The ID Algorithm Reformulated via Fixing

Thomas Richardson
University of Washington

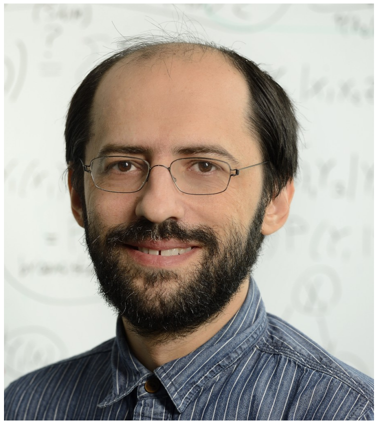
Simons Causal Bootcamp Day 3.4
20 January 2022

Joint work with Robin J. Evans, Ilya Shpitser, James M. Robins

Collaborators

Robin Evans
(Oxford)

James Robins
(Harvard)



Ilya Shpitser
(Johns Hopkins)

Outline

- Part One: A Complete Identification Algorithm for Intervention Distributions in DAGs with Latent Variables
- (Not Covered Today) Part Two: The Nested Markov Model

Part One: A Complete Identification Algorithm

- The general identification problem for DAGs with unobserved variables
- Simple examples
- Tian's Algorithm
- Formulation in terms of 'Fixing' operation

Intervention distributions (I)

Given a causal DAG $\mathcal{G}(V)$ with distribution:

$$p(V) = \prod_{v \in V} p(v \mid \mathrm{pa}(v))$$

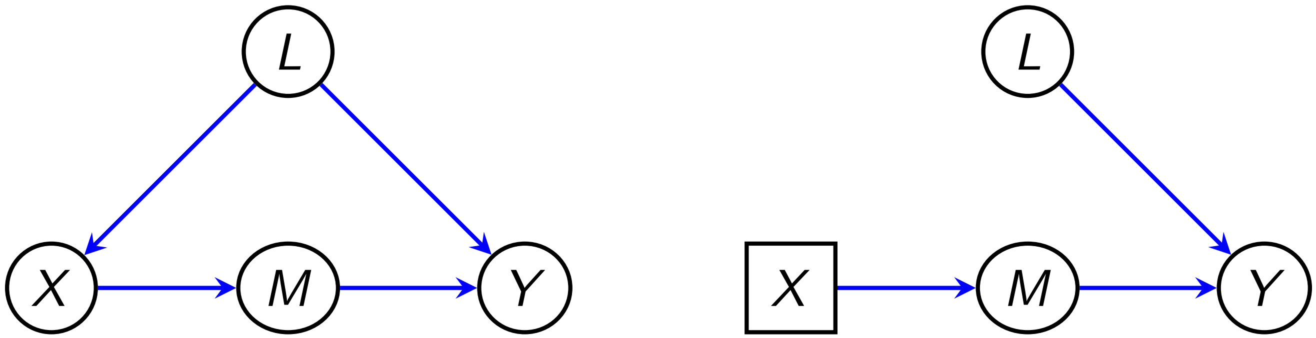
where $\mathrm{pa}(v) = \{x \,|\, x \to v\}$;

Intervention distribution on X:

$$p(V \setminus X \mid \mathrm{do}(X = \mathbf{x})) \quad = \quad \prod_{v \in V \setminus X} p(v \mid \mathrm{pa}(v)).$$

here on the RHS a variable in X occurring in $\mathrm{pa}(v)$, for some $v \in V \setminus X$, takes the corresponding value in $\mathbf{x}$.

Example



$$p(X, L, M, Y) = p(L)\ p(X \mid L)\ p(M \mid X) p(Y \mid L, M)$$

$$p(L, M, Y \mid \mathrm{do}(X = \tilde{x})) = p(L) \quad \times \quad p(M \mid \tilde{x}) p(Y \mid L, M)$$

Intervention distributions (II)

Given a causal DAG $\mathcal{G}$ with distribution:

$$p(V) = \prod_{v \in V} p(v \mid \mathrm{pa}(v))$$

we wish to compute an intervention distribution via truncated factorization:

$$p(V \setminus X \mid \mathrm{do}(X = \mathbf{x})) \quad = \quad \prod_{v \in V \setminus X} p(v \mid \mathrm{pa}(v)).$$

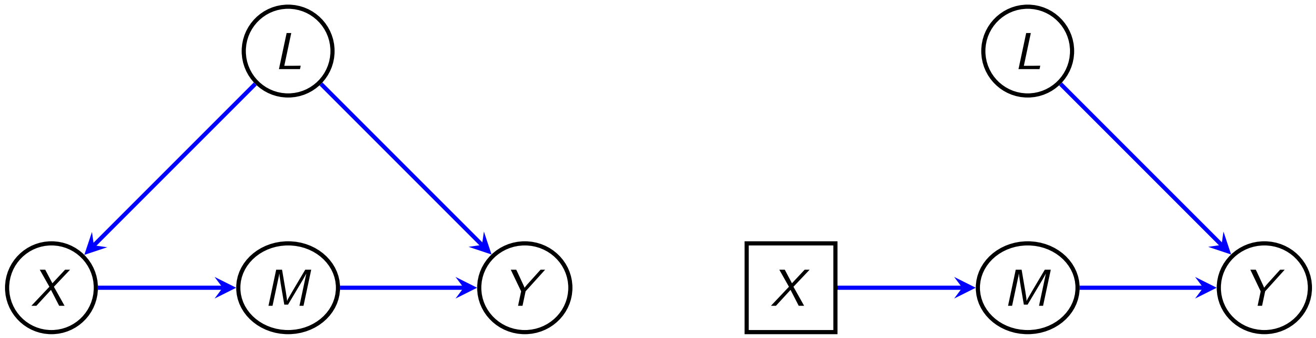
Hence if we are interested in $Y \subset V \setminus X$ then we simply marginalize:

$$p(Y \mid \mathrm{do}(X = \mathbf{x})) \quad = \quad \sum_{w \in V \setminus (X \cup Y)} \quad \prod_{v \in V \setminus X} p(v \mid \mathrm{pa}(v)).$$

('g-computation' formula of Robins (1986); see also Spirtes *et al.* 1993.)

Note: $p(Y \mid \mathrm{do}(X = \mathbf{x}))$ is a sum over a product of terms $p(v \mid \mathrm{pa}(v))$.

Example



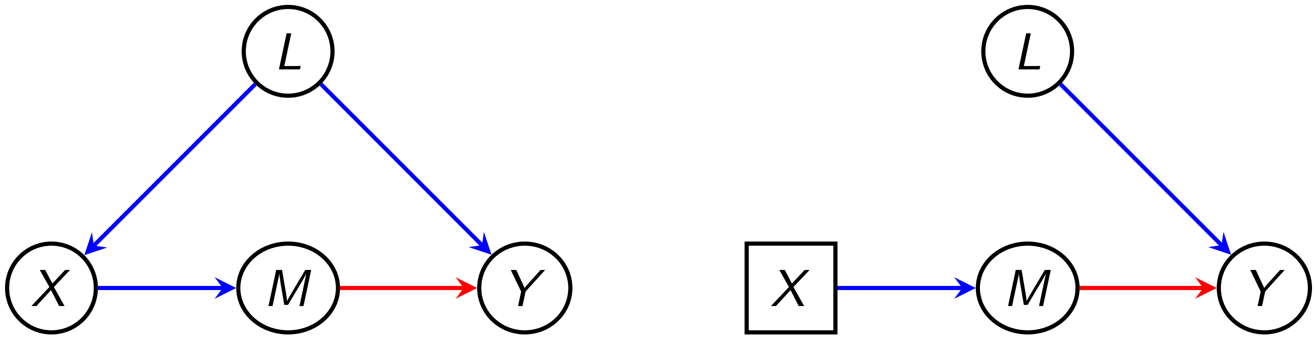
$$p(X, L, M, Y) = p(L)p(X \mid L)p(M \mid X)p(Y \mid L, M)$$

$$p(L, M, Y \mid \mathrm{do}(X = \tilde{x})) = p(L)p(M \mid \tilde{x})p(Y \mid L, M)$$

$$p(Y \mid \mathrm{do}(X = \tilde{x})) = \sum_{l,m} p(L = l)p(M = m \mid \tilde{x})p(Y \mid L = l, M = m)$$

Note that $p(Y \mid \mathrm{do}(X = \tilde{x})) \neq p(Y \mid X = \tilde{x})$.

Special case: no effect of M on Y



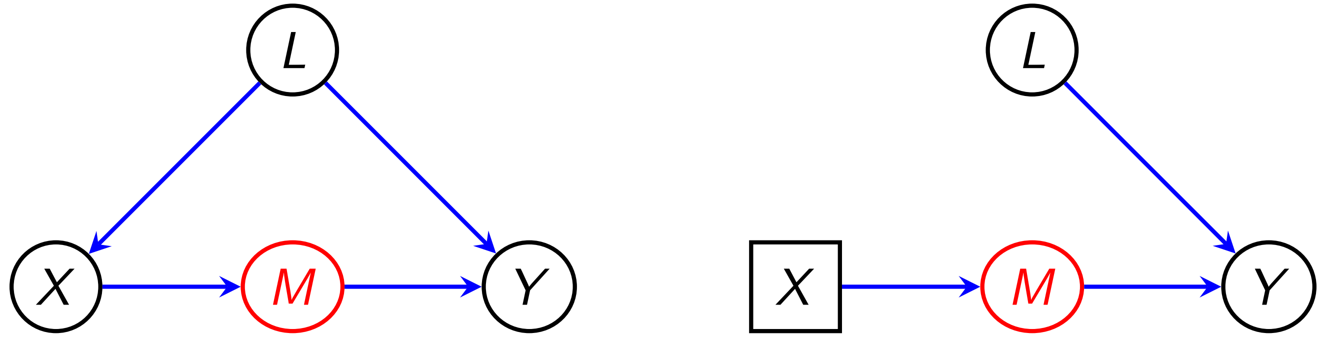
$$p(X, L, M, Y) = p(L)p(X \mid L)p(M \mid X)p(Y \mid L, M)$$

$$p(L, M, Y \mid \text{do}(X = \tilde{x})) = p(L)p(M \mid \tilde{x})p(Y \mid L)$$

$$\begin{aligned} p(Y \mid \text{do}(X = \tilde{x})) &= \sum_{l,m} p(L = l)p(M = m \mid \tilde{x})p(Y \mid L = l) \\ &= \sum_{l} p(L = l)p(Y \mid L = l) \\ &= p(Y) \neq P(Y \mid \tilde{x}) \end{aligned}$$

since $X \not\perp\!\!\!\perp Y$. ‘Correlation is not Causation’.

Example with M unobserved



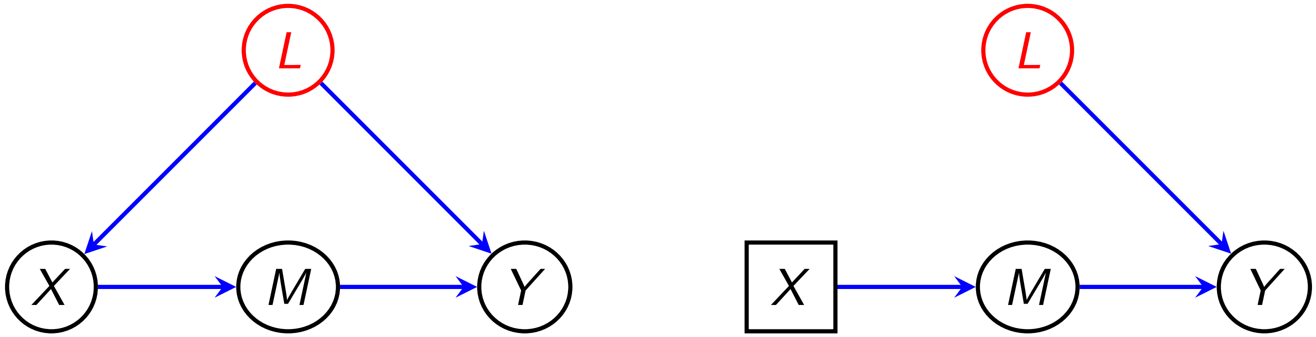
$$\begin{aligned} p(Y \mid \text{do}(X=\tilde{x})) &= \sum_{l,m} p(L=l)p(M=m \mid \tilde{x})p(Y \mid L=l, M=m) \\ &= \sum_{l,m} p(L=l)p(M=m \mid \tilde{x}, L=l)p(Y \mid L=l, M=m, X=\tilde{x}) \\ &= \sum_{l,m} p(L=l)p(Y, M=m \mid L=l, X=\tilde{x}) \\ &= \sum_{l} p(L=l)p(Y \mid L=l, X=\tilde{x}). \end{aligned}$$

Here we have used that $M \perp\!\!\!\perp L \mid X$ and $Y \perp\!\!\!\perp X \mid L, M$.

$\Rightarrow$ can find $p(Y \mid \text{do}(X=\tilde{x}))$ even if M not observed.

This is an example of the 'back door formula', aka 'standardization'.

Example with L unobserved

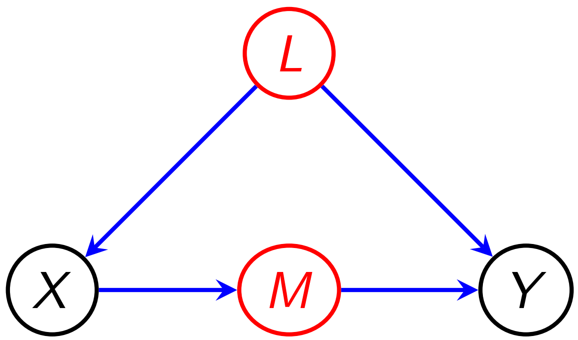


$$
\begin{aligned}
p(Y \mid \mathrm{do}(X = \tilde{x}))& \\
&= \sum_m p(M = m \mid \mathrm{do}(X = \tilde{x})) p(Y \mid \mathrm{do}(M = m)) \\
&= \sum_m p(M = m \mid X = \tilde{x}) p(Y \mid \mathrm{do}(M = m)) \\
&= \sum_m p(M = m \mid X = \tilde{x}) \left(\sum_{x^*} p(X = x^*) p(Y \mid M = m, X = x^*) \right)
\end{aligned}
$$

$\Rightarrow$ can find $p(Y \mid \mathrm{do}(X = \tilde{x}))$ even if L not observed.

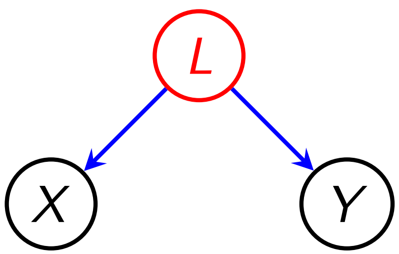
This is an example of the 'front door formula' of Pearl (1995).

But with *both* L and M unobserved....



...we are out of luck!

Given $P(X, Y)$, absent further assumptions we cannot distinguish:



General Identification Question

Given: a latent DAG $\mathcal{G}(O \cup H)$, where O are observed, H are hidden, and disjoint subsets $X, Y \subseteq O$.

Q: Is $p(Y \mid \text{do}(X))$ identified given $p(O)$?

A: Provide either an identifying formula that is a function of $p(O)$

or report that $p(Y \mid \text{do}(X))$ is not identified.

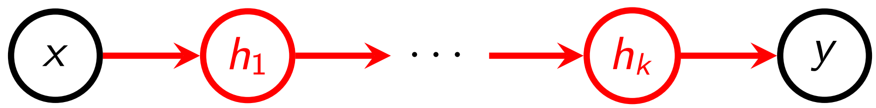
Motivations:

- Characterize which interventions can be identified without parametric assumptions;
- Understand which functionals of the observed margin have a causal interpretation;

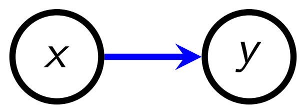
Latent Projection

Can preserve conditional independences and causal coherence with latents using paths. DAG $\mathcal{G}$ on vertices $V = O \dot{\cup} H$, define **latent projection** as follows: (Verma and Pearl, 1992)

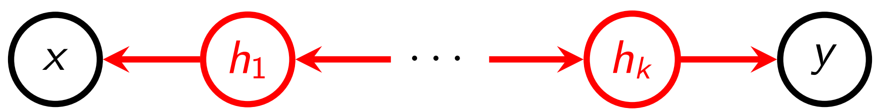
Whenever there is a path of the form



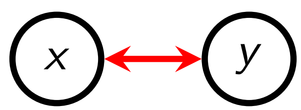
add



Whenever there is a path of the form

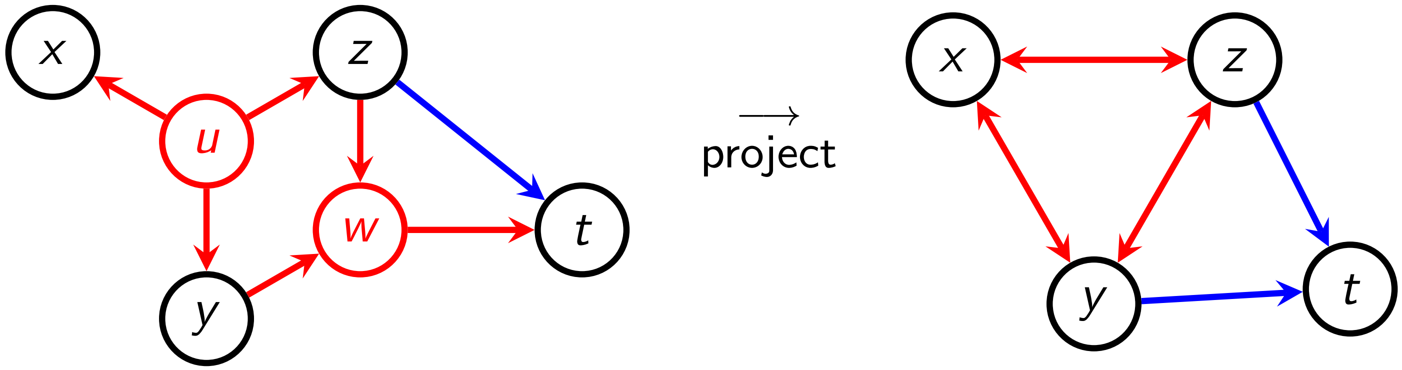


add



Then remove all latent variables H from the graph.

ADMGs



Latent projection leads to an **acyclic directed mixed graph** (ADMG)

Can read off independences with d/m-separation.

The projection preserves the (algebraic*) causal structure; Verma and Pearl (1992).

* Some information relating to inequality constraints is lost.

‘Conditional’ Acyclic Directed Mixed Graphs

An ‘conditional’ acyclic directed mixed graph (CADMG) is a bi-partite graph $\mathcal{G}(V, W)$, used to represent structure of a distribution over V, indexed by W, for example $P(V \mid \text{do}(W))$.

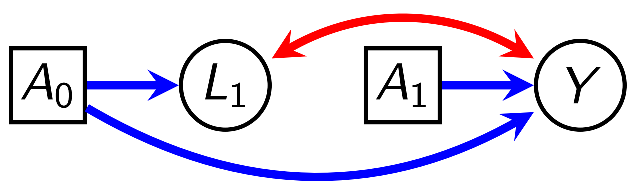
We require:

(i) The induced subgraph of $\mathcal{G}$ on V is an ADMG;

(ii) The induced subgraph of $\mathcal{G}$ on W contains no edges;

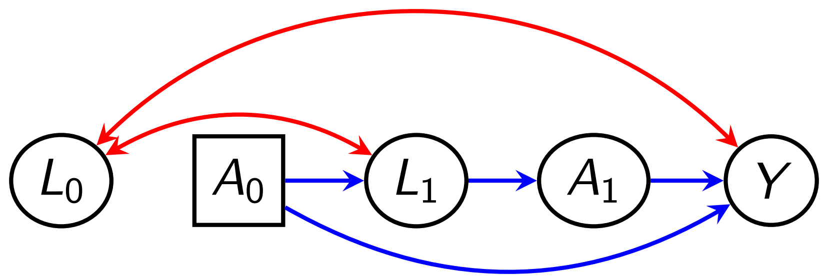
(iii) Edges between vertices in W and V take the form $w \to v$.

We represent V with circles, W with squares:



Here $V = \{L_1, Y\}$ and $W = \{A_0, A_1\}$.

Ancestors and Descendants



In a CADMG $\mathcal{G}(V, W)$ for $v \in V$, let the set of *ancestors* , *descendants* of v be:

$$\mathrm{an}_{\mathcal{G}}(v) = \{a \mid a \to \cdots \to v \text{ or } a = v \text{ in } \mathcal{G}, a \in V \cup W\},$$

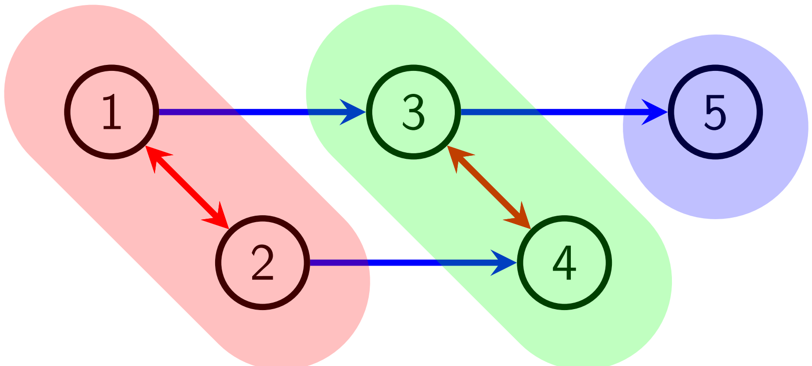
$$\mathrm{de}_{\mathcal{G}}(v) = \{d \mid d \leftarrow \cdots \leftarrow v \text{ or } d = v \text{ in } \mathcal{G}, d \in V \cup W\},$$

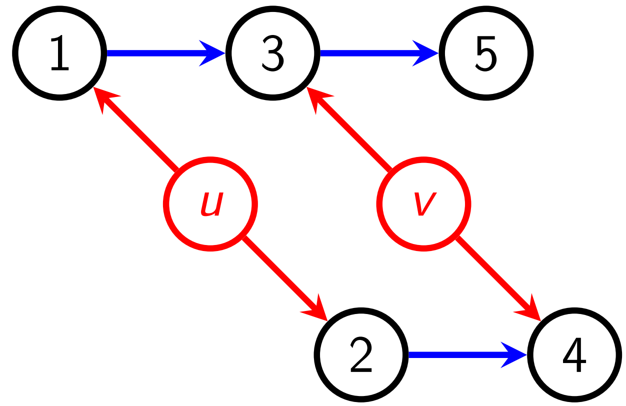
In the example above:

$$\text{an}(y) = \{a_0, l_1, a_1, y\}.$$

Districts

Define a **district** in a C/ADMG to be maximal sets connected by bi-directed edges:





$$\sum_{u,v} p(u)\, p(x_1 \mid u)\, p(x_2 \mid u) \quad p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v) \quad p(x_5 \mid x_3)$$

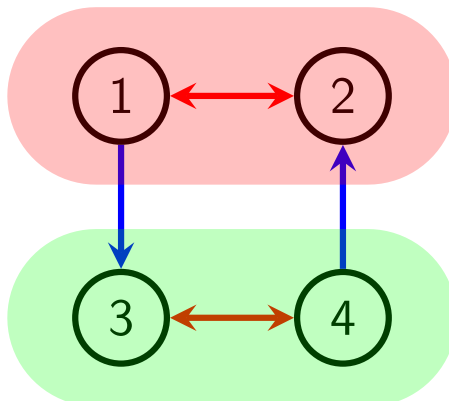
$$= \sum_u p(u)\, p(x_1 \mid u)\, p(x_2 \mid u) \sum_v p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)$$

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

$$= \prod_i q_{D_i}(x_{D_i} \mid x_{\mathrm{pa}(D_i) \setminus D_i})$$

Districts are called ‘c-components’ by Tian.

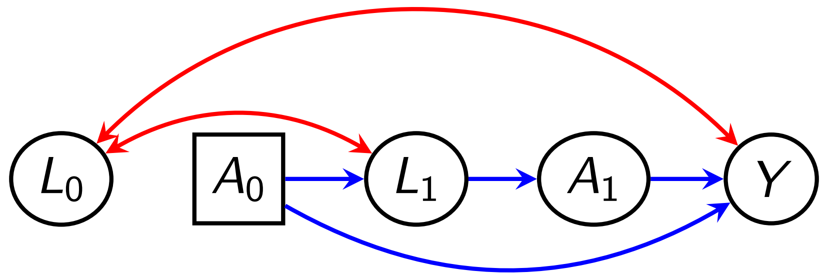
Edges between districts



There is no ordering on vertices such that parents of a district precede every vertex in the district.

(Cannot form a ‘chain graph’ ordering.)

Notation for Districts



In a CADMG $\mathcal{G}(V, W)$ for $v \in V$, the district of v is:

$$\text{dis}_{\mathcal{G}}(v) = \{d \mid d \leftrightarrow \cdots \leftrightarrow v \text{ or } d = v \text{ in } \mathcal{G}, d \in V\}.$$

Only variables in V are in districts.

In example above:

$$\text{dis}(y) = \{l_0, l_1, y\}, \quad \text{dis}(a_1) = \{a_1\}.$$

We use $\mathcal{D}(\mathcal{G})$ to denote the set of districts in $\mathcal{G}$.

In example $\mathcal{D}(\mathcal{G}) = \{\ \{l_0, l_1, y\}, \{a_1\}\ \}$.

Tian's ID algorithm for identifying $P(Y \mid \mathbf{do}(X))$

Jin Tian

(A) Re-express the query as a sum over a product of intervention distributions on districts:

$$p(Y \mid \text{do}(X)) = \sum \prod_i p(D_i \mid \text{do}(\text{pa}(D_i) \setminus D_i)).$$

(B) Check whether each term: $p(D_i \mid \text{do}(\text{pa}(D_i) \setminus D_i))$ is identified.

This is clearly sufficient for identifiability.

Necessity follows from results of Shpitser (2006); see also Huang and Valtorta (2006).

(A) Decomposing the query

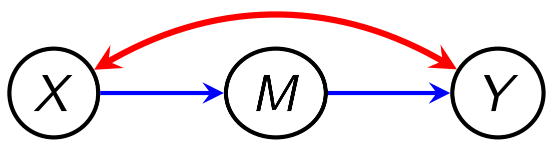
1. Remove edges into X:
   Let $\mathcal{G}[V \setminus X]$ denote the graph formed by removing edges with an arrowhead into X.
2. Restrict to variables that are (still) ancestors of Y:
   Let $T = \text{an}_{\mathcal{G}[V \setminus X]}(Y)$
   be vertices that lie on directed paths between X and Y (after cutting edges into X). Equivalently, T are variables on 'proper causal paths' from X to Y.
   Let $\mathcal{G}^*$ be formed from $\mathcal{G}[V \setminus X]$ by removing vertices not in T.
3. Find the districts:
   Let $D_1, \ldots, D_s$ be the districts in $\mathcal{G}^*$.

Then:

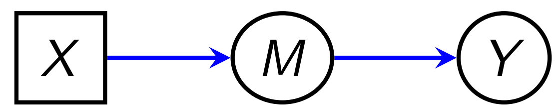
$$P(Y \mid \text{do}(X)) = \sum_{T \setminus (X \cup Y)} \prod_{D_i} p(D_i \mid \text{do}(\text{pa}(D_i) \setminus D_i)).$$

Example: front door graph

$\mathcal{G}$ $\qquad\qquad$ $\mathcal{G}_{[V \setminus \{X\}]} = \mathcal{G}^*$



$p(Y \mid \text{do}(X))$



$T = \{X, M, Y\}$

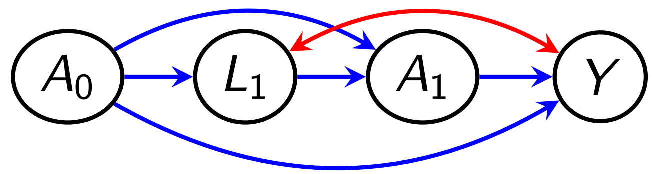
Districts in $T \setminus \{X\}$ are $D_1 = \{M\}$, $D_2 = \{Y\}$.

$$p(Y \mid \text{do}(X)) = \sum_M p(M \mid \text{do}(X)) p(Y \mid \text{do}(M))$$

Example: Sequentially randomized trial

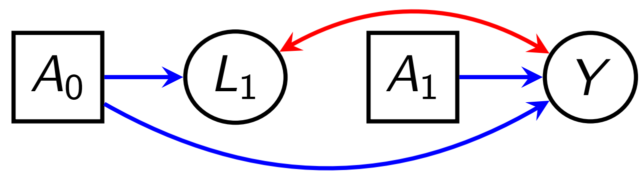
A_1 is randomized; A_2 is randomized conditional on L, A_1;

$\mathcal{G}$



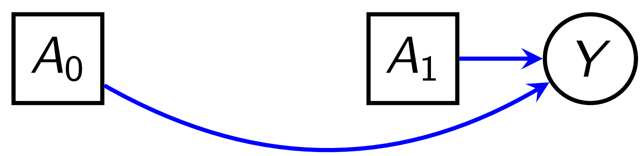
$$p(Y \mid \text{do}(A_0, A_1))$$

$\mathcal{G}_{[V \setminus \{A_0, A_1\}]}$



$$T = \{A_0, A_1, Y\}$$

$\mathcal{G}^*$



$$D_1 = \{Y\}$$

(Here the decomposition is trivial since there is only one district and no summation.)

(B) Finding if $P(D \mid \mathbf{do}(\mathrm{pa}(D) \setminus D))$ is identified

Idea: Find an ordering $r_1, \ldots, r_p$ of $O \setminus D$ such that:

If $P(O \setminus \{r_1, \ldots, r_{t-1}\} \mid \mathrm{do}(r_1, \ldots, r_{t-1}))$ is identified

Then $P(O \setminus \{r_1, \ldots, r_t\} \mid \mathrm{do}(r_1, \ldots, r_t))$ is also identified.

Sufficient for identifiability of $P(D \mid \mathrm{do}(\mathrm{pa}(D) \setminus D))$, since:

$P(O)$ is identified

$D = O \setminus \{r_1, \ldots, r_p\}$, so
$P(O \setminus \{r_1, \ldots, r_p\} \mid \mathrm{do}(r_1, \ldots, r_p)) = P(D \mid \mathrm{do}(\mathrm{pa}(D) \setminus D))$.

Such a vertex r_t will said to be 'fixable', given that we have already 'fixed' $r_1, \ldots, r_{t-1}$:

'fixing' differs formally from 'do'/cutting edges since the latter does not preserve identifiability in general.

To do:

- Give a graphical characterization of 'fixability';
- Construct the identifying formula.

The set of fixable vertices

Given a CADMG $\mathcal{G}(V, W)$ we define the set of fixable vertices,

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

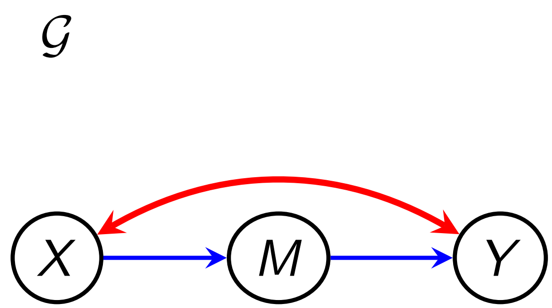
In words, a vertex $v \in V$ is fixable in $\mathcal{G}$ if there is no (proper) descendant of v that is in the same district as v in $\mathcal{G}$.

Thus v is fixable if there is no vertex $y \neq v$ such that

$$v \leftrightarrow \cdots \leftrightarrow y \quad \text{and} \quad v \rightarrow \cdots \rightarrow y \quad \text{in } \mathcal{G}.$$

Note that the set of fixable vertices is a subset of V, and contains at least one vertex from each district in $\mathcal{G}$.

Example: Front door graph

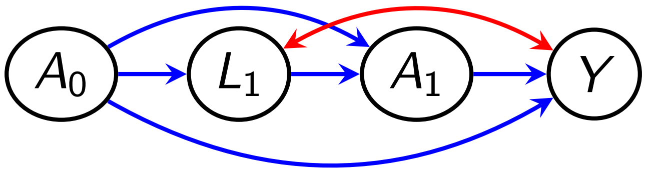


$F(\mathcal{G}) = \{M, Y\}$

X is not fixable since Y is a descendant of X and

Y is in the same district as X

Example: Sequentially randomized trial



Here $F(\mathcal{G}) = \{A_0, A_1, Y\}$.

L_1 is not fixable since Y is a descendant of L_1 and

Y is in the same district as L_1.

The *graphical* operation of fixing vertices

Given a CADMG $\mathcal{G}(V, W, E)$, for every $r \in F(\mathcal{G})$ we associate a transformation ϕ_r on the pair $(\mathcal{G}, P(X_V \mid X_W))$:

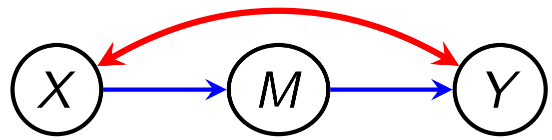
$$\phi_r(\mathcal{G}) \equiv \mathcal{G}^\dagger(V \setminus \{r\}, W \cup \{r\}),$$

where in $\mathcal{G}^\dagger$ we remove from $\mathcal{G}$ any edge that has an arrowhead at r.

The operation of 'fixing r' simply transfers r from 'V' to 'W', and removes edges $r \leftrightarrow$ or $r \leftarrow$.

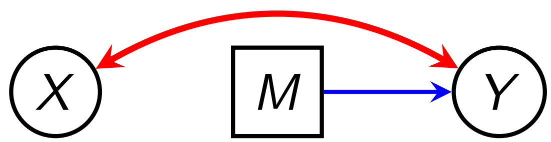
Example: front door graph

$\mathcal{G}$



$F(\mathcal{G}) = \{M, Y\}$

$\phi_M(\mathcal{G})$

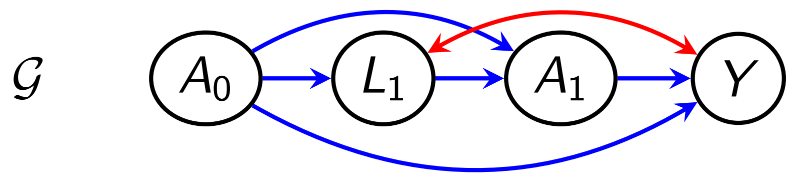


$F(\phi_M(\mathcal{G})) = \{X, Y\}$

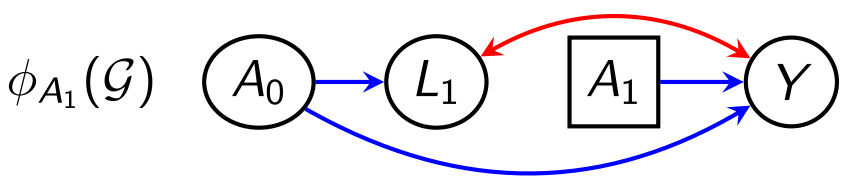
Note that X was not fixable in $\mathcal{G}$,

but it is fixable in $\phi_M(\mathcal{G})$ after fixing M.

Example: Sequentially randomized trial



Here $F(\mathcal{G}) = \{A_0, A_1, Y\}$.



Notice $F(\phi_{A_1}(\mathcal{G})) = \{A_0, L_1, Y\}$.

Thus L_1 was not fixable prior to fixing A_1,

but L_1 is fixable in $\phi_{A_1}(\mathcal{G})$ after fixing A_1.

The *probabilistic* operation of fixing vertices

Given a distribution $P(V \mid W)$ we associate a transformation:

$$\phi_r(P(V \mid W); \mathcal{G}) \quad \equiv \quad \frac{P(V \mid W)}{P(r \mid \text{mb}_{\mathcal{G}}(r))}.$$

Here
$\text{mb}_{\mathcal{G}}(r) = \{y \neq r \mid (r \leftarrow y) \text{ or } (r \leftrightarrow \circ \cdots \circ \leftrightarrow y) \text{ or } (r \leftrightarrow \circ \cdots \circ \leftrightarrow \circ \leftarrow y)\}$.

In words: *we divide by the conditional distribution of r given the other vertices in the district containing r, and the parents of the vertices in that district.*

It can be shown that if r is fixable in $\mathcal{G}$ then:

$$\phi_r(P(V \mid \text{do}(W)); \mathcal{G}) = P(V \setminus \{r\} \mid \text{do}(W \cup \{r\})).$$

as required.

Note: If r is fixable in $\mathcal{G}$ then $\text{mb}_{\mathcal{G}}(r)$ is the 'Markov blanket' of r in $\text{an}_{\mathcal{G}}(\text{dis}_{\mathcal{G}}(r))$.

Unifying Marginalizing and Conditioning

Some special cases:

- If $\text{mb}_{\mathcal{G}}(r) = (V \cup W) \setminus \{r\}$ then fixing corresponds to marginalizing:

$$\phi_r(P(V \mid W); \mathcal{G}) = \frac{P(V \mid W)}{P(r \mid (V \cup W) \setminus \{r\})} = P(V \setminus \{r\} \mid W)$$

- If $\text{mb}_{\mathcal{G}}(r) = W$ then fixing corresponds to ordinary conditioning:

$$\phi_r(P(V \mid W); \mathcal{G}) = \frac{P(V \mid W)}{P(r \mid W)} = P(V \setminus \{r\} \mid W \cup \{r\})$$

- In the general case fixing corresponds to re-weighting, so

$$\phi_r(P(V \mid W); \mathcal{G}) = P^*(V \setminus \{r\} \mid W \cup \{r\}) \neq P(V \setminus \{r\} \mid W \cup \{r\})$$

Having a single operation simplifies the identification algorithm.

Composition of fixing operations

We use $\circ$ to indicate composition of operations in the natural way.

If s is fixable in $\mathcal{G}$ and then r is fixable in $\phi_s(\mathcal{G})$ (after fixing s) then:

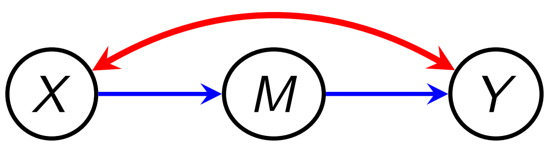
$$\phi_r \circ \phi_s(\mathcal{G}) \equiv \phi_r(\phi_s(\mathcal{G}))$$

$$\phi_r \circ \phi_s(P(V \mid W); \mathcal{G}) \equiv \phi_r\left(\phi_s\left(P(V \mid W); \mathcal{G}\right); \phi_s(\mathcal{G})\right)$$

Back to step (B) of identification

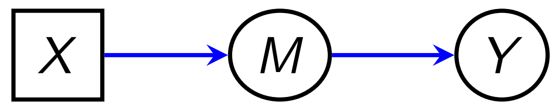
Recall our goal is to identify $P(D \mid \mathrm{do}(\mathrm{pa}(D) \setminus D))$, for the districts D in $\mathcal{G}^*$:

$\mathcal{G}$



$p(Y \mid \mathrm{do}(X))$

$\mathcal{G}_{[V \setminus \{X\}]} = \mathcal{G}^*$



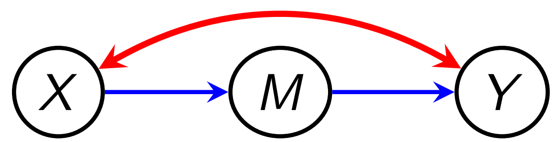
$T = \{X, M, Y\}$

Districts in $T \setminus \{X\}$ are $D_1 = \{M\}$, $D_2 = \{Y\}$.

$$p(Y \mid \mathrm{do}(X)) = \sum_M p(M \mid \mathrm{do}(X)) p(Y \mid \mathrm{do}(M))$$

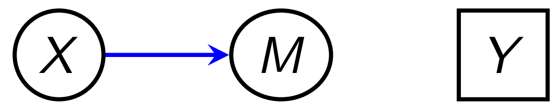
Example: front door graph: $D_1 = \{M\}$

$\mathcal{G}$



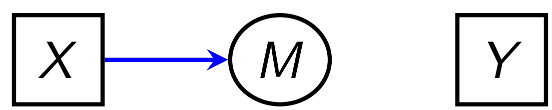
$F(\mathcal{G}) = \{M, Y\}$

$\phi_Y(\mathcal{G})$



$F(\phi_Y(\mathcal{G})) = \{X, M\}$

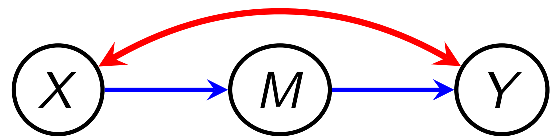
$\phi_X \circ \phi_Y(\mathcal{G})$



This proves that $p(M \mid \text{do}(X))$ is identified.

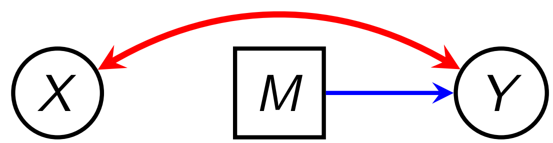
Example: front door graph: $D_2 = \{Y\}$

$\mathcal{G}$



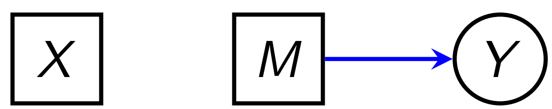
$F(\mathcal{G}) = \{M, Y\}$

$\phi_M(\mathcal{G})$



$F(\phi_M(\mathcal{G})) = \{X, Y\}$

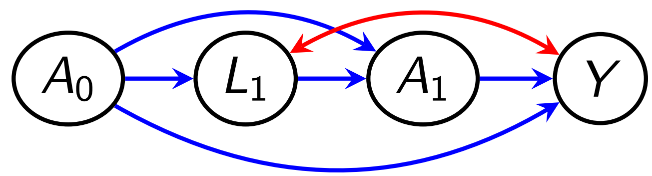
$\phi_X \circ \phi_M(\mathcal{G})$



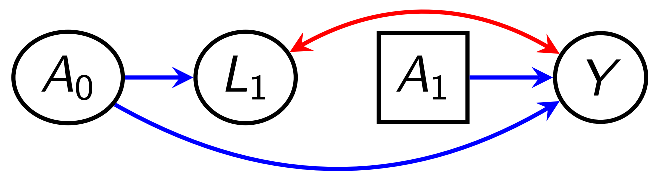
This proves that $p(Y \mid \text{do}(M))$ is identified.

Example: Sequential Randomization

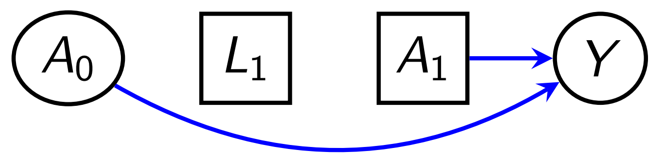
$\mathcal{G}$



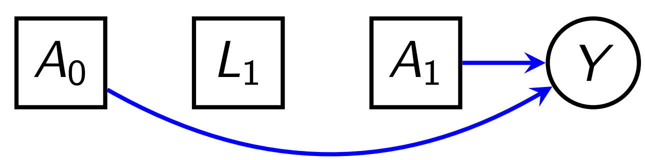
$\phi_{A_1}(\mathcal{G})$



$\phi_{L_1} \circ \phi_{A_1}(\mathcal{G})$



$\phi_{A_0} \circ \phi_{L_1} \circ \phi_{A_1}(\mathcal{G})$



This establishes that $P(Y \mid \mathrm{do}(A_0, A_1))$ is identified.

Review: Tian's ID algorithm via fixing

(A) Re-express the query as a sum over a product of intervention distributions on districts:

$$p(Y \mid \text{do}(X)) = \sum \prod_i p(D_i \mid \text{do}(\text{pa}(D_i) \setminus D_i)).$$

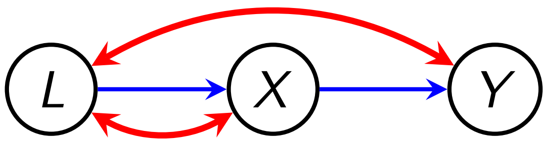
- Cut edges into X;
- Restrict to vertices that are (still) ancestors of Y;
- Find the set of districts $D_1, \ldots, D_p$.

(B) Check whether each term: $p(D_i \mid \text{do}(\text{pa}(D_i) \setminus D_i))$ is identified:

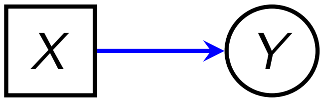
- Iteratively find a vertex that r_t that is fixable in $\phi_{r_{t-1}} \circ \cdots \circ \phi_{r_1}(\mathcal{G})$, with $r_t \notin D_i$;
- If no such vertex exists then $P(D_i \mid \text{do}(\text{pa}(D_i) \setminus D_i))$ is not identified.

Not identified example

$\mathcal{G}$



$\mathcal{G}^*$



Suppose we wish to find $p(Y \mid \text{do}(X))$.

There is one district $D = \{Y\}$ in $\mathcal{G}^*$.

But since the only fixable vertex in $\mathcal{G}$ is Y, we see that $p(Y \mid \text{do}(X))$ is not identified.

Reachable subgraphs of an ADMG

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

$$w_1 \in F(\mathcal{G}^*) \text{ and for } j = 2, \ldots, k,$$
$$w_j \in F(\phi_{w_{j-1}} \circ \cdots \circ \phi_{w_1}(\mathcal{G}^*)).$$

Thus a subgraph is reachable if, under some ordering, each of the vertices in W may be fixed, first in $\mathcal{G}^*$, and then in $\phi_{w_1}(\mathcal{G}^*)$, then in $\phi_{w_2}(\phi_{w_1}(\mathcal{G}^*))$, and so on.

Invariance to orderings

In general, there may exist multiple sequences that fix a set W, however, they all result in both the same graph and distribution.

This is a consequence of the following:

Lemma

Let $\mathcal{G}(V, W)$ be a CADMG with $r, s \in \mathbb{F}(\mathcal{G})$, and let $q_V(V \mid W)$ be Markov w.r.t. $\mathcal{G}$, and further (a) $\phi_r(q_V; \mathcal{G})$ is Markov w.r.t. $\phi_r(\mathcal{G})$; and (b) $\phi_s(q_V; \mathcal{G})$ is Markov w.r.t. $\phi_s(\mathcal{G})$. Then

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

Consequently, if $\mathcal{G}(V, W)$ is reachable from $\mathcal{G}(V \cup W)$ then $\phi_V(p(V, W); \mathcal{G})$ is uniquely defined.

Intrinsic sets

A set D is said to be *intrinsic* if it forms a *district* in a *reachable* subgraph. If D is intrinsic in $\mathcal{G}$ then $p(D \mid \text{do}(\text{pa}(D) \setminus D))$ is identified.

Let $\mathcal{I}(\mathcal{G})$ denote the intrinsic sets in $\mathcal{G}$.

Theorem

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

$$p(Y \mid \text{do}_{\mathcal{G}(H \cup V)}(A)) = \sum_{Y^* \setminus Y} \prod_{D \in \mathcal{D}(\mathcal{G}(V)_{Y^*})} \phi_{V \setminus D}(p(V); \mathcal{G}(V)). \qquad (*)$$

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

Thus $p(D \mid \text{do}(pa(D) \setminus D))$ for intrinsic D play the same role as $P(v \mid \text{do}(\text{pa}(v))) = p(v \mid \text{pa}(v))$ in the simple fully observed case.

Shpitser+R+Robins (2012) give an efficient algorithm for computing $(*)$.

Intrinsic sets and 'hedges'

Shpitser (2006) provided a characterization in terms of graphical structures called 'hedges' of those interventional distributions that were *not* identified.

It may be shown that if a $\leftrightarrow$-connected set is *not* intrinsic then there exists a hedge, hence we have:

$\leftrightarrow$-connected set S is intrinsic iff $p(S \mid \mathrm{do}(\mathrm{pa}(S) \setminus S))$ is identified.

It follows that intrinsic sets may thus also be defined in terms of the *non-existence* of a hedge.

Part Two: The Nested Markov Model

1. Deriving constraints via fixing
2. The Nested Markov Model
3. Finer Factorizations
4. Discrete Parameterization
5. Testing and Fitting
6. Completeness

Identification and Nested Markov model references

- Evans, R. J. (2015). Margins of discrete Bayesian networks. arXiv preprint:1501.02103.
- Evans, R. J. and Richardson, T. S. (2015). Smooth, identifiable supermodels of discrete DAG models with latent variables. arXiv:1511.06813.
- Evans, R.J. and Richardson, T.S. (2014). Markovian acyclic directed mixed graphs for discrete data. Annals of Statistics vol. 42, No. 4, 1452-1482.
- Richardson, T.S., Evans, R. J., Robins, J. M. and Shpitser, I. (2017). Nested Markov properties for acyclic directed mixed graphs. arXiv:1701.06686.
- Richardson, T.S. (2003). Markov Properties for Acyclic Directed Mixed Graphs. The Scandinavian Journal of Statistics, March 2003, vol. 30, no. 1, pp. 145-157(13).
- Shpitser, I., Evans, R.J., Richardson, T.S., Robins, J.M. (2014). Introduction to Nested Markov models. Behaviormetrika, vol. 41, No.1, 2014, 3–39.
- Shpitser, I., Richardson, T.S. and Robins, J.M. (2011). An efficient algorithm for computing interventional distributions in latent variable causal models. In Proceedings of UAI-11.
- Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. Twenty-First National Conference on Artificial Intelligence.
- Tian, J. (2002) Studies in Causal Reasoning and Learning, CS PhD Thesis, UCLA.
- Tian, J. and Pearl, J. (2002a). A general identification condition for causal effects. In Proceedings of AAAI-02.
- Tian, J. and J. Pearl (2002b). On the testable implications of causal models with hidden variables. In Proceedings of UAI-02.

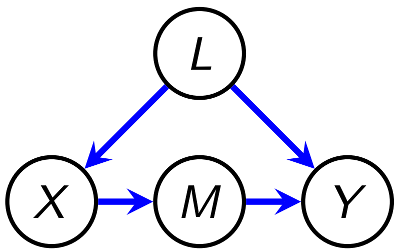
Parameterization & Completeness References

(Including earlier work on the ordinary Markov model.)

- Evans, R.J. – Margins of discrete Bayesian networks. *Annals of Statistics*, 2018.
- Evans, R.J. and Richardson, T.S. – Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 2019.
- Evans, R.J. and Richardson, T.S. – Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, 2014.
- Shpitser, I., Richardson, T.S. and Robins, J.M. An efficient algorithm for computing interventional distributions in latent variable causal models. *UAI*, 2011.
- Shpitser, I., Richardson, T.S., Robins, J.M. and Evans, R.J. – Parameter and structure learning in nested Markov models. *UAI*, 2012.
- Shpitser, I., Evans, R.J., Richardson, T.S. and Robins, J.M. – Sparse nested Markov models with log-linear parameters. *UAI*, 2013.
- Shpitser, I., Evans, R.J., Richardson, T.S. and Robins, J.M. – Introduction to Nested Markov Models. *Behaviormetrika*, 2014.
- Shpitser, I., Evans, R.J., and Richardson, T.S. – Acyclic Linear SEMs Obey the Nested Markov Property. *UAI*, 2018.
- Spirtes, P., Glymour, G., Scheines, R. – *Causation Prediction and Search*, 2nd Edition, MIT Press, 2000.

Motivation

- So far we have shown how to estimate interventional distributions separately, but no guarantee these estimates are coherent.
- We also may have multiple identifying expressions: which one should we use?



$p(Y \mid do(X))$
front door?
back door?
does it matter?

- We can test constraints separately, but ultimately don't have a way to check if the model is a good one.
- Being able to evaluate a likelihood would allow lots of standard inference techniques (e.g. LR, Bayesian).
- Even better, if model can be shown smooth we get nice asymptotics for free.

All this suggests we should define a model which we can parameterize.