Acyclic Linear SEMs Obey the Nested Markov Property

Ilya Shpitser
Department of Computer Science
Johns Hopkins University
ilyas@cs.jhu.edu

Robin J. Evans
Department of Statistics
University of Oxford
evans@stats.ox.ac.uk

Thomas S. Richardson
Department of Statistics
University of Washington
thomasr@u.washington.edu

Abstract

The conditional independence structure induced on the observed marginal distribution by a hidden variable directed acyclic graph (DAG) may be represented by a graphical model represented by mixed graphs called maximal ancestral graphs (MAGs). This model has a number of desirable properties, in particular the set of Gaussian distributions can be parameterized by viewing the graph as a path diagram. Models represented by MAGs have been used for causal discovery [22], and identification theory for causal effects [28].

In addition to ordinary conditional independence constraints, hidden variable DAGs also induce generalized independence constraints. These constraints form the nested Markov property [20]. We first show that acyclic linear SEMs obey this property. Further we show that a natural parameterization for all Gaussian distributions obeying the nested Markov property arises from a generalization of maximal ancestral graphs that we call maximal arid graphs (MArG). We show that every nested Markov model can be associated with a MArG; viewed as a path diagram this MArG parametrizes the Gaussian nested Markov model. This leads directly to methods for ML fitting and computing BIC scores for Gaussian nested models.

1 INTRODUCTION

Causal models associated with graphs have a long history in statistics, beginning with the seminal work of Wright in pedigree analysis [27], Haavelmo's work on simultaneous equations in econometrics [13] and the more recent synthesis of earlier work into a general causal modeling framework due to Pearl [17]. Causal graphical models are widely used in a variety of disciplines, with many theoretical developments and applications.

An important parametric subclass of causal graphical models are the linear structural equation models with correlated errors (SEMs). In fact, Wright and to some extent Haavelmo's work was originally within the SEM class. SEMs are defined over a class of mixed graphs containing directed ($\rightarrow$) edges representing direct causation, and bidirected ($\leftrightarrow$) edges representing correlated errors. Mixed graphs of this type without directed cycles—an assumption that rules out cyclic causation—are called *acyclic directed mixed graphs* (ADMGs).

Given an ADMG $\mathcal{G}$, the *linear structural equation model with correlated errors (SEM)* associated with $\mathcal{G}$ is formally defined as the set of multivariate normal distributions with covariance matrices of the form

$$\Sigma = (I - B)^{-T}\Omega(I - B)^{-1},$$

where $\omega_{ij} = \omega_{ji} = 0$ unless $i \leftrightarrow j$ exists in $\mathcal{G}$, and $b_{ij} = 0$ unless $i \rightarrow j$ exists in $\mathcal{G}$. The matrix Ω—and therefore Σ—is assumed to be positive definite. We denote this set by $\mathcal{P}_{\text{sem}}(\mathcal{G})$, and the set of Gaussians with arbitrary covariances $\mathcal{N}$.

It is easy to show that this model is equivalent to assuming that each variable X_i is a linear function of its parents with coefficients b_{ji} together with an additive error term. The error terms are assumed to have a multivariate normal distribution with covariance matrix given by Ω. If $\Omega = I$, error terms are uncorrelated and the SEM corresponds to a directed acyclic graph (DAG).

Elements of $\mathcal{P}_{\text{sem}}(\mathcal{G})$ are known to obey the global Markov property for $\mathcal{G}$ given by a criterion called *m-separation* [15, 19, 23]; this is the natural extension of d-separation to mixed graphs—see the Appendix for a definition. This criterion implies that absences of certain edges in $\mathcal{G}$ correspond to conditional independences in elements of $\mathcal{P}_{\text{sem}}(\mathcal{G})$. Densities that obey this global Markov property are said to be in the *ordinary Markov model* of $\mathcal{G}$, a set of densities we denote $\mathcal{P}_o(\mathcal{G})$ [9].

Hence $\mathcal{P}_{\mathrm{sem}}(\mathcal{G}) \subseteq \mathcal{P}_o(\mathcal{G}) \cap \mathcal{N}$; that is elements of the SEM for $\mathcal{G}$ are multivariate normal and are in the ordinary Markov model of $\mathcal{G}$.

Given a DAG $\mathcal{D}$ with observed variables O and hidden variables H, a simple operation, called the latent projection [26], maps it to an ADMG $\mathcal{G}$ with only observed variables O, such that d-separation applied to any variables in O in $\mathcal{D}$, and m-separation applied to $\mathcal{G}$ yields the same set of conditional independence relations on O. Thus, distributions Markov relative to a hidden variable DAG yield marginal distributions in $\mathcal{P}_o(\mathcal{G})$ for a latent projection $\mathcal{G}$.

However, more recent work has shown that these marginal distributions also obey certain *generalized conditional independence constraints*, sometimes called *Verma constraints* [6]. These define a model known as the *nested Markov model*, also associated with $\mathcal{G}$, and denoted $\mathcal{P}_n(\mathcal{G})$ [20]; generally $\mathcal{P}_n(\mathcal{G}) \subseteq \mathcal{P}_o(\mathcal{G})$, since the nested model implies all the constraints of m-separation. In this paper we show that distributions in the SEM for $\mathcal{G}$ also obey the additional constraints of the nested Markov model, so $\mathcal{P}_{\mathrm{sem}}(\mathcal{G}) \subseteq \mathcal{P}_n(\mathcal{G}) \cap \mathcal{N}$.

Although well-studied, general SEMs possess many complexities that make them potentially difficult to work with. The models may not be everywhere identifiable, and may contain singularities that prevent convergence of fitting algorithms [5]. No general distributional equivalence result is available for SEMs; see [25] for recent developments. In addition, while characterization of identifiability of causal effects is known for nonparametric structural equations [14, 21], a similar result is not known for SEMs despite decades of work [1, 2, 3, 4, 8, 11, 12, 24].

It is known that SEMs are everywhere identified if and only if they are associated with ADMGs in a special class [7]; in this paper we call this class *arid graphs*. We show that in a further subclass called *maximal arid graphs* (MArGs), it is the case that $\mathcal{P}_{\mathrm{sem}}(\mathcal{G}) = \mathcal{P}_n(\mathcal{G}) \cap \mathcal{N}$. Moreover, we show that restricting to maximal arid graphs is without loss of generality in the sense that for any ADMG $\mathcal{G}$, there exists a maximal arid graph $\mathcal{G}^\dagger$ such that $\mathcal{P}_n(\mathcal{G}) = \mathcal{P}_n(\mathcal{G}^\dagger)$. We also provide an algorithm for obtaining this *maximal arid projection* from $\mathcal{G}$, and show that $\mathcal{G}^\dagger$ has the same ancestral relations as $\mathcal{G}$.

Our results immediately imply that the nested Markov model over multivariate normal densities is a curved exponential family of known dimension, and is everywhere identifiable. They also imply that Gaussian nested models can be fitted efficiently with existing algorithms, such as RICF [5], applied to the SEM associated with $\mathcal{G}^\dagger$. Conversely, our results imply that every SEM obeys all the generalized independence constraints implied by $\mathcal{P}_n(\mathcal{G})$.

MArGs form a natural subclass of ADMGs for the purposes of nested Markov model search methods, which could be used for causal discovery. This would be a more powerful alternative to model search with maximal ancestral graphs (MAGs), since nested models are more fine-grained and therefore make more unambiguous causal information available. Though the results in this paper make significant progress towards causal structure learning with nested Markov models, more work is required. In particular, a natural next step would be to fully describe equivalence classes of nested Markov models, and develop a constraint based model search algorithm that is akin to the FCI algorithm [22], but that also takes generalized conditional independence constraints into account.

The remainder of the paper is organized as follows. Section 2 gives some preliminary definitions, including that of acyclic directed mixed graphs (ADMGs). In Section 3 we define the nested Markov model associated with ADMGs formally, including the central notion of 'fixing'. Section 4 shows that the class of nested models can be represented, without loss of generality, by the class of maximal arid graphs. Section 5 characterizes fixing in terms of zeroes of SEM parameters; this leads to the result in Section 6, which shows that for maximal arid graphs, the nested model and SEM coincide. We conclude with an example in Section 7, and discussion in Section 8. The proofs of certain results that are not essential to the presentation are deferred to the Appendix.

2 PRELIMINARIES

In this paper, we consider mixed graphs with directed ($\rightarrow$) and bidirected ($\leftrightarrow$) arrows connecting pairs of distinct vertices. There is at most one edge of each type between any pair of vertices, and we forbid directed cycles (i.e. sequences of the form $v_1 \rightarrow v_2 \rightarrow \cdots \rightarrow v_k \rightarrow v_1$ for $k \geq 2$). Graphs in this class are called *acyclic directed mixed graphs* (ADMGs). ADMGs may contain *bows*, where both $a \rightarrow b$ and $a \leftrightarrow b$, but this is the only circumstance in which more than one edge may be present between two vertices. See Fig. 2, 3 and 4 for examples of ADMGs.

We will use standard genealogical terminology for relations between vertices. Given a vertex v in an ADMG $\mathcal{G}$ with a vertex set V, define the sets of *parents*, *children*, *ancestors*, *descendants*, and *siblings* of v as

$$\begin{aligned}
\mathrm{pa}_{\mathcal{G}}(v) &\equiv \{w : w \rightarrow v \text{ in } \mathcal{G}\} \\
\mathrm{ch}_{\mathcal{G}}(v) &\equiv \{w : v \rightarrow w \text{ in } \mathcal{G}\} \\
\mathrm{an}_{\mathcal{G}}(v) &\equiv \{w : w = v \text{ or } w \rightarrow \cdots \rightarrow v \text{ in } \mathcal{G}\} \\
\mathrm{de}_{\mathcal{G}}(v) &\equiv \{w : w = v \text{ or } v \rightarrow \cdots \rightarrow w \text{ in } \mathcal{G}\} \\
\mathrm{sib}_{\mathcal{G}}(v) &\equiv \{w : w \leftrightarrow v \text{ in } \mathcal{G}\}.
\end{aligned}$$

respectively. Define also the *non-descendants* of v to be $\mathrm{nd}_{\mathcal{G}}(v) \equiv V \setminus \mathrm{de}_{\mathcal{G}}(v)$. The definitions apply disjunctively to sets, e.g. for a set of vertices $W \subseteq V$, $\mathrm{pa}_{\mathcal{G}}(W) \equiv \bigcup_{w \in W} \mathrm{pa}_{\mathcal{G}}(w)$. In addition, we define the *district* of v to be the set

$$\mathrm{dis}_{\mathcal{G}}(v) \equiv \{w : w \leftrightarrow \ldots \leftrightarrow v \text{ in } \mathcal{G}\}.$$

The set of districts of an ADMG $\mathcal{G}$, which we denote by $\mathcal{D}(\mathcal{G})$, always partitions the set of vertices in $\mathcal{G}$.

An internal vertex v on a path is a *collider* (on the path) if both adjacent edges have an arrowhead at v. A path from w to v in $\mathcal{G}$ is called a *collider path* if every internal vertex is a collider on the path. For example $w \rightarrow z \leftrightarrow m \leftarrow v$ is a collider path, while $w \rightarrow z \rightarrow m \rightarrow v$ is not.

Given an ADMG $\mathcal{G}$, and a subset S of vertices V in $\mathcal{G}$, the *induced subgraph* $\mathcal{G}_S$ is the graph with vertex set S, and those edges in $\mathcal{G}$ between elements in S. A set S is called *bidirected-connected* in $\mathcal{G}$ if $\mathcal{D}(\mathcal{G}_S)$ contains a single set.

3 NESTED MARKOV MODELS

Nested Markov models are a class of graphical models associated with ADMGs, and defined by generalized independence constraints. We consider random variables $X_V \equiv (X_v : v \in V)$ taking values in the product space $\mathfrak{X}_V = \times_{v \in V} \mathfrak{X}_v$, for finite dimensional sets $\mathfrak{X}_v$. For any $A \subseteq V$ we denote the subset $(X_v : v \in A)$ by X_A.

A *kernel* $q_V(x_V \,|\, x_W)$ is a collection of densities over X_V, indexed by $x_W \in \mathfrak{X}_W$. Conditional densities are kernels, but not all kernels are obtained by conditioning; we give some examples later. Conditioning and marginalization are defined in the usual way in kernels.

A joint density $p(x_V)$ over X_V is said to be *nested Markov* with respect to an ADMG $\mathcal{G}$ if it obeys certain independence constraints in kernels derived from $p(x_V)$ using a 'fixing' operation. These constraints are implied by the m-separation criterion applied to *conditional ADMGs (CADMGs)* obtained from $\mathcal{G}$ by an analogous fixing operation. We now define these terms, and the nested Markov model, precisely.

A CADMG $\mathcal{G}(V, W)$ is an ADMG with a set of *random* vertices V and *fixed* vertices W, with the property that $\mathrm{sib}_{\mathcal{G}}(w) \cup \mathrm{pa}_{\mathcal{G}}(w) = \emptyset$ for every $w \in W$. An example can be found in Fig. 1(b); note that we depict fixed vertices with rectangular nodes, and random vertices with round nodes. Vertices V in a CADMG correspond to random variables, as in standard graphical models, while vertices in W correspond to variables that were fixed to a specific value by some operation, such as conditioning or causal interventions. The genealogical relations in Section 2 generalize in a straightforward way to CADMGs by ignoring the distinction between V and W; the only exception is that districts are only defined for random vertices, so $\mathcal{D}(\mathcal{G}(V, W))$ partitions V.

3.1 Fixing

A vertex $r \in V$ is said to be *fixable* in a CADMG $\mathcal{G}(V, W)$ if $\mathrm{dis}_{\mathcal{G}}(r) \cap \mathrm{de}_{\mathcal{G}}(r) = \emptyset$. Given a CADMG $\mathcal{G}(V, W)$, and a fixable $r \in V$, the fixing operation $\phi_r(\mathcal{G})$ yields a new CADMG $\widetilde{\mathcal{G}}(V \setminus \{r\}, W \cup \{r\})$ obtained from $\mathcal{G}(V, W)$ by removing all edges of the form $\rightarrow r$ and $\leftrightarrow r$, and keeping all other edges. Given a kernel $q_V(x_V \,|\, x_W)$ associated with a CADMG $\mathcal{G}(V, W)$, and a fixable $r \in V$, the fixing operation $\phi_r(q_V; \mathcal{G})$ yields a new kernel

$$\tilde{q}_{V \setminus \{r\}}(x_{V \setminus \{r\}} \,|\, x_W, x_r) \equiv \frac{q_V(x_V \,|\, x_W)}{q_V(x_r \,|\, x_{\mathrm{nd}_{\mathcal{G}}(r)})}.$$

A sequence $r_1, \ldots, r_k$ of vertices in V is said to be fixable if r_1 is fixable in $\mathcal{G}$, r_2 is fixable in $\phi_{r_1}(\mathcal{G})$, etc. A result in [20] states that for any $p(x_V) \in \mathcal{P}_n(\mathcal{G})$, two valid fixing sequences on the same set of variables R yield the same CADMG and kernel. We therefore unambiguously define

$$\phi_R(\mathcal{G}) \equiv \phi_{r_k}(\ldots \phi_{r_2}(\phi_{r_1}(\mathcal{G})) \ldots),$$

and similarly the kernel $\phi_R(p; \mathcal{G})$.

If a fixing sequence exists for a set $R \subseteq V$ in $\mathcal{G}(V, W)$, we say $V \setminus R$ is a *reachable set*. Such a set is called *intrinsic* if the vertices in $V \setminus R$ are bidirected-connected (so that $\mathcal{D}(\phi_R(\mathcal{G}))$ has a single element). We denote the collections of reachable and intrinsic sets in $\mathcal{G}$ respectively by $\mathcal{R}(\mathcal{G})$ and $\mathcal{I}(\mathcal{G})$.

For any $v \in V$, such that $\mathrm{ch}_{\mathcal{G}}(v) = \emptyset$, the *Markov blanket* of v in a CADMG $\mathcal{G}(V, W)$ is defined as

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

this is the set of vertices that are connected to v by collider paths. For brevity, we will denote $\mathrm{mb}_{\phi_{V \setminus S}(\mathcal{G})}(v)$ by $\mathrm{mb}_{\mathcal{G}}(v, S)$.

We are now ready to define the nested Markov model $\mathcal{P}_n(\mathcal{G})$. Given an ADMG $\mathcal{G}$, let $\prec$ be any topological ordering on V. A distribution $p(x_V)$ is in the nested Markov model associated with $\mathcal{G}$ if, for each intrinsic $S \subseteq V$ and $\prec$-maximal $v \in S$,

$$X_v \perp\!\!\!\perp X_{V \setminus (\{v\} \cup \mathrm{mb}_{\mathcal{G}}(v,S))} \mid X_{\mathrm{mb}_{\mathcal{G}}(v,S)}$$

holds in $\phi_{V \setminus S}(p(x_V); \mathcal{G})$. This is known as the *ordered local Markov property* for nested models. As a consequence, under the nested model, fixing r

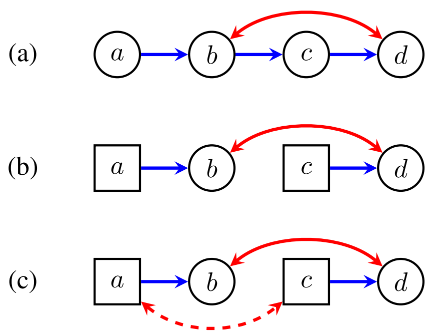

Figure 1: (a) An ADMG $\mathcal{G}$ that is not ancestral; (b) a CADMG obtained from $\mathcal{G}$ in (a) by fixing a and c; (c) the graph obtained from (b) that is used to check conditional independence statements associated with $\phi_{\{a,c\}}(p(a,b,c,d);\mathcal{G})$.

within any $R \in \mathcal{R}(\mathcal{G})$ may be redefined as dividing by $q_R(x_r \,|\, x_{\text{mb}_{\mathcal{G}}(r,R)})$, instead of dividing by $q_R(x_r \,|\, x_{\text{nd}_{\mathcal{G}}(r)})$ (see section 2.11 in [20]). Nested models can be equivalently defined by a *global nested Markov property* obtained by applying the m-separation criterion to each reachable graph $\phi_{V \setminus S}(\mathcal{G})$ after adding bidirected edges between all pairs of fixed vertices; adding these bidirected edges ensures no independences are implied between vertices in W. These m-separations imply independences in the kernel $\phi_{V \setminus S}(p;\mathcal{G})$; see [20].

Example 1. Consider the ADMG in Fig. 1(a). The vertices a, c and d all satisfy the condition of being fixable, but b does not since d is both a descendant of, and in the same district as, b. The CADMG $\mathcal{G}(\{b,d\},\{a,c\})$ obtained after fixing a and c is shown in Fig. 1(b). Notice that fixing c removes the edge $b \to c$, but that the edge $c \to d$ is preserved. Applying m-separation to the graph shown in Fig. 1 (c), obtained from Fig. 1 (b) by connecting a, c by a bidirected edge, yields

$$X_d \perp\!\!\!\perp X_a \mid X_c \text{ in } \phi_{\{a,c\}}(p(x_{\{a,b,c,d\}});\mathcal{G}).$$

In addition, one can see easily that if an edge $a \to d$ had been present in the original graph, then we would not have obtained this m-separation.

4 ARID GRAPHS

The main result of this section is that the nested Markov model associated with *any* ADMG $\mathcal{G}$ can be associated, without loss of generality, with a closely related *maximal arid graph (MArG)* $\mathcal{G}^{\dagger}$.

Arid graphs lack certain structures called C-trees [21] (aka convergent arborescences [7]) that present difficulties for identifiability. As a result, any linear SEM associated with an arid graph is everywhere identifiable [7]. In addition, maximal arid graphs are analogous to (but a strict superset of) maximal ancestral graphs (MAGs), a class of ADMGs also used for causal discovery.

The section proceeds as follows. In Section 4.1 we define the reachable closure of a set which is the smallest reachable superset of that set. These structures will be used in the proofs of our results, and to define C-trees and (maximal) arid graphs in Section 4.2. In Section 4.3 we define a projection operation which constructs, for any ADMG, its maximal arid graph counterpart. In Section 4.4, we show a number of useful graphical properties remain invariant between the original ADMG, and its maximal arid graph, leading to the proof of our main result in Section 4.5.

4.1 Reachable Closures

For a CADMG $\mathcal{G}(V,W)$, a (reachable) subset $C \subseteq V$ is called a *reachable closure* for $S \subseteq C$ if the set of fixable vertices in $\phi_{V \setminus C}(\mathcal{G})$ is a subset of S. Every set S in $\mathcal{G}$ has a reachable closure.

Proposition 2. *If $A, B \in \mathcal{R}(\mathcal{G})$, then $A \cap B \in \mathcal{R}(\mathcal{G})$.*

Proof. This follows from the fact that if a vertex is fixable, it remains fixable after fixing other vertices (see Lemma 27 of [20]). □

Proposition 3. *For any set of random vertices S in a CADMG $\mathcal{G}$, there is a unique reachable closure.*

Proof. Assume there are two such distinct closures W_1, W_2. Since both W_1 and W_2 are reachable, so is $W_1 \cap W_2$, by Proposition 2. Since $S \subseteq W_1$ and $S \subseteq W_2$, $S \subseteq W_1 \cap W_2$. Consider a fixing sequence σ_1 for $V \setminus W_1$. Then there exists a fixing sequence σ_2 for $V \setminus (W_1 \cap W_2)$ which contains σ_1 as a prefix. Note that W_1 being reachable implies that $W_1 \not\subseteq W_2$, by the same argument as in the proof of Proposition 2; hence σ_2 is non-empty. But this implies W_1 is not a reachable closure for S, since the next element in σ_2 after the σ_1 prefix cannot lie in S. This is a contradiction. □

In light of Proposition 3, we denote the unique reachable closure of a set S in $\mathcal{G}$ by $\langle S \rangle_{\mathcal{G}}$. By definition $\langle S \rangle_{\mathcal{G}} \in \mathcal{R}(\mathcal{G})$ for any S, and if $S \in \mathcal{R}(\mathcal{G})$ then $\langle S \rangle_{\mathcal{G}} = S$. To avoid clutter, if $S = \{s\}$, we write $\langle \{s\} \rangle_{\mathcal{G}}$ as $\langle s \rangle_{\mathcal{G}}$.

Proposition 4. *Let $A \subseteq B$ with B a reachable set; then $\langle A \rangle_{\phi_{V \setminus B}(\mathcal{G})} = \langle A \rangle_{\mathcal{G}}$.*

Lemma 5. $\langle S \rangle_{\mathcal{G}} \subseteq S \cup \text{pa}_{\mathcal{G}}(\langle S \rangle_{\mathcal{G}})$.

Proof. If $s \in \langle S \rangle_{\mathcal{G}} \setminus S$ then s has a child in $\langle S \rangle_{\mathcal{G}}$ since otherwise s is fixable, which is a contradiction. □

4.2 C-Trees and Arid Graphs

For any $v \in V$ in an ADMG $\mathcal{G}$, the induced subgraph $\mathcal{G}_{\langle v \rangle}$ is called a *v-rooted C-tree* [21] or an *arborescence converging on v* [7]. These subgraphs are particularly important because of the following result.

Theorem 6 ([7], Theorem 2)**.** *The SEM for an ADMG $\mathcal{G}$ is everywhere identifiable if and only if $\langle v \rangle_{\mathcal{G}} = \{v\}$ for all $v \in V$.*

In other words, the SEM parameterization is identifiable everywhere if and only if $\mathcal{G}$ does not contain any non-trivial converging arborescences. We call such graphs 'arid', since they do not contain such 'trees'.

Definition 7. An ADMG $\mathcal{G}$ is called *arid* if for every vertex v in $\mathcal{G}$, $\langle v \rangle_{\mathcal{G}} = \{v\}$.

Arid ADMGs are "DAG-like," in the sense that in any DAG $\mathcal{G}$, it is also the case that $\langle v \rangle_{\mathcal{G}} = \{v\}$. The central result of this section is that the nested Markov model $\mathcal{P}_n(\mathcal{G})$, a statistical model with desirable properties, may be associated without loss of generality with arid graphs.

The ordinary Markov model $\mathcal{P}_o(\mathcal{G})$ has previously been associated with another special class of ADMGs called *ancestral graphs* in [18].

Definition 8. An ADMG $\mathcal{G}$ is called *ancestral* if for every vertex v in $\mathcal{G}$, $\mathrm{sib}_{\mathcal{G}}(v) \cap \mathrm{an}_{\mathcal{G}}(v) = \emptyset$.

Arid graphs may be viewed as a strict generalization of ancestral graphs of [18], due to the following property of C-trees.

Proposition 9. *If $\langle a \rangle_{\mathcal{G}}$ contains more than one element, then there exists $b \in \langle a \rangle_{\mathcal{G}}$ with $a \leftrightarrow b$ in $\mathcal{G}$.*

Proof. If no such element exists then every element in $\mathrm{pa}_{\mathcal{G}}(a) \cap \langle a \rangle_{\mathcal{G}}$ is fixable in $\phi_{V \setminus \langle a \rangle_{\mathcal{G}}}(\mathcal{G})$, which is a contradiction. □

Proposition 10. *Ancestral graphs are arid.*

Proof. Follows immediately by the contrapositive application of Proposition 9 and $\langle a \rangle_{\mathcal{G}} \subseteq \mathrm{an}_{\mathcal{G}}(a)$. □

Proposition 11. *An arid graph with at least two vertices contains at least two fixable vertices.*

Proof. Since the graph is acyclic, there is some childless v that is therefore fixable. Since the graph is arid, $\langle v \rangle_{\mathcal{G}} = \{v\} \subset V$, and so there is also some other vertex that can be fixed to make $\{v\}$ reachable. □

4.3 Maximal Arid Projection

To prove that $\mathcal{P}_n(\mathcal{G})$ can always be associated with an arid graph, we define the *maximal arid projection* operation which, for every ADMG $\mathcal{G}$, yields a closely related graph $\mathcal{G}^\dagger$ that is arid, and ultimately show that $\mathcal{G}$ and $\mathcal{G}^\dagger$ yield the same nested model.

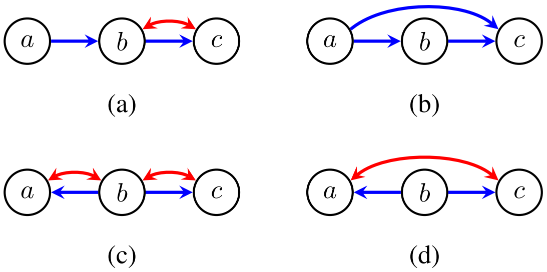

Figure 2: Graphs illustrating maximal arid projection. The graphs (a) and (c) are not arid, but have maximal arid projections given by (b) and (d) respectively.

In this section we define this projection operation and derive several of its properties, culminating in a proof that the projection and fixing operations commute. We first need a preliminary definition.

Definition 12. A pair of vertices $a \neq b$ in an ADMG $\mathcal{G}$ is *densely connected* if either $a \in \mathrm{pa}_{\mathcal{G}}(\langle b \rangle_{\mathcal{G}})$, or $b \in \mathrm{pa}_{\mathcal{G}}(\langle a \rangle_{\mathcal{G}})$, or $\langle \{a, b\} \rangle_{\mathcal{G}}$ is a bidirected-connected set.

A CADMG $\mathcal{G}$ is called *maximal* if every pair of densely connected vertices in $\mathcal{G}$ are adjacent.

Densely connected vertex pairs form the nested Markov analogue of *inducing paths* [26]. Just as the existence of an inducing path between a pair of vertices prevents m-separation by any set, so does the existence of dense connectedness between a pair of vertices prevents m-separation by any set within any CADMG corresponding to a reachable set. In effect, a densely connected pair cannot be made independent, by any combination of conditioning and fixing operations.

Definition 13. For a CADMG $\mathcal{G}$, we define the *maximal arid projection* of $\mathcal{G}$, denoted $\mathcal{G}^\dagger$, to be the graph that shares the vertex sets V, W with $\mathcal{G}$, and that contains the following edges:

- for $b \in V$, the edge $a \to b$ exists in $\mathcal{G}^\dagger$ if $a \in \mathrm{pa}_{\mathcal{G}}(\langle b \rangle_{\mathcal{G}})$,
- for $a, b \in V$, the edge $a \leftrightarrow b$ exists in $\mathcal{G}^\dagger$ if neither $a \in \mathrm{pa}_{\mathcal{G}}(\langle b \rangle_{\mathcal{G}})$, nor $b \in \mathrm{pa}_{\mathcal{G}}(\langle a \rangle_{\mathcal{G}})$, but $\langle \{a, b\} \rangle_{\mathcal{G}}$ is a bidirected-connected set.

Fig. 2 provides some elementary examples of the maximal arid projection. In each of (a) and (c) we have a dense inducing path between the vertices a and c. For (a) we insert the edge $a \to c$ to represent this (yielding (b)), while in (c) we add $a \leftrightarrow c$ (yielding (d)). In each case the bow arcs are replaced by directed edges.

We provide several results characterizing the output of the maximal arid projection operation, first noting that pairs of vertices adjacent in $\mathcal{G}$ are also adjacent in $\mathcal{G}^\dagger$.

Proposition 14.

(i) If $a \in \mathrm{pa}_\mathcal{G}(b)$, then $a \in \mathrm{pa}_{\mathcal{G}^\dagger}(b)$.

(ii) If $a \in \mathrm{sib}_\mathcal{G}(b)$, then either $a \in \mathrm{pa}_{\mathcal{G}^\dagger}(b)$ or $a \in \mathrm{sib}_{\mathcal{G}^\dagger}(b)$ or $b \in \mathrm{pa}_{\mathcal{G}^\dagger}(a)$.

Ancestral relationships are also preserved in $\mathcal{G}^\dagger$.

Proposition 15. *$a \in \mathrm{an}_\mathcal{G}(b)$ if and only if $a \in \mathrm{an}_{\mathcal{G}^\dagger}(b)$.*

Proof. If $a \in \mathrm{an}_\mathcal{G}(b)$, then $a \in \mathrm{an}_{\mathcal{G}^\dagger}(b)$ follows by an inductive application of Proposition 14(i). If $a \in \mathrm{an}_{\mathcal{G}^\dagger}(b)$, then fix a directed path $a \to w_1 \to \cdots w_k \to b$ in $\mathcal{G}^\dagger$. Each directed edge on this path from c to d is due to $c \in \mathrm{pa}_\mathcal{G}(\langle d \rangle_\mathcal{G})$ being true. But since every element $\langle d \rangle_\mathcal{G}$ is an ancestor of d in $\mathcal{G}$, this implies the existence of a directed path from c to d in $\mathcal{G}$. Thus, there is a directed path from a to b in $\mathcal{G}$. □

Proposition 16. *If $\mathcal{G}$ is a (C)ADMG, then so is $\mathcal{G}^\dagger$.*

Proof. Acyclicity of $\mathcal{G}^\dagger$ follows from Proposition 15; in addition, in a CADMG it is clear from the definition that no arrowheads are introduced into W. □

If $\mathcal{G}$ is acyclic then $\mathcal{G}^\dagger$ is simple, i.e. contains at most one edge between each pair of vertices, so if $\mathcal{G}$ is an ADMG then $\mathcal{G}^\dagger$ is an example of a *bow-free acyclic path diagram* (BAP) [5, 16].

Proposition 17. *$\mathcal{D}(\mathcal{G}^\dagger)$ is a sub-partition of $\mathcal{D}(\mathcal{G})$. Further, for any S reachable in $\mathcal{G}$ and $\mathcal{G}^\dagger$, $\mathcal{D}(\phi_{V\setminus S}(\mathcal{G}^\dagger))$ forms a sub-partition of $\mathcal{D}(\phi_{V\setminus S}(\mathcal{G}))$.*

Note that it will follow from Theorem 19 that if S is reachable in $\mathcal{G}$ then it is also reachable in $\mathcal{G}^\dagger$.

Lemma 18. *Let v be fixable in $\mathcal{G}$. For any $a, b \in V$ there is a directed path from a to b in $\mathcal{G}$ with no intermediate vertex being v, if and only if there is such a path in $\mathcal{G}^\dagger$.*

Theorem 19. *If S is reachable in an ADMG $\mathcal{G}$, then it is also reachable in $\mathcal{G}^\dagger$ via the same fixing sequence. In this case, $(\phi_{V\setminus S}(\mathcal{G}))^\dagger = \phi_{V\setminus S}(\mathcal{G}^\dagger)$.*

Corollary 20. *$\langle S \rangle_{\mathcal{G}^\dagger} \subseteq \langle S \rangle_\mathcal{G}$ for any set S.*

Proposition 21. *$\mathcal{G}^\dagger$ is a maximal arid graph.*

4.4 Invariance Results In Maximal Arid Projections

A key result will be that the nested Markov model associated with a maximal arid projection is the same as that for the original graph, and this will be proven by showing that the Markov blankets in the two graphs are the same.

Lemma 22. *Suppose that $w \in \mathrm{pa}_\mathcal{G}(\langle v \rangle_\mathcal{G})$, and that $\langle \{v, w\} \rangle_\mathcal{G}$ is bidirected-connected. Then $\langle \{v, w\} \rangle_\mathcal{G} = \langle v \rangle_\mathcal{G}$ and in particular $w \in \langle v \rangle_\mathcal{G}$.*

Lemma 23. *If $v, w \in V$ are connected by a collider path π in $\mathcal{G}$ then they are connected by a collider path $\pi^\dagger$ in $\mathcal{G}^\dagger$ that uses a subset of the internal vertices of π. In addition, if π starts with an edge $v \to$, then so does $\pi^\dagger$.*

This follows by definition of $\mathcal{G}^\dagger$, and properties of closures of sets of vertices of size 1 and 2. A detailed proof is in the Appendix.

As an example of this result, notice that the path $t \to x \leftrightarrow bp \leftrightarrow y$ in Fig. 4 (a) is replaced by $t \to bp \leftrightarrow y$ in the maximal arid projection in Fig. 4 (b).

We provide a partial converse.

Lemma 24. *If $v, w \in V$ are connected by a collider path $\pi^\dagger$ in $\mathcal{G}^\dagger$, then they are also connected by a collider path in $\mathcal{G}$.*

Proof. $\pi^\dagger$ is of the form $\to\leftrightarrow \cdots \leftrightarrow\leftarrow$ (possibly without the directed edges). Each $\leftrightarrow$ represents a bidirected path in $\mathcal{G}$. A directed edge in $\mathcal{G}^\dagger$, say $v \to t$, represents an edge $v \to s$ for $s \in \langle t \rangle_\mathcal{G}$. Since $\langle t \rangle_\mathcal{G}$ is bidirected-connected, there is a path of the form $v \to\leftrightarrow \cdots \leftrightarrow t$ in $\mathcal{G}$. Concatenating these paths (and possibly shortening) gives another collider path. □

Theorem 25. *Let S be a reachable set in $\mathcal{G}$. Then $\mathrm{mb}_\mathcal{G}(v, S) = \mathrm{mb}_{\mathcal{G}^\dagger}(v, S)$.*

Proof. First note that v is childless in $\phi_{V\setminus S}(\mathcal{G})$ if and only if it is so in $\phi_{V\setminus S}(\mathcal{G}^\dagger)$, so the statement is well-defined. Theorem 19 shows that it is enough to show this for $S = V$. The result is then a direct consequence of Lemmas 23 and 24, since the Markov blanket is just the set of vertices connected to v by collider paths. □

Proposition 26. *There is a one-to-one correspondence between intrinsic sets in $\mathcal{G}$ and in $\mathcal{G}^\dagger$.*

Remark 27. The set H is referred to by [10] as the *recursive head* associated with S. A consequence of the argument in the proof above is that the discrete parameterization given by [10] is identical for $\mathcal{G}$ and $\mathcal{G}^\dagger$.

An important result is that fixing corresponds to the same probabilistic operation in $\mathcal{G}$ and $\mathcal{G}^\dagger$.

Proposition 28. *If $S \in \mathcal{R}(\mathcal{G})$, then any fixing sequence σ for $V \setminus S$ valid in $\mathcal{G}$ consists of the same set of fixing operations when applied to $p(x_V)$ using $\mathcal{G}$ and when applied to $p(x_V)$ using $\mathcal{G}^\dagger$.*

Proof. Recall that fixing is division of $\phi_W(p(x_V); \mathcal{G}) \equiv q_V(x_V \,|\, x_W)$ by $q_V(x_v \,|\, x_{\mathrm{mb}_\mathcal{G}(v,S) \cup W})$. By Theorem 25, $\mathrm{mb}_\mathcal{G}(v, S) = \mathrm{mb}_{\mathcal{G}^\dagger}(v, S)$, and a simple induction gives the result. □

Thus, for any $S \in \mathcal{R}(\mathcal{G}^\dagger)$, q_S is well defined without specifying the particular sequence of fixing operations in $\mathcal{G}^\dagger$. Some fixing sequences may be valid in $\mathcal{G}^\dagger$ but not in $\mathcal{G}$; however the kernels reached in $\mathcal{G}^\dagger$ are related to those reachable in $\mathcal{G}$ by the following result.

Lemma 29. *Suppose $S \in \mathcal{I}(\mathcal{G}^\dagger)$, and let $S^\dagger = \langle S \rangle_{\mathcal{G}}$. Let v be the maximal element of S. Any independences involving the full conditional of v hold in q_S if and only if they hold in $q_{S^\dagger}$.*

Proof. For simplicity assume $S^\dagger = V$, so we write $\mathcal{G}$ and $\mathcal{G}^\dagger$ in place of $\phi_{V\setminus S^\dagger}(\mathcal{G})$ and $\phi_{V\setminus S^\dagger}(\mathcal{G}^\dagger)$ respectively. This is justified by Theorem 19.

Suppose that $s \in S^\dagger \setminus S$ is fixable in $\mathcal{G}^\dagger$ but not in $\mathcal{G}$. Then $s \in \text{dis}_{\mathcal{G}}(v)$ and is an ancestor of some $r \in S$ such that $\text{ch}_{\mathcal{G}}(r) = \emptyset$ (in both $\mathcal{G}$ and $\mathcal{G}^\dagger$).

The vertices r and v are connected by a collider path in $\mathcal{G}$, and so also are in $\mathcal{G}^\dagger$. Further, since they have no children in $\mathcal{G}$, they also have no children in $\mathcal{G}^\dagger$, and these paths are therefore made up entirely of bidirected edges in both graphs; in other words, r and v are in the same district in both graphs. Since s is fixable in $\mathcal{G}^\dagger$ but not in $\mathcal{G}$, and since ancestor relations are preserved, it follows that s is in the same district as r and v in $\mathcal{G}$, but not in $\mathcal{G}^\dagger$.

Fixing s involves division by $q_{S^\dagger}(x_s \mid x_{\text{mb}_{\mathcal{G}^\dagger}(s)})$. Since v is in a different district to s and has no children, then $v \notin \text{mb}_{\mathcal{G}^\dagger}(s)$, and so by Lemma 10 of [20] we have $q_{S^\dagger}(x_v \mid x_{S^\dagger\setminus\{v\}}, x_W) = q_{S^\dagger\setminus\{s\}}(x_v \mid x_{S^\dagger\setminus\{v\}}, x_W)$. Any further vertices in $S^\dagger \setminus S$ are also not in the same district as v for the same reason, so v never appears in their Markov blankets and hence this is also the same as the full conditional $q_S(x_v \mid x_{S\setminus\{v\}}, x_{W\cup(S^\dagger\setminus S)})$. The result follows. □

4.5 Any ADMG And Its Maximal Arid Projection Define The Same Nested Model

We are now ready to state and prove the main result of this section.

Theorem 30. $\mathcal{P}_n(\mathcal{G}) = \mathcal{P}_n(\mathcal{G}^\dagger)$.

Proof. Let $\prec$ be a topological order and consider any pair $S, S^\dagger$ as defined in Proposition 26. We will show that the corresponding independences for the ordered local nested Markov property are equivalent. Let v be the $\prec$-maximal element of S (and therefore of $S^\dagger$). Then the two independences are

$$\left(X_v \perp\!\!\!\perp X_{V\setminus(\text{mb}_{\mathcal{G}}(v,S)\cup\{v\})} \mid X_{\text{mb}_{\mathcal{G}}(v,S)}\right)$$

in $\phi_{V\setminus S}(p(x_V); \mathcal{G})$, and

$$\left(X_v \perp\!\!\!\perp X_{V\setminus(\text{mb}_{\mathcal{G}^\dagger}(v,S^\dagger)\cup\{v\})} \mid X_{\text{mb}_{\mathcal{G}^\dagger}(v,S^\dagger)}\right)$$

in $\phi_{V\setminus S^\dagger}(p(x_V); \mathcal{G}^\dagger)$. Since—as follows from the proof of Proposition 26—we have $\text{mb}_{\mathcal{G}}(v, S) = \text{mb}_{\mathcal{G}^\dagger}(v, S^\dagger)$, it only remains to bridge the difference between the kernels. But this is an independence on the full conditional of $\phi_{V\setminus S^\dagger}(p(x_V); \mathcal{G})$, so by Lemma 29, it holds in that kernel if and only if it holds in $\phi_{V\setminus S}(p(x_V); \mathcal{G}^\dagger)$. □

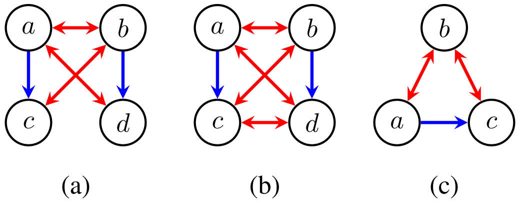

Figure 3: (a) A graph with a non-nested SEM constraint, and (b) A nested Markov equivalent graph. (c) A graph in which the parameter ω_{bc} is not identifiable after any fixing.

5 THE FIXING OPERATION IN STRUCTURAL EQUATION MODELS

If v is fixable in an ADMG $\mathcal{G}$, the kernel $q_{V\setminus\{v\}}$ resulting from fixing v is obtained by dividing $p(x_V)$ by the conditional distribution $p(x_v \mid x_{\text{nd}(v)})$. Hence

$$\begin{aligned} &q_{V\setminus\{v\}}(x_{V\setminus\{v\}} \mid x_v) \\ &\quad \equiv p(x_{\text{nd}(v)}) \cdot p(x_{\text{de}(v)\setminus\{v\}} \mid x_{\text{nd}(v)\cup\{v\}}), \end{aligned}$$

and therefore $q_{V\setminus\{v\}}$ preserves both the marginal distribution of $X_{\text{nd}(v)}$ and the conditional distribution of $X_{\text{de}(v)\setminus\{v\}}$ given $X_{\text{nd}(v)\cup\{v\}}$.

Remark 31. A Gaussian kernel $q(x_S \mid x_{V\setminus S})$ is parameterized via a set of means $E[x_S \mid x_{V\setminus S}]$ indexed by $x_{V\setminus S}$ and variances $\text{Cov}[x_S \mid x_{V\setminus S}]$. There is a distribution naturally associated with $q(x_S \mid x_{V\setminus S})$ given by:

$$p^*_S(x_V) \equiv q_S(x_S \mid x_{V\setminus S}) \prod_{v\in V\setminus S} q^*_v(x_v),$$

where $q^*_v(x_v)$ is an arbitrary marginal distribution.

In what follows we will consider kernels $q_S(x_S \mid x_{V\setminus S})$ derived from a mean zero Gaussian distribution $p(x_V)$, hence parametrized via $\text{Cov}[x_V]$. We will then take $q^*_v(x_v)$ to be the univariate normal distribution $p(x_v)$. It then follows that the Gaussian distribution $p^*_S(x_V)$ corresponding to $q_S(x_S \mid x_{V\setminus S})$ will also be parametrized via a covariance matrix $\text{Cov}^*_S[x_V]$.

Proposition 32. *Every conditional independence that holds in q_S also holds in p^*_S.*

We now show that fixing in the linear SEM corresponds to setting all coefficients corresponding to incoming edges to zero, but keeping all other parameters constant.

Lemma 33. *Let v be fixable in an ADMG $\mathcal{G}$. Then in an SEM corresponding to $\mathcal{G}$, setting $b_{wv} = \omega_{vw} = \omega_{wv} = 0$ for all $w \neq v$ is equivalent to dividing by $p(x_v \mid x_{\mathrm{nd}(v)})/q_v^*(x_v)$.*

Proof. We show that setting parameters to zero as indicated leaves the marginal distribution $p(x_{\mathrm{nd}(v)})$ and the conditional distribution $p(x_{\mathrm{de}(v)\setminus\{v\}} \mid x_{\mathrm{nd}(v)}, x_v)$ unchanged. The former follows easily from the trek rule (see, for example, [7]), since no edge in any trek between non-descendants of v is altered. Similarly, since we choose $\mathrm{Var}\, X_v = \omega_{vv}$ (see Remark 31), and the only trek from v to itself in the fixed graph is the trivial trek; hence ω_{vv} is also preserved.

It remains to show that the same holds for the conditional distribution of the strict descendants given $\{v\} \cup \mathrm{nd}(v)$. To see this, note that it is equivalent to check that the concentration $k_{ij} = (\Sigma^{-1})_{ij}$ remains the same whenever either i or j is a descendant of v. Without loss of generality we may assume that j has no descendants (by marginalizing anything which is not an ancestor of i, j, v). Then we have

$$K = \Sigma^{-1} = (I - B)^T \Omega^{-1} (I - B);$$

denote $\omega^{ij} = \left(\Omega^{-1}\right)_{ij}$. By definition of the model, Ω is a block-diagonal matrix with blocks corresponding to districts in the graph $\mathcal{G}$, and therefore so is Ω^{-1}. Hence k_{ij} (including the case $i = j$) can be written as

$$k_{ij} = \sum_{d=1}^{p} b_{id}\omega^{dj} + \omega^{ij} \quad = \sum_{d \in \mathrm{dis}(j)} b_{id}\omega^{dj} + \omega^{ij}, \quad (1)$$

(corresponding to paths of the form $i \to d \leftrightarrow \cdots \leftrightarrow j$ and $i \leftrightarrow \cdots \leftrightarrow j$ respectively). We claim none of the quantities in (1) are modified by setting the parameters b_{wv} and $\omega_{vw} = \omega_{wv}$ to zero.

Suppose for a contradiction that b_{id} for $d \in \mathrm{dis}_{\mathcal{G}}(j)$ is one of the parameters set to zero; this could happen only if $d = v$. Now, if i is the descendant of v then this would imply a cycle, and if j is the descendant of v then $d \in \mathrm{dis}_{\mathcal{G}}(j)$ implies v is not fixable which is also a contradiction. Hence b_{id} is not set to zero.

Next consider ω^{dj}. If d, j are in different districts then $\omega^{dj} = 0$. Since Ω^{-1} is block diagonal, this parameter will only change if v is in the same district as j and d. If j is a descendant of v then $v \notin \mathrm{dis}_{\mathcal{G}}(j)$ since v is fixable. If i is a descendant of v then so is d, and therefore $v \notin \mathrm{dis}(j) = \mathrm{dis}(d)$ for the same reason. Therefore the quantities ω^{ij}, ω^{dj} all remain unchanged, as they are a function only of the block of Ω corresponding to $\mathrm{dis}(j)$. Hence if either i or j is a descendant of v, then none of the terms in (1) is changed by setting $b_{wv} = \omega_{vw} = \omega_{wv} = 0$ for all $w \neq v$. □

Thus in the context of a linear SEM fixing v corresponds to setting the parameters b_{wv} and $\omega_{vw} = \omega_{wv}$ to zero. We have the following result as a direct consequence:

Theorem 34. *Let $\mathcal{G}$ be an ADMG then $\mathcal{P}_{\mathrm{sem}}(\mathcal{G}) \subseteq \mathcal{P}_n(\mathcal{G}) \cap \mathcal{N}$.*

Recall that $\mathcal{N}$ is the set of multivariate Gaussian distributions with positive definite covariance matrix.

6 ARID SEMS REPRESENT ALL GAUSSIAN NESTED MODELS

The Gaussian nested Markov model associated with an ADMG $\mathcal{G}$ is exactly the linear SEM corresponding to the maximal arid projection $\mathcal{G}^\dagger$ of $\mathcal{G}$:

Theorem 35. *Let $\mathcal{G}$ be an ADMG. Then $\mathcal{P}_{\mathrm{sem}}(\mathcal{G}^\dagger) = \mathcal{P}_n(\mathcal{G}) \cap \mathcal{N}$.*

Proof. By Theorem 34 $\mathcal{P}_{\mathrm{sem}}(\mathcal{G}^\dagger) \subseteq \mathcal{P}_n(\mathcal{G}) \cap \mathcal{N}$. Further, by Theorem 30, $\mathcal{P}_n(\mathcal{G}) = \mathcal{P}_n(\mathcal{G}^\dagger)$. Thus it suffices to prove that $\mathcal{P}_n(\mathcal{G}) \cap \mathcal{N} \subseteq \mathcal{P}_{\mathrm{sem}}(\mathcal{G})$ where $\mathcal{G}$ is maximal and arid.

In order to facilitate our inductive argument, we extend the definitions of $\mathcal{P}_n(\mathcal{G})$ and $\mathcal{P}_{\mathrm{sem}}(\mathcal{G})$ and the result to CADMGs and kernels. Specifically, if $\mathcal{G}(V, W)$ is a CADMG and $\prec$ is a topological ordering on V, then kernel $q_V \in \mathcal{P}_n(\mathcal{G})$ if, for each intrinsic $S \subseteq V$ and $\prec$-maximal $v \in S$,

$$X_v \perp\!\!\!\perp X_{V \setminus (\{v\} \cup \mathrm{mb}_{\mathcal{G}}(v,S))} \mid X_{\mathrm{mb}_{\mathcal{G}}(v,S)}$$

holds in $\phi_{V\setminus S}(q_V(x_V \mid x_W); \mathcal{G})$. Similarly, $\mathcal{P}_{\mathrm{sem}}(\mathcal{G})$ represents a SEM where, if there is an edge $w \to v$ with $w \in W$, $v \in V$ then the equation for X_v contains $b_{wv}x_w$ as a summand. We now claim that if $\mathcal{G}$ is a CADMG then $\mathcal{P}_{\mathrm{sem}}(\mathcal{G}^\dagger) = \mathcal{P}_n(\mathcal{G}) \cap \mathcal{N}$, where $\mathcal{N}$ is the set of Gaussian kernels. This is clearly sufficient.

Suppose $p \in \mathcal{P}_n(\mathcal{G}) \cap \mathcal{N}$, where $\mathcal{G}(V, W)$ is a CADMG with topological ordering $\prec$.

If $|V| = 1$ then the result follows by regression on $X_{\mathrm{pa}_{\mathcal{G}}(v)}$. Otherwise we proceed by induction on $|V|$. Let v be the maximal vertex under $\prec$.

For any fixable w in $\mathcal{G}$ we obtain by the induction hypothesis that $q_{V\setminus\{w\}} \in \mathcal{P}_n(\phi_w(\mathcal{G})) = \mathcal{P}_{\mathrm{sem}}(\phi_w(\mathcal{G}))$. Hence we can identify parameters for edges not involving v by fixing v. Any directed edge parameter b_{ij} can be identified provided j is not fixed; if $|V| \geq 2$ then, since $\mathcal{G}$ is arid, it contains at least two fixable vertices by Proposition 11. Hence we can identify every b_{ij} in this manner. [Since valid fixings commute, all such results will agree.]

Similarly we can identify any bidirected edge this way except possibly ω_{wv} if w and v are the only two fixable vertices in $\mathcal{G}$. In this case, $\mathcal{G}$ contains only one district and every vertex is an ancestor of v or w.

Since we have identified every other parameter, let $\tilde{p}_\gamma$ be the distribution obtained from all the other parameters with $\omega_{vw} = \gamma$. Then by construction, $\tilde{p}_\gamma$ and p have the same margins over $V \setminus \{v\}$ and $V \setminus \{w\}$. If we can choose γ so that the covariance σ_{vw} matches that in p, then $p = \tilde{p}_\gamma$. It is not hard to see that σ_{vw} is a monotonic function of ω_{vw}, so the only restriction is on the positive definiteness of the relevant covariance matrices. Since the set of positive definite matrices is convex, the set of valid ω_{vw} is an interval; in addition, Ω is positive definite if and only if Σ is positive definite. Hence, by the intermediate value theorem there exists a γ that maps to the appropriate σ_{vw}. □

Example 36. To see the difficulty with the induction in the previous proof, consider the graph in Fig. 3 (c). There are two fixable vertices, b and c, and in either case the fixing corresponds to marginalizing over the corresponding random variable. This means that, in either case, the edge parameter ω_{bc} is not identifiable. Every other parameter can be identified inductively, and we may finally use σ_{bc} to identify ω_{bc}.

7 EXAMPLE

Consider the following simplified medical trial to examine the effect of diet and exercise on diabetes, adapted from [5]. At baseline, patients are randomly assigned to perform t hours of exercise in a week, but actually perform x hours. At the end of the week their blood pressure (bp) is measured, this is assumed to depend upon x, but also to be confounded with it by lifestyle factors. In the second phase of the trial, patients are assigned to lose Δbmi kilograms in weight; the value of Δbmi is random, but for ethical reasons depends linearly on x and bp. Finally, at the end of the trial, triglyceride levels (y) are measured, which is used to diagnose diabetes; these are assumed to be correlated with blood pressure, and dependent on exercise and weight loss. This causal structure naturally yields the ADMG shown in Fig. 4(a).

Though a perfectly reasonable causal description of the model, Fig. 4(a) contains a bow and therefore the associated model is non-smooth and not everywhere identifiable. Performing maximal arid projection gives the graph $\mathcal{G}^\dagger$ in Fig. 4(b), which gives an SEM that induces a curved exponential family and is nested Markov equivalent to the SEM corresponding to the original graph. Note that the resulting graph is not an ancestral graph; indeed $\mathcal{G}^\dagger$ preserves more of the structure than the corresponding MAG.

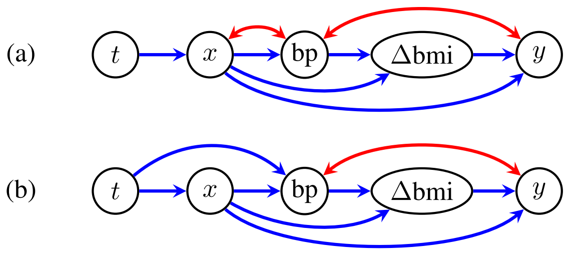

Figure 4: (a) A graph representing a clinical trial for interventions in diabetes; the associated SEM is non-smooth. (b) A nested Markov equivalent graph whose SEM represents a curved exponential family.

8 DISCUSSION

We have presented a subclass of ADMGs—the maximal arid graphs (MArGs)—that fully represents the class of nested Markov models. We have shown that any linear SEM associated with a MArG is precisely equal to the class of Gaussian densities in the nested Markov model for that MArG.

We remark that the results on arid graphs we derived in Section 4 are completely non-parametric, and apply not only to the Gaussian models that we study here, but to any model; we showed that any nested Markov equivalence class contains a MArG which, since MArGs are maximal and simple graphs, is a canonical representative of the class to use in search procedures within the set of nested models. In this sense MArGs serve a similar role to MAGs for scoring-based searches for ordinary Markov models corresponding to ADMGs.

Acknowledgments

The authors would like to thank the American Institute of Mathematics for supporting this research via the SQuaRE program. Evans was supported by an EPRSC First Grant. Shpitser was supported in part by NIH grants R01 AI104459-01A1 and R01 AI127271-01A1. Richardson was supported by the U.S. Office of Naval Research grant N00014-15-1-2672.

References

[1] C. Brito and J. Pearl. A graphical criterion for the identification of causal effects in linear models. In *Eighteenth National Conference on Artificial intelligence*, pages 533–538, 2002.

[2] C. Brito and J. Pearl. Graphical condition for identification in recursive SEM. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 47–54, 2006.

[3] B. Chen. Identification and overidentification of linear structural equation models. In *Advances in Neural Information Processing Systems*, volume 29, pages 1579–1587, 2016.

[4] B. Chen, J. Tian, and J. Pearl. Testable implications of linear structural equation models. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2424–2430, 2014.

[5] M. Drton, M. Eichler, and T. S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(Oct):2329–2348, 2009.

[6] M. Drton, C. Fox, and A. Käufl. Comments on: Sequences of regressions and their independencies. *Test*, 21(2):255–261, 2012.

[7] M. Drton, R. Foygel, and S. Sullivant. Global identifiability of linear structural equation models. *Annals of Statistics*, 39(2):865–886, 2011.

[8] M. Drton and L. Weihs. Generic identifiability of linear structural equation models by ancestor decomposition. *Scand. J. Statist.*, 2016.

[9] R. J. Evans and T. S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, pages 1–30, 2014.

[10] R. J. Evans and T. S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 2018. (to appear).

[11] R. Foygel, J. Draisma, and M. Drton. Half-trek criterion for generic identifiability of linear structural equation models. *Annals of Statistics*, 40(3):1682–1713, 2012.

[12] L. D. Garcia-Puente, S. Spielvogel, and S. Sullivant. Identifying causal effects with computer algebra. In *Proceedings of the Twenty-sixth Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2010.

[13] T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12, 1943.

[14] Y. Huang and M. Valtorta. Pearl's calculus of interventions is complete. In *Twenty Second Conference On Uncertainty in Artificial Intelligence*, 2006.

[15] J. T. Koster. On the validity of the Markov interpretation of path diagrams of gaussian structural equations systems with correlated errors. *Scandinavian Journal of Statistics*, 26(3):413–431, 1999.

[16] C. Nowzohour, M. H. Maathuis, R. J. Evans, and P. Bühlmann. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11(2):5342–5374, 2017.

[17] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

[18] T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30:962–1030, 2002.

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

[20] T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. Working paper, `https://arxiv.org/abs/1701.06686v2`, 2017.

[21] I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto, 2006.

[22] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer Verlag, New York, second edition, 2001.

[23] P. Spirtes, T. S. Richardson, C. Meek, R. Scheines, and C. Glymour. Using path diagrams as a structural equation modeling tool. *Sociological Methods & Research*, 27(2):182–225, 1998.

[24] J. Tian. Parameter identification in a class of linear structural equation models. In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1970–1975. AAAI Press, Palo Alto, CA, 2009.

[25] T. van Ommen and J. M. Mooij. Algebraic equivalence of linear structural equation models. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI-17)*, 2017.

[26] T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (UAI-90)*, 1990.

[27] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

[28] J. Zhang. Generalized do-calculus with testable causal assumptions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 667–674, 2007.