

Latent-Free Equivalent mDAGs

Robin J. Evans

June 3, 2023

Abstract

We show that the marginal model for a discrete directed acyclic graph (DAG) with hidden variables is distributionally equivalent to another fully observable DAG model if and only if it does not induce any non-trivial inequality constraints.

1 Introduction

The *marginal model* of a directed acyclic graph (DAG) model with latent variables is defined simply as the set of distributions that are realizable as margins over the observed variables, from those joint distributions that are Markov with respect to the whole graph and where no restrictions are placed on the state-space of the latents. It was shown by Evans (2016) that we can represent this class of models using a collection of hypergraphs known as mDAGs (standing for *marginal* DAGs).

Much is known about the properties of these models. For example, in the discrete and Gaussian cases the models are semi-algebraic, meaning that the equalities and inequalities that define them are all polynomials in the joint probabilities or covariance matrix respectively. The equality constraints in the discrete case are understood (Evans, 2018), and there are methods for finding (in principle) all inequality constraints as well (Wolfe et al., 2019; Navascués and Wolfe, 2020). However, it is still an important open problem to determine whether or not two marginal models are equivalent.

A specific question that may be of interest in this respect, is whether or not the marginal model of a DAG with observed variables V and latent variables L is *distributionally equivalent* to another DAG over only V . The question of understanding *distributional equivalence classes* of models is a significant open problem, and is a critical component of causal model search. We cannot hope to choose between two models from data if they are distributionally equivalent, so any contribution to understanding when this occurs is extremely useful. In addition, for the purpose of finding the most efficient influence function in semiparametric statistics, for example, this is much easier if the model is known to be (equivalent to) a DAG model, because the tangent cone can be easily decomposed into pieces that correspond to each variable conditional precisely upon its parents (Tsiatis, 2006, Section 4.4).

We show in this paper that, if the observed variables are all discrete, this is true if and only if the marginal model does not induce any inequality constraints, beyond those already implied by the required equality constraints and the necessity of probabilities being non-negative. This main result is stated in the following theorem; note that $\mathcal{M}(\mathcal{G})$ denotes the collection of distributions that satisfy the marginal Markov property (Definition 2.3) with respect to the mDAG \mathcal{G} (Definition 2.1).

Theorem 1.1. *Let \mathcal{G} be an mDAG with vertices V , inducing a model $\mathcal{M}(\mathcal{G})$ over a collection of discrete random variables X_V . Then there exists a DAG \mathcal{H} such that $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{H})$ if and only if $\mathcal{M}(\mathcal{G})$ is described entirely by probability distributions that satisfy a finite number of equality constraints.*

The ‘only if’ direction is trivial, since DAG models do not imply any inequalities, and are defined entirely by a finite list of ordinary conditional independences.

In Section 2 we present necessary concepts relating to DAGs and mDAGs, including distributional equivalence. In Section 3 we introduce the ‘nested’ Markov model, and show that any model with a non-trivial nested constraint can be reduced to a model with only standard conditional independences that are not consistent with any DAG. In Section 4 we prove our main result, and in Section 5 we consider possible extensions to continuous random variables.

2 Basics concepts for DAGs and mDAGs

We consider mixed (hyper)graphs with one set of vertices V , and (up to) two edge sets \mathcal{D} and \mathcal{B} ; the set \mathcal{D} contains ordered pairs of vertices, and \mathcal{B} is a simplicial complex over the set V .

Definition 2.1. In a *directed graph* $\mathcal{G} = (V, \mathcal{D})$, if $(v, w) \in \mathcal{D}$ then we write $v \rightarrow w$ and say that v is a *parent* of w , and w a *child* of v . The set of parents of w in \mathcal{G} is denoted by $\text{pa}_{\mathcal{G}}(w)$. A *directed walk with length k* is a sequence of vertices v_0, \dots, v_k such that each v_i is a parent of v_{i+1} . A directed graph is said to be *acyclic* if there are no directed walks of length $k \geq 1$ from any vertex back to itself; we call such an object a *directed acyclic graph* (DAG).

An *mDAG* is a DAG (V, \mathcal{D}) together with a simplicial complex \mathcal{B} over V . We refer to the entries of \mathcal{B} as *bidirected faces*, and the maximal entries as *bidirected facets*. If a face contains two vertices we may also call it a *bidirected edge*.

An example of an mDAG consisting of a DAG with 4 edges and the bidirected facets $\{a, b\}$, $\{a, c, e\}$ and $\{d, f\}$ is shown in Figure 1(i). Note that we use blue to draw directed edges, and red for the bidirected facets.

2.1 Marginal models

We first define what it means for a distribution to be Markov with respect to a DAG.

Definition 2.2. A distribution p over random variables X_V is said to be *Markov* with respect to a directed acyclic graph \mathcal{G} if there is a topological ordering \prec of V such that

$$X_v \perp\!\!\!\perp X_{\text{pre}(v; \prec) \setminus \text{pa}(v)} \mid X_{\text{pa}(v)} \text{ under } p$$

for each $v \in V$, where $\text{pre}_{\mathcal{G}}(v; \prec) = \{w \in V : w \prec v\}$.

Note that we omit the subscripts on operators when they are themselves written in a subscript and the meaning is clear. We remark that if Definition 2.2 holds for one topological ordering, then it can be shown using standard implications of conditional independences that it holds for every other topological ordering (Lauritzen et al., 1990).

Let \mathcal{G} be an mDAG, and let $\bar{\mathcal{G}}$ denote the *canonical DAG* for \mathcal{G} . That is, we replace each bidirected facet B with a latent variable that has the set of children B ; see Figure 1(ii) for the canonical DAG associated with the mDAG in 1(i). We colour the edges similarly to in the mDAG: if an edge is between two observed vertices it is blue, and otherwise it is red.

Definition 2.3. We define the *marginal model* for \mathcal{G} as the set of distributions that can be obtained as a margin over the observed variables in \mathcal{G} under a distribution that is Markov with respect to $\bar{\mathcal{G}}$. This set of distributions is denoted $\mathcal{M}(\mathcal{G})$.

This model is defined in Evans (2016), and its properties and the sufficiency of mDAGs for representing such models are laid out more fully in that paper. We remark that the state-space of the latent variables is in principle arbitrary, but that a uniform random variable on $(0, 1)$ always has sufficiently large cardinality. A result of Rosset et al. (2018) shows that if all the variables are discrete with a finite state-space, then there is a corresponding finite bound on the cardinality of the latent variables.

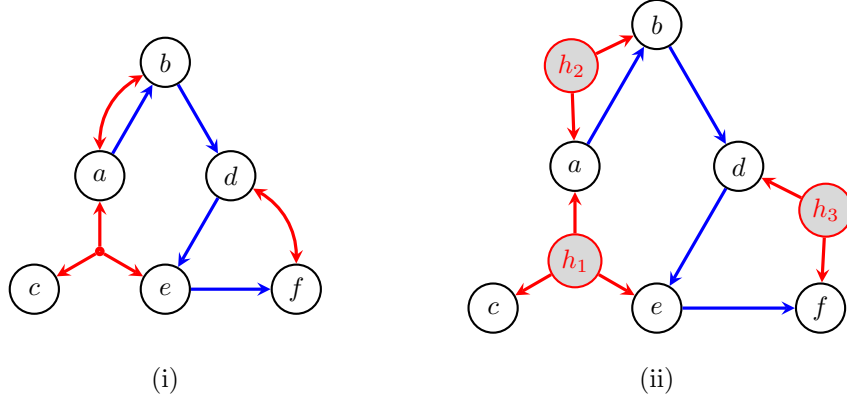


Figure 1: (i) An mDAG and (ii) its canonical DAG. Note that (i) is also the latent projection of (ii) over $\{a, b, c, d, e, f\}$.

2.2 Distributional and Markov equivalence

Given an mDAG, one can read off the conditional independences that are satisfied by distributions that are Markov to it using *m-separation*. For readers familiar with d-separation in directed graphs, it is essentially the same; like d-separation it is based on whether there is an *open path* between two variables, or whether all such paths are *blocked*. The only modification is that the definition of a collider and non-collider has to be expanded to take account of bidirected facets. The full definition is given in Appendix A.1.

Definition 2.4. We say that two mDAGs \mathcal{G} and \mathcal{G}' are *distributionally equivalent* if $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{G}')$. The ordinary conditional independences implied by $\mathcal{M}(\mathcal{G})$ are used to define the *ordinary Markov* model for \mathcal{G} . We say that two mDAGs are *ordinary Markov equivalent* if they imply the same set of conditional independences (i.e. they exhibit the same collection of m-separations.)

If two graphs are distributionally equivalent then they are also ordinary Markov equivalent. See Proposition A.13 for a comparison between these two models, as well as the ‘nested’ Markov model (see Section 3).

Example 2.5. Consider the mDAG \mathcal{G} shown in Figure 2(i). We can see that $a \perp_m c \mid b$ and so therefore if $p \in \mathcal{M}(\mathcal{G})$ it holds that $X_a \perp\!\!\!\perp X_c \mid X_b$. Note that there is no way to m-separate a and d in this graph, because there is a directed path via b and c , and if we condition on either of these vertices then a path $a \rightarrow b \leftrightarrow d$ will be opened up.

In fact there *is* a constraint between X_a and X_d , but it is only revealed after fixing the vertex c (see Sections 3 and A.6 for more detail); this yields the graph in (ii), which shows that now $d \perp_m a \mid c$, so there is a *nested* constraint: $X_d \perp\!\!\!\perp X_a \mid X_c$ after fixing $X_c \mid X_b$.

In addition, the model implied by the graph in Figure 2(ii) contains an inequality constraint, being the Clauser-Horne-Shimony-Holt (CHSH) inequality (Clauser et al., 1969). This says that if (for example) all four variables take values in $\{-1, +1\}$, then

$$\begin{aligned} -2 \leq & \mathbb{E}[X_b X_d \mid X_a = -1, X_c = +1] + \mathbb{E}[X_b X_d \mid X_a = +1, X_c = -1] \\ & + \mathbb{E}[X_b X_d \mid X_a = -1, X_c = -1] - \mathbb{E}[X_b X_d \mid X_a = +1, X_c = +1] \leq 2. \end{aligned} \quad (*)$$

Note however that a distribution exists satisfying the two independences given, but for which this quantity in $(*)$ attains the value 4: set $P(X_b = -X_d = \pm 1) = \frac{1}{2}$ if $X_a = X_c = +1$, and $P(X_b = X_d = \pm 1) = \frac{1}{2}$ otherwise, and one can verify that the two independences mentioned are satisfied, and that each term in $(*)$ has the value $+1$. In this sense the inequalities are *non-trivial*, because they are not implied by any of the equality constraints.

Remark 2.6. We can also read off some inequalities using a generalization of m-separation called *e-separation* (Evans, 2012); this involves first deleting a set of variables D , and then checking for

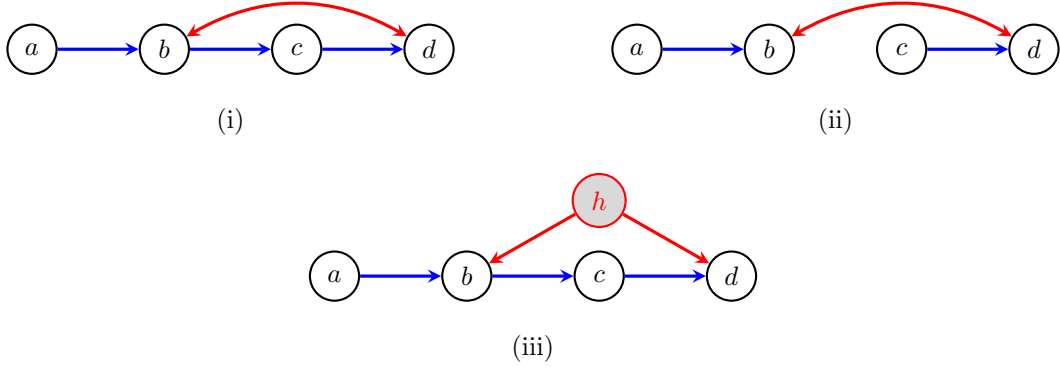


Figure 2: (i) An mDAG exhibiting all three kinds of constraint: a conditional independence ($X_a \perp\!\!\!\perp X_c \mid X_b$), a nested conditional independence ($X_d \perp\!\!\!\perp X_a \mid X_c$ after fixing $X_c \mid X_b$), and an inequality constraint (see Example 2.5). (ii) The graph from (i) after c has been fixed. (iii) The canonical DAG for the mDAG in (i).

m-separation in the resulting graph. If $A \perp_m B \mid C$ in the graph \mathcal{G} after removing the vertices D (and all edges incident to vertices in D) we denote it by $A \perp_e B \mid C \not\parallel D$; see Appendix A.2 for more details on the resulting constraints. The *instrumental inequality* of Pearl (1995) can be read off using this criterion, although $(*)$ cannot.

2.3 Equivalence

Here we give some examples of (non-)equivalence of the marginal models for different mDAGs. Consider the graphs in Figure 3, which are all ordinary Markov equivalent; the mDAGs in (i) and (ii) can be shown to be equivalent to the DAG in (iii).

For (i), note that the only constraint in (iii) is that $X_a, X_c \perp\!\!\!\perp X_d$. This can clearly be achieved by (i) just by setting the implied latent variable to tell a and c what values they each take, and then pass this information onto b . Since X_a and X_c are determined jointly, this clearly allows any distribution such that the constraint holds to be attained in the model for the graph in (i).

For (ii), first note that it is clearly equivalent for the (implied) latent variable between b and d to simply contain the value of X_d . Hence the edge between b and d can be the same as in (i) and (iii). Then, similarly, the latent variable between a and c can just contain X_a , so again we can replace it with a directed edge as in (iii). Now, for the final bidirected edge between b and c , note that b needs to know what value c will take; this can be arranged by making the latent variable be a map telling X_c what to do for each value of X_a . If this information is passed to b , then (since it can see X_a directly) it can compute what X_c must be. Hence, we obtain equivalence between the two models.

The graph in (iv) is *not* equivalent to the other three, because the induced subgraph over $\{a, b, c\}$ implies an inequality constraint (Fritz, 2012; Evans, 2016).

3 Nested Markov model

As we have seen in Example 2.5, there are two types of *equality* constraint that can be obtained in an mDAG. The first is an ordinary conditional independence, and the second is a (strictly) *nested* constraint, which is a conditional independence that arises only after probabilistically ‘fixing’ some of the other variables. We now define this operation more formally.

Definition 3.1. The *Markov blanket* of a vertex v in an mDAG \mathcal{G} is the set of (other) vertices w that can be reached by a walk whose internal vertices are all *colliders*; that is $w \rightarrow v$ or $w \leftrightarrow \dots \leftrightarrow v$

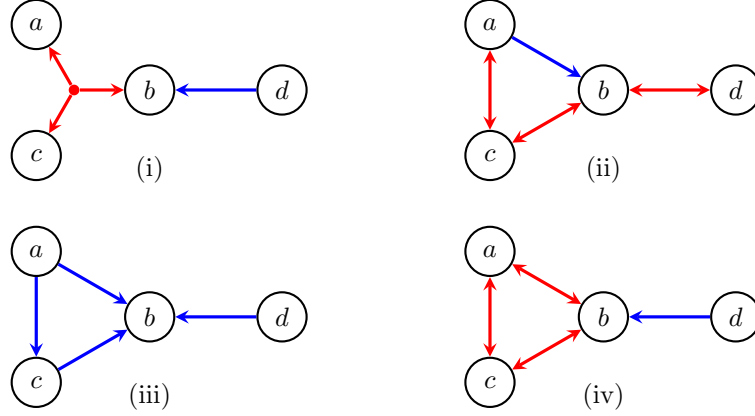


Figure 3: (i) and (ii) mDAGs that are equivalent to the DAG (iii). (iv) is an mDAG that is *not* equivalent to any DAG.

or $w \rightarrow \leftrightarrow \dots \leftrightarrow v$ (where $w \leftrightarrow v$ is shorthand for v, w being contained in the same bidirected facet.) We denote this set by $\text{mb}_{\mathcal{G}}(v)$.

We say a vertex v is *fixable* (in \mathcal{G}) if it has no strict descendants (i.e. vertices that can be reached by a directed walk from v) that can also be reached by walks over only bidirected edges.

Given such a vertex, we can *fix* it in the graph by removing all incoming edges (whether directed or bidirected), but keeping any directed edges oriented *out* of v ; let this new graph be \mathcal{G}^* . Probabilistically, we compute

$$p^*(x_V) = \frac{p^*(x_v)}{p(x_v | x_{\text{mb}(v)})} \cdot p(x_V),$$

where $p^*(x_v)$ is an arbitrary strictly positive marginal density over \mathfrak{X}_v .

Results from Richardson et al. (2023) tell us that if p is in the marginal model for \mathcal{G} , then p^* will be in the marginal model for \mathcal{G}^* . Hence any non-trivial constraints we deduce on p^* must also apply to p .

3.1 Nested models are not DAG-like

In this section we show that any non-trivial nested constraint in an mDAG \mathcal{G} will imply that the conditional independence model after fixing cannot be represented by any DAG.

Proposition 3.2. *Suppose that fixing a vertex v from an mDAG \mathcal{G} leads directly to a non-trivial nested constraint. Then the conditional independence model implied by \mathcal{G} after the fixing is not faithfully represented by any DAG model.*

Proof. When we fix v we multiply by $p^*(x_v)/p(x_v | x_{\text{mb}(v)})$, and so we artificially introduce the independence $X_v \perp\!\!\!\perp X_{\text{mb}(v)}$ by performing the fixing. Let the new constraint be $X_A \perp\!\!\!\perp X_B | X_C$, where each of A , B and C are chosen to be inclusion minimal; that is, if any vertex is removed from A or B then the independence also held in some form before the fixing (possibly with a different conditioning set), and if from C then the required m-separation no longer holds in the new graph.

Since the new constraint $X_A \perp\!\!\!\perp X_B | X_C$ is non-trivial, it cannot have been induced just by deleting paths through v , so there exists a path π from $a \in A$ to $b \in B$ *not* through v , that was previously open given $A' \cup B' \cup C$, but is now blocked (here $A' = A \setminus \{a\}$ and $B' = B \setminus \{b\}$). Hence there is a set of colliders S on π that were ancestors of v in \mathcal{G} , but are not after the fixing, and hence no longer ancestors of things in $A \cup B \cup C$.

Then choose $D = A' \cup B' \cup C \cup S'$, where S' is a maximal subset such that $a \perp_m b \mid A' \cup B' \cup C \cup S'$ in \mathcal{G}^* , but not if we add in another element $s \in S \setminus S'$. Clearly $S \setminus S' \neq \emptyset$ from the discussion in the previous paragraph. Now we can apply Proposition A.15 to obtain the result. \square

4 mDAGs without nested constraints

Now, we need only prove that models whose equality constraints are equivalent to those of an mDAG model (and not ordinary Markov equivalent to a conditional DAG model) will induce some sort of non-trivial inequality in their marginal model. We can do this by assuming that we consider the ‘final’ fixing that to reveal a non-trivial nested constraint, and then look at the independence model that this induces.

4.1 Partial ancestral graphs

If \mathcal{G} does not have any nested constraints, then we consider its *partial ancestral graph* (PAG) $[\mathcal{G}]$, which represents precisely the ordinary conditional independence constraints implied by \mathcal{G} . There is a one-to-one correspondence between PAGs and conditional independence models induced by mDAGs (Richardson and Spirtes, 2002, 2003). PAGs are ordinary mixed graphs (i.e. they do not contain hyper-edges) with three edge markings: a tail, an arrowhead and a circle; a circle means that at least one *maximal ancestral graph* (MAG) in the equivalence class has a tail mark here, and at least one has an arrowhead. See Figures 4 and 5 for some examples. More details about MAGs and PAGs are given in Appendices A.4 and A.5. The crucial fact here is that the conditional independence structure of any mDAG can always be represented by a MAG, and therefore by a PAG.

Proposition 4.1. *Suppose that $\mathcal{P} = [\mathcal{G}]$. Then the conditional independence structure of \mathcal{G} is the same as that of a DAG if and only if \mathcal{P} does not have any bidirected edges.*

Proof. We know from Lemma 3.3.4 of Zhang (2006) that a PAG can always be oriented to a MAG in such a way that it does not introduce any additional bidirected edges. Hence, if there are none to start with, the model is ordinary Markov equivalent to a DAG.

For the converse, note that if it were false that would imply that the edge is bidirected in every Markov equivalent MAG, which contradicts the existence of a Markov equivalent DAG. \square

Now, since the PAG represents *invariant* edges (i.e. ones that are the same in all members of the equivalence class), the graph is ordinary Markov equivalent to a DAG if and only if there are no bidirected edges in its PAG.

We say that a collider path $\langle v_0, \dots, v_k \rangle$ is *locally* unshielded if there is no edge between v_i and v_{i+2} for any $i = 0, \dots, k-2$.

Proposition 4.2. *Suppose that \mathcal{G} contains no non-trivial nested constraints, and that there is a bidirected edge in $\mathcal{P} = [\mathcal{G}]$. Then a non-trivial inequality constraint is induced over the distributions in $\mathcal{M}(\mathcal{G})$.*

Proof. There are two reasons that a bidirected edge can be included in a PAG. Either there is a locally unshielded collider path of length 3 from (say) a to d (see Figure 4), or there is a discriminating path of length at least 3 (see Figure 5). In the first case, the PAG must have an induced subgraph of one of the forms in Figure 4. The graphs in (i) and (ii) induce the CHSH inequality $(*)$ (Bell, 1964; Clauser et al., 1969).

For (iii) and (iv), consider the submodel in which all information about X_d is contained as part of X_c . This means that X_d must obtain all its information from the latent it shares with X_c , since either X_c or X_d is marginally independent of all other variables. Hence we can remove the edge

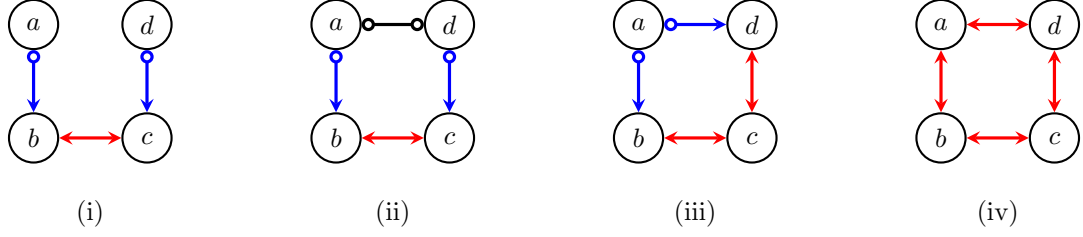


Figure 4: Up to symmetry, the four possible induced subgraphs of a PAG containing a locally unshielded collider path of length 3 from a to d .

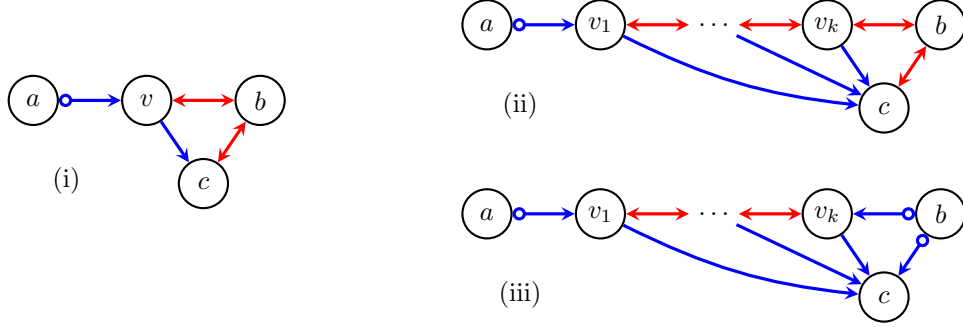


Figure 5: Discriminating paths for b : (i) a path of length 3 (containing a bidirected edge); and (ii)–(iii) two possible configurations of length $k + 2$.

between a and d , and then note that the mDAGs become distributionally equivalent to Figure 4(i). Hence this submodel induces the CHSH inequality, and so the whole distribution also satisfies an inequality.

On the other hand, suppose that there is no locally unshielded collider path of length $k \geq 3$ but there is a discriminating path $\langle a, v_1, \dots, v_k, b, c \rangle$ with $k \geq 1$. In fact, by Proposition B.1, if these conditions are satisfied, then there will also be one an induced subgraph that looks like Figure 5(i). Note that the m-separations for this subgraph imply that $X_a \perp\!\!\!\perp X_b$ and $X_a \perp\!\!\!\perp X_c \mid X_v$; a distribution over binary variables that satisfies both of these constraints would be to have $X_a + X_b + X_c \equiv 0$ (where \equiv denotes equality modulo 2), and $P(X_v = 0) = 1$. However, there is also an e-separation constraint between a and $\{b, c\}$ if we delete v , and the corresponding inequality constraint is *not* satisfied by this distribution. Hence, there is indeed a non-trivial inequality. \square

The proof technique used for the graphs in Figures 4(iii) and (iv) is known as the ‘Fritz trick’*, because it is a generalization of the approach that Tobias Fritz uses in Proposition 2.13 of Fritz (2012).

4.2 Proof of the main result

We now have enough information to prove our main result.

Proof of Theorem 1.1. From the results in Section 3 we know that if there is a non-trivial nested constraint, then the set of ordinary independences induced after a final fixing are not ones that can be represented faithfully by a DAG model.

Then for such models, as well as other models without nested conditional independences, we can always be represent the conditional independence structure by a partial ancestral graph. If there is a necessary bidirected edge then this induces a non-trivial inequality constraint (Proposition 4.2).

*This is a term coined by members of the Perimeter Institute, including Elie Wolfe.

Since Proposition 4.1 tells us that the presence of a bidirected edge in the PAG implies there is no DAG that can represent the equivalence class, this proves that not having a marginal model that is not Markov equivalent to a DAG implies the existence of a non-trivial inequality.

For the converse the result is trivial, since DAG models are defined by the finite list of independences in Definition 2.2. \square

Now we have proven our main result. Marginal DAG models can be categorized into several classes: (i) those which are distributionally equivalent to a DAG; (ii) those with additional inequality constraints only; and (iii) graphs with non-DAG-like conditional independences or (iv) graphs with nested conditional independences, both of which induce inequalities.

5 Extension to the continuous case

One obvious question for an extension to this paper is to ask whether or not the result also holds in the case of variables that are not discrete. Bell inequalities (i.e. ones analogous to the CHSH inequality) are known to hold even if all the variables are continuous (Cavalcanti et al., 2007), and indeed hold on arbitrary discretizations of such variables.

However, there are obstacles to generalizing this result to the continuous case. The first is that the results of Evans (2018) only apply to models where all the observed variables are discrete. Another is that results of Rosset et al. (2018) enable one to show that the model is semi-algebraic if observed variables have a finite state-space, so for continuous (or even countably infinite) state-spaces we would need an analogous condition. The final problem is that e-separation results require the distribution of the variables deleted to have at least one atom, even if the other variables are continuous. Indeed, it is an open question whether inequalities are contained in models such as the one induced by the mDAG in Figure 5(i) when X_v is continuous. The Shannon-cone of this model does not induce any non-trivial entropic inequalities in that case, for example (Chaves et al., 2014).

Acknowledgements

We thank Richard Guo for suggesting the problem and Elie Wolfe for conjecturing the result. This work was largely completed while the author was a visiting researcher at the Simons Institute in Berkeley, California.

References

- J. S. Bell. On the Einstein-Podolsky-Rosen paradox. *Physics*, 1(3):195, 1964.
- E. G. Cavalcanti, C. J. Foster, M. D. Reid, and P. D. Drummond. Bell inequalities for continuous-variable correlations. *Physical Review Letters*, 99(21):210405, 2007.
- R. Chaves, L. Luft, T. O. Maciel, D. Gross, D. Janzing, and B. Schölkopf. Inferring latent structures via information inequalities. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI-14)*, 2014.
- T. Claassen and I. G. Bucur. Greedy equivalence search in the presence of latent confounders. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, 2022.
- J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt. Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, 23(15):880, 1969.
- R. J. Evans. Graphical methods for inequality constraints in marginalized DAGs. In *Machine Learning for Signal Processing*. IEEE, 2012.

- R. J. Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43: 625–648, 2016.
- R. J. Evans. Margins of discrete Bayesian networks. *Annals of Statistics*, 46(6A):2623–2656, 2018.
- T. Fritz. Beyond Bell’s theorem: correlation scenarios. *New Journal of Physics*, 14(10):103001, 2012.
- S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.
- M. Navascués and E. Wolfe. The inflation technique completely solves the causal compatibility problem. *Journal of Causal Inference*, 8(1):70–91, 2020.
- J. Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 435–443, 1995.
- T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.
- T. S. Richardson and P. Spirtes. Causal inference via ancestral graph models. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, chapter 3, pages 83–105. OUP, 2003.
- T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. *Annals of Statistics (accepted)*, 2023. arXiv:1701.06686.
- D. Rosset, N. Gisin, and E. Wolfe. Universal bound on the cardinality of local hidden variables in networks. *Quantum Information and Computation*, 18(11-12):0910–0926, 2018.
- A. A. Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.
- E. Wolfe, R. W. Spekkens, and T. Fritz. The inflation technique for causal inference with latent variables. *Journal of Causal Inference*, 7(2), 2019.
- J. Zhang. *Causal inference and reasoning in causally insufficient systems*. PhD thesis, Carnegie Mellon University, 2006.
- J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 16–17(172):1873–1896, 2008.

A Definitions for mDAGs

A.1 Basic definitions and m-separation

Let $\mathcal{G} = (V, \mathcal{D}, \mathcal{B})$ be a mixed (hyper-)graph with directed edges \mathcal{D} and bidirected simplicial complex \mathcal{B} .

Definition A.1. A *path* in \mathcal{G} is a sequence of edges and (distinct) vertices $\langle v_0, e_1, v_1, e_2, \dots, e_k, v_k \rangle$, such that $v_{i-1}, v_i \in e_i$ for $i = 1, \dots, k$. A path is *directed* if each e_i is $v_{i-1} \rightarrow v_i$. The *length* of the path is k (the number of edges in it), and this can be zero.

Definition A.2. Given a vertex $v \in V$ in an mDAG \mathcal{G} we define

$$\begin{aligned} \text{pa}_{\mathcal{G}}(v) &= \{w : w \rightarrow v \text{ in } \mathcal{G}\} \\ \text{an}_{\mathcal{G}}(v) &= \{w : w \rightarrow \dots \rightarrow v \text{ in } \mathcal{G} \text{ or } w = v\} \\ \text{and} \quad \text{de}_{\mathcal{G}}(v) &= \{w : v \rightarrow \dots \rightarrow w \text{ in } \mathcal{G} \text{ or } w = v\} \end{aligned}$$

to be respectively the *parents*, *ancestors* and *descendants* of v .

We use $v \leftrightarrow w$ as a shorthand to denote that v and w are contained within some bidirected facet. Then define

$$\begin{aligned} \text{sib}_{\mathcal{G}}(v) &= \{w : w \leftrightarrow v \text{ in } \mathcal{G}\} \\ \text{and} \quad \text{dis}_{\mathcal{G}}(v) &= \{w : w \leftrightarrow \dots \leftrightarrow v \text{ in } \mathcal{G} \text{ or } w = v\} \end{aligned}$$

to be the *siblings* and *district* of v respectively. Siblings of v are vertices for which a latent ‘parent’ is shared, and the districts are easily identified as maximal connected red components in the graph.

Definition A.3. Given a path π of length k , an internal vertex v_i (i.e. not v_0 or v_k) is said to be a *collider* on the path if the adjacent edges e_i, e_{i+1} have arrowheads at v_i . Otherwise an internal vertex is a *non-collider*.

A path from a to b is said to be *open* given a set C if no non-colliders on the path are in C , and any collider is in the set of vertices that can reach C via a directed path (possibly of length zero). Otherwise the path is *blocked*.

We say that sets of vertices A and B are *m-separated* given a set C if every path from any $a \in A$ to any $b \in B$ is blocked by C . We denote this by $A \perp_m B \mid C$.

A.2 Random variables and constraints

We consider random variables $X_V = (X_v)_{v \in V}$ taking values in a finite-dimensional Cartesian product space $\mathcal{X}_V := \times_{v \in V} \mathcal{X}_v$.

Definition A.4. A distribution p is said to satisfy the *global Markov property* for an mDAG \mathcal{G} if whenever A, B, C are disjoint subsets of the vertices of \mathcal{G} and $A \perp_m B \mid C$, we have the corresponding conditional independence $X_A \perp\!\!\!\perp X_B \mid X_C$ under p .

We can extend m-separation to *e-separation* (or *extended m-separation*) by first deleting some variables and their incident edges, and then checking for m-separations among what remains.

Definition A.5. We say sets of vertices A and B are *e-separated* given a set C and after deletion of D if every path from any $a \in A$ to any $b \in B$ is either blocked by C or passes through a node in D . We denote this by $A \perp_e B \mid C \not\parallel D$.

Then a result from Evans (2012) tells us that an e-separation will induce (at least) an inequality constraint on p .

Theorem A.6. Suppose that a distribution p lies in the marginal model of an mDAG \mathcal{G} , and that the e-separation $A \perp_e B \mid C \not\parallel D$ holds in \mathcal{G} , where X_D takes values in a finite set. Then, for every $x_D \in \mathcal{X}_D$, we have that there exists a distribution p^{x_D} such that:

$$p(y_{V \setminus D}, x_D) = p^{x_D}(y_{V \setminus D}, x_D) \quad \text{for all } y_{V \setminus D} \in \mathcal{X}_{V \setminus D},$$

and $X_A \perp\!\!\!\perp X_B \mid X_C$ under p^{x_D} .

If we consider a distribution in which $X_D = x_D$ for some arbitrary state $x_D \in \mathcal{X}_D$ with very high probability, it is clear that this induces (at least) an inequality constraint. See Evans (2012) for further details.

A.3 Latent projection

Given an mDAG \mathcal{G} with vertices $V \dot{\cup} L$ where V and L are disjoint, the *latent projection* of \mathcal{G} over V is given by the mDAG with vertices V and edges within V given by:

- $a \rightarrow b$ whenever there is a directed walk in \mathcal{G} from a to b and any other (internal) vertices on the path are in L ;
- B is a bidirected face if there exists a *source* such that there is a directed path from the source down to each $b \in B$ and every variable on that path (other than b) is in L .

Here a ‘source’ is either a single bidirected face or a variable that is contained in L . See Evans (2016) for some examples.

A.4 Maximal Ancestral Projection for mDAGs

For this section we consider only ordinary mixed graphs (i.e. without any hyper-edges) that contain both bidirected and directed edges.

Definition A.7. An ordinary mixed graph is *ancestral* if its directed part is acyclic, and no vertex is an ancestor of any of its siblings; it is *maximal* if every pair of vertices that are not adjacent satisfy an m-separation or a nested constraint. Note that ancestral graphs are, by definition, simple.

For an mDAG \mathcal{G} , the *maximal ancestral projection* \mathcal{G}^* includes edges

- $a \rightarrow b$ if $a \in \text{an}_{\mathcal{G}}(b)$; and
- $a \leftrightarrow b$ if there is no ancestral relation in \mathcal{G} ;

for any pair of vertices a, b that cannot be m-separated in \mathcal{G} .

The crucial fact about a maximal ancestral projection is that it always induces precisely the same m-separations as the original mDAG did (Richardson and Spirtes, 2002; Evans, 2016).

A.5 Partial Ancestral Graphs

Given the maximal ancestral projection of an mDAG, one can consider all these projections for all mDAGs over the same set of vertices that are ordinary Markov equivalent to one another. We can denote this equivalence class $[\mathcal{G}]$. Then the *partial ancestral graph* $\mathcal{P} = [\mathcal{G}]$ is the unique graph that:

- has the same skeleton as the maximal ancestral projection of any element of $[\mathcal{G}]$;
- has an arrowhead (respectively tail) in any position for which the maximal ancestral projection of every element of the equivalence class has an arrowhead (resp. tail);
- has a circle at the end of any other edge.

More details about PAGs can be found in Richardson and Spirtes (2003) and Zhang (2006, 2008).

A.6 Nested Models and Fixing

Definition A.8. A vertex is said to be *fixable* if it has no (strict) descendants within its own district; that is, if $\deg(v) \cap \text{dis}_{\mathcal{G}}(v) = \{v\}$.

Note that a vertex v is fixable in \mathcal{G} precisely when, given a distribution p that is nested Markov with respect to \mathcal{G} , we can identify the distribution that would result if we *intervened* to fix the value of $X_v = x_v$ from p (Richardson et al., 2023).

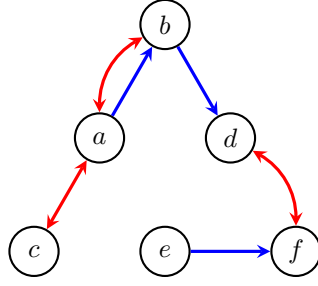


Figure 6: A conditional mDAG obtained by fixing e from Figure 1(a).

Definition A.9. Given an mDAG \mathcal{G} and a vertex v that is fixable, the *Markov blanket* of v is given by

$$\text{mb}_{\mathcal{G}}(v) := (\text{dis}_{\mathcal{G}}(v) \setminus \{v\}) \cup \text{pa}_{\mathcal{G}}(\text{dis}_{\mathcal{G}}(v)).$$

For an arbitrary set V , let $\mathcal{P}(V)$ denote the *power set* of V ; that is, the collection of all subsets of V .

Definition A.10. Let $\mathcal{G} = (V, \mathcal{D}, \mathcal{B})$ be an mDAG. Then if we can *fix* a vertex $v \in V$ we obtain a new graph \mathcal{G}^* with vertices V , and edges obtained by taking precisely those edges in $\mathcal{B} \cap \mathcal{P}(V \setminus \{v\})$ and $\mathcal{D} \cap (V \times (V \setminus \{v\}))$.

In other words, when we fix we remove (or reduce) any edges that have arrowheads at the vertex that has been fixed.

Definition A.11. We also associate a fixing operation to the distribution. If we fix v from \mathcal{G} , then we replace p with p^* , given by

$$p^*(x_V) = \frac{p^*(x_v)}{p(x_v \mid x_{\text{mb}(v)})} \cdot p(x_V).$$

In other words, we remove any dependence of X_v on its Markov blanket.

Example A.12. Consider the mDAG in Figure 1(i) and notice that e is fixable; after fixing it we obtain the graph in Figure 6. Whereas previously there was no set that could m-separate b and f , in spite of them not being adjacent, notice that now they are m-separated conditionally upon e . This is an example of a non-trivial *nested constraint*.

Results relating to the nested model

Let the set of distributions that are ordinary Markov with respect to an mDAG \mathcal{G} be denoted $\mathcal{O}(\mathcal{G})$, and those that are nested Markov be denoted $\mathcal{N}(\mathcal{G})$.

Proposition A.13. Suppose that \mathcal{G} is an mDAG. Then:

$$p \in \mathcal{M}(\mathcal{G}) \implies p \in \mathcal{N}(\mathcal{G}) \implies p \in \mathcal{O}(\mathcal{G}).$$

In other words, *distributional equivalence is a stronger requirement than nested equivalence, which is in turn a stronger requirement than ordinary equivalence.*

We now provide some results that are used in the proof of Proposition 3.2.

Lemma A.14. Suppose that in an mDAG \mathcal{G} we have $a \perp_m b \mid D$ but $a \not\perp_m b \mid D \cup \{s\}$. Then there is a valid topological ordering in which s comes after a , b and every element in D .

Proof. Suppose not. Then there is a path from a to b that is blocked by D but becomes open when we also condition on s . This implies that there is a collider that has s as a descendant, but no other element of D . (If there are multiple colliders, then reduce to one by taking the directed path from the first collider and the final collider to s , and use whichever vertex is the one at which these paths meet.) By the supposition that $a \perp_m b \mid D$ there is no directed path from s to any element of D .

Now, if s is an ancestor of b we can take the path from a to the collider, then follow the directed path from here to s and then to b . Clearly this path is open without conditioning on s , so we reach a contradiction. \square

Proposition A.15. *Consider an independence model \mathcal{I} such that:*

$$\begin{array}{ll} v \perp\!\!\!\perp s & [\mathcal{I}] \\ a \perp\!\!\!\perp b \mid D & [\mathcal{I}] \\ a \not\perp\!\!\!\perp b \mid D \cup \{s\} & [\mathcal{I}], \end{array}$$

where $v \in \{a, b\}$. Then there is no DAG that faithfully represents the independence model \mathcal{I} .

If D is chosen to be inclusion minimal such that $a \perp\!\!\!\perp b \mid D$ holds, then $v \in D$ is also not allowed by any faithful DAG independence model.

Proof. From Lemma A.14 we know that none of a , b or D are necessarily descendants of s . In this case, choose a particular DAG such that s comes after a , b and D in the chosen topological ordering (say $<$).

Then the only way in which \mathcal{I} could hold with a factorization that represents a DAG is if we can divide the predecessors of s under $<$ into two sets $S \cup T$, and we have $S \cup \{s\} \perp_m T$, with $v \in T$. In this case, if either a or b is in T then conditioning on s cannot make them dependent conditional on any subset that m-separates them. If $a, b \notin T$ but some $d \in D \cap T$, then the m-separation between a and b would hold given $D \setminus \{d\}$, which contradicts the minimality of D . Either way, we obtain the result. \square

B Other results

Proposition B.1. *Suppose that there is an mDAG with no locally unshielded collider path of length at least 3, but that does have a discriminating path from a to c for b of length at least 4. Then there also exists an induced subgraph isomorphic to Figure 5(i).*

Proof. If there is a discriminating path $\langle a, v_1, \dots, v_k, b, c \rangle$ with $k \geq 2$, then clearly either there is a locally unshielded collider path of length at least 3, or there is a directed edge between the vertices v_{k-1} and b , or between v_{k-2} and v_k . In the latter case, the colliders $\langle v_{k-2}, v_{k-1}, v_k \rangle$ and $\langle v_{k-1}, v_k, b \rangle$ must have discriminating paths of a strictly lower order of their own (Claassen and Bucur, 2022). Hence we can consider a lower order discriminating path, and by induction we will eventually reach a first-order discriminating path. In this case, if $k \geq 2$ then $a * \rightarrow v_1 \leftrightarrow v_2 \leftarrow * v_3$ (where possibly $v_3 = b$) will be a locally unshielded collider path of length 3, or we will have a discriminating path that looks like Figure 5(i). \square