \setminus S})$ does not depend upon $x_{(\mathrm{pa}_{\mathcal{G}}(S) \setminus S) \setminus (\mathrm{pa}_{\mathcal{G}}(D') \setminus D')}$, or equivalently $x_{(\mathrm{pa}_{\mathcal{G}}(S) \setminus S) \setminus \mathrm{pa}_{\mathcal{G}}(D')}$.

Translated into a statement of conditional independence on kernels we have that $X_{D'}$ is independent of $X_{(\mathrm{pa}_{\mathcal{G}}(S) \setminus S) \setminus \mathrm{pa}_{\mathcal{G}}(D')}$ (conditional on $X_{\mathrm{pa}_{\mathcal{G}}(D') \setminus D'}$) in $q_S(x_{D'} \mid x_W)$. This constraint is added on line 6 of Algorithm 2, with the conditional statement on line 5 checking that the independence is not vacuous.

Step (A2) of Tian's algorithm potentially adds another constraint associated with the set D'. We consider $Q[D']$ associated with the subgraph $\phi_{O \setminus D'}(\mathcal{G}(O))$; if this graph has more than one district and $v \in E \in \mathcal{D}(\phi_{O \setminus D'}(\mathcal{G}(O)))$, then $Q[D'] / \sum_{x_v} Q[D']$ is a function only of $x_{\mathrm{mb}(v,E)}$. In our notation and using Lemma 64 it is clear that

$$Q[D'] / \sum_{x_v} Q[D'] = q_{D'}(x_v \mid x_{(D' \cup \mathrm{pa}_{\mathcal{G}}(D')) \setminus \{v\}}) = q_{D'}(x_v \mid x_{(E \cup \mathrm{pa}_{\mathcal{G}}(E)) \setminus \{v\}}).$$

Thus the constraint added by line 10 of Algorithm 2 is therefore

$$X_v \perp\!\!\!\perp X_{(D' \cup \mathrm{pa}_{\mathcal{G}}(D')) \setminus (\mathrm{mb}_{\mathcal{G}}(v,E) \cup \{v\})} \mid X_{\mathrm{mb}_{\mathcal{G}}(v,E)} \, [q_S],$$

---

[2]Tian and Pearl (2002b) use D for sets that are closed under descendants, *not* to indicate a district.

by definition of $\text{mb}_{\mathcal{G}}(v, E)$. Finally, Tian's algorithm is called recursively with v, $Q[E]$ and the corresponding subgraph of $\mathcal{G}$. In our notation, given that the previous invocation was with v, $q_S(x_S \mid x_W)$ and the corresponding CADMG $\mathcal{G}(S, W)$, the new invocation is with v, $\phi_{S \setminus E}(q_S(x_S \mid W); \mathcal{G}(S, W))$, and $\phi_{S \setminus E}(\mathcal{G}(S, W))$. This invocation is done on line 12 of Algorithm 2.

The next two lemmas justify our use of fixing and kernels in our translation of Tian's algorithm into Algorithms 1 and 2.

Lemma 63. *Let $\mathcal{G}'(H' \cup S)$ be a DAG obtained at some step of Tian's algorithm from the original DAG $\mathcal{G}(H \cup O)$, where O and S are the sets of observable vertices in $\mathcal{G}$ and $\mathcal{G}'$, respectively. Then:*

(a) *S is reachable in $\mathcal{G}(O)$.*

(b) *For any subset $D \subseteq S$ closed under descendants in $\mathcal{G}'(S)$, $S \setminus D$ is ancestral in $\phi_{O \setminus S}(\mathcal{G}(O))$.*

(c) *A district E in $\mathcal{G}'(S)$ is a district in $\phi_{O \setminus S}(\mathcal{G}(O))$.*

Proof. We prove this by induction on the recursive structure of Tian's algorithm.

We first establish that (a) holds in the base case. Note that when Tian's algorithm calls step (A2) from step (A1), it is with sets $S_i \equiv S$ which are districts in $\mathcal{G}(T)$, where each T is a subset of O that is ancestral in $\mathcal{G}(O)$. Thus, these sets S are reachable by definition.

Assume we are in step (A2) with a reachable set S. Then for any subset $D \subseteq S$ closed under descendants in $\mathcal{G}'(S)$, $S \setminus D$ is ancestral in $\mathcal{G}'(S)$. By Lemma 52, $\mathcal{G}'(S) = \phi_{O \setminus S}(\mathcal{G}(O))$, which implies (b). All subsequent recursive calls in step (A2) correspond to districts E_i in $\mathcal{G}'(S \setminus D)$, for some D above. By Lemma 52, all such E_i are districts of $\phi_{O \setminus (S \setminus D)}(\mathcal{G}(O))$ which establishes (c). It follows from (b) and (c) that any set arising in recursive applications of (A2) are reachable from $\mathcal{G}(O)$, thus establishing (a). $\qquad\square$

The next result inductively establishes a correspondence between q-factors and kernels.

58

Algorithm 1 The constraint-finding algorithm in (Tian and Pearl, 2002b) expressed in the CADMG and kernel notation used in this manuscript. v is an element of O.

Input : $\mathcal{G}(O)$: an ADMG over a vertex set O,

$\qquad p(x_O)$: a density over x_O,

$\qquad \prec$: a total topological ordering on O.

Output : A list of constraints on $p(x_O)$ implied by $\mathcal{G}(O)$.

1: **procedure** FIND-CONSTRAINTS($\mathcal{G}(O), p(x_O), \prec$)

2: $\qquad \mathbf{L} \leftarrow \{\}$

3: $\qquad$ **for all** $v \in O$ **do**

4: $\qquad\qquad T \leftarrow \mathrm{pre}_{\mathcal{G}(O), \prec}(v) \cup \{v\}$;

5: $\qquad\qquad$ **Let** $S \in \mathcal{D}(\mathcal{G}(T))$ s.t. $v \in S$;

6: $\qquad\qquad$ **if** $T \setminus (\mathrm{mb}(v, T) \cup \{v\}) \neq \emptyset$ **then**

7: $\qquad\qquad\qquad \mathbf{L} \leftarrow \mathbf{L} \cup$ "$(X_v \perp\!\!\!\perp X_{T \setminus (\mathrm{mb}(v,T) \cup \{v\})} \mid X_{\mathrm{mb}(v,T)}) [q_T]$ ";

8: $\qquad\qquad$ **end if**

9: $\qquad\qquad \mathbf{L} \leftarrow \mathbf{L} \cup$ NODE-CONSTRAINTS($v, \phi_{T \setminus S}(\mathcal{G}(T)), \phi_{T \setminus S}(p(x_T); \mathcal{G}(T))$);

10: $\qquad$ **end for**

11: $\qquad$ **return L**.

12: **end procedure**

Algorithm 2 A subroutine of Algorithm 1 which finds constraints associated with a particular vertex.

Input : $\mathcal{G}(S, W)$: a CADMG,

$\quad\quad\quad\quad v$ a vertex in S with no children in $\mathcal{G}$,

$\quad\quad\quad\quad q_S(x_S \mid x_W)$: a kernel associated with $\mathcal{G}$,

Output : A list of constraints on $q(x_S \mid x_W)$.

1: **procedure** NODE-CONSTRAINTS$(v, \mathcal{G}(S, W), q_S)$

2: $\quad\quad \mathbf{L} \leftarrow \{\}$;

3: $\quad\quad$ **for** every $\emptyset \subset D \subset S$ closed under descendants in $\mathcal{G}$, s.t. $v \notin D$ **do**

4: $\quad\quad\quad\quad$ Let $D' \leftarrow S \setminus D$;

5: $\quad\quad\quad\quad$ **if** $(\mathrm{pa}_{\mathcal{G}}(S) \setminus S) \setminus \mathrm{pa}_{\mathcal{G}}(D') \neq \emptyset$ **then**

6: $\quad\quad\quad\quad\quad\quad \mathbf{L} \leftarrow \mathbf{L} \cup$ "$(X_{D'} \perp\!\!\!\perp X_{(\mathrm{pa}_{\mathcal{G}}(S) \setminus S) \setminus \mathrm{pa}_{\mathcal{G}}(D')} \mid X_{\mathrm{pa}_{\mathcal{G}}(D') \setminus D'}) [q_{D'}]$";

7: $\quad\quad\quad\quad$ **end if**

8: $\quad\quad\quad\quad$ Let $E \in \mathcal{D}(\phi_{S \setminus D'}(\mathcal{G}))$, s.t. $v \in E$;

9: $\quad\quad\quad\quad$ **if** $|\mathcal{D}(\phi_{S \setminus D'}(\mathcal{G}))| > 1$ **and** $(D' \cup \mathrm{pa}_{\mathcal{G}}(D')) \setminus (\mathrm{mb}_{\mathcal{G}}(v, E) \cup \{v\}) \neq \emptyset$

$\quad\quad\quad$ **then**

10: $\quad\quad\quad\quad\quad\quad \mathbf{L} \leftarrow \mathbf{L} \cup$ "$(X_v \perp\!\!\!\perp X_{(D' \cup \mathrm{pa}_{\mathcal{G}}(D')) \setminus (\mathrm{mb}_{\mathcal{G}}(v, E) \cup \{v\})} \mid X_{\mathrm{mb}_{\mathcal{G}}(v, E)}) [q_{D'}]$";

11: $\quad\quad\quad\quad$ **end if**

12: $\quad\quad\quad\quad \mathbf{L} \leftarrow \mathbf{L} \cup$ NODE-CONSTRAINTS$(v, \phi_{S \setminus E}(\mathcal{G}), \phi_{S \setminus E}(q_S; \mathcal{G}))$;

13: $\quad\quad$ **end for**

14: $\quad\quad$ **return L**

15: **end procedure**

Lemma 64. *Let $Q[S]$ be a q-factor obtained from $p(x_V)$ by Tian's algorithm. Assume $Q[S]$ is equal to a kernel $q_S(x_S \mid x_W)$, and corresponds to the DAG $\mathcal{G}'(H' \cup S)$ obtained from the original DAG $\mathcal{G}(H \cup O)$. Then:*

(a) *For any subset $D \subseteq S$ closed under descendants in $\mathcal{G}'$, $\sum_{x_D} Q[S] = \phi_D(q_S(x_S \mid x_W); \phi_{O \setminus S}(\mathcal{G}(O)))$.*

(b) *For any district E in $\mathcal{G}'$, the q-factor $Q[E] = \phi_{S \setminus E}(q_S(x_S \mid x_W); \phi_{O \setminus S}(\mathcal{G}(O)))$.*

Proof. This follows inductively by definition of fixing on kernels, and Lemma 1 and Lemma 2 in (Tian and Pearl, 2002b). $\qquad\square$

We are now ready to prove the main result of this section. For an ADMG $\mathcal{G}(O)$, let $\mathcal{P}_t(\mathcal{G}, O, \prec)$ be the set of densities $p(x_O)$ in which the list of constraints found by Algorithm 1 holds. We will show that this statistical model is identical to the nested Markov model $\mathcal{P}^n(\mathcal{G}(O))$.

Theorem 65. *Let $\mathcal{G}(O)$ be an ADMG over vertex set O. Then*

$$\mathcal{P}^n_m(\mathcal{G}(O)) \subseteq \mathcal{P}_t(\mathcal{G}, O, \prec).$$

Proof. It suffices to show that every constraint found by Algorithms 1 and 2, if given a graph $\mathcal{G}(O)$ as one of the inputs, is implied by some constraint given by the global nested Markov property for $\mathcal{G}(O)$.

All constraints found in Algorithm 1 on line 6 are ordinary conditional independence constraints. Moreover, they are easily seen to follow from the m-separation criterion, which forms a part of the global nested Markov property (since the sets of nodes T in which these constraint are found are all reachable in $\mathcal{G}(O)$).

Consider some D' obtained during some recursive call of Algorithm 2. By Lemma 63, D' is ancestral in a CADMG corresponding to a set reachable in $\mathcal{G}(O)$. Therefore D' is itself reachable, so the nested global Markov property implies that the kernel $q_{D'}(x_{D'} \mid x_W) = \phi_{O \setminus D'}(p(x_O); \mathcal{G}(O))$ is Markov with respect to $\phi_{O \setminus D'}(\mathcal{G}(O))$.

61

If $(\mathrm{pa}_{\mathcal{G}}(S)\backslash S)\backslash(\mathrm{pa}_{\mathcal{G}}(D')\backslash D')$ is non-empty, D' is m-separated from $(\mathrm{pa}_{\mathcal{G}}(S)\backslash S) \setminus (\mathrm{pa}_{\mathcal{G}}(D') \setminus D')$ by $\mathrm{pa}_{\mathcal{G}}(D') \setminus D'$ in the graph $\mathcal{G}(D',W)^{|W}$ obtained from $\mathcal{G}(D',W) = \phi_{O\backslash D'}(\mathcal{G})$ in the usual way (here $W = O \setminus D'$). This implies that D' is m-separated from the smaller set $(\mathrm{pa}_{\mathcal{G}}(S) \setminus S) \setminus (D \cup \mathrm{pa}_{\mathcal{G}}(D') \setminus D')$ by $\mathrm{pa}_{\mathcal{G}}(D') \setminus D'$ in $\mathcal{G}(D',W)^{|W}$, which is precisely the constraint on line 6.

Similarly, if the preconditions on line 9 hold, v is m-separated from a non-empty set $\mathrm{mb}_{\mathcal{G}}(v,D')\backslash\mathrm{mb}_{\mathcal{G}}(v,E)$ given $\mathrm{mb}_{\mathcal{G}}(v,E)$ in $\mathcal{G}(D',W)^{|W}$. This directly implies the constraint on line 10. □

Before showing the other direction, we need to show that Algorithm 2 reaches all intrinsic sets. Let $\mathcal{G}(O)$ be an ADMG. Denote by $\mathcal{T}(\mathcal{G}(O))$ the set of subsets of O such that the corresponding CADMG and kernel are arguments to some call of Algorithm 2, if Algorithm 1 is invoked with $\mathcal{G}(O)$.

Lemma 66. *Let $\mathcal{G}(O)$ be an ADMG. Then*

$$\mathcal{T}(\mathcal{G}(O)) = \mathcal{I}(\mathcal{G}(O)).$$

Proof. $\mathcal{T}(\mathcal{G}(O)) \subseteq \mathcal{I}(\mathcal{G}(O))$ follows from Lemma 63 and the fact that every S associated with arguments $\mathcal{G}(S,W)$ and $q_S(x_S \mid x_W)$ formed a district in the graph associated with the called subroutine.

To show $\mathcal{I}(\mathcal{G}(O)) \subseteq \mathcal{T}(\mathcal{G}(O))$, let S be intrinsic and assume for contradiction that the algorithm never reaches S. Since S is bidirected-connected this means that Algorithm 2 is called for some $S' \supset S$ but that there is no strict ancestral subset D' of S' in $\mathcal{G}(S',W)$ which contains S. Since it was called from Algorithm 2, S' is a single district, and since no ancestral subset contains S it is the case that every $d \in S'$ is an ancestor of some element of S. But then the only fixable vertices in S' are also in S; this contradicts the reachability of S. □

We now prove the main result of this section.

Theorem 67. *Let $\mathcal{G}(O)$ be an ADMG with a vertex set O. Then*

$$\mathcal{P}_t(\mathcal{G},O,\prec) \subseteq \mathcal{P}_l^n(\mathcal{G}(O),\prec).$$

Proof. The ordered local nested property for $\mathcal{G}(O)$ and $\prec$ has at most a single constraint for each $S \in \mathcal{I}(\mathcal{G}(O))$, involving the $\prec$-maximal element v of S. This constraint is that

$$X_v \perp\!\!\!\perp X_{O \setminus (\mathrm{mb}(v,S) \cup \{v\})} \mid X_{\mathrm{mb}(v,S)} \; [q_S]. \tag{45}$$

We will show that all such constraints are implied by those found by Algorithms 1 and 2 by a double induction on the sequence of calls made by the algorithm. In the outer induction, Algorithm 1 is called in $\prec$-order on subgraphs $\mathcal{G}(\{v\} \cup \mathrm{pre}_{\mathcal{G},\prec}(v))$, so (letting $T = \mathrm{pre}_{\mathcal{G},\prec}(v)$) we can assume by the induction hypothesis that $p(x_T)$ satisfies the local nested Markov property for $\mathcal{G}(T)$. The base case for this is trivial, since any distribution satisfies the local nested Markov property for a graph with one vertex. Throughout the proof we will ignore trivial independences of the form $X_A \perp\!\!\!\perp X_\emptyset \mid X_C$.

For the second, inner induction, we work on the sequence of calls made within one iteration of the 'do' routine from lines 4-9 of Algorithm 1, for a particular $v \in O$ and $T = \{v\} \cup \mathrm{pre}_{\mathcal{G},\prec}(v)$. The base case is the constraint from line 7 associated with the intrinsic set $S \in \mathcal{D}(\mathcal{G}(T))$ containing v. The independence is $X_v \perp\!\!\!\perp X_{T \setminus (\mathrm{mb}(v,T) \cup \{v\})} \mid X_{\mathrm{mb}(v,T)}$ in q_T, which also holds in p since T is an ancestral margin of $\mathcal{G}$ and all the variables in the independence are contained in T.

Since $\mathrm{mb}(v,T) = \mathrm{mb}(v,S)$, we conclude that $X_v \perp\!\!\!\perp X_{T \setminus (\mathrm{mb}(v,S) \cup \{v\})} \mid X_{\mathrm{mb}(v,S)}$ in q_T. Since $O \setminus (\mathrm{mb}(v,T) \cup \{v\})$ is partitioned into $T \setminus (\mathrm{mb}(v,T) \cup \{v\})$ and $O \setminus T$, an inductive application of Proposition 13 implies that

$$X_v \perp\!\!\!\perp X_{O \setminus (\mathrm{mb}(v,S) \cup \{v\})} \mid X_{\mathrm{mb}(v,S)} \quad [q_T]; \tag{46}$$

note that q_T does not depend on the vertices in $O \setminus T$ as these are fixed and occur after T under $\prec$.

An inductive application of Proposition 14 implies (46) holds not only in q_T, but also in $q_S \equiv \phi_{T \setminus S}(q_T; \mathcal{G}(T))$. This establishes the base case, namely that for an intrinsic set $S \in \mathcal{D}(\mathcal{G}(T))$, (45) is implied by constraints found by Algorithms 1 and 2.

63

We now consider the inductive case, namely (45) for all intrinsic sets which are not districts in $\mathcal{G}(T)$ for any v. We know by Lemma 66 that all such intrinsic sets are visited by Algorithm 2. Consider a set $E \in \mathcal{I}(\mathcal{G}(O))$ such that v is the $\prec$-greatest element of E, and let S^* be the intrinsic set whose own recursive call made the recursive call corresponding to E. By the inductive hypothesis (45) is implied for S^* by constraints found by Algorithms 1 and 2. We now claim that

$$X_v \perp\!\!\!\perp X_{\mathrm{mb}(v,S^*) \setminus \mathrm{mb}(v,E)} \mid X_{\mathrm{mb}(v,E)} \quad [q_E] \tag{47}$$

is sufficient for the local nested Markov constraint (45) applied to the intrinsic set E.

Proof of claim: Let D be the set considered in the recursive call of Algorithm 2 corresponding to S, where E is the district in $\phi_{O \setminus (S \setminus D)}(\mathcal{G})$ containing v. Recall that, by construction, v has no children in S. Note that if $p(x_O) \in \mathcal{P}^n(\mathcal{G})$, then

$$X_D \perp\!\!\!\perp X_{O \setminus (\mathrm{mb}(v,S) \cup \{v\})} \mid X_{\mathrm{mb}(v,S) \setminus D} \quad [q_{S \setminus \{v\}}] \tag{48}$$

holds, since the corresponding m-separation statement holds in $\tilde{\mathcal{G}}^{|W}$, where $\tilde{\mathcal{G}} = \phi_{O \setminus (S \setminus \{v\})}(\mathcal{G})$.

In fact, since D, $S \setminus \{v\}$ and $\mathrm{mb}(v,S)$ are subsets of $T \setminus \{v\}$, it suffices to establish $p(x_{T \setminus \{v\}}) \in \mathcal{P}^n(\mathcal{G}(T \setminus \{v\}))$ to obtain

$$X_D \perp\!\!\!\perp X_{T \setminus (\mathrm{mb}(v,S) \cup \{v\})} \mid X_{\mathrm{mb}(v,S) \setminus D} \quad [q_{S \setminus \{v\}}],$$

from which we can conclude (48) by an inductive application of Lemma 11.

By the outer induction hypothesis we have already shown that $p(x_{T \setminus \{v\}}) \in \mathcal{P}^n(\mathcal{G}(T \setminus \{v\}))$. By the inner induction hypothesis we also have

$$X_v \perp\!\!\!\perp X_{O \setminus (\mathrm{mb}(v,S) \cup \{v\})} \mid X_{\mathrm{mb}(v,S)} \quad [q_S]. \tag{49}$$

We know v is the $\prec$-maximal element of S, so $q_{S \setminus \{v\}}$ is an ordinary margin of q_S, and by Proposition 13 we can use the graphoid axiom

64

of contraction with (48) and (49) to obtain

$$X_{\{v\}\cup D} \perp\!\!\!\perp X_{O\setminus(\mathrm{mb}(v,S)\cup\{v\})} \mid X_{\mathrm{mb}(v,S)\setminus D} \quad [q_S]$$

$$\implies \quad X_v \perp\!\!\!\perp X_{O\setminus(\mathrm{mb}(v,S)\cup\{v\})} \mid X_{\mathrm{mb}(v,S)\setminus D}. \quad [q_S]. \qquad (50)$$

An inductive application of Proposition 13 implies

$$X_v \perp\!\!\!\perp X_{O\setminus((\mathrm{mb}(v,S)\cup\{v\})\setminus D)} \mid X_{\mathrm{mb}(v,S)\setminus D} \qquad [q_{S\setminus D}]$$

$$\implies \quad X_v \perp\!\!\!\perp X_{O\setminus(\mathrm{mb}(v,S)\cup\{v\})} \mid X_{\mathrm{mb}(v,S)} \qquad [q_{S\setminus D}]$$

by the graphoid axiom of decomposition. We can now use Proposition 14 inductively to further fix $(S\setminus D)\setminus E$ and conclude that

$$X_v \perp\!\!\!\perp X_{O\setminus(\mathrm{mb}(v,S)\cup\{v\})} \mid X_{\mathrm{mb}(v,S)} \quad [q_E]. \qquad (51)$$

Finally, we use the graphoid axiom of contraction to conclude from (47) and (51) that

$$X_v \perp\!\!\!\perp X_{O\setminus(\mathrm{mb}(v,E)\cup\{v\})} \mid X_{\mathrm{mb}(v,E)} \quad [q_E]. \qquad (52)$$

holds and the claim is proved.

All that remains is to show that (47) is implied by Algorithms 1 and 2. Let $D' = S^*\setminus D$ be the non-empty set found on line 4, such that $E \in \mathcal{D}(\phi_{O\setminus(S\setminus D)}(\mathcal{G}))$. The constraints added on lines 6 and 10 are:

$$X_{D'} \perp\!\!\!\perp X_{(\mathrm{pa}_\mathcal{G}(S)\setminus S)\setminus \mathrm{pa}_\mathcal{G}(D')} \mid X_{\mathrm{pa}_\mathcal{G}(D')\setminus D'} \quad [q_{D'}],$$

$$X_v \perp\!\!\!\perp X_{(D'\cup \mathrm{pa}_\mathcal{G}(D'))\setminus(\mathrm{mb}_\mathcal{G}(v,E)\cup\{v\})} \mid X_{\mathrm{mb}_\mathcal{G}(v,E)} \quad [q_{D'}]. \qquad (53)$$

By the axioms of weak union and then contraction with (53) we have

$$X_v \perp\!\!\!\perp X_{(\mathrm{pa}_\mathcal{G}(S)\setminus S)\setminus \mathrm{pa}_\mathcal{G}(D')} \mid X_{(D'\cup \mathrm{pa}(D'))\setminus\{v\}} \quad [q_{D'}],$$

$$X_v \perp\!\!\!\perp X_{[(\mathrm{pa}_\mathcal{G}(S)\setminus S)\cup D'\cup \mathrm{pa}_\mathcal{G}(D')]\setminus(\mathrm{mb}_\mathcal{G}(v,E)\cup\{v\})} \mid X_{\mathrm{mb}_\mathcal{G}(v,E)} \quad [q_{D'}].$$

By an application of Lemma 11,

$$X_v \perp\!\!\!\perp X_D \mid X_{(\mathrm{pa}_\mathcal{G}(S)\setminus S)\cup(D'\cup \mathrm{pa}_\mathcal{G}(D'))\setminus\{v\}} \quad [q_{D'}].$$

65

Using contraction again with the fact that $(\mathrm{pa}_{\mathcal{G}}(S) \setminus S) \cup D' \cup \mathrm{pa}_{\mathcal{G}}(D') \cup D = \mathrm{pa}_{\mathcal{G}}(S) \cup S = \mathrm{mb}_{\mathcal{G}}(v, S) \cup \{v\}$, and the fact that D and $\mathrm{mb}_{\mathcal{G}}(v, E)$ are disjoint, we have

$$X_v \perp\!\!\!\perp X_{\mathrm{mb}_{\mathcal{G}}(v,S) \setminus \mathrm{mb}_{\mathcal{G}}(v,E)} \mid X_{\mathrm{mb}_{\mathcal{G}}(v,E)} \quad [q_{D'}]$$

To show (47), we must also show the same constraint holds in q_E, which follows by an inductive application of Proposition 14. This completes the proof. $\qquad\square$

The main result of this section is an immediate corollary of Theorems 65, 67, and 44.

Corollary 68.

$$\mathcal{P}_t(\mathcal{G}, O, \prec) = \mathcal{P}^n(\mathcal{G}(O)).$$

4.5   Connections with r-Factorization

Shpitser et al. (2011) used constraints in causal DAG models with latent variables to construct a variable elimination (VE) algorithm for evaluating causal queries $p(x_Y \mid \mathrm{do}(x_A))$ in a computationally efficient manner. This algorithm used an older definition called the 'r-factorization property'. The nested Markov model r-factorizes which implies that the VE algorithm applies to these models as well.

Theorem 69. *If $p(x_V) \in \mathcal{P}^n(\mathcal{G}(V))$, then $p(x_V)$ r-factorizes with respect to $\mathcal{G}$ and $\{\phi_{V \setminus C}(p(x_V); \mathcal{G}) \mid C \in \mathcal{I}(\mathcal{G})\}$.*

Proof. This follows directly from the definition of r-factorization, the definition of $\mathcal{P}^n(\mathcal{G})$ and Theorem 38. $\qquad\square$

5   Summary

We have defined a novel statistical model which represents equality (but not inequality) constraints in marginals of DAG models, including the Verma con-

straint. Though this model represents constraints found in marginal distributions, it does not itself model latent variables explicitly. We call this model the nested Markov model, and it is represented by means of an acyclic directed mixed graph (ADMG). Our model is 'nested' because it is defined on sets of graphs and kernels derived recursively from the original marginal distribution and ADMG via a fixing operation. The fixing operation unifies certain marginalization, conditioning, and applications of the g-formula. Central to our model definition is the fact that any two valid sequences of fixing operations that fix the same set of nodes give the same result. We have characterized our model via Markov properties and a factorization. We have also shown a close connection between our model and a constraint enumeration algorithm for marginals of causal DAG models given in (Tian and Pearl, 2002b), and used the fixing operation to characterize all identifiable causal effects in hidden variable DAG models using a one line formula (41).

6 Acknowledgements

References

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

Bell, J. S. (1964). On the Einstein-Podolsky-Rosen paradox. *Physics 1*(3), 195–200.

Clauser, J. F., M. A. Horne, A. Shimony, and R. A. Holt (1969). Proposed experiment to test local hidden-variable theories. *Physical Review Letters 23*(15), 880.

Constantinou, P. (2013). *Conditional Independence and Applications in Statistical Causality*. Ph. D. thesis, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge.

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

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Ser. B 41*, 1–31.

Evans, R. J. (2012). Graphical methods for inequality constraints in marginalized DAGs. In *Machine Learning for Signal Processing (MLSP)*.

Evans, R. J. (2015). Margins of discrete Bayesian networks. arXiv preprint:1501.02103.

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

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

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

Koster, J. (2002). Marginalizing and conditioning in graphical models. *Bernoulli 8*, 817–840.

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

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 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 435–443.

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

Pearl, J. and T. S. Verma (1991). A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pp. 441–452.

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

Richardson, T. S. and P. Spirtes (2002). Ancestral graph Markov models. *Ann. Statist. 30*, 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. (1999). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In C. Glymour and G. Cooper (Eds.), *Computation, Causation, and Discovery*, pp. 349–405. Cambridge, MA: MIT Press.

Shpitser, I., R. J. Evans, T. S. Richardson, and J. Robins (2014). An introduction to nested Markov models. *Behaviormetrika 41*(1), 3–39.

Shpitser, I. and J. Pearl (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Twenty-First National Conference on Artificial Intelligence*.

Shpitser, I. and J. Pearl (2007). Complete identification methods for the causal hierarchy. Technical Report R–336, UCLA Cognitive Systems Laboratory.

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

Shpitser, I., T. S. Richardson, and J. M. Robins (2009). Testing edges by truncations. In *International Joint Conference on Artificial Intelligence*, Volume 21, pp. 1957–1963.

Shpitser, I., T. S. Richardson, and J. M. Robins (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.

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

Spirtes, P., C. Glymour, R. Scheines, C. Meek, S. Fienberg, and E. Slate (1992). Prediction and experimental design with graphical causal models. Technical Report 524, Department of Philosophy, Carnegie Mellon University.

Strotz, R. and H. Wold (1960). Recursive versus non-recursive systems: An attempt at synthesis. *Econometrica 28*(2), 417–427.

Studený, M. (1992). Conditional independence relations have no finite complete characterization. In *Information Theory, Statistical Decision Functions and Random Processes. Transactions of the 11th Prague Conference vol. B*, pp. 377–396. Kluwer, Dordrecht - Boston - London (also Academia, Prague).

Tian, J. and J. Pearl (2002a). A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, pp. 567–573.

Tian, J. and J. Pearl (2002b). On the testable implications of causal models with hidden variables. In *Proceedings of UAI-02*, pp. 519–527.

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

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

7 Appendix

Proposition 3 *If $\mathcal{G}(V \dot\cup L)$ is a DAG then $\mathcal{G}(V)$ is an ADMG.*

Proof: It follows directly from the construction that if $v \to v'$ in $\mathcal{G}(V)$ then $v \in \mathrm{an}_{\mathcal{G}}(v')$. The presence of a directed cycle in $\mathcal{G}(V)$ would imply a directed cycle in $\mathcal{G}$, which is a contradiction. $\qquad\qquad\square$

Proposition 4 *Let $\mathcal{G}(V \dot\cup L)$ be a DAG. For disjoint subsets $A, B, C \subseteq V$, (C may be empty), A is d-separated from B given C in $\mathcal{G}$ if and only if A is m-*

70

separated from B *given* C *in* $\mathcal{G}(V)$.

Proof: For every path π in $\mathcal{G}$, by Definition 6 there is a corresponding path π^* in $\mathcal{G}(V)$ consisting of a subsequence of the vertices on π, such that if a vertex v is a collider (non-collider) on π^* then it is a collider (non-collider) on π. It follows from this that d-connection in $\mathcal{G}$ implies m-connection in $\mathcal{G}(V)$. Conversely, by Definition 6 for each edge ϵ^* with endpoints e and f on π^* in $\mathcal{G}(V)$ there is a unique path μ_{ϵ^*} with endpoints e and f in $\mathcal{G}$ such that there is an arrowhead at e (f) on ϵ^* if and only if the edge on μ_{ϵ^*} with e (f) as an endpoint has an arrowhead at e (f). It then follows from Lemma 3.3.1 in (Spirtes et al., 1993) that if there is a path m-connecting a and b given C in $\mathcal{G}(V)$ then there is a path d-connecting a and b given C in $\mathcal{G}$. The result then follows. □

Theorem 19 $\mathcal{P}_f^c(\mathcal{G}) = \mathcal{P}_l^c(\mathcal{G}, \prec) = \mathcal{P}_m^c(\mathcal{G}) = \mathcal{P}_a^c(\mathcal{G})$.

To prove this result we will need a number of intermediate results.

Lemma 70. *(Lemma 3 in Richardson (2003)) For a CADMG* $\mathcal{G}(V, W, E)$, *suppose* $\boldsymbol{\mu}$ *is a path which m-connects* x *and* y *given* Z *in* $\mathcal{G}^{|W}$. *Then the sequence of non-colliders on* $\boldsymbol{\mu}$ *form a path connecting* x *and* y *in* $(\mathcal{G}_{\mathrm{an}(\{x,y\} \cup Z)})^a$.

Proof: Every vertex on an m-connecting path is either an ancestor of a collider, and hence Z, or an ancestor of an endpoint. Thus all the vertices on $\boldsymbol{\mu}$ are in $\mathcal{G}_{\mathrm{an}(\{x,y\} \cup Z)}$. Suppose that w_i and w_{i+1} ($1 \leq i \leq k-1$) are the successive non-colliders on $\boldsymbol{\mu}$. The subpath $\boldsymbol{\mu}(w_i, w_{i+1})$ consists entirely of colliders, hence w_i and w_{i+1} are adjacent in $(\mathcal{G}_{\mathrm{an}(\{x,y\} \cup Z)})^a$. Similarly w_1 and w_k are adjacent to x and y, respectively, in $(\mathcal{G}_{\mathrm{an}(\{x,y\} \cup Z)})^a$. □

Lemma 71. *For a CADMG* $\mathcal{G}(V, W, E)$, *suppose* $\boldsymbol{\mu}$ *is a path which m-connects* x *and* y *given* Z *in* $\mathcal{G}^{|W}$. *Then the sequence of non-colliders on* $\boldsymbol{\mu}$ *form a path connecting* x *and* y *in* $(\mathcal{G}_{\mathrm{an}(\{x,y\} \cup Z)})^a$.

Proof: Every vertex on an m-connecting path is either an ancestor of a collider, and hence Z, or an ancestor of an endpoint. Thus all the vertices on $\boldsymbol{\mu}$ are in $\mathcal{G}_{\mathrm{an}(\{x,y\} \cup Z)}$. Suppose that w_i and w_{i+1} ($1 \leq i \leq k-1$) are the successive

71

non-colliders on $\boldsymbol{\mu}$. The subpath $\boldsymbol{\mu}(w_i, w_{i+1})$ consists entirely of colliders, hence w_i and w_{i+1} are adjacent in $(\mathcal{G}_{\mathrm{an}(\{x,y\}\cup Z)})^a$. Similarly w_1 and w_k are adjacent to x and y, respectively, in $(\mathcal{G}_{\mathrm{an}(\{x,y\}\cup Z)})^a$. $\qquad\square$

Theorem 72. *If $\mathcal{G}$ is a CADMG then*

$$\mathcal{P}_m^c(\mathcal{G}) = \mathcal{P}_a^c(\mathcal{G}).$$

This proof follows that of Theorem 1 in Richardson (2003).

Proof: $\quad \mathcal{P}_m^c(\mathcal{G}) \subseteq \mathcal{P}_a^c(\mathcal{G})$

We proceed by showing that if X and Y are m-connected given Z in $\mathcal{G}^{|W}$ then X and Y are not separated by Z in $(\mathcal{G}_{\mathrm{an}(X\cup Y\cup Z)})^a$. If X and Y are m-connected given Z in $\mathcal{G}^{|W}$ then there are vertices $x \in X$, $y \in Y$ such that there is a path $\boldsymbol{\mu}$ which m-connects x and y given Z in $\mathcal{G}^{|W}$. By Lemma 70 the non-colliders on $\boldsymbol{\mu}$ form a path $\boldsymbol{\mu}^*$ connecting x and y in $(\mathcal{G}_{\mathrm{an}(X\cup Y\cup Z)})^a$. Since $\boldsymbol{\mu}$ is m-connecting, no non-collider is in Z hence no vertex on $\boldsymbol{\mu}^*$ is in Z. Thus X and Y are not separated by Z in $(\mathcal{G}_{\mathrm{an}(X\cup Y\cup Z)})^a$. $\qquad\square$

Proof: $\quad \mathcal{P}_a^c(\mathcal{G}) \subseteq \mathcal{P}_m^c(\mathcal{G})$

We show that if X and Y are not separated by Z in $(\mathcal{G}_{\mathrm{an}(X\cup Y\cup Z)})^a$ then X and Y are m-connected given Z in $\mathcal{G}^{|W}$. If X and Y are not separated by Z in $(\mathcal{G}_{\mathrm{an}(X\cup Y\cup Z)})^a$ then there are vertices $x \in X$, $y \in Y$ such that there is a minimal path $\boldsymbol{\pi}$ between x and y in $(\mathcal{G}_{\mathrm{an}(X\cup Y\cup Z)})^a$ on which no vertex is in Z. Our strategy is to replace each augmented edge on $\boldsymbol{\pi}$ with a corresponding collider path in $\mathcal{G}^{|W}$ and replace the other edges on $\boldsymbol{\pi}$ with the corresponding edges in $\mathcal{G}$ (choosing arbitrarily if there is more than one). It follows from Lemma 2 in Richardson (2003) that the resulting sequence of edges form a path from x to y in $\mathcal{G}^{|W}$, which we denote $\boldsymbol{\nu}$. Further, any non-collider on $\boldsymbol{\nu}$ is a vertex on $\boldsymbol{\pi}$ and hence not in Z. Finally, since all vertices in $\boldsymbol{\nu}$ are in $\mathcal{G}_{\mathrm{an}(X\cup Y\cup Z)}$ it follows that every collider is in $\mathrm{an}(X \cup Y \cup Z)$. Thus by Lemma 1 in Richardson (2003) there exist vertices $x^* \in X$ and $y^* \in Y$ which are m-connected given Z in $\mathcal{G}^{|W}$, hence X and Y are m-connected given Z. $\qquad\square$

Lemma 73. *If $\mathcal{G}(V,W,E)$ is a CADMG, x is a vertex in an ancestral set $A \subseteq V \cup W$ and $ch_{\mathcal{G}}(x) \cap A = \emptyset$, then the induced subgraph of the moral graph $(\mathcal{G}_A)^a$ on the set $\{x\} \cup mb(x,A)$ is always a clique. In addition, if $y \mathbin{\text{---}} x$ in $(\mathcal{G}_A)^a$ then $y \in mb(x,A)$.*

Proof: (Cf. proof of Theorem 4 in Richardson (2003))

$y \mathbin{\text{---}} x$ in $(\mathcal{G}_A)^a$ if and only if x is collider connected to y in $\mathcal{G}_A^{|W}$. Since $ch_{\mathcal{G}}(x) \cap A = \emptyset$, the vertex adjacent to x on any collider path is in $\mathrm{sp}_{\mathcal{G}_A^{|W}}(x) \cup \mathrm{pa}_{\mathcal{G}_A^{|W}}(x)$. Consequently a collider path to x in $\mathcal{G}_A^{|W}$ takes one of three forms:

(a) $y \to x \quad \Leftrightarrow \quad y \in \mathrm{pa}_{\mathcal{G}_A^{|W}}(x)$;

(b) $y \leftrightarrow w \leftrightarrow \cdots \leftrightarrow x \quad \Leftrightarrow \quad y \in \mathrm{dis}_{\mathcal{G}_A^{|W}}(x) \setminus \{x\}$;

(c) $y \to w \leftrightarrow \cdots \leftrightarrow x \quad \Leftrightarrow \quad y \in \mathrm{pa}_{\mathcal{G}_A^{|W}}(\mathrm{dis}_{\mathcal{G}_A^{|W}}(x) \setminus \{x\})$.

It then follows from the definition of a Markov blanket that y is collider connected to x in $\mathcal{G}_A^{|W}$ if and only if $y \in mb(x,A)$. This establishes that if $y \mathbin{\text{---}} x$ in $(\mathcal{G}_A)^a$ then $y \in \mathrm{mb}(x,A)$.

Suppose that $u, v \in \mathrm{mb}(x,A)$, with $u \neq v$. Then there are collider paths $\boldsymbol{\nu}_{ux}, \boldsymbol{\nu}_{vx}$ in $\mathcal{G}_A^{|W}$. Traversing the path $\boldsymbol{\nu}_{ux}$ from u to x, let w be the first vertex which is also on $\boldsymbol{\nu}_{vx}$; such a vertex is guaranteed to exist since x is common to both paths. Concatenating the subpaths $\boldsymbol{\nu}_{ux}(u,w)$ and $\boldsymbol{\nu}_{vx}(v,w)$ forms a collider path connecting u and w in $\mathcal{G}_A^{|W}$. (If $w = x$, this follows from $ch_{\mathcal{G}}(x) \cap A = \emptyset$.) Hence $u \mathbin{\text{---}} v$ in $(\mathcal{G}_A)^a$, proving the first claim. Finally, if $A \subseteq W$ then $\mathcal{G}_A$ is a complete graph containing bi-directed edges and so it is easy to see that the claim holds for $x \in A \subseteq W$. □

Theorem 74. *If $\mathcal{G}(V,W,E)$ is a CADMG and $\prec$ is a consistent ordering then*

$$\mathcal{P}_l^c(\mathcal{G}, \prec) = \mathcal{P}_f^c(\mathcal{G}).$$

Proof: $\mathcal{P}_l^c(\mathcal{G}, \prec) \subseteq \mathcal{P}_f^c(\mathcal{G})$

Fix an ancestral set $A \subseteq V \cup W$. We have:

$$q_V(x_{A\cap V} \mid x_W) = \prod_{d\in D\in \mathcal{D}(\mathcal{G}_{A\cup W})} q_V(x_d \mid x_{A\cap \mathrm{pre}_{\mathcal{G},\prec}(d)\setminus\{d\}})$$

$$= \prod_{d\in D\in \mathcal{D}(\mathcal{G}_{A\cup W})} q_V(x_d \mid x_{\mathrm{mb}(d, A\cap \mathrm{pre}_{\mathcal{G},\prec}(d))}).$$

The first line is by the chain rule of probabilities and the fact that $\mathcal{D}(\mathcal{G}_{A\cup W})$ is a partition of nodes in $\mathcal{G}_{A\cup W}$, the second by the ordered local Markov property. This is sufficient for the conclusion. □

Proof: $\mathcal{P}_f^c(\mathcal{G}) \subseteq \mathcal{P}_l^c(\mathcal{G}, \prec)$

Let $V = \{w_1, \ldots, w_n\}$ be a numbering of the vertices such that $w_i \prec w_j$ if and only if $i < j$, so $\mathrm{pre}_{\mathcal{G},\prec}(w_k) = \{w_1, \ldots, w_k\}$.

The proof is by induction on the sequence of ordered vertices. For $k = 1$, there is nothing to show. Assume the inductive hypothesis holds for $j < k$, and fix an ancestral set $A \subseteq \mathrm{pre}_{\mathcal{G},\prec}(w_k)$. We have:

$$q_V(x_{A\cap V} \mid x_W) = \prod_{d\in D\in \mathcal{D}(\mathcal{G}_{A\cup W})} q_V(x_d \mid x_{A\cap \mathrm{pre}_{\mathcal{G},\prec}(d)})$$

$$q_V(x_{A\cap V} \mid x_W) = \prod_{d\in D\in \mathcal{D}(\mathcal{G}_{A\cup W})} q_V(x_d \mid x_{\mathrm{mb}(d,\ A\cap \mathrm{pre}_{\mathcal{G},\prec}(d))}).$$

The first line is, again, by chain rule of probabilities and the fact that $\mathcal{D}(\mathcal{G})$ is a partition of nodes in $\mathcal{G}$, and the second is by the Markov factorization.

$$\prod_{D\in\mathcal{D}(\mathcal{G}_{A\cup W})} \prod_{d\in D} q_V(x_d \mid x_{A\cap \mathrm{pre}_{\mathcal{G},\prec}(d)}) =$$

$$\prod_{D\in\mathcal{D}(\mathcal{G}_{A\cup W})} \prod_{d\in D} q_V(x_d \mid x_{\mathrm{mb}(d,\ A\cap \mathrm{pre}_{\mathcal{G},\prec}(d))}).$$

The inductive hypothesis holds for all nodes in $\mathrm{pre}_{\mathcal{G}, \prec}(w_k) \setminus \{w_k\}$. This allows us to cancel all terms from the above equality, except the terms $q_V(x_{w_k} \mid x_{\mathrm{pre}_{\mathcal{G}, \prec}(w_k)}) = q_V(x_{w_k} \mid x_{\mathrm{mb}(w_k,\ A \cap \mathrm{pre}_{\mathcal{G}, \prec}(w_k))})$ and $q_V(x_{w_k} \mid x_{\mathrm{mb}(w_k,A)})$, which immediately establishes our conclusion. □

Theorem 75. *If $\mathcal{G}(V, W, E)$ is a CADMG and $\prec$ is a consistent ordering then*

$$\mathcal{P}_a^c(\mathcal{G}) = \mathcal{P}_l^c(\mathcal{G}, \prec).$$

Proof:

We first show that $\mathcal{P}_l^c(\mathcal{G}, \prec) \subseteq \mathcal{P}_a^c(\mathcal{G})$. The proof is similar to Proposition 5 in Lauritzen *et al.* (1990), and Theorem 2 in Richardson (2003).

Let $V = \{w_1, \ldots, w_n\}$ be a numbering of the vertices such that $w_i \prec w_j$ if and only if $i < j$, so $\mathrm{pre}_{\mathcal{G}, \prec}(w_k) = \{w_1, \ldots, w_k\}$. Let $P \in \mathcal{P}_l^c(\mathcal{G}, \prec)$. The proof is by induction on the sequence of ordered vertices. The inductive hypothesis is that if $W \dot\cup Y \dot\cup Z \subseteq \{w_1, \ldots, w_k\}$ and W is separated from Y by Z in $(\mathcal{G}_{\mathrm{an}(W \cup Y \cup Z)})^a$ then $X_W \perp\!\!\!\perp X_Y \mid X_Z$ in P.

For $k = 1$ there is nothing to show. Suppose that the induction hypothesis holds for $j < k$. Let $\mathcal{H} \equiv (\mathcal{G}_{\mathrm{an}(W \cup Y \cup Z)})^a$. If W is separated from Y by Z in $\mathcal{H}$ and $v \in \mathrm{an}(W \cup Y \cup Z) \setminus (W \cup Y \cup Z)$ then in $\mathcal{H}$ either v is separated from Y by Z, or v is separated from W by Z (or both). Hence we may always extend W and Y, so that $\mathrm{an}(W \cup Y \cup Z) = W \cup Y \cup Z$, and thus need only consider this case. If $w_k \notin (W \cup Y \cup Z)$ then $W \cup Y \cup Z \subseteq \mathrm{pre}_{\mathcal{G}, \prec}(w_j)$ for some $w_j \prec w_k$ hence the required independence follows directly from the induction hypothesis. Thus we suppose that $w_k \in (W \cup Y \cup Z) \subseteq \mathrm{pre}_{\mathcal{G}, \prec}(w_k)$. Let $A \equiv W \cup Y \cup Z$. Since A is ancestral and w_k has no children in A, the local Markov property implies that

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

There are now three cases to consider: (i) $w_k \in W$; (ii) $w_k \in Y$; (iii) $w_k \in Z$.

(i) Since $(\mathcal{G}_{\mathrm{an}(Y \cup Z \cup (W \setminus \{w_k\}))})^a$ contains a subset of the edges in $\mathcal{H}$, $W \setminus \{w_k\}$ is separated from Y by Z in $(\mathcal{G}_{\mathrm{an}(Y \cup Z \cup (W \setminus \{w_k\}))})^a$. If $W \neq \{w_k\}$ then

75

$X_{W \setminus \{w_k\}} \perp\!\!\!\perp X_Y \mid X_Z$ in P by the induction hypothesis. It is thus sufficient to prove that $X_{w_k} \perp\!\!\!\perp X_Y \mid X_{Z \cup (W \setminus \{w_k\})}$; this also covers the case where $W = \{w_k\}$. Since the vertices in $\{w_k\} \cup \mathrm{mb}(w_k, A)$ form a clique in $\mathcal{H}$ it follows that $\{w_k\} \cup \mathrm{mb}(w_k, A) \subseteq W \cup Z$, so $Y \subseteq A \setminus (\{w_k\} \cup \mathrm{mb}(w_k, A))$. Thus

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

$$\Rightarrow \quad X_{w_k} \perp\!\!\!\perp X_Y \mid X_{Z \cup (W \setminus \{w_k\})} \ [P]$$

(ii) Similar to case (i).

(iii) Since, by hypothesis, $A = \mathrm{an}(A) \subseteq \mathrm{pre}_{\mathcal{G}, \prec}(w_k)$, $w_k \notin \mathrm{an}(W \cup Y \cup (Z \setminus \{w_k\}))$. Thus the vertex w_k is not in $(\mathcal{G}_{\mathrm{an}(W \cup Y \cup (Z \setminus \{w_k\}))})^a$, and this graph contains a subset of the edges in $\mathcal{H}$. Hence W is separated from Y given $Z \setminus \{w_k\}$ in $(\mathcal{G}_{\mathrm{an}(W \cup Y \cup (Z \setminus \{w_k\}))})^a$. So by the induction hypothesis $X_W \perp\!\!\!\perp X_Y \mid X_{Z \setminus \{w_k\}}$. It is then sufficient to prove that either $X_{w_k} \perp\!\!\!\perp X_Y \mid X_{W \cup (Z \setminus \{w_k\})}$ or $X_{w_k} \perp\!\!\!\perp X_W \mid X_{Y \cup (Z \setminus \{w_k\})}$ in P. Since by Lemma 73 $\{w_k\} \cup \mathrm{mb}(w_k, A)$ forms a clique in $(\mathcal{G}_A)^a$ it follows that either

$$\{w_k\} \cup \mathrm{mb}(w_k, A) \ \subseteq \ W \cup Z \quad \text{or}$$

$$\{w_k\} \cup \mathrm{mb}(w_k, A) \ \subseteq \ Y \cup Z.$$

Suppose the former. In this case by the ordered local Markov property

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

$$\Rightarrow \quad X_{w_k} \perp\!\!\!\perp X_Y \mid X_{W \cup (Z \setminus \{w_k\})} \ [P]$$

If the latter then $X_{w_k} \perp\!\!\!\perp X_W \mid X_{Y \cup (Z \setminus \{w_k\})} \ [P]$. $\quad\square$

Proof: $\mathcal{P}_a^c(\mathcal{G}) \subseteq \mathcal{P}_\ell^c(\mathcal{G}, \prec)$

By Lemma 73 every vertex adjacent to x in $(\mathcal{G}_A)^a$ is in $\mathrm{mb}(x, A)$, hence x is separated from $A \setminus (\mathrm{mb}(x, A) \cup \{x\})$ by $\mathrm{mb}(x, A)$ in $(\mathcal{G}_A)^a$. Hence if $P \in \mathcal{P}_a^c(\mathcal{G})$ then $P \in \mathcal{P}_\ell^c(\mathcal{G}, \prec)$. $\quad\square$