Nested Markov Models

Thomas S. Richardson

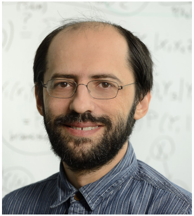
Department of Statistics
University of Washington

Center for Causal Discovery
University of Pittsburgh
20 April 2017

Collaborators

Robin Evans
(Oxford)

James Robins
(Harvard)



Ilya Shpitser
(Johns Hopkins)

Outline

- Part One: Non-parametric Identification
- Part Two: The Nested Markov Model

Part One: Non-parametric identification

- 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:

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

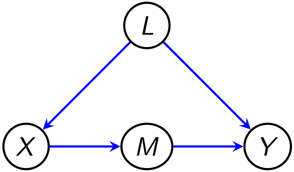
where $\text{pa}(v) = \{x \,|\, x \rightarrow v\}$;

Intervention distribution on X:

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

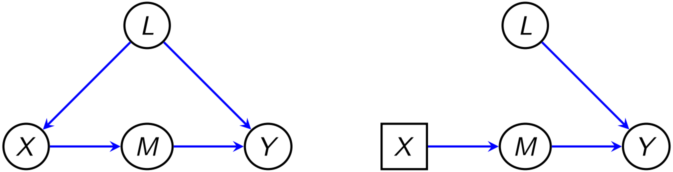
here on the RHS a variable in X occurring in $\text{pa}(v)$, for some $v \in V \setminus X$, takes the corresponding value in $\mathbf{x}$.

Example



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

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.)

Intervention distributions (II)

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

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

we wish to compute an intervention distribution via truncated factorization:

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

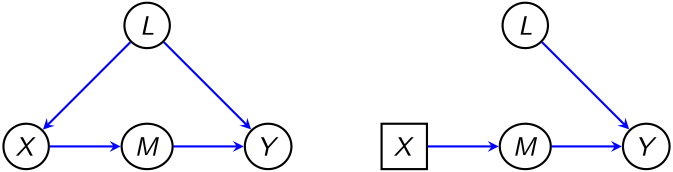
Hence if we are interested in $Y \subset V \setminus X$ then we simply marginalize:

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

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

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

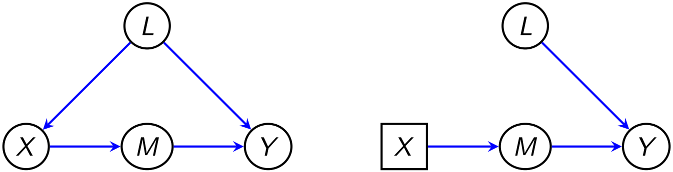
Example



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

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

Example

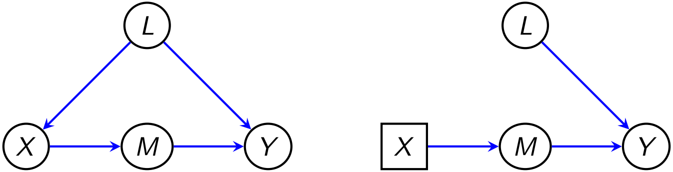


$$p(X, L, M, Y) = p(L)p(X \mid L)p(M \mid X)p(Y \mid L, M)$$

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

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

Example



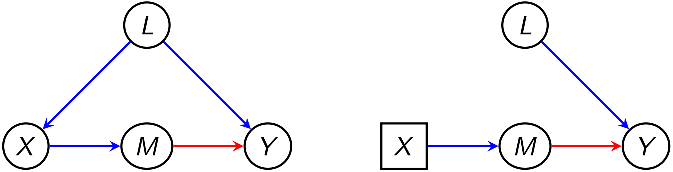
$$p(X, L, M, Y) = p(L)p(X \mid L)p(M \mid X)p(Y \mid L, M)$$

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

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

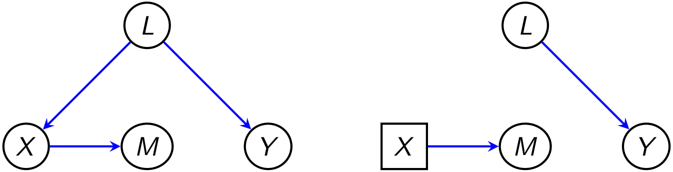
Note that $p(Y \mid \text{do}(X = \tilde{x})) \neq p(Y \mid X = \tilde{x})$.

Special case: no effect of M on Y



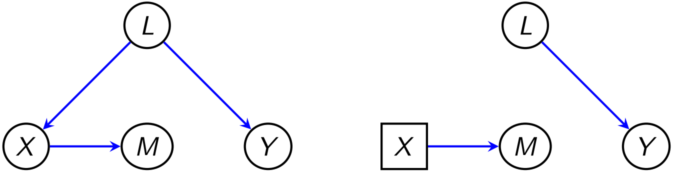
$$p(X, L, M, Y) = p(L)p(X \mid L)p(M \mid X)p(Y \mid L, M)$$

Special case: no effect of M on Y



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

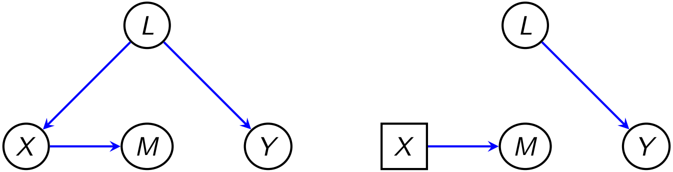
Special case: no effect of M on Y



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

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

Special case: no effect of M on Y

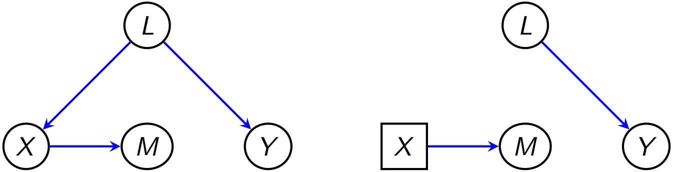


$$p(X, L, M, Y) = p(L)p(X \mid L)p(M \mid X)p(Y \mid L)$$

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

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

Special case: no effect of M on Y

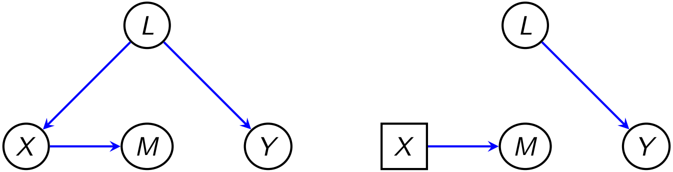


$$p(X, L, M, Y) = p(L)p(X \mid L)p(M \mid X)p(Y \mid L)$$

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

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

Special case: no effect of M on Y



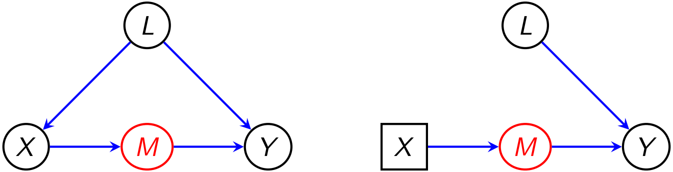
$$p(X, L, M, Y) = p(L)p(X \mid L)p(M \mid X)p(Y \mid L)$$

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

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

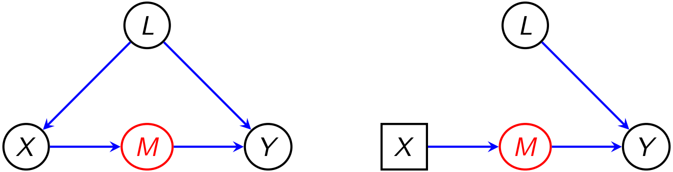
since $X \not\perp\!\!\!\perp Y$. 'Correlation is not Causation'.

Example with M unobserved



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

Example with M unobserved

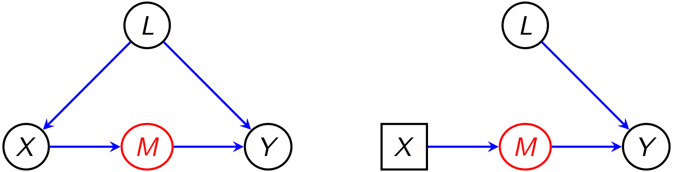


$$p(Y \mid \text{do}(X = \tilde{x})) = \sum_{l,m} p(L = l) p(M = m \mid \tilde{x}) p(Y \mid L = l, M = m)$$

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

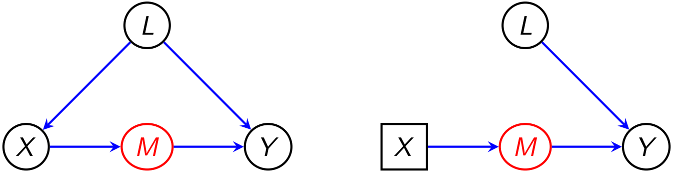
Here we have used that $M \perp\!\!\!\perp L \mid X$ and $Y \perp\!\!\!\perp X \mid L, M$.

Example with M unobserved



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

Example with M unobserved

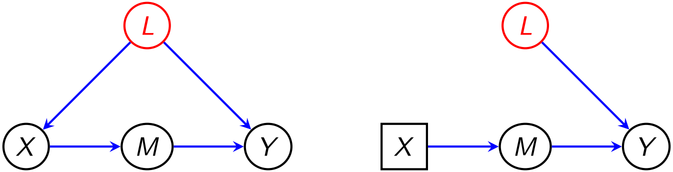


$$
\begin{aligned}
p(Y \mid \mathrm{do}(X=\tilde{x})) &= \sum_{l,m} p(L=l)p(M=m \mid \tilde{x})p(Y \mid L=l, M=m) \\
&= \sum_{l,m} p(L=l)p(M=m \mid \tilde{x}, L=l)p(Y \mid L=l, M=m, X=\tilde{x}) \\
&= \sum_{l,m} p(L=l)p(Y, M=m \mid L=l, X=\tilde{x}) \\
&= \sum_{l} p(L=l)p(Y \mid L=l, X=\tilde{x}).
\end{aligned}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

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

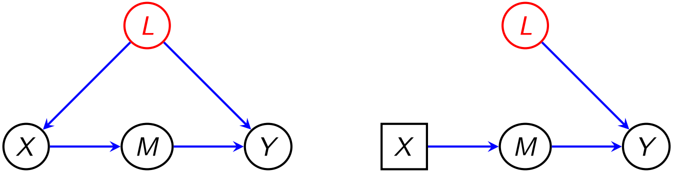
This is an example of the ‘back door formula’, aka ‘standardization’.

Example with L unobserved



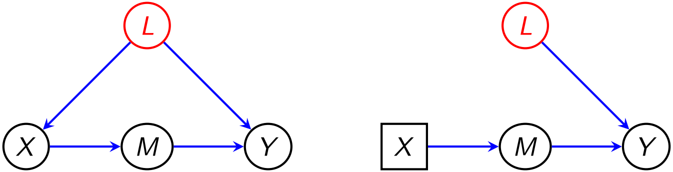
$p(Y \mid \mathrm{do}(X = \tilde{x}))$

Example with L unobserved



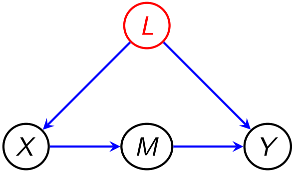
$$
p(Y \mid \mathrm{do}(X = \tilde{x})) \\
= \sum_m p(M = m \mid \mathrm{do}(X = \tilde{x})) p(Y \mid \mathrm{do}(M = m))
$$

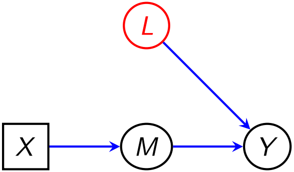
Example with L unobserved



$$
\begin{aligned}
p(Y \mid \mathrm{do}(X = \tilde{x})) \\
&= \sum_m p(M = m \mid \mathrm{do}(X = \tilde{x}))p(Y \mid \mathrm{do}(M = m)) \\
&= \sum_m p(M = m \mid X = \tilde{x})p(Y \mid \mathrm{do}(M = m))
\end{aligned}
$$

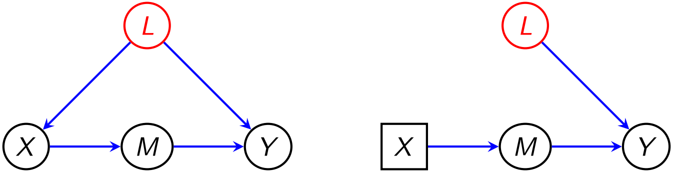
Example with L unobserved





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

Example with L unobserved

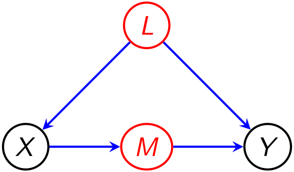


$$
\begin{aligned}
&p(Y \mid \mathrm{do}(X = \tilde{x})) \\
&\qquad = \sum_m p(M = m \mid \mathrm{do}(X = \tilde{x}))p(Y \mid \mathrm{do}(M = m)) \\
&\qquad = \sum_m p(M = m \mid X = \tilde{x})p(Y \mid \mathrm{do}(M = m)) \\
&\qquad = \sum_m p(M = m \mid X = \tilde{x}) \left(\sum_{x^*} p(X = x^*)p(Y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

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

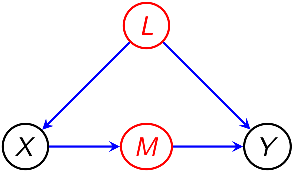
This is an example of the 'front door formula' of Pearl (1995).

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



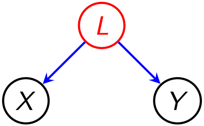
...we are out of luck!

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 $p(O)$?

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.

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

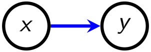
Can preserve conditional independences and causal coherence with latents using paths. DAG $\mathcal{G}$ on vertices $V = O \dot{\cup} H$, define **latent projection** as follows: (Verma and Pearl, 1992)

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

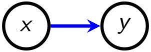


Latent Projection

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)

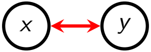
Whenever there is a path of the form

add



Whenever there is a path of the form

add

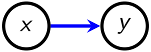


Latent Projection

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)

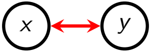
Whenever there is a path of the form

add



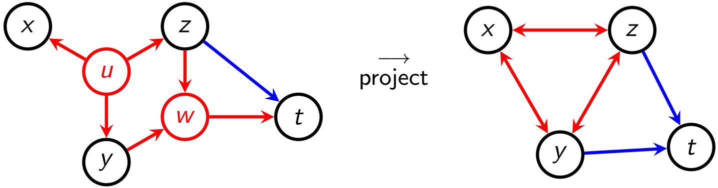
Whenever there is a path of the form

add

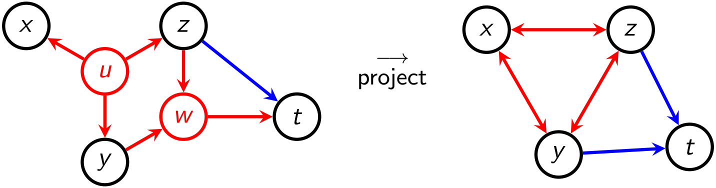


Then remove all latent variables H from the graph.

ADMGs

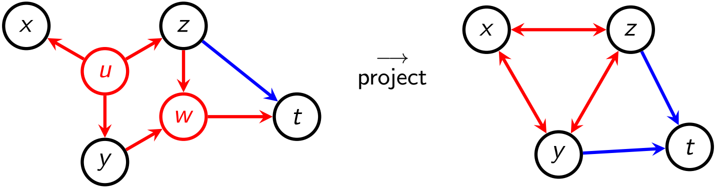


ADMGs



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

ADMGs



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

Can read off independences with d/m-separation.

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

‘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 $P(V \mid \mathrm{do}(W))$.

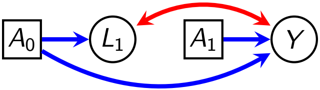
We require:

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

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

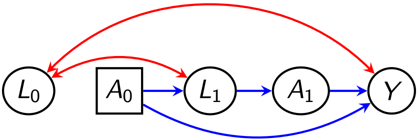
(iii) Edges between vertices in W and V take the form $w \rightarrow v$.

We represent V with circles, W with squares:



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

Ancestors and Descendants

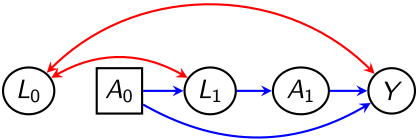


In a CADMG $\mathcal{G}(V, W)$ for $v \in V$, let the set of *ancestors* , *descendants* of v be:

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

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

Ancestors and Descendants



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

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

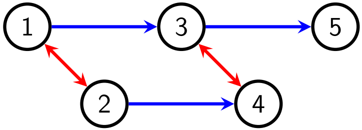
$$\text{de}_{\mathcal{G}}(v) = \{d \mid d \leftarrow \cdots \leftarrow v \text{ or } d = v \text{ in } \mathcal{G}, d \in V \cup W\},$$

In the example above:

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

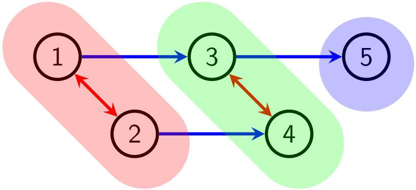
Districts

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



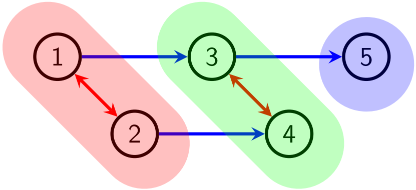
Districts

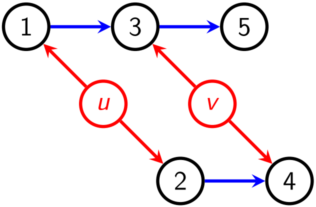
Define a **district** in a C/ADMG to be maximal sets connected by bi-directed edges:



Districts

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

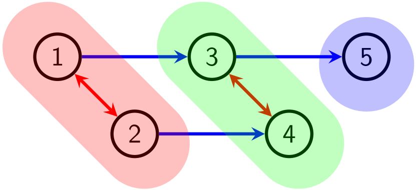


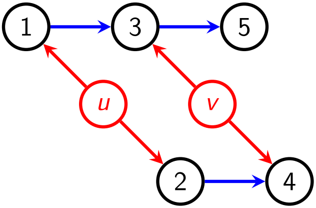


$$\sum_{u,v} p(u)\, p(x_1 \mid u)\, p(x_2 \mid u) \quad p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v) \quad p(x_5 \mid x_3)$$

Districts

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

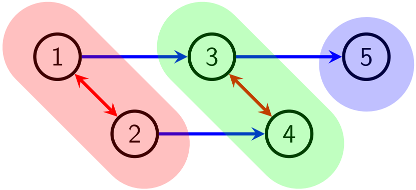


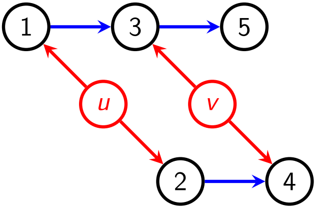


$$\sum_{u,v} p(u)\, p(x_1 \mid u)\, p(x_2 \mid u) \quad p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v) \quad p(x_5 \mid x_3)$$

Districts

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



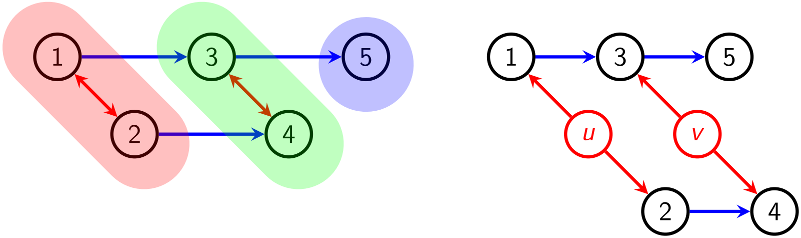


$$\sum_{u,v} p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)$$

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

Districts

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



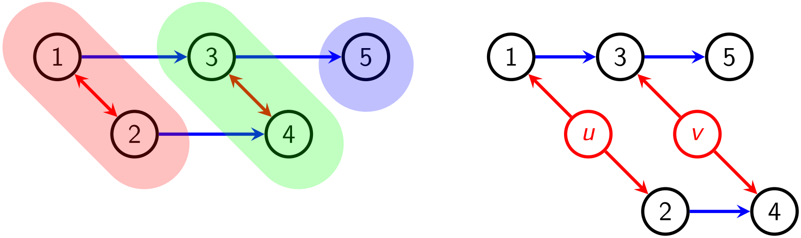
$$\sum_{u,v} p(u)\, p(x_1 \mid u)\, p(x_2 \mid u) \; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v) \; p(x_5 \mid x_3)$$

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

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

Districts

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



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

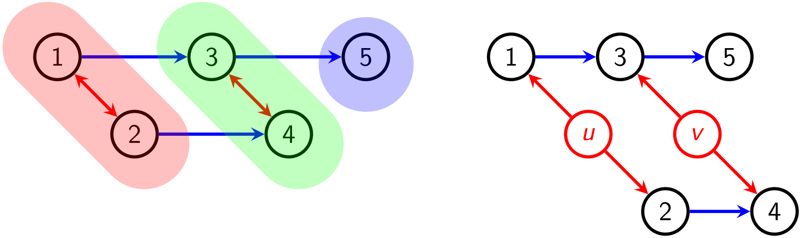
$$= \sum_{u} p(u)\,p(x_1 \mid u)\,p(x_2 \mid u) \sum_{v} p(v)\,p(x_3 \mid x_1, v)\,p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)$$

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

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

Districts

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



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

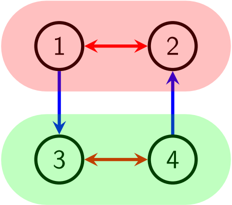
$$= \sum_{u} p(u)\, p(x_1 \mid u)\, p(x_2 \mid u) \sum_{v} p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)$$

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

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

Districts are called 'c-components' by Tian.

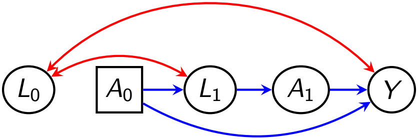
Edges between districts



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

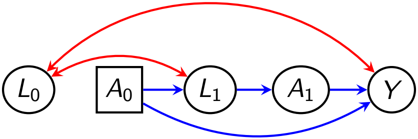


In a CADMG $\mathcal{G}(V, W)$ for $v \in V$, the district of v is:

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

Only variables in V are in districts.

Notation for Districts



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

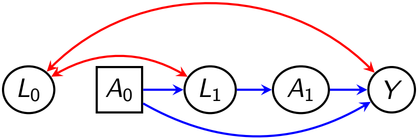
$$\text{dis}_{\mathcal{G}}(v) = \{d \mid d \leftrightarrow \cdots \leftrightarrow v \text{ or } d = v \text{ in } \mathcal{G}, d \in V\}.$$

Only variables in V are in districts.

In example above:

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

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:

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

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.

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.

(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).

(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).
   Let $\mathcal{G}^*$ be formed from $\mathcal{G}[V \setminus X]$ by removing vertices not in T .

(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).
   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 $\mathcal{G}^*$.

(A) Decomposing the query

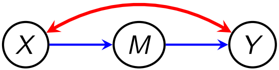
1. Remove edges into X:
   Let $\mathcal{G}[V \setminus X]$ denote the graph formed by removing edges with an arrowhead into X.
2. Restrict to variables that are (still) ancestors of Y:
   Let $T = \text{an}_{\mathcal{G}[V \setminus X]}(Y)$
   be vertices that lie on directed paths between X and Y (after cutting edges into X).
   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

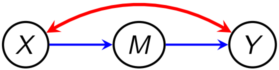
$\mathcal{G}$



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

Example: front door graph

$\mathcal{G}$



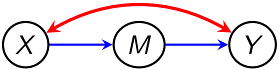
$p(Y \mid \mathrm{do}(X))$

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

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

Example: front door graph

$\mathcal{G}$



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

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

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

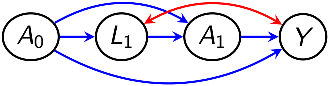
Districts in $T \setminus \{X\}$ are $D_1 = \{M\}$, $D_2 = \{Y\}$.

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

Example: Sequentially randomized trial

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

$\mathcal{G}$

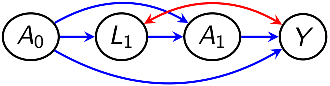


$p(Y \mid \text{do}(A_0, A_1))$

Example: Sequentially randomized trial

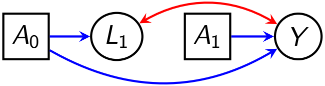
A_1 is randomized; A_2 is randomized conditional on L, A_1;

$\mathcal{G}$



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

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

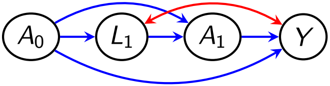


$$T = \{A_0, A_1, Y\}$$

Example: Sequentially randomized trial

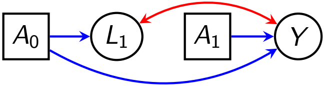
A_1 is randomized; A_2 is randomized conditional on L, A_1;

$\mathcal{G}$



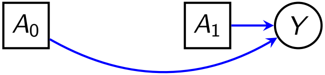
$$p(Y \mid \text{do}(A_0, A_1))$$

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



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

$\mathcal{G}^*$

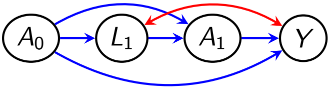


$$D_1 = \{Y\}$$

Example: Sequentially randomized trial

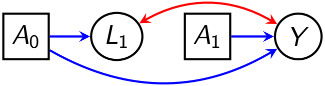
A_1 is randomized; A_2 is randomized conditional on L, A_1;

$\mathcal{G}$



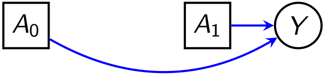
$p(Y \mid \mathrm{do}(A_0, A_1))$

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



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

$\mathcal{G}^*$



$D_1 = \{Y\}$

(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.

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

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

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

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

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

$P(O)$ is identified

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

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

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

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

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

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

$P(O)$ is identified

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

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.

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

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

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

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

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

$P(O)$ is identified

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

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

'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 $\mathcal{G}$.

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 $\mathcal{G}$.

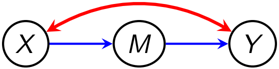
Thus v is fixable if there is no vertex $y \neq v$ such that

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

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

$\mathcal{G}$

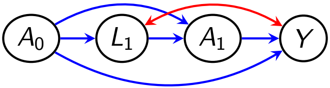


$F(\mathcal{G}) = \{M, Y\}$

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 *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 $(\mathcal{G}, P(X_V \mid X_W))$:

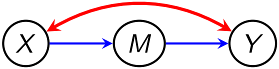
$$\phi_r(\mathcal{G}) \equiv \mathcal{G}^\dagger(V \setminus \{r\}, W \cup \{r\}),$$

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 $r \leftarrow$.

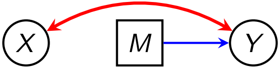
Example: front door graph

$\mathcal{G}$



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

$\phi_M(\mathcal{G})$

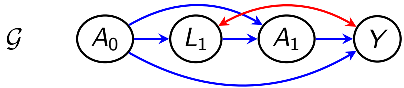


$F(\phi_M(\mathcal{G})) = \{X, Y\}$

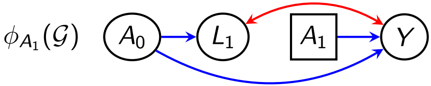
Note that X was not fixable in $\mathcal{G}$,

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:

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

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

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.*

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

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

Unifying Marginalizing and Conditioning

Some special cases:

- If $\mathrm{mb}_{\mathcal{G}}(r) = (V \cup W) \setminus \{r\}$ 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 $\mathrm{mb}_{\mathcal{G}}(r) = W$ 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:

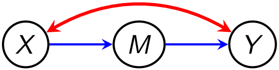
$$\phi_r \circ \phi_s(\mathcal{G}) \equiv \phi_r(\phi_s(\mathcal{G}))$$

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

Back to step (B) of identification

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

$\mathcal{G}$



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

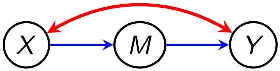
Districts in $T \setminus \{X\}$ are $D_1 = \{M\}$, $D_2 = \{Y\}$.

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

Back to step (B) of identification

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

$\mathcal{G}$



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

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

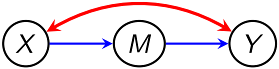
$T = \{X, M, Y\}$

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

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

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

$\mathcal{G}$



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

$\phi_Y(\mathcal{G})$

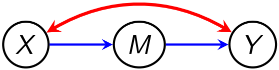
$F(\phi_Y(\mathcal{G})) = \{X, M\}$

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

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

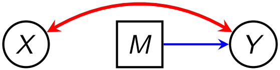
Example: front door graph: $D_2 = \{Y\}$

$\mathcal{G}$



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

$\phi_M(\mathcal{G})$



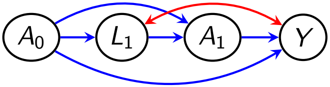
$F(\phi_M(\mathcal{G})) = \{X, Y\}$

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

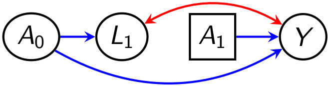
This proves that $p(Y \mid \mathrm{do}(M))$ is identified.

Example: Sequential Randomization

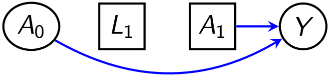
$\mathcal{G}$



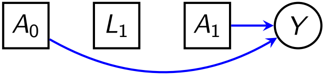
$\phi_{A_1}(\mathcal{G})$



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



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



This establishes that $P(Y \mid \text{do}(A_0, A_1))$ 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 $D_1, \dots, D_p$.

Review: Tian's ID algorithm via fixing

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

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

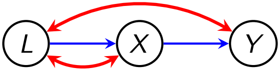
- Cut edges into X;
- Restrict to vertices that are (still) ancestors of Y;
- Find the set of districts $D_1, \ldots, D_p$.

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

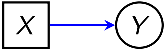
- Iteratively find a vertex that r_t that is fixable in $\phi_{r_{t-1}} \circ \cdots \circ \phi_{r_1}(\mathcal{G})$, with $r_t \notin D_i$;
- If no such vertex exists then $P(D_i \mid \mathrm{do}(\mathrm{pa}(D_i) \setminus D_i))$ is not identified.

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 $\mathcal{G}^*$.

Not identified example

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.

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

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

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 $\mathrm{(a)}$ $\phi_r(q_V; \mathcal{G})$ is Markov w.r.t. $\phi_r(\mathcal{G})$; and $\mathrm{(b)}$ $\phi_s(q_V; \mathcal{G})$ 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 $\mathcal{G}(H \cup V)$.

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.

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 $p(D \mid \mathrm{do}(\mathrm{pa}(D) \setminus D))$ 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 $Y^* = \mathrm{an}_{\mathcal{G}(V)_{V \setminus A}}(Y)$. Then if $\mathcal{D}(\mathcal{G}(V)_{Y^*}) \subseteq \mathcal{I}(\mathcal{G}(V))$,

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

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

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

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

Part Two: The Nested Markov Model

1. Motivation

2. Deriving constraints via fixing

3. The Nested Markov Model

4. Finer Factorizations

5. Discrete Parameterization

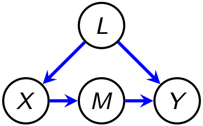
6. Testing and Fitting

7. Completeness

Outline

Motivation

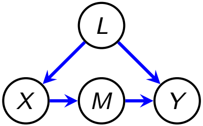
- So far we have shown how to estimate interventional distributions separately, but no guarantee these estimates are coherent.
- We also may have multiple identifying expressions: which one should we use?



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

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?

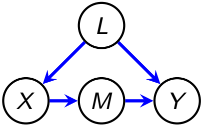


$p(Y \mid do(X))$
front door?
back door?
does it matter?

- We can test constraints separately, but ultimately don't have a way to check if the model is a good one.

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?

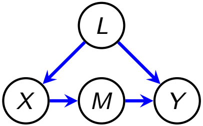


$p(Y \mid do(X))$
front door?
back door?
does it matter?

- 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).

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?

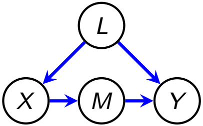


$p(Y \mid do(X))$
front door?
back door?
does it matter?

- 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.

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.

Outline

Deriving constraints via fixing

Let $p(O)$ be the observed margin from a DAG with latents $\mathcal{G}(O \cup H)$,

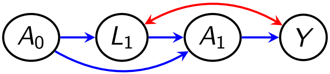
Idea: If $r \in O$ is fixable then $\phi_r(p(O); \mathcal{G})$ will obey the Markov property for the graph $\phi_r(\mathcal{G})$.

...and this can be iterated.

This gives non-parametric constraints that are not independences, that are implied by the latent DAG.

Example: The 'Verma' Constraint

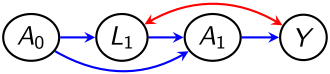
$\mathcal{G}$



This graph implies no conditional independences on $P(A_0, L_1, A_1, Y)$.

Example: The 'Verma' Constraint

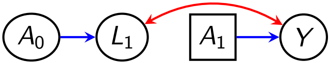
$\mathcal{G}$



This graph implies no conditional independences on $P(A_0, L_1, A_1, Y)$.

But since $F(\mathcal{G}) = \{A_0, A_1, Y\}$, we may construct:

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



$$\phi_{A_1}(p(A_0, L_1, A_1, Y)) \quad = \quad p(A_0, L_1, A_1, Y)/p(A_1 \mid A_0, L_1)$$

$$A_0 \perp\!\!\!\perp Y \mid A_1 \qquad [\phi_{A_1}(p(A_0, L_1, A_1, Y); \mathcal{G})]$$

Outline

The nested Markov model

These independences may be used to define a graphical model:

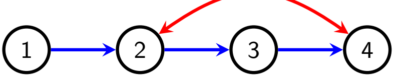
Definition

$p(V)$ obeys the *global nested Markov property* for $\mathcal{G}$ if for all reachable sets R, the kernel $\phi_{V\setminus R}(p(V);\mathcal{G})$ obeys the global Markov property for $\phi_{V\setminus R}(\mathcal{G})$.

This is a 'generalized' Markov property since it is defined by conditional independence in re-weighted distributions (obtained via fixing).

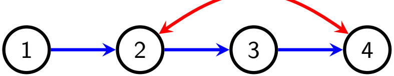
We will use $\mathcal{N}(\mathcal{G})$ to indicate the set of distributions obeying this property.

Notation



Note that we can potentially reach the same district by different methods: e.g. marginalize 4, fix 1, 2 or reverse.

Notation



Note that we can potentially reach the same district by different methods: e.g. marginalize 4, fix 1, 2 or reverse.

Theorem (R,Evans, Shpitser, Robins, 2017)

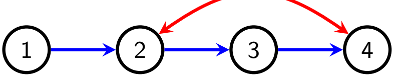
For a positive distribution $p \in \mathcal{N}(\mathcal{G})$ and vertices v_1, v_2 fixable in $\mathcal{G}$,

$$(\phi_{v_1} \circ \phi_{v_2})(p) = (\phi_{v_2} \circ \phi_{v_1})(p).$$

Hence, the order of fixing doesn't matter.

This is another way of saying that all identifying expressions for a causal quantity will be the same.

Notation



Note that we can potentially reach the same district by different methods: e.g. marginalize 4, fix 1, 2 or reverse.

Theorem (R,Evans, Shpitser, Robins, 2017)

For a positive distribution $p \in \mathcal{N}(\mathcal{G})$ and vertices v_1, v_2 fixable in $\mathcal{G}$,

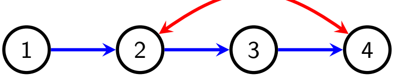
$$(\phi_{v_1} \circ \phi_{v_2})(p) = (\phi_{v_2} \circ \phi_{v_1})(p).$$

Hence, the order of fixing doesn't matter.

This is another way of saying that all identifying expressions for a causal quantity will be the same.

For any reachable R this justifies the (unambiguous) notation $\phi_{V \setminus R}$.

Notation



Note that we can potentially reach the same district by different methods: e.g. marginalize 4, fix 1, 2 or reverse.

Theorem (R,Evans, Shpitser, Robins, 2017)

For a positive distribution $p \in \mathcal{N}(\mathcal{G})$ and vertices v_1, v_2 fixable in $\mathcal{G}$,

$$(\phi_{v_1} \circ \phi_{v_2})(p) = (\phi_{v_2} \circ \phi_{v_1})(p).$$

Hence, the order of fixing doesn't matter.

This is another way of saying that all identifying expressions for a causal quantity will be the same.

For any reachable R this justifies the (unambiguous) notation $\phi_{V \setminus R}$.

For $p \in \mathcal{N}(\mathcal{G})$, let

$$\mathcal{G}[R] \equiv \phi_{V \setminus R}(\mathcal{G}) \qquad\qquad q_R \equiv \phi_{V \setminus R}(p).$$

be respectively, the graph and distribution where $V \setminus R$ is fixed.

Reachable CADMGs

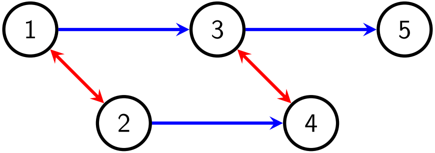
Note that $\mathcal{G}[R]$ is always just the CADMG with:

- random vertices R,
- fixed vertices $\text{pa}_{\mathcal{G}}(R) \setminus R$,
- induced edges from $\mathcal{G}$ among R and of the form $\text{pa}_{\mathcal{G}}(R) \to R$.

Reachable CADMGs

Note that $\mathcal{G}[R]$ is always just the CADMG with:

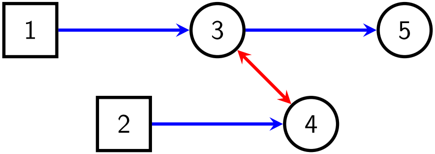
- random vertices R,
- fixed vertices $\text{pa}_{\mathcal{G}}(R) \setminus R$,
- induced edges from $\mathcal{G}$ among R and of the form $\text{pa}_{\mathcal{G}}(R) \to R$.



Reachable CADMGs

Note that $\mathcal{G}[R]$ is always just the CADMG with:

- random vertices R,
- fixed vertices $\mathrm{pa}_{\mathcal{G}}(R) \setminus R$,
- induced edges from $\mathcal{G}$ among R and of the form $\mathrm{pa}_{\mathcal{G}}(R) \to R$.

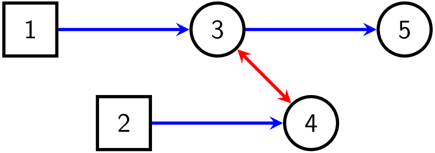


Graph shown is $\mathcal{G}[\{3, 4, 5\}]$.

Reachable CADMGs

Note that $\mathcal{G}[R]$ is always just the CADMG with:

- random vertices R,
- fixed vertices $\text{pa}_{\mathcal{G}}(R) \setminus R$,
- induced edges from $\mathcal{G}$ among R and of the form $\text{pa}_{\mathcal{G}}(R) \to R$.

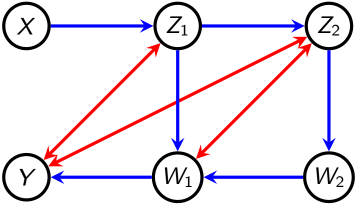


Graph shown is $\mathcal{G}[\{3, 4, 5\}]$.

Also recall that if there is an underlying causal DAG then $p(x_V)$ then:

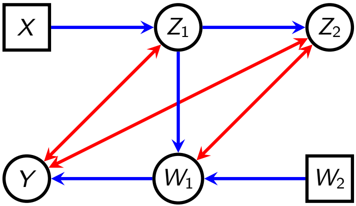
$$q_R(x_R \mid x_{\text{pa}(R)\setminus R}) = p(x_R \mid do(x_{V\setminus R})).$$

Example



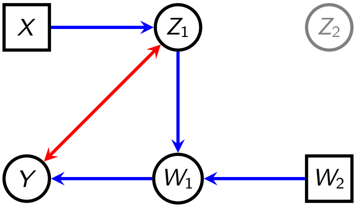
$p(x, y, w_1, w_2, z_1, z_2)$

Example



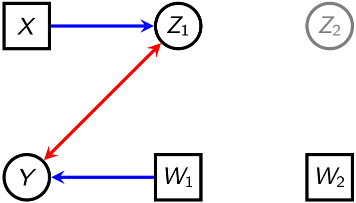
$$q_{yw_1z_1z_2}(y, w_1, z_1, z_2 \mid x, w_2) = \frac{p(x, y, w_1, w_2, z_1, z_2)}{p(x)p(w_2 \mid z_2)}$$

Example



$$q_{yw_1z_1z_2}(y, w_1, z_1, z_2 \mid x, w_2) = \frac{p(x, y, w_1, w_2, z_1, z_2)}{p(x)p(w_2 \mid z_2)}$$

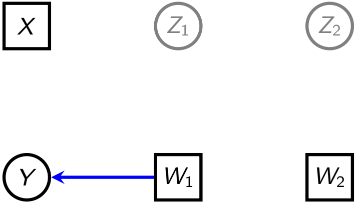
Example



$$q_{yw_1z_1z_2}(y, w_1, z_1, z_2 \mid x, w_2) = \frac{p(x, y, w_1, w_2, z_1, z_2)}{p(x)p(w_2 \mid z_2)}$$

$$q_{yz_1}(y, z_1 \mid x, w_1) = \frac{q_{yw_1z_1z_2}(y, w_1, z_1 \mid x, w_2)}{q_{yw_1z_1z_2}(w_1 \mid x, w_2)}$$

Example



$$q_{yw_1z_1z_2}(y, w_1, z_1, z_2 \mid x, w_2) = \frac{p(x, y, w_1, w_2, z_1, z_2)}{p(x)p(w_2 \mid z_2)}$$

$$q_{yz_1}(y, z_1 \mid x, w_1) = \frac{q_{yw_1z_1z_2}(y, w_1, z_1 \mid x, w_2)}{q_{yw_1z_1z_2}(w_1 \mid x, w_2)}$$

and $q_{yz_1}(y \mid x, w_1)$ doesn't depend upon x.

Nested Markov Model

Various equivalent formulations:

Factorization into Districts.
For each reachable R in $\mathcal{G}$,

$$q_R(x_R \,|\, x_{\mathrm{pa}(R)\setminus R}) = \prod_{D \in \mathcal{D}(\mathcal{G}[R])} f_D(x_{D \cup \mathrm{pa}(D)})$$

some functions f_D.

Nested Markov Model

Various equivalent formulations:

Factorization into Districts.
For each reachable R in $\mathcal{G}$,

$$q_R(x_R \,|\, x_{\mathrm{pa}(R)\setminus R}) = \prod_{D \in \mathcal{D}(\mathcal{G}[R])} f_D(x_{D \cup \mathrm{pa}(D)})$$

some functions f_D.

Weak Global Markov Property.
For each reachable R in $\mathcal{G}$,

$$A \text{ m-separated from } B \text{ by } C \text{ in } \mathcal{G}[R] \implies X_A \perp\!\!\!\perp X_B \,|\, X_C [q_R].$$

Nested Markov Model

Various equivalent formulations:

Factorization into Districts.
For each reachable R in $\mathcal{G}$,

$$q_R(x_R \,|\, x_{\mathrm{pa}(R)\setminus R}) = \prod_{D \in \mathcal{D}(\mathcal{G}[R])} f_D(x_{D \cup \mathrm{pa}(D)})$$

some functions f_D.

Weak Global Markov Property.
For each reachable R in $\mathcal{G}$,

$$A \text{ m-separated from } B \text{ by } C \text{ in } \mathcal{G}[R] \implies X_A \perp\!\!\!\perp X_B \,|\, X_C [q_R].$$

Ordered Local Markov Property.
For every intrinsic S and v maximal in S under some topological ordering,

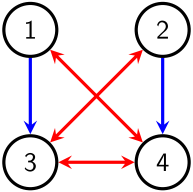
$$X_v \perp\!\!\!\perp X_{V \setminus \mathrm{mb}_{\mathcal{G}[S]}(v)} \,|\, X_{\mathrm{mb}_{\mathcal{G}[S]}(v)} \, [q_S].$$

Theorem. These are all equivalent.

Outline

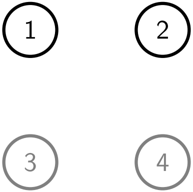
Heads and Tails

As established, we can factorize a graph into districts; however, finer factorizations are possible.



Heads and Tails

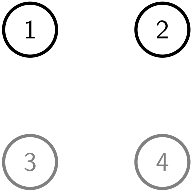
As established, we can factorize a graph into districts; however, finer factorizations are possible.



In the graph above, there is a single district, but $X_1 \perp\!\!\!\perp X_2$.

Heads and Tails

As established, we can factorize a graph into districts; however, finer factorizations are possible.

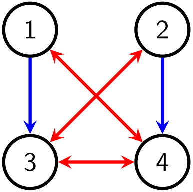


In the graph above, there is a single district, but $X_1 \perp\!\!\!\perp X_2$.
So could factorize this as

$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= p(x_1, x_2) p(x_3, x_4 \mid x_1, x_2) \\ &= p(x_1) p(x_2) p(x_3, x_4 \mid x_1, x_2). \end{aligned}$$

Heads and Tails

As established, we can factorize a graph into districts; however, finer factorizations are possible.



In the graph above, there is a single district, but $X_1 \perp\!\!\!\perp X_2$.
So could factorize this as

$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= p(x_1, x_2) p(x_3, x_4 \mid x_1, x_2) \\ &= p(x_1) p(x_2) p(x_3, x_4 \mid x_1, x_2). \end{aligned}$$

Note that the vertices $\{3, 4\}$ can't be d-separated from one another.

Heads and Tails

Definition

The **recursive head** associated with intrinsic set S is $H \equiv S \setminus \mathrm{pa}_{\mathcal{G}}(S)$.
The **tail** is $\mathrm{pa}_{\mathcal{G}}(S)$.

Heads and Tails

Definition

The **recursive head** associated with intrinsic set S is $H \equiv S \setminus \mathrm{pa}_{\mathcal{G}}(S)$.
The **tail** is $\mathrm{pa}_{\mathcal{G}}(S)$.

Recall that the Markov blanket for a fixable vertex is the whole intrinsic set and its parents $S \cup \mathrm{pa}_{\mathcal{G}}(S) = H \cup T$. So the head cannot be further divided:

$$p(x_S \mid x_{\mathrm{pa}(S) \setminus S}) = p(x_H \mid x_T) \cdot p(x_{S \setminus H} \mid x_{\mathrm{pa}(S) \setminus S}).$$

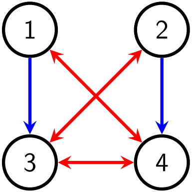
Heads and Tails

Definition

The **recursive head** associated with intrinsic set S is $H \equiv S \setminus \text{pa}_{\mathcal{G}}(S)$.
The **tail** is $\text{pa}_{\mathcal{G}}(S)$.

Recall that the Markov blanket for a fixable vertex is the whole intrinsic set and its parents $S \cup \text{pa}_{\mathcal{G}}(S) = H \cup T$. So the head cannot be further divided:

$$p(x_S \mid x_{\text{pa}(S)\setminus S}) = p(x_H \mid x_T) \cdot p(x_{S\setminus H} \mid x_{\text{pa}(S)\setminus S}).$$



But vertices in $S \setminus H$ may factorize:

$$\begin{aligned} &p(x_1, x_2, x_3, x_4) \\ &= p(x_3, x_4 \mid x_1, x_2) p(x_1, x_2) \end{aligned}$$

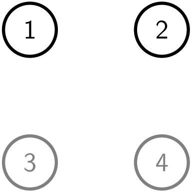
Heads and Tails

Definition

The **recursive head** associated with intrinsic set S is $H \equiv S \setminus \mathrm{pa}_{\mathcal{G}}(S)$. The **tail** is $\mathrm{pa}_{\mathcal{G}}(S)$.

Recall that the Markov blanket for a fixable vertex is the whole intrinsic set and its parents $S \cup \mathrm{pa}_{\mathcal{G}}(S) = H \cup T$. So the head cannot be further divided:

$$p(x_S \mid x_{\mathrm{pa}(S) \setminus S}) = p(x_H \mid x_T) \cdot p(x_{S \setminus H} \mid x_{\mathrm{pa}(S) \setminus S}).$$



But vertices in $S \setminus H$ may factorize:

$$\begin{aligned} &p(x_1, x_2, x_3, x_4) \\ &= p(x_3, x_4 \mid x_1, x_2) p(x_1, x_2) \\ &= p(x_3, x_4 \mid x_1, x_2) p(x_1) p(x_2). \end{aligned}$$

Factorizations

Recursively define a partition of reachable sets as follows. If R has multiple districts,

$$[R]_{\mathcal{G}} \equiv [D_1]_{\mathcal{G}} \cup \cdots \cup [D_k]_{\mathcal{G}};$$

Factorizations

Recursively define a partition of reachable sets as follows. If R has multiple districts,

$$[R]_{\mathcal{G}} \equiv [D_1]_{\mathcal{G}} \cup \cdots \cup [D_k]_{\mathcal{G}};$$

else R is intrinsic with head H, so

$$[R]_{\mathcal{G}} \equiv \{H\} \cup [R \setminus H]_{\mathcal{G}}.$$

Factorizations

Recursively define a partition of reachable sets as follows. If R has multiple districts,

$$[R]_{\mathcal{G}} \equiv [D_1]_{\mathcal{G}} \cup \cdots \cup [D_k]_{\mathcal{G}};$$

else R is intrinsic with head H, so

$$[R]_{\mathcal{G}} \equiv \{H\} \cup [R \setminus H]_{\mathcal{G}}.$$

Theorem (Head Factorization Property)

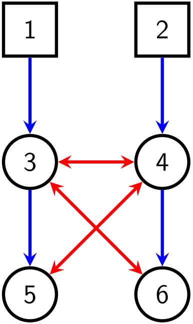
p obeys the nested Markov property for $\mathcal{G}$ if and only if for every reachable set R,

$$q_R(x_R \,|\, x_{\text{pa}(R)\setminus R}) = \prod_{H \in [R]_{\mathcal{G}}} q_H(x_H \,|\, x_T).$$

Here $q_H \equiv q_{S(H)}$ is density associated with intrinsic set for H.
(Recursive heads are in one-to-one correspondence with intrinsic sets.)

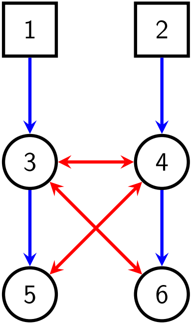
Heads and Tails

Recall, intrinsic sets are reachable districts:



Heads and Tails

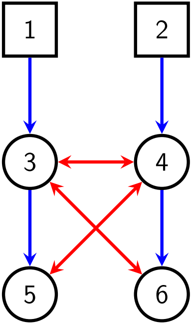
Recall, intrinsic sets are reachable districts:



intrinsic set	I	$\{3, 4, 5, 6\}$
recursive head	H	$\{5, 6\}$
tail	T	$\{1, 2, 3, 4\}$

Heads and Tails

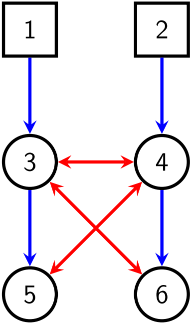
Recall, intrinsic sets are reachable districts:



intrinsic set	I	$\{3, 4, 5, 6\}$
recursive head	H	$\{5, 6\}$
tail	T	$\{1, 2, 3, 4\}$
intrinsic set	I	$\{3, 4\}$
recursive head	H	$\{3, 4\}$
tail	T	$\{1, 2\}$

Heads and Tails

Recall, intrinsic sets are reachable districts:



intrinsic set	I	$\{3,4,5,6\}$
recursive head	H	$\{5,6\}$
tail	T	$\{1,2,3,4\}$
intrinsic set	I	$\{3,4\}$
recursive head	H	$\{3,4\}$
tail	T	$\{1,2\}$

So

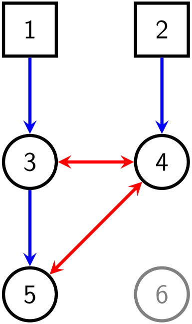
$$[\{3,4,5,6\}]_{\mathcal{G}} = \{\{3,4\},\{5,6\}\}.$$

Factorization:

$$q_{3456}(x_{3456} \mid x_{12}) = q_{56}(x_{56} \mid x_{1234}) \cdot q_{34}(x_{34} \mid x_{12})$$

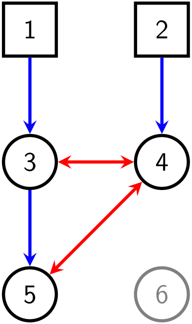
Heads and Tails

What if we fix 6 first?



Heads and Tails

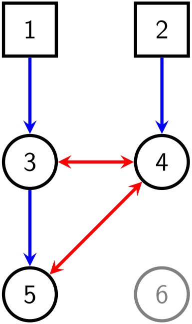
What if we fix 6 first?



intrinsic set	I	$\{3, 4, 5\}$
recursive head	H	$\{4, 5\}$
tail	T	$\{1, 2, 3\}$

Heads and Tails

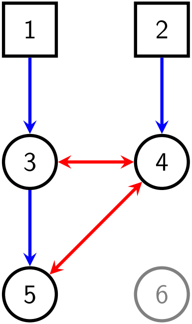
What if we fix 6 first?



intrinsic set	I	$\{3, 4, 5\}$
recursive head	H	$\{4, 5\}$
tail	T	$\{1, 2, 3\}$
intrinsic set	I	$\{3\}$
recursive head	H	$\{3\}$
tail	T	$\{1\}$

Heads and Tails

What if we fix 6 first?



intrinsic set	I	$\{3,4,5\}$
recursive head	H	$\{4,5\}$
tail	T	$\{1,2,3\}$
intrinsic set	I	$\{3\}$
recursive head	H	$\{3\}$
tail	T	$\{1\}$

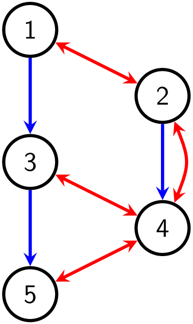
So

$$[\{3,4,5\}]_{\mathcal{G}} = \{\{3\},\{4,5\}\}.$$

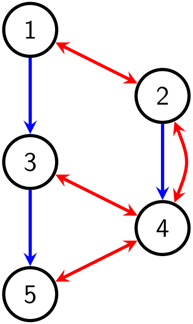
Factorization:

$$q_{345}(x_{345} \mid x_{12}) = q_{45}(x_{45} \mid x_{123}) \cdot q_3(x_3 \mid x_1)$$

Heads and Tails

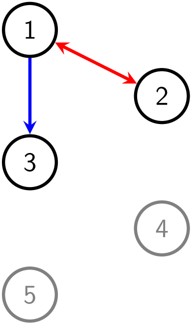


Heads and Tails



intrinsic set	I	$\{1,2,3,4,5\}$
recursive head	H	$\{4,5\}$
tail	T	$\{1,2,3\}$

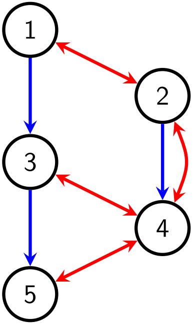
Heads and Tails



intrinsic set	I	$\{1,2,3,4,5\}$
recursive head	H	$\{4,5\}$
tail	T	$\{1,2,3\}$

intrinsic set	I	$\{1,2\}$
recursive head	H	$\{1,2\}$
tail	T	$\emptyset$
intrinsic set	I	$\{3\}$
recursive head	H	$\{3\}$
tail	T	$\{1\}$

Heads and Tails



intrinsic set	I	$\{1, 2, 3, 4, 5\}$
recursive head	H	$\{4, 5\}$
tail	T	$\{1, 2, 3\}$
intrinsic set	I	$\{1, 2\}$
recursive head	H	$\{1, 2\}$
tail	T	$\emptyset$
intrinsic set	I	$\{3\}$
recursive head	H	$\{3\}$
tail	T	$\{1\}$

Factorization:

$$q_{12345}(x_{12345}) = q_{45}(x_{45} \mid x_{123}) \cdot q_3(x_3 \mid x_1) \cdot q_{12}(x_{12}).$$

Outline

Parameterizations

Let $\mathcal{M}$ be a model (i.e. collection of probability distributions).

A **parameterization** is a continuous bijective map

$$\theta : \mathcal{M} \to \Theta$$

with continuous inverse, where Θ is an open subset of $\mathbb{R}^d$.

Parameterizations

Let $\mathcal{M}$ be a model (i.e. collection of probability distributions).

A **parameterization** is a continuous bijective map

$$\theta : \mathcal{M} \to \Theta$$

with continuous inverse, where Θ is an open subset of $\mathbb{R}^d$.

If θ, θ^{-1} are twice differentiable then this is a **smooth parameterization**.

Parameterizations

Let $\mathcal{M}$ be a model (i.e. collection of probability distributions).

A **parameterization** is a continuous bijective map

$$\theta : \mathcal{M} \rightarrow \Theta$$

with continuous inverse, where Θ is an open subset of $\mathbb{R}^d$.

If θ, θ^{-1} are twice differentiable then this is a **smooth parameterization**.

We will assume all variables are binary; this extends easily to the general categorical / discrete case.

Parameterization

Say binary distribution p *parameterized according to* $\mathcal{G}$ if[1]

$$p(x_V \mid x_W) = \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} \theta_H(x_T),$$

for some parameters $\theta_H(x_T)$ where $O = \{v : x_v = 0\}$.

[1] The definition of $[\cdot]_{\mathcal{G}}$ has to be extended to arbirary sets; see appendix.

Parameterization

Say binary distribution p *parameterized according to* $\mathcal{G}$ if[1]

$$p(x_V \mid x_W) = \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} \theta_H(x_T),$$

for some parameters $\theta_H(x_T)$ where $O = \{v : x_v = 0\}$.
Note: there is no need to assume that $\theta_H(x_T) \in [0, 1]$, this comes for free if $p(x_V \mid x_W) \geq 0$.

If suitable causal interpretation of model exists,

$$\begin{aligned}\theta_H(x_T) = q_S(0_H \mid x_T) &= p(0_H \mid x_{S \setminus H}, do(x_{T \setminus S})) \\ &\neq p(0_H \mid x_T).\end{aligned}$$

[1] The definition of $[\cdot]_{\mathcal{G}}$ has to be extended to arbirary sets; see appendix.

Parameterization

Say binary distribution p *parameterized according to* $\mathcal{G}$ if[1]

$$p(x_V \,|\, x_W) = \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} \theta_H(x_T),$$

for some parameters $\theta_H(x_T)$ where $O = \{v : x_v = 0\}$.
Note: there is no need to assume that $\theta_H(x_T) \in [0, 1]$, this comes for free if $p(x_V \,|\, x_W) \geq 0$.

If suitable causal interpretation of model exists,

$$\begin{aligned} \theta_H(x_T) = q_S(0_H \,|\, x_T) &= p(0_H \,|\, x_{S \setminus H}, do(x_{T \setminus S})) \\ &\neq p(0_H \,|\, x_T). \end{aligned}$$

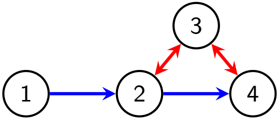
Theorem (Evans and Richardson, 2015)

p is parameterized according to $\mathcal{G}$ if and only if it recursively factorizes according to $\mathcal{G}$ (so $p \in \mathcal{N}(\mathcal{G})$).

---

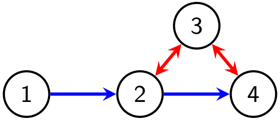
[1] The definition of $[\cdot]_{\mathcal{G}}$ has to be extended to arbirary sets; see appendix.

Probabilities



Example: how do we calculate $p(1_1, 0_2, 1_3, 1_4)$?

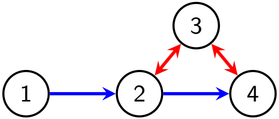
Probabilities



Example: how do we calculate $p(1_1, 0_2, 1_3, 1_4)$?
First,

$$p(1_1, 0_2, 1_3, 1_4) = q_1(1_1) \cdot q_{234}(0_2, 1_3, 1_4 \mid 1_1).$$

Probabilities

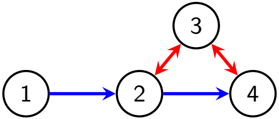


Example: how do we calculate $p(1_1, 0_2, 1_3, 1_4)$?
First,

$$p(1_1, 0_2, 1_3, 1_4) = q_1(1_1) \cdot q_{234}(0_2, 1_3, 1_4 \mid 1_1).$$

Then $q_1(1_1) = 1 - q_1(0_1) = 1 - \theta_1$.

Probabilities



Example: how do we calculate $p(1_1, 0_2, 1_3, 1_4)$?
First,

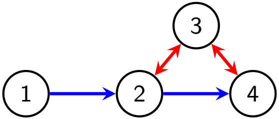
$$p(1_1, 0_2, 1_3, 1_4) = q_1(1_1) \cdot q_{234}(0_2, 1_3, 1_4 \mid 1_1).$$

Then $q_1(1_1) = 1 - q_1(0_1) = 1 - \theta_1$.

For the district $\{2, 3, 4\}$ get

$$\begin{aligned} &q_{234}(0_2, 1_3, 1_4 \mid x_1) \\ &= q_{234}(0_2 \mid x_1) - q_{234}(0_{23} \mid x_1) - q_{234}(0_{24} \mid x_1) + q_{234}(0_{234} \mid x_1) \end{aligned}$$

Probabilities



Example: how do we calculate $p(1_1, 0_2, 1_3, 1_4)$?
First,

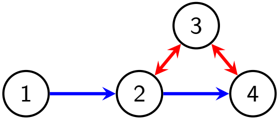
$$p(1_1, 0_2, 1_3, 1_4) = q_1(1_1) \cdot q_{234}(0_2, 1_3, 1_4 \mid 1_1).$$

Then $q_1(1_1) = 1 - q_1(0_1) = 1 - \theta_1$.

For the district $\{2, 3, 4\}$ get

$$\begin{aligned} &q_{234}(0_2, 1_3, 1_4 \mid x_1) \\ &= q_{234}(0_2 \mid x_1) - q_{234}(0_{23} \mid x_1) - q_{234}(0_{24} \mid x_1) + q_{234}(0_{234} \mid x_1) \\ &= \theta_2(x_1) - \theta_{23}(x_1) - \theta_2(x_1)\theta_4(0_2) + \theta_2(x_1)\theta_{34}(x_1, 0_2). \end{aligned}$$

Probabilities



Example: how do we calculate $p(1_1, 0_2, 1_3, 1_4)$?
First,

$$p(1_1, 0_2, 1_3, 1_4) = q_1(1_1) \cdot q_{234}(0_2, 1_3, 1_4 \mid 1_1).$$

Then $q_1(1_1) = 1 - q_1(0_1) = 1 - \theta_1$.

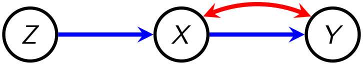
For the district $\{2, 3, 4\}$ get

$$\begin{aligned} &q_{234}(0_2, 1_3, 1_4 \mid x_1) \\ &= q_{234}(0_2 \mid x_1) - q_{234}(0_{23} \mid x_1) - q_{234}(0_{24} \mid x_1) + q_{234}(0_{234} \mid x_1) \\ &= \theta_2(x_1) - \theta_{23}(x_1) - \theta_2(x_1)\theta_4(0_2) + \theta_2(x_1)\theta_{34}(x_1, 0_2). \end{aligned}$$

Putting this all together:

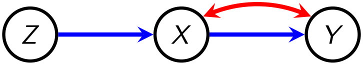
$$\begin{aligned} &p(1_1, 0_2, 1_3, 1_4) \\ &= \{1 - \theta_1\} \{\theta_2(1) - \theta_{23}(1) - \theta_2(1)\theta_4(0) + \theta_2(1)\theta_{34}(1, 0)\}. \end{aligned}$$

Example 1



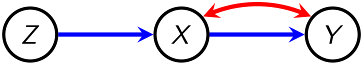
Intrinsic Sets	Z	X, Y	X

Example 1



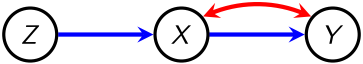
Intrinsic Sets	Z	X, Y	X
Heads	Z	Y	X

Example 1



Intrinsic Sets	Z	X, Y	X
Heads	Z	Y	X
Tails	$\emptyset$	Z, X	Z

Example 1



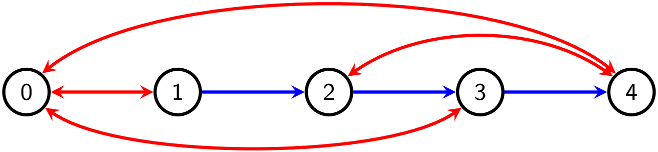
Intrinsic Sets	Z	X, Y	X
Heads	Z	Y	X
Tails	$\emptyset$	Z, X	Z

So parameterization is just

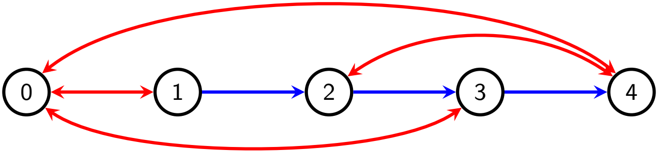
$$p(z = 0), \qquad p(x = 0 \mid z) \qquad p(y = 0 \mid x, z).$$

Model is saturated.

Example 2

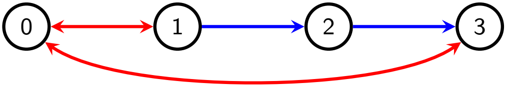


Example 2



$$p(0_0, 1_1, 1_2, 0_3, 0_4) = p(0_0, 1_1, 1_2, 0_3) \cdot q_4(0_4 \mid 0_0, 1_1, 1_2, 0_3)$$

Example 2

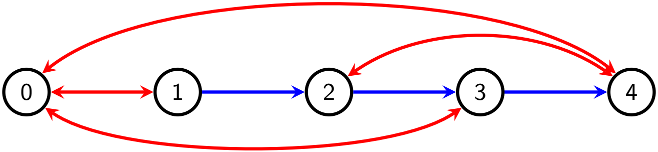


$$p(0_0, 1_1, 1_2, 0_3, 0_4) = p(0_0, 1_1, 1_2, 0_3) \cdot q_4(0_4 \mid 0_0, 1_1, 1_2, 0_3)$$
$$p(0_0, 1_1, 1_2, 0_3) = q_2(1_2 \mid 1_1) \cdot q_{013}(0_0, 1_1, 0_3 \mid 1_2)$$

Example 2

$$
\begin{aligned}
p(0_0, 1_1, 1_2, 0_3, 0_4) &= p(0_0, 1_1, 1_2, 0_3) \cdot q_4(0_4 \mid 0_0, 1_1, 1_2, 0_3) \\
p(0_0, 1_1, 1_2, 0_3) &= q_2(1_2 \mid 1_1) \cdot q_{013}(0_0, 1_1, 0_3 \mid 1_2) \\
q_{013}(0_0, 1_1, 0_3 \mid 1_2) &= q_{03}(0_0, 0_3 \mid 1_2) - q_{013}(0_0, 0_1, 0_3 \mid 1_2) \\
&= \theta_{03}(1) - \theta_{013}(1)
\end{aligned}
$$

Example 2



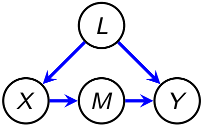
$$
\begin{aligned}
p(0_0, 1_1, 1_2, 0_3, 0_4) &= p(0_0, 1_1, 1_2, 0_3) \cdot q_4(0_4 \mid 0_0, 1_1, 1_2, 0_3) \\
p(0_0, 1_1, 1_2, 0_3) &= q_2(1_2 \mid 1_1) \cdot q_{013}(0_0, 1_1, 0_3 \mid 1_2) \\
q_{013}(0_0, 1_1, 0_3 \mid 1_2) &= q_{03}(0_0, 0_3 \mid 1_2) - q_{013}(0_0, 0_1, 0_3 \mid 1_2) \\
&= \theta_{03}(1) - \theta_{013}(1)
\end{aligned}
$$

so

$$
p(0_0, 1_1, 1_2, 0_3, 0_4) = \{1 - \theta_2(1)\} \{\theta_{03}(1) - \theta_{013}(1)\} \cdot \theta_4(0, 1, 1, 0).
$$

Motivation

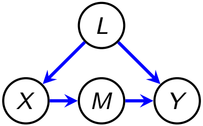
- So far we have shown how to estimate interventional distributions separately, but no guarantee these estimates are coherent.
- We also may have multiple identifying expressions: which one should we use?



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

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?

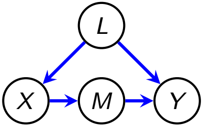


$p(Y \mid do(X))$
front door?
back door?
does it matter?

- We can test constraints separately, but ultimately don't have a way to check if the model is a good one.

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?

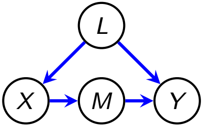


$p(Y \mid do(X))$
front door?
back door?
does it matter?

- 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).

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?

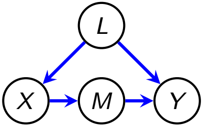


$p(Y \mid do(X))$
front door?
back door?
does it matter?

- 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.

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.

Outline

Exponential Families

Theorem

Let $\mathcal{N}(\mathcal{G})$ be the collection of binary distributions that recursively factorize according to $\mathcal{G}$. Then $\mathcal{N}(\mathcal{G})$ is a curved exponential family of dimension

$$d(\mathcal{G}) = \sum_{H \in \mathcal{H}(\mathcal{G})} 2^{|\operatorname{tail}(H)|}.$$

(This extends in the obvious way to finite discrete distributions.)

Exponential Families

Theorem

Let $\mathcal{N}(\mathcal{G})$ be the collection of binary distributions that recursively factorize according to $\mathcal{G}$. Then $\mathcal{N}(\mathcal{G})$ is a curved exponential family of dimension

$$d(\mathcal{G}) = \sum_{H \in \mathcal{H}(\mathcal{G})} 2^{|\operatorname{tail}(H)|}.$$

(This extends in the obvious way to finite discrete distributions.)

This justifies classical statistical theory:

- likelihood ratio tests have asymptotic χ^2-distribution;
- BIC as Laplace approximation of marginal likelihood.

Exponential Families

Theorem

Let $\mathcal{N}(\mathcal{G})$ be the collection of binary distributions that recursively factorize according to $\mathcal{G}$. Then $\mathcal{N}(\mathcal{G})$ is a curved exponential family of dimension

$$d(\mathcal{G}) = \sum_{H \in \mathcal{H}(\mathcal{G})} 2^{|\operatorname{tail}(H)|}.$$

(This extends in the obvious way to finite discrete distributions.)

This justifies classical statistical theory:

- likelihood ratio tests have asymptotic χ^2-distribution;
- BIC as Laplace approximation of marginal likelihood.

(Shpitser et al., 2013) give an alternative log-linear parametrization.

Algorithms for Model Search

Can, for example, use greedy edge replacement for a score-based approach (Evans and Richardson, 2010).

Shpitser et al. (2011) developed efficient algorithms for evaluating probabilities in the alternating sum.

Algorithms for Model Search

Can, for example, use greedy edge replacement for a score-based approach (Evans and Richardson, 2010).

Shpitser et al. (2011) developed efficient algorithms for evaluating probabilities in the alternating sum.

Currently no equivalent of PC algorithm for full nested model.

Can use FCI algorithm (Spirtes at al., 2000) for **ordinary Markov models** associated with ADMG (conditional independences only), in general this is a supermodel of the nested model (see Evans and Richardson, 2014).

Open Problems:

- Nested Markov equivalence;
- Constraint based search;
- Gaussian parametrization.

Outline

Completeness

Could the nested Markov property be further refined?

[2] and we are in the relative interior of the model space.

Completeness

Could the nested Markov property be further refined? No and Yes.

[2] and we are in the relative interior of the model space.

Completeness

Could the nested Markov property be further refined? No and Yes.

Theorem (Evans, 2015)

The constraints implied by the nested Markov model are algebraically equivalent to causal model with latent variables (with suff. large latent state-space).

[2] and we are in the relative interior of the model space.

Completeness

Could the nested Markov property be further refined? No and Yes.

Theorem (Evans, 2015)

The constraints implied by the nested Markov model are algebraically equivalent to causal model with latent variables (with suff. large latent state-space).

'Algebraically equivalent' = 'of the same dimension'.

So if the latent variable model is correct[2], fitting the nested model is asymptotically equivalent fitting the LV model.

[2] and we are in the relative interior of the model space.

Completeness

Could the nested Markov property be further refined? No and Yes.

Theorem (Evans, 2015)

The constraints implied by the nested Markov model are algebraically equivalent to causal model with latent variables (with suff. large latent state-space).

'Algebraically equivalent' = 'of the same dimension'.

So if the latent variable model is correct[2], fitting the nested model is asymptotically equivalent fitting the LV model.

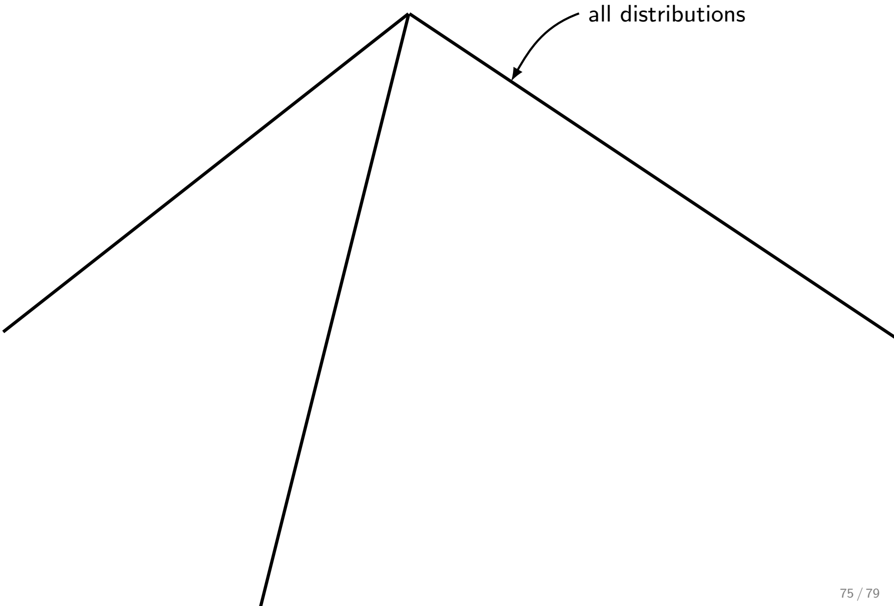
However, there are additional **inequality constraints**. e.g. Instrumental inequalities, CHSH inequalities etc.,

Potentially unsatisfactory as may not be a causal model corresponding to our inferred parameters.

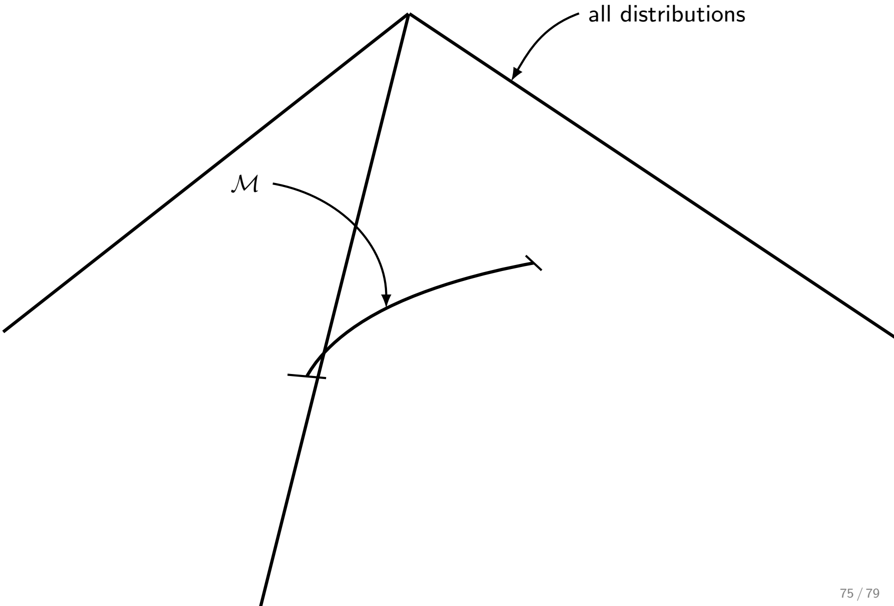
---

[2] and we are in the relative interior of the model space.

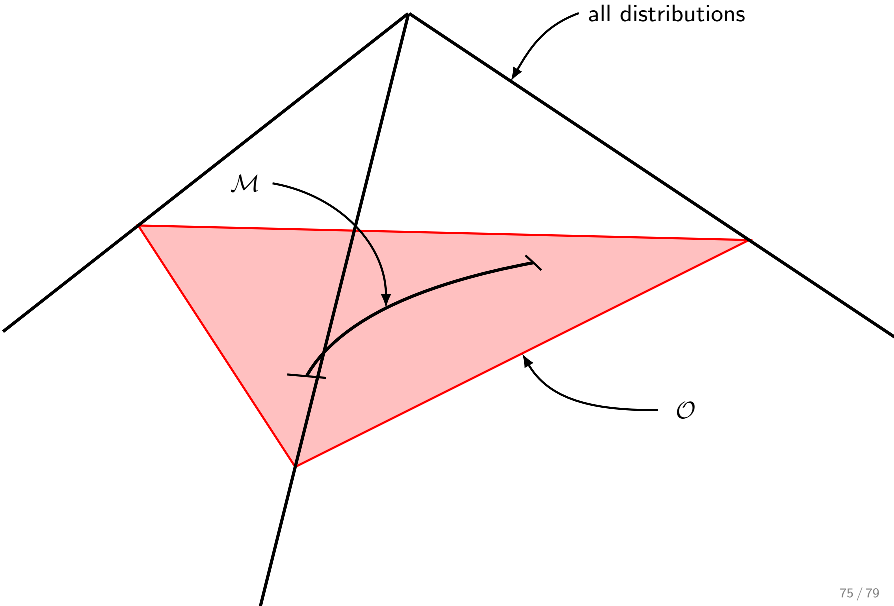
Big Picture



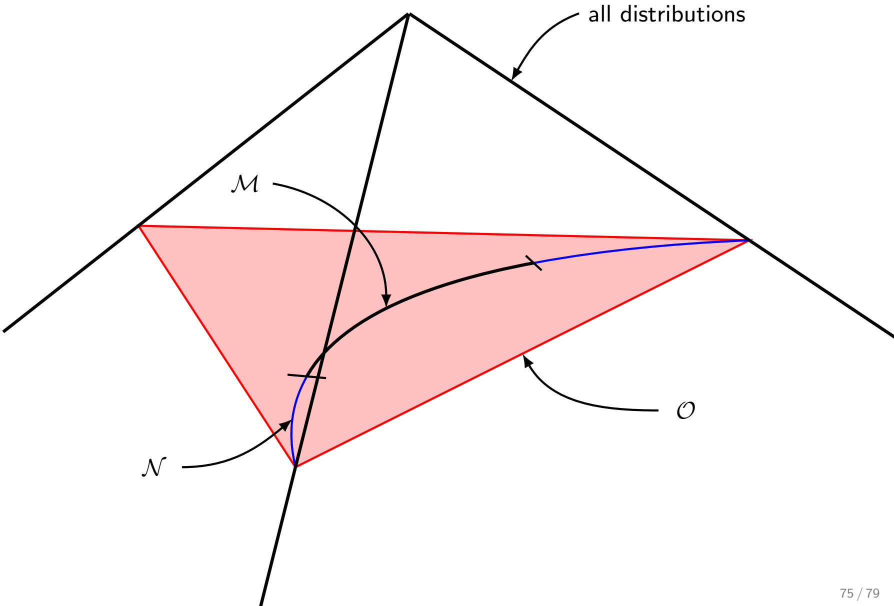
Big Picture



Big Picture



Big Picture



More on the nested Markov model

- Evans (2015) shows that the nested Markov model implies all *algebraic* constraints arising from the corresponding DAG with latent variables;
- A parameterization for discrete variables is given by Evans + R (2015), via an extension of the Möbius parametrization;

More on the nested Markov model

- Evans (2015) shows that the nested Markov model implies all *algebraic* constraints arising from the corresponding DAG with latent variables;
- A parameterization for discrete variables is given by Evans + R (2015), via an extension of the Möbius parametrization;
- In general a latent DAG model may also imply inequalities not captured by the nested Markov model: cf. the CHSH / Bell inequalities in quantum mechanics;

More on the nested Markov model

- Evans (2015) shows that the nested Markov model implies all *algebraic* constraints arising from the corresponding DAG with latent variables;
- A parameterization for discrete variables is given by Evans + R (2015), via an extension of the Möbius parametrization;
- In general a latent DAG model may also imply inequalities not captured by the nested Markov model: cf. the CHSH / Bell inequalities in quantum mechanics;
- The nested model may also be defined by constraints resulting from an algorithm given in (Tian, 2002b).

Future Work

- Characterizing nested Markov equivalence;
- Methods for inferring graph structure.

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 References

(Including earlier work on the ordinary Markov model.)

- Evans, R.J. and Richardson, T.S. – Maximum likelihood fitting of acyclic directed mixed graphs to binary data. *UAI*, 2010.
- 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.
- Spirtes, P., Glymour, G., Scheines, R. – *Causation Prediction and Search*, 2nd Edition, MIT Press, 2000.

Inequality References

- Bonet, B. – Instrumentality tests revisited, *UAI*, 2001.
- Cai Z., Kuroki, M., Pearl, J. and Tian, J. – Bounds on direct effects in the presence of confounded intermediate variables, *Biometrics*, 64(3):695–701, 2008.
- Evans, R.J. – Graphical methods for inequality constraints in marginalized DAGs, *MLSP*, 2012.
- Evans, R.J. – Margins of discrete Bayesian networks, *arXiv:1501.02103*, 2015.
- Kang, C. and Tian, J. – Inequality Constraints in Causal Models with Hidden Variables, *UAI*, 2006.
- Pearl, J. – On the testability of causal models with latent and instrumental variables, *UAI*, 1995.

Partition Function for General Sets

Let $\mathcal{I}(\mathcal{G})$ be the intrinsic sets of $\mathcal{G}$. Define a partial ordering $\prec$ on $\mathcal{I}(\mathcal{G})$ by $S_1 \prec S_2$ if and only if $S_1 \subset S_2$. This induces an isomorphic partial ordering on the corresponding recursive heads.

For any $B \subseteq V$ let

$$\Phi_{\mathcal{G}}(B) = \{H \subseteq B \mid H \text{ maximal under } \prec \text{ among heads contained in } B\};$$
$$\phi_{\mathcal{G}}(B) = \bigcup_{H \in \Phi_{\mathcal{G}}(B)} H.$$

So $\Phi_{\mathcal{G}}(B)$ is the 'maximal heads' in B, $\phi_{\mathcal{G}}(B)$ is their union.

Define (recursively)

$$[\emptyset]_{\mathcal{G}} \equiv \emptyset$$
$$[B]_{\mathcal{G}} \equiv \Phi_{\mathcal{G}}(B) \cup [\phi_{\mathcal{G}}(B)]_{\mathcal{G}}.$$

Then $[B]_{\mathcal{G}}$ is a partition of B.

d-Separation

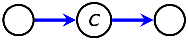
A **path** is a sequence of edges in the graph; vertices may not be repeated.

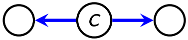
d-Separation

A **path** is a sequence of edges in the graph; vertices may not be repeated.

A path from v to w is **blocked** by $C \subseteq V \setminus \{v, w\}$ if either

(i) any non-collider is in C:



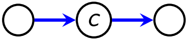


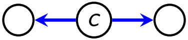
d-Separation

A **path** is a sequence of edges in the graph; vertices may not be repeated.

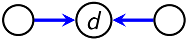
A path from v to w is **blocked** by $C \subseteq V \setminus \{v, w\}$ if either

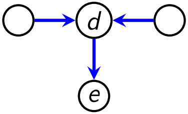
(i) any non-collider is in C:





(ii) or any collider is not in C, nor has descendants in C:





d-Separation

A **path** is a sequence of edges in the graph; vertices may not be repeated.

A path from v to w is **blocked** by $C \subseteq V \setminus \{v, w\}$ if either

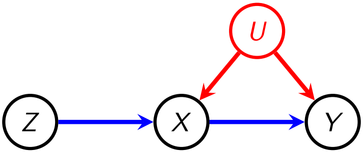
(i) any non-collider is in C:

(ii) or any collider is not in C, nor has descendants in C:

Two vertices v and w are **d-separated** given $C \subseteq V \setminus \{v, w\}$ if **all** paths are blocked.

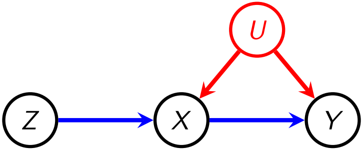
The IV Model

Assume four variable DAG shown, but U unobserved.



The IV Model

Assume four variable DAG shown, but U unobserved.



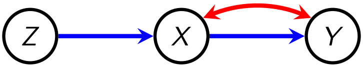
Marginalized DAG model

$$p(z, x, y) = \int p(u)\, p(z)\, p(x \mid z, u)\, p(y \mid x, u)\, du$$

Assume all observed variables are discrete; no assumption made about latent variables.

The IV Model

Assume four variable DAG shown, but U unobserved.



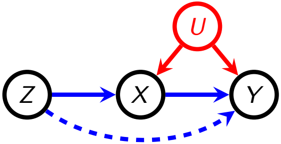
Marginalized DAG model

$$p(z, x, y) = \int p(u)\, p(z)\, p(x \mid z, u)\, p(y \mid x, u)\, du$$

Assume all observed variables are discrete; no assumption made about latent variables.

Nested Markov property gives saturated model, so true model of full dimension.

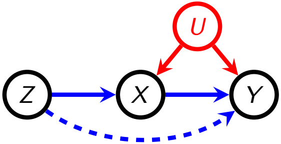
Instrumental Inequalities



The assumption $Z \not\to Y$ is important.
Can we check it?

Instrumental Inequalities

The assumption $Z \not\to Y$ is important.
Can we check it?

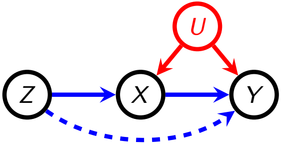


Pearl (1995) showed that if the observed variables are discrete,

$$\max_x \sum_y \max_z P(X = x, Y = y \mid Z = z) \leq 1. \qquad (*)$$

This is the **instrumental inequality**, and can be empirically tested.

Instrumental Inequalities



The assumption $Z \not\to Y$ is important.
Can we check it?

Pearl (1995) showed that if the observed variables are discrete,

$$\max_x \sum_y \max_z P(X = x, Y = y \mid Z = z) \leq 1. \qquad (*)$$

This is the **instrumental inequality**, and can be empirically tested.

If Z, X, Y are binary, then $(*)$ defines the marginalized DAG model (Bonet, 2001). e.g.

$$P(X = x, Y = 0 \mid Z = 0) + P(X = x, Y = 1 \mid Z = 1) \leq 1$$