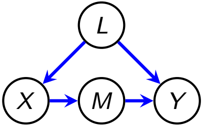
Non-parametric causal models II.

Robin Evans
www.stats.ox.ac.uk/∼evans
and
Thomas Richardson
www.stat.washington.edu/∼tsr

UAI Tutorial
12th July 2015

Model

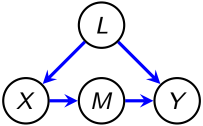
- So far we have shown how to estimate interventional distributions separately, but no guarantee these estimates are coherent.
- We also may have multiple identifying expressions: which one should we use?



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

Model

- 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?

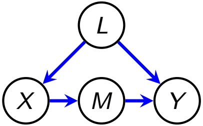


$p(Y \mid do(X))$
front door?
back door?
does it matter?

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

Model

- 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?

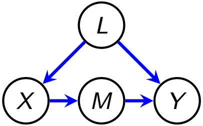


$p(Y \mid do(X))$
front door?
back door?
does it matter?

- 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).

Model

- 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?

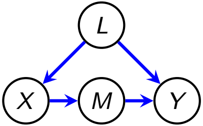


$p(Y \mid do(X))$
front door?
back door?
does it matter?

- 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.

Model

- 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?

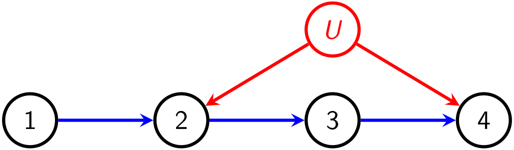
- We can test constraints separately, but ultimately don't have a way to check if the model is a good one.
- Being able to evaluate a likelihood would allow lots of standard inference techniques (e.g. LR, Bayesian).
- Even better, if model can be shown smooth we get nice asymptotics for free.

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

Outline

Outline

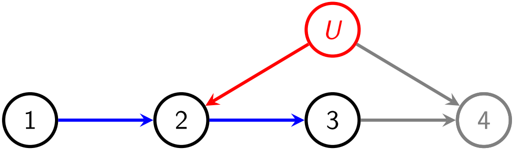
Ancestral Sets



$$p(x_1, x_2, x_3, x_4)$$
$$= \sum_u p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(x_4 \mid x_3, u)$$

.

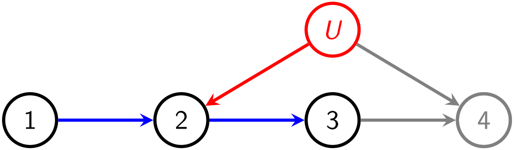
Ancestral Sets



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

.

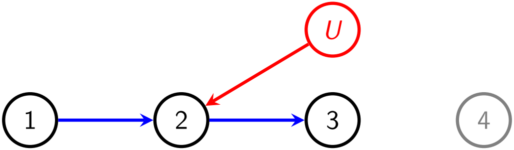
Ancestral Sets



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

.

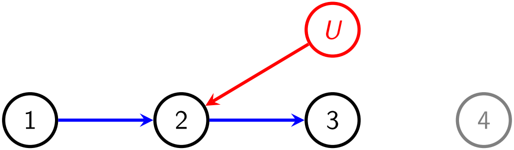
Ancestral Sets



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

.

Ancestral Sets



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

.

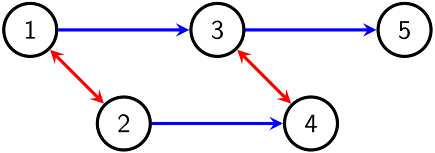
Ancestral Sets

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

Density has form corresponding to ancestral sub-graph.

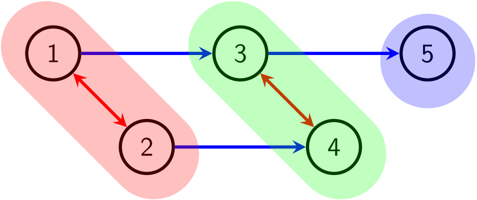
Factorization into Districts

District is a maximal set connected by latent variables / bidirected edges:



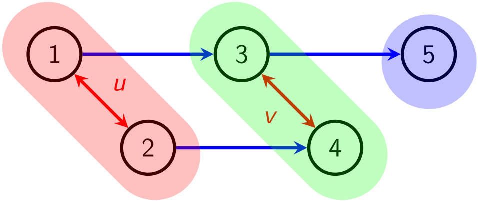
Factorization into Districts

District is a maximal set connected by latent variables / bidirected edges:



Factorization into Districts

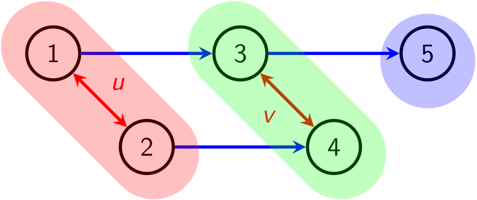
District is a maximal set connected by latent variables / bidirected edges:



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

Factorization into Districts

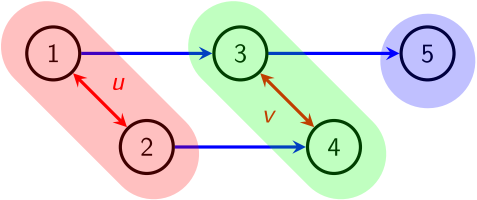
District is a maximal set connected by latent variables / bidirected edges:



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

Factorization into Districts

District is a maximal set connected by latent variables / bidirected edges:

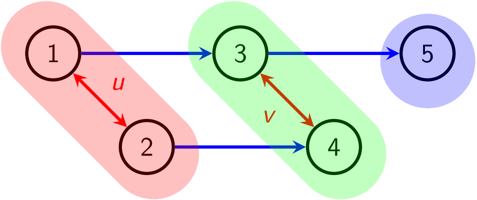


$$\sum_{u,v} p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)$$

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

Factorization into Districts

District is a maximal set connected by latent variables / bidirected edges:



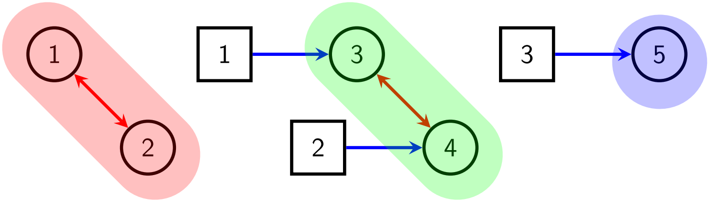
$$\sum_{u,v} \boxed{p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)}\ \boxed{p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)}\ \boxed{p(x_5 \mid x_3)}$$

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

$$= \boxed{q_{12}(x_1, x_2)} \cdot \boxed{q_{34}(x_3, x_4 \mid x_1, x_2)} \cdot \boxed{q_5(x_5 \mid x_3)}.$$

Factorization into Districts

District is a maximal set connected by latent variables / bidirected edges:



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

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

$$= q_{12}(x_1, x_2) \cdot q_{34}(x_3, x_4 \mid x_1, x_2) \cdot q_5(x_5 \mid x_3).$$

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

Each q_D piece should come from the model based on district subgraph and its parents ($\mathcal{G}[D]$).

Axiomatic Approach

We use these two rules to define our model.

Say (conditional) probability distribution p **recursively factorizes** according to CADMG $\mathcal{G}$ and write $p \in \mathcal{N}(\mathcal{G})$ if:

Axiomatic Approach

We use these two rules to define our model.

Say (conditional) probability distribution p **recursively factorizes** according to CADMG $\mathcal{G}$ and write $p \in \mathcal{N}(\mathcal{G})$ if:

1. **Ancestrality.**
$$\sum_{x_v} p(x_V \mid x_W) \in \mathcal{N}(\mathcal{G}_{-v})$$
for each childless $v \in V$.

Axiomatic Approach

We use these two rules to define our model.

Say (conditional) probability distribution p **recursively factorizes** according to CADMG $\mathcal{G}$ and write $p \in \mathcal{N}(\mathcal{G})$ if:

1. **Ancestrality.**
$$\sum_{x_v} p(x_V \mid x_W) \in \mathcal{N}(\mathcal{G}_{-v})$$
for each childless $v \in V$.

2. **Factorization into districts.**
$$p(x_V \mid x_W) = \prod_D q_D(x_D \mid x_{\mathrm{pa}(D) \setminus D})$$
for districts D, where $q_D \in \mathcal{N}(\mathcal{G}[D])$.

Axiomatic Approach

We use these two rules to define our model.

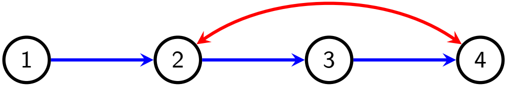
Say (conditional) probability distribution p **recursively factorizes** according to CADMG $\mathcal{G}$ and write $p \in \mathcal{N}(\mathcal{G})$ if:

1. **Ancestrality.**
$$\sum_{x_v} p(x_V \mid x_W) \in \mathcal{N}(\mathcal{G}_{-v})$$
for each childless $v \in V$.
2. **Factorization into districts.**
$$p(x_V \mid x_W) = \prod_D q_D(x_D \mid x_{\mathrm{pa}(D) \setminus D})$$
for districts D, where $q_D \in \mathcal{N}(\mathcal{G}[D])$.

Note that one can iterate between 1 and 2.

This defines the **nested Markov model** $\mathcal{N}(\mathcal{G})$.

Verma Example



X_4 childless,

Verma Example

X_4 childless, so if $p \in \mathcal{N}(\mathcal{G})$, then

$$p(x_1, x_2, x_3) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_2),$$

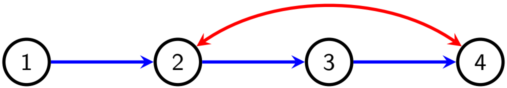
Verma Example

X_4 childless, so if $p \in \mathcal{N}(\mathcal{G})$, then

$$p(x_1, x_2, x_3) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_2),$$

and therefore $X_1 \perp\!\!\!\perp X_3 \mid X_2$.

Verma Example



Axiom 2:

$$p(x_1, x_2, x_3, x_4) = q_1(x_1) \cdot q_3(x_3 \mid x_2) \cdot q_{24}(x_2, x_4 \mid x_1, x_3).$$

Verma Example

Axiom 2:

$$p(x_1, x_2, x_3, x_4) = q_1(x_1) \cdot q_3(x_3 \mid x_2) \cdot q_{24}(x_2, x_4 \mid x_1, x_3).$$

Can consider the district $\{2, 4\}$ and factor q_{24}...

Verma Example

Axiom 2:

$$p(x_1, x_2, x_3, x_4) = q_1(x_1) \cdot q_3(x_3 \mid x_2) \cdot q_{24}(x_2, x_4 \mid x_1, x_3).$$

Can consider the district $\{2, 4\}$ and factor q_{24}...
and then marginalize X_2.

We see that $X_1 \perp\!\!\!\perp X_3, X_4 \, [q_{24}]$.

Verma Example

Axiom 2:

$$p(x_1, x_2, x_3, x_4) = q_1(x_1) \cdot q_3(x_3 \mid x_2) \cdot q_{24}(x_2, x_4 \mid x_1, x_3).$$

Can consider the district $\{2, 4\}$ and factor q_{24}...
and then marginalize X_2.

We see that $X_1 \perp\!\!\!\perp X_3, X_4 \, [q_{24}]$.

This places a non-trivial constraint on p.

Relationship to Fixing

Could also recursively define a model $\mathcal{N}'$ by fixing:

$$p \in \mathcal{N}'(\mathcal{G}) \implies \phi_v(p) \in \mathcal{N}'(\phi_v(\mathcal{G}))$$

for any v fixable in $\mathcal{G}$.

Relationship to Fixing

Could also recursively define a model $\mathcal{N}'$ by fixing:

$$p \in \mathcal{N}'(\mathcal{G}) \implies \phi_v(p) \in \mathcal{N}'(\phi_v(\mathcal{G}))$$

for any v fixable in $\mathcal{G}$.

The pair of operations used in recursive factorization is less rich than those allowed by fixing, but...

Theorem

The recursive factorization and fixing models are identical: $\mathcal{N}(\mathcal{G}) = \mathcal{N}'(\mathcal{G})$.

Relationship to Fixing

Could also recursively define a model $\mathcal{N}'$ by fixing:

$$p \in \mathcal{N}'(\mathcal{G}) \implies \phi_v(p) \in \mathcal{N}'(\phi_v(\mathcal{G}))$$

for any v fixable in $\mathcal{G}$.

The pair of operations used in recursive factorization is less rich than those allowed by fixing, but...

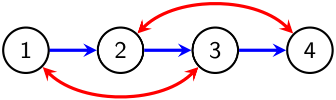
Theorem

The recursive factorization and fixing models are identical: $\mathcal{N}(\mathcal{G}) = \mathcal{N}'(\mathcal{G})$.

The recursive factorization model is useful for parameterization proofs.

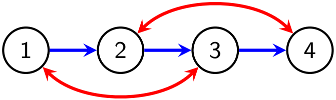
Relationship to Fixing

Recall that to 'fix' a vertex, it must not have children in its district.
Equivalent to splitting, marginalizing, and then pasting back together.

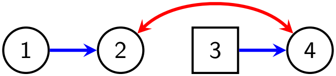


Relationship to Fixing

Recall that to ‘fix’ a vertex, it must not have children in its district.
Equivalent to splitting, marginalizing, and then pasting back together.

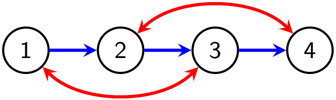


fix $\rightarrow$

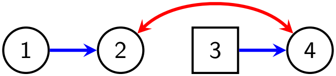


Relationship to Fixing

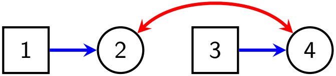
Recall that to ‘fix’ a vertex, it must not have children in its district.
Equivalent to splitting, marginalizing, and then pasting back together.

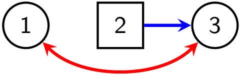


fix $\rightarrow$



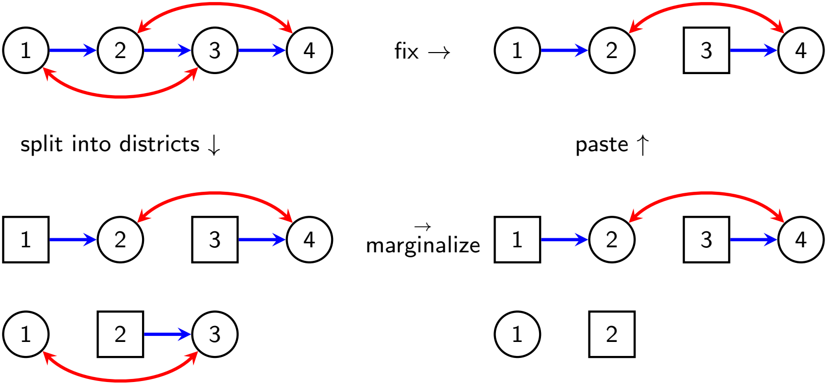
split into districts $\downarrow$



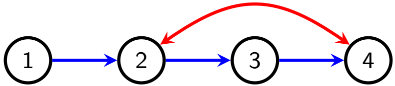


Relationship to Fixing

Recall that to 'fix' a vertex, it must not have children in its district.
Equivalent to splitting, marginalizing, and then pasting back together.

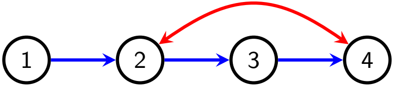


Notations



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

Notations



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

Theorem (Richardson, Shpitser, Robins, 201x)

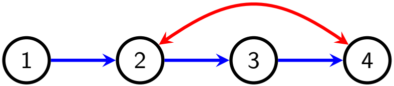
For a positive distribution $p \in \mathcal{N}(\mathcal{G})$ and vertices v_1, v_2 that are fixable in $\mathcal{G}$,

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

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.

Notations



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

Theorem (Richardson, Shpitser, Robins, 201x)

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

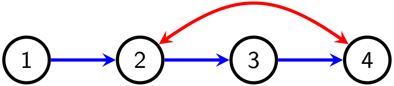
$$(\phi_{v_1} \circ \phi_{v_2})(p) = (\phi_{v_2} \circ \phi_{v_1})(p).$$

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 $\phi_{V \setminus R}$.

Notations



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

Theorem (Richardson, Shpitser, Robins, 201x)

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

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

Reachable CADMGs

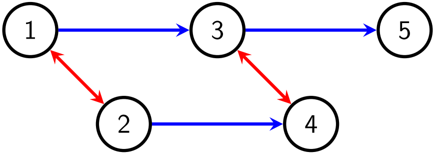
Note that $\mathcal{G}[R]$ is always just the CADMG with:

- 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:

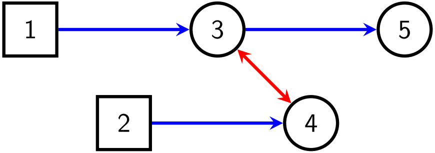
- random vertices R,
- fixed vertices $\text{pa}_{\mathcal{G}}(R) \setminus R$,
- induced edges from $\mathcal{G}$ among R and of the form $\text{pa}_{\mathcal{G}}(R) \to R$.



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 $\mathrm{pa}_{\mathcal{G}}(R) \to R$.

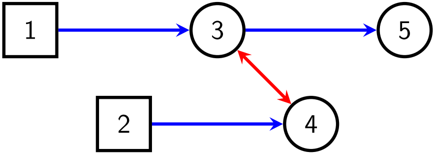


Graph shown is $\mathcal{G}[\{3, 4, 5\}]$.

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 $\text{pa}_{\mathcal{G}}(R) \to R$.

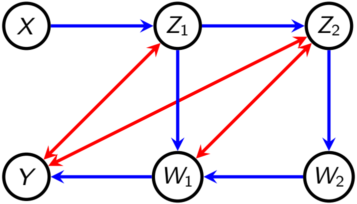


Graph shown is $\mathcal{G}[\{3, 4, 5\}]$.

Also recall that

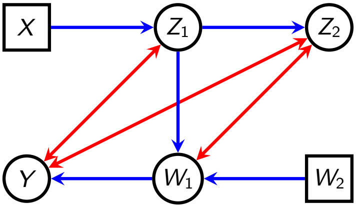
$$q_R(x_R \mid x_{\text{pa}(R) \setminus R}) = p(x_R \mid do(x_{V \setminus R})).$$

Example



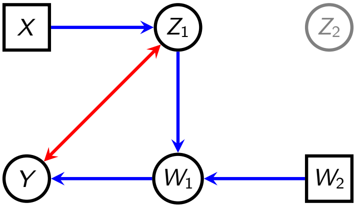
$p(x, y, w_1, w_2, z_1, z_2)$

Example



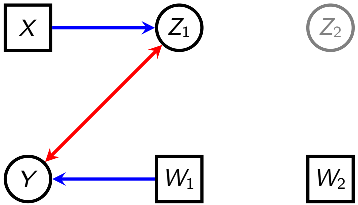
$$q_{yw_1z_1z_2}(y, w_1, z_1, z_2 \mid x, w_2) = \frac{p(x, y, w_1, w_2, z_1, z_2)}{p(x)p(w_2 \mid z_2)}$$

Example



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

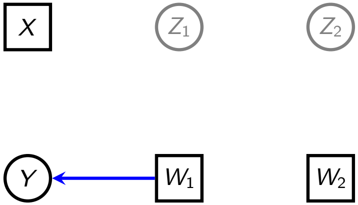
Example



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

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

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 $\mathcal{G}$,

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

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 $\mathcal{G}$,

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

some functions f_D.

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

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

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

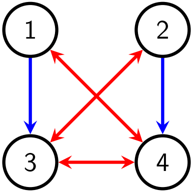
$$X_v \perp\!\!\!\perp X_{V \setminus \mathrm{mb}_{\mathcal{G}[S]}(v)} \mid X_{\mathrm{mb}_{\mathcal{G}[S]}(v)} \, [q_S].$$

Theorem. These are all equivalent.

Outline

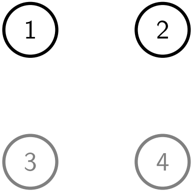
Heads and Tails

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



Heads and Tails

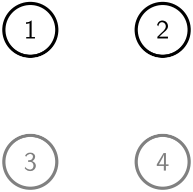
As established, we can factorize a graph into districts; however, finer factorizations are possible.



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.

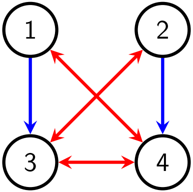


In the graph above, there is a single district, but $X_1 \perp\!\!\!\perp X_2$.
So could factorize this as

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

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 $H \equiv S \setminus \text{pa}_{\mathcal{G}}(S)$.
The **tail** is $\text{pa}_{\mathcal{G}}(S)$.

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:

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

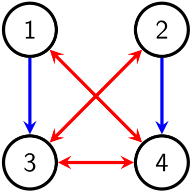
Heads and Tails

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:

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



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

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

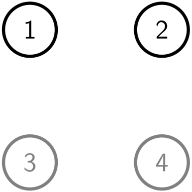
Heads and Tails

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)

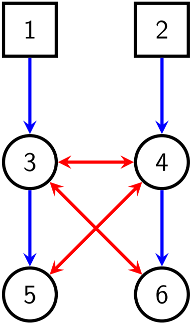
p obeys the nested Markov property for $\mathcal{G}$ if and only if for every reachable set R,

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

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.)

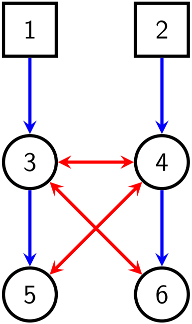
Heads and Tails

Recall, intrinsic sets are reachable districts:



Heads and Tails

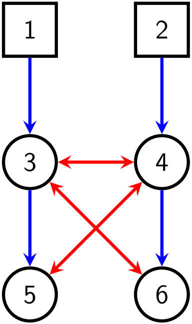
Recall, intrinsic sets are reachable districts:



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

Heads and Tails

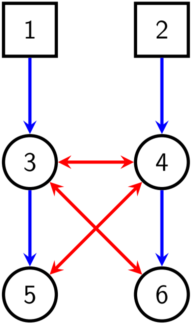
Recall, intrinsic sets are reachable districts:



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

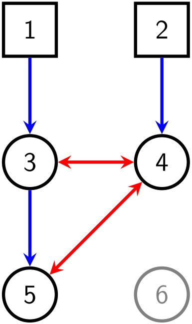
$$[\{3, 4, 5, 6\}]_{\mathcal{G}} = \{\{3, 4\}, \{5, 6\}\}.$$

Factorization:

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

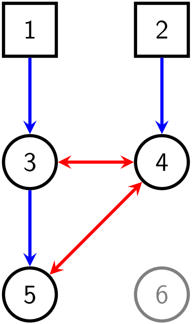
Heads and Tails

What if we fix 6 first?



Heads and Tails

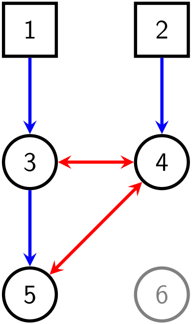
What if we fix 6 first?



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

Heads and Tails

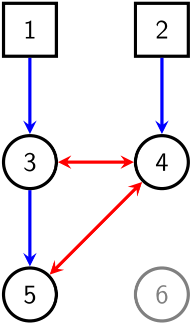
What if we fix 6 first?



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	$\{1\}$

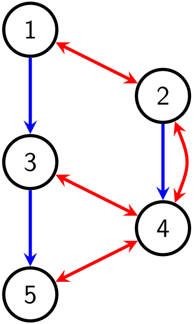
So

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

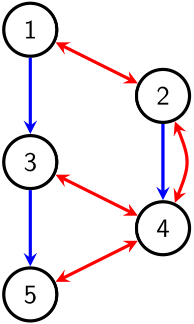
Factorization:

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

Heads and Tails

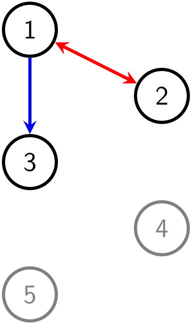


Heads and Tails



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

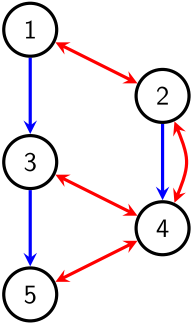
Heads and Tails



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

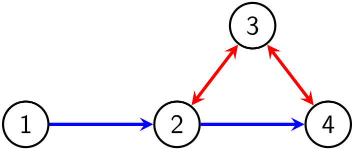
$$\theta : \mathcal{M} \to \Theta$$

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**.

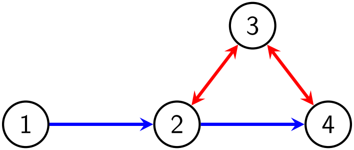
We will assume all variables are binary; this extends easily to the general categorical / discrete case.

Factorization into Districts



We'd like a parametrization which exhibits the axioms directly.
Then all reachable subgraphs will be taken care of too.

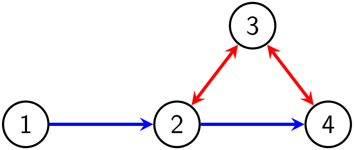
Factorization into Districts



We'd like a parametrization which exhibits the axioms directly.
Then all reachable subgraphs will be taken care of too.

The Game: proceed inductively to explicitly construct θ, and assume all reachable sub-graphs can be parameterized .

Factorization into Districts



We'd like a parametrization which exhibits the axioms directly.
Then all reachable subgraphs will be taken care of too.

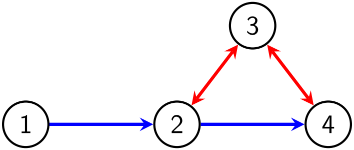
The Game: proceed inductively to explicitly construct θ, and assume all reachable sub-graphs can be parameterized.

If $\mathcal{G}$ has multiple districts $\mathcal{D}$, then by Axiom 1

$$p(x_V \mid x_W) = \prod_{D \in \mathcal{D}(\mathcal{G})} q_D(x_D \mid x_{\mathrm{pa}(D) \setminus D});$$

so parameterize each q_D according to $\mathcal{G}[D]$ separately (parameter cut).

Factorization into Districts



We'd like a parametrization which exhibits the axioms directly.
Then all reachable subgraphs will be taken care of too.

The Game: proceed inductively to explicitly construct θ, and assume all reachable sub-graphs can be parameterized.

If $\mathcal{G}$ has multiple districts $\mathcal{D}$, then by Axiom 1

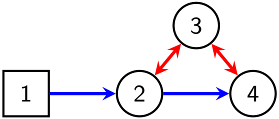
$$p(x_V \mid x_W) = \prod_{D \in \mathcal{D}(\mathcal{G})} q_D(x_D \mid x_{\mathrm{pa}(D) \setminus D});$$

so parameterize each q_D according to $\mathcal{G}[D]$ separately (parameter cut).
E.g.

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

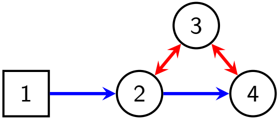
Note for a DAG this is usual CPTs.

Marginalization



To satisfy Axiom 2, we'd like ancestral margins of $p(x_{234} \mid x_1)$ to factorize according to CADMG.

Marginalization

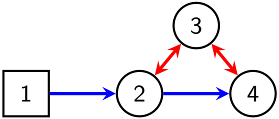


To satisfy Axiom 2, we'd like ancestral margins of $p(x_{234} \mid x_1)$ to factorize according to CADMG.

$$p(x_2, x_3, 1_4 \mid x_1) + p(x_2, x_3, 0_4 \mid x_1) = p(x_2, x_3 \mid x_1)$$

and $p(x_{23} \mid x_1)$ should by parameterized according to $\mathcal{G}_{-4}$.

Marginalization



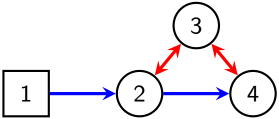
To satisfy Axiom 2, we'd like ancestral margins of $p(x_{234} \,|\, x_1)$ to factorize according to CADMG.

$$p(x_2, x_3, 1_4 \,|\, x_1) + p(x_2, x_3, 0_4 \,|\, x_1) = p(x_2, x_3 \,|\, x_1)$$

and $p(x_{23} \,|\, x_1)$ should by parameterized according to $\mathcal{G}_{-4}$.
Can use this to define probabilities where some entries of head are 1.

$$p(x_2, x_3, 1_4 \,|\, x_1) = \boxed{p(x_2, x_3 \,|\, x_1)} - p(x_2, x_3, 0_4 \,|\, x_1).$$

Marginalization



To satisfy Axiom 2, we'd like ancestral margins of $p(x_{234} \mid x_1)$ to factorize according to CADMG.

$$p(x_2, x_3, 1_4 \mid x_1) + p(x_2, x_3, 0_4 \mid x_1) = p(x_2, x_3 \mid x_1)$$

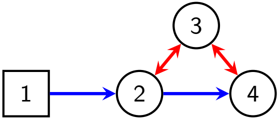
and $p(x_{23} \mid x_1)$ should by parameterized according to $\mathcal{G}_{-4}$.
Can use this to define probabilities where some entries of head are 1.

$$p(x_2, x_3, 1_4 \mid x_1) = p(x_2, x_3 \mid x_1) - p(x_2, x_3, 0_4 \mid x_1).$$

Repeat until all vertices in recursive head are 0; e.g.

$$\begin{aligned} &p(x_2, 1_3, 1_4 \mid x_1) \\ &= p(x_2 \mid x_1) - p(x_2, 0_3 \mid x_1) - p(x_2, 0_4 \mid x_1) + p(x_2, 0_3, 0_4 \mid x_1). \end{aligned}$$

Marginalization



To satisfy Axiom 2, we'd like ancestral margins of $p(x_{234} \mid x_1)$ to factorize according to CADMG.

$$p(x_2, x_3, 1_4 \mid x_1) + p(x_2, x_3, 0_4 \mid x_1) = p(x_2, x_3 \mid x_1)$$

and $p(x_{23} \mid x_1)$ should by parameterized according to $\mathcal{G}_{-4}$.
Can use this to define probabilities where some entries of head are 1.

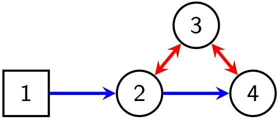
$$p(x_2, x_3, 1_4 \mid x_1) = p(x_2, x_3 \mid x_1) - p(x_2, x_3, 0_4 \mid x_1).$$

Repeat until all vertices in recursive head are 0; e.g.

$$\begin{aligned} &p(x_2, 1_3, 1_4 \mid x_1) \\ &= p(x_2 \mid x_1) - p(x_2, 0_3 \mid x_1) - p(x_2, 0_4 \mid x_1) + p(x_2, 0_3, 0_4 \mid x_1). \end{aligned}$$

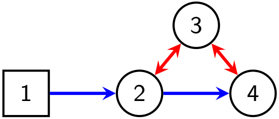
So every term represents an ancestral sub-graph, except for final term where every variable in the recursive head is 0.

Example



Now, how to deal with $p(x_2, 0_3, 0_4 \mid x_1)$?
We're now 'stuck' precisely when we get a full head of 0s.

Example

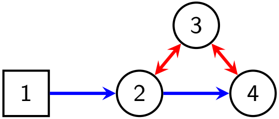


Now, how to deal with $p(x_2, 0_3, 0_4 \mid x_1)$?
We're now 'stuck' precisely when we get a full head of 0s.

We can use out finer factorization once:

$$p(x_2, 0_3, 0_4 \mid x_1) = p(0_3, 0_4 \mid x_1, x_2) \cdot \boxed{p(x_2 \mid x_1)}$$

Example

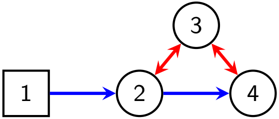


Now, how to deal with $p(x_2, 0_3, 0_4 \mid x_1)$?
We're now 'stuck' precisely when we get a full head of 0s.

We can use out finer factorization once:

$$\begin{aligned} p(x_2, 0_3, 0_4 \mid x_1) &= p(0_3, 0_4 \mid x_1, x_2) \cdot \boxed{p(x_2 \mid x_1)} \\ &\equiv \theta_{34}(x_1, x_2) \cdot p(x_2 \mid x_1). \end{aligned}$$

Example



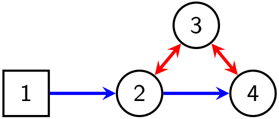
Now, how to deal with $p(x_2, 0_3, 0_4 \mid x_1)$?
We're now 'stuck' precisely when we get a full head of 0s.

We can use out finer factorization once:

$$\begin{aligned} p(x_2, 0_3, 0_4 \mid x_1) &= p(0_3, 0_4 \mid x_1, x_2) \cdot p(x_2 \mid x_1) \\ &\equiv \theta_{34}(x_1, x_2) \cdot p(x_2 \mid x_1). \end{aligned}$$

Have a collection of parameters $\theta_{34}(x_1, x_2)$ associated with the head $H = \{3, 4\}$ conditional upon the tail $\{1, 2\}$.

Example



Now, how to deal with $p(x_2, 0_3, 0_4 \mid x_1)$?
We're now 'stuck' precisely when we get a full head of 0s.

We can use out finer factorization once:

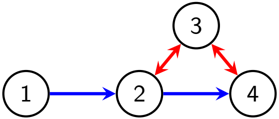
$$\begin{aligned} p(x_2, 0_3, 0_4 \mid x_1) &= p(0_3, 0_4 \mid x_1, x_2) \cdot p(x_2 \mid x_1) \\ &\equiv \theta_{34}(x_1, x_2) \cdot p(x_2 \mid x_1). \end{aligned}$$

Have a collection of parameters $\theta_{34}(x_1, x_2)$ associated with the head $H = \{3, 4\}$ conditional upon the tail $\{1, 2\}$.

Generally parameters are

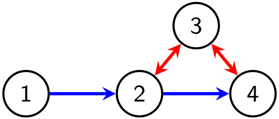
$$\theta_H(x_T) \equiv q_H(0_H \mid x_T), \qquad \text{for all heads } H, x_T.$$

Probabilities



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

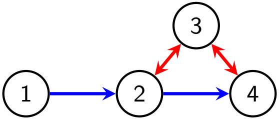
Probabilities



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

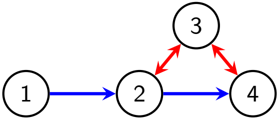


Example: how do we calculate $p(1_1, 0_2, 1_3, 1_4)$?
First,

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

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,

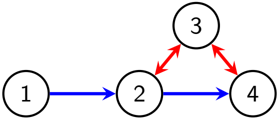
$$p(1_1, 0_2, 1_3, 1_4) = q_1(1_1) \cdot q_{234}(0_2, 1_3, 1_4 \mid 1_1).$$

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,

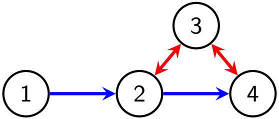
$$p(1_1, 0_2, 1_3, 1_4) = q_1(1_1) \cdot q_{234}(0_2, 1_3, 1_4 \mid 1_1).$$

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:

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

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 $q_H(x_T)$ 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 $q_H(x_T)$ 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]

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

for some parameters $q_H(x_T)$ 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,

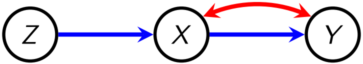
$$\begin{aligned}\theta_H(x_T) = q_S(0_H \mid x_T) &= p(0_H \mid x_{S \setminus H}, do(x_{T \setminus S})) \\ &\neq p(0_H \mid x_T).\end{aligned}$$

Theorem (Evans and Richardson, forthcoming)

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

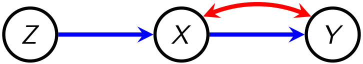
[1] The definition of $[\cdot]_{\mathcal{G}}$ has to be extended to arbirary sets; see appendix.

Example 1



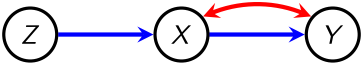
Intrinsic Sets	Z	X, Y	X

Example 1



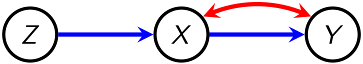
Intrinsic Sets	Z	X, Y	X
Heads	Z	Y	X

Example 1



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

Example 1



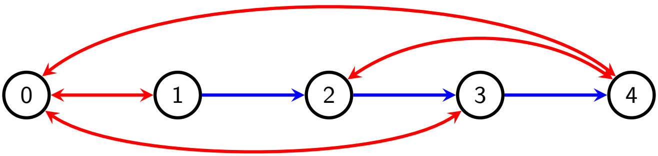
Intrinsic Sets	Z	X, Y	X
Heads	Z	Y	X
Tails	$\emptyset$	Z, X	Z

So parameterization is just

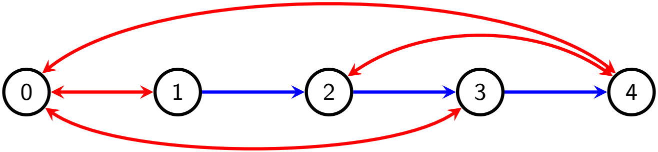
$$p(z = 0), \qquad p(x = 0 \mid z) \qquad p(y = 0 \mid x, z).$$

Model is saturated.

Example 2

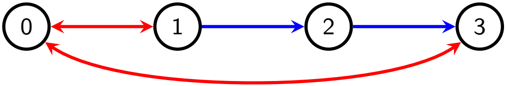


Example 2



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

Example 2

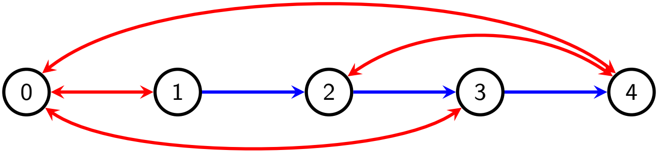


$$p(0_0, 1_1, 1_2, 0_3, 0_4) = p(0_0, 1_1, 1_2, 0_3) \cdot q_4(0_4 \mid 0_0, 1_1, 1_2, 0_3)$$
$$p(0_0, 1_1, 1_2, 0_3) = q_2(1_2 \mid 1_1) \cdot q_{013}(0_0, 1_1, 0_3 \mid 1_2)$$

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



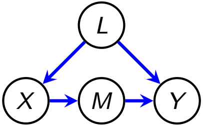
$$
\begin{aligned}
p(0_0, 1_1, 1_2, 0_3, 0_4) &= p(0_0, 1_1, 1_2, 0_3) \cdot q_4(0_4 \mid 0_0, 1_1, 1_2, 0_3) \\
p(0_0, 1_1, 1_2, 0_3) &= q_2(1_2 \mid 1_1) \cdot q_{013}(0_0, 1_1, 0_3 \mid 1_2) \\
q_{013}(0_0, 1_1, 0_3 \mid 1_2) &= q_{03}(0_0, 0_3 \mid 1_2) - q_{013}(0_0, 0_1, 0_3 \mid 1_2) \\
&= \theta_{03}(1)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

so

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

Model

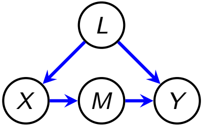
- So far we have shown how to estimate interventional distributions separately, but no guarantee these estimates are coherent.
- We also may have multiple identifying expressions: which one should we use?



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

Model

- 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?

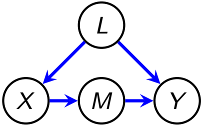


$p(Y \mid do(X))$
front door?
back door?
does it matter?

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

Model

- 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?

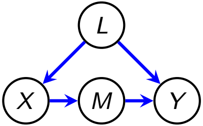


$p(Y \mid do(X))$
front door?
back door?
does it matter?

- 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).

Model

- 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?

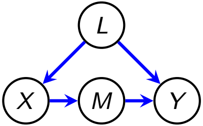


$p(Y \mid do(X))$
front door?
back door?
does it matter?

- 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.

Model

- 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.

Can also parameterize as GLM response model (Shpitser et al., 2013).

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 mixed graphical models** (conditional independences only), which is generally a supermodel of nested (see Evans and Richardson, 2014).

Parameterization References

Evans – Graphs for margins of Bayesian networks, *arXiv:1408.1809*, 2014.

Evans and Richardson – Maximum likelihood fitting of acyclic directed mixed graphs to binary data. *UAI*, 2010.

Evans and Richardson – Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, 2014.

Shpitser, Richardson, Robins. An efficient algorithm for computing interventional distributions in latent variable causal models. *UAI*, 2011.

Shpitser, Richardson, Robins and Evans – Parameter and structure learning in nested Markov models. *UAI*, 2012.

Shpitser, Evans, Richardson, and Robins – Sparse nested Markov models with log-linear parameters. *UAI*, 2013.

Shpitser, Evans, Richardson, and Robins – Introduction to Nested Markov Models. *Behaviormetrika*, 2014.

Spirtes, Glymour, Scheines – *Causation Prediction and Search*, 2nd Edition, MIT Press, 2000.

Outline

Completeness

How do we know there isn't a 'third' axiom we could invoke?

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

Completeness

How do we know there isn't a 'third' axiom we could invoke?

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

How do we know there isn't a 'third' axiom we could invoke?

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

How do we know there isn't a 'third' axiom we could invoke?

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'.

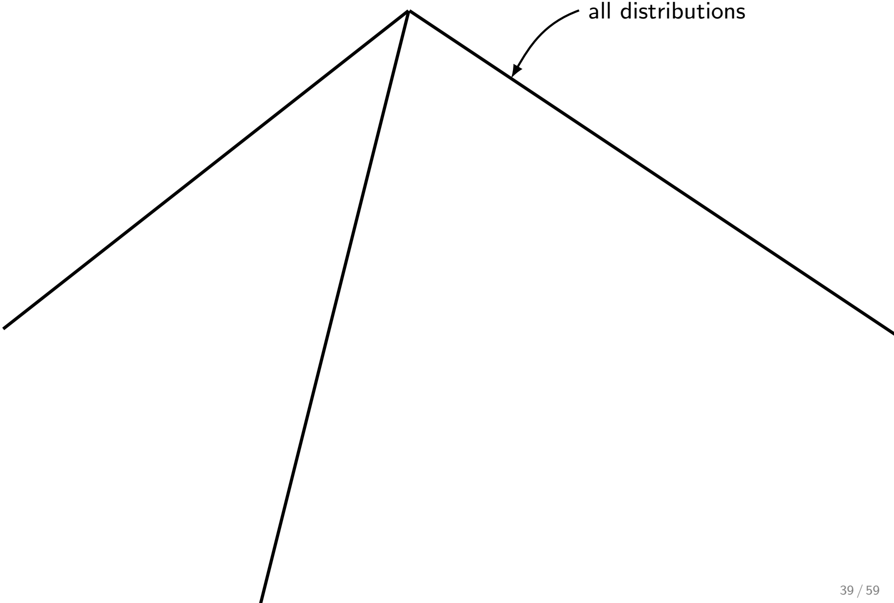
So if the latent variable model is correct[2], fitting the nested model is asymptotically equivalent fitting the LV model.

However, there are additional **inequality constraints**.

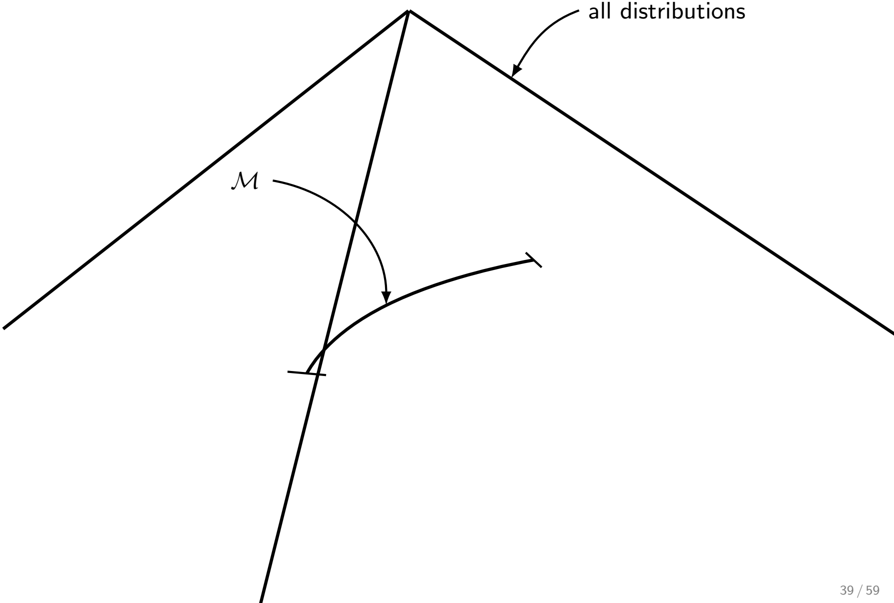
Potentially unsatisfactory as may not be a causal model corresponding to our inferred parameters.

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

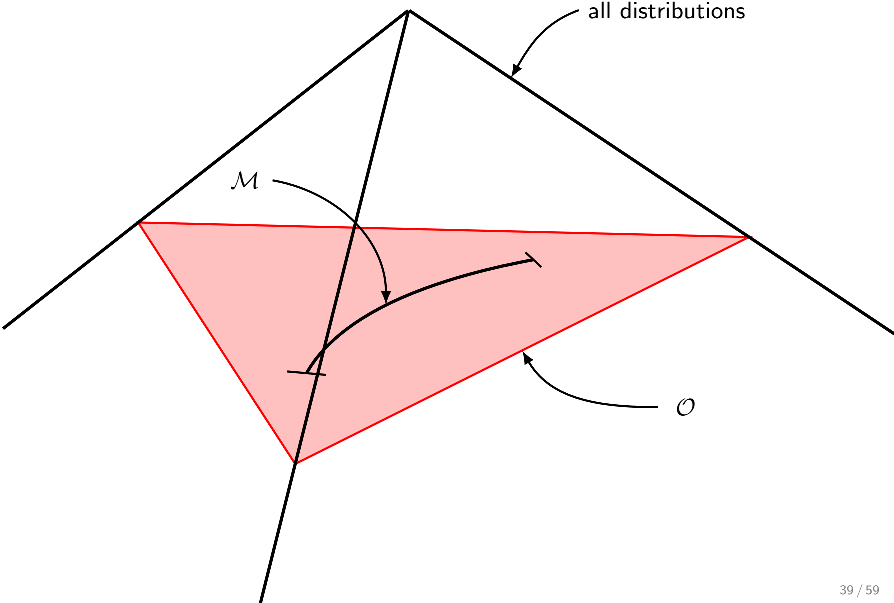
Getting the Picture



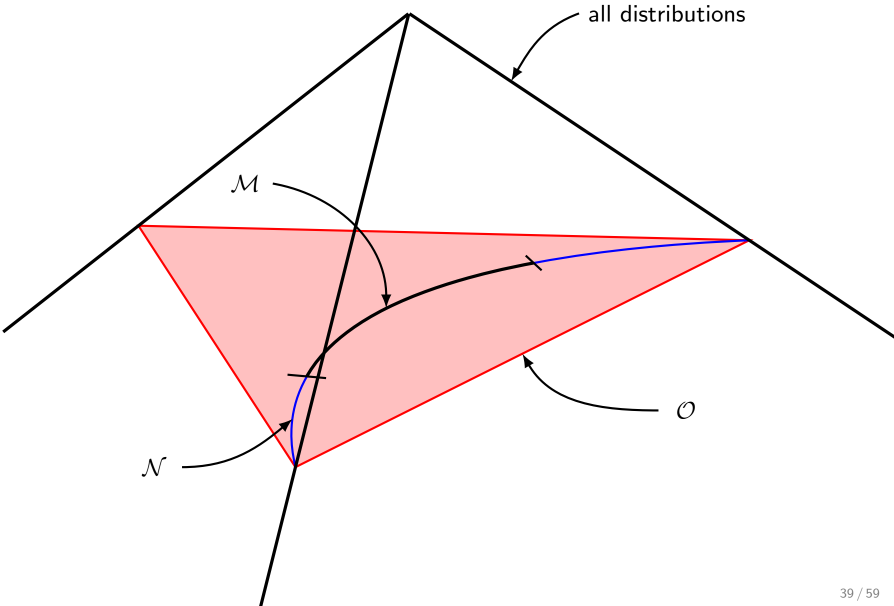
Getting the Picture



Getting the Picture

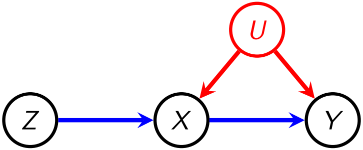


Getting the Picture



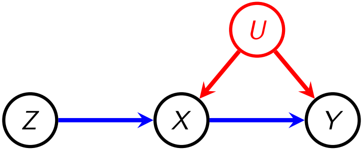
The IV Model

Assume four variable DAG shown, but U unobserved.



The IV Model

Assume four variable DAG shown, but U unobserved.



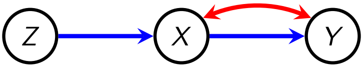
Marginalized DAG model

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

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

The IV Model

Assume four variable DAG shown, but U unobserved.



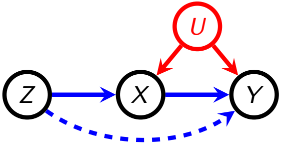
Marginalized DAG model

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

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

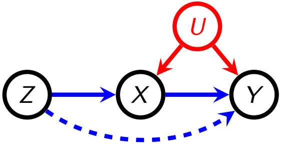
Nested Markov property gives saturated model, so true model of full dimension.

Instrumental Inequalities



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

Instrumental Inequalities



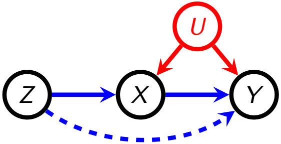
The assumption $Z \not\to Y$ is important.
Can we check it?

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.

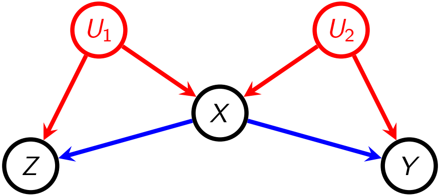
If Z, X, Y are binary, then (**??**) defines the marginalized DAG model (Bonet, 2001). e.g.

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

The Problem

Question

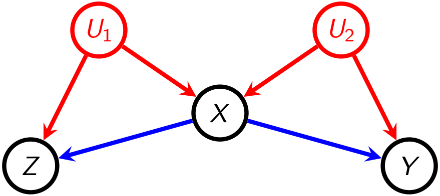
How can we determine if a general marginalized DAG model induces inequality constraints?



The Problem

Question

How can we determine if a general marginalized DAG model induces inequality constraints?

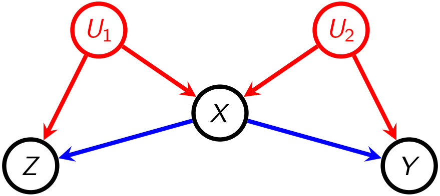


Pearl's proof of the instrumental inequality does not obviously generalize.

The Problem

Question

How can we determine if a general marginalized DAG model induces inequality constraints?



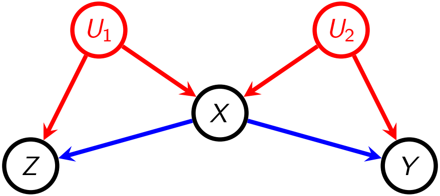
Pearl's proof of the instrumental inequality does not obviously generalize.

Computational linear algebra only works without adjacent latent variables.
Also very computationally intensive.

The Problem

Question

How can we determine if a general marginalized DAG model induces inequality constraints?

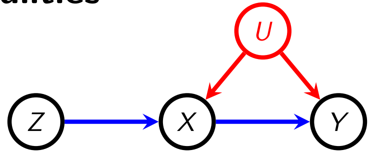


Pearl's proof of the instrumental inequality does not obviously generalize.

Computational linear algebra only works without adjacent latent variables. Also very computationally intensive.

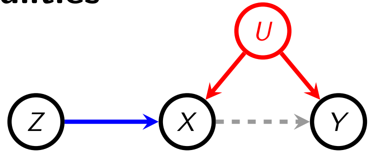
Finding complete bounds in general is currently intractably hard.

Derivation of Inequalities



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

Derivation of Inequalities



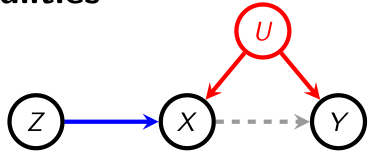
Have: $$p(x, y \mid z) = \int p(u)\, p(x \mid z, u)\, p(y \mid x, u)\, du.$$

Construct a **fictitious distribution** p^*:

$$p^*(x, y \mid z) = \int p(u)\, p(x \mid z, u)\, p(y \mid x = 0, u)\, du.$$

Now Y behaves as though $X = 0$ regardless of X's actual value.

Derivation of Inequalities



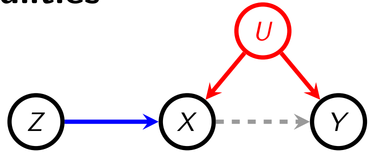
Have: $$p(x, y \mid z) = \int p(u)\, p(x \mid z, u)\, p(y \mid x, u)\, du.$$

Construct a **fictitious distribution** p^*:

$$p^*(x, y \mid z) = \int p(u)\, p(x \mid z, u)\, p(y \mid x = 0, u)\, du.$$

Now Y behaves as though $X = 0$ regardless of X's actual value.
Causally, we can think of this as an **intervention** severing $X \to Y$.

Derivation of Inequalities



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

Construct a **fictitious distribution** p^*:

$$p^*(x, y \mid z) = \int p(u)\, p(x \mid z, u)\, p(y \mid x = 0, u)\, du.$$

Now Y behaves as though $X = 0$ regardless of X's actual value.
Causally, we can think of this as an **intervention** severing $X \to Y$.

Can't observe p^* but:

- **Consistency:** $p(0, y \mid z) = p^*(0, y \mid z)$ for each z, y; and
- **Independence:** $Y \perp\!\!\!\perp Z$ under p^*.

Derivation of Inequalities

For each $x = \xi$ we require p^*_ξ:

$$p_\xi(\xi, y \mid z) = p^*_\xi(\xi, y \mid z) \text{ for each } y, z, \qquad\qquad Y \perp\!\!\!\perp Z\,[p^*_\xi].$$

Does such distributions exist?

Derivation of Inequalities

For each $x = \xi$ we require p^*_ξ:

$$p_\xi(\xi, y \mid z) = p^*_\xi(\xi, y \mid z) \text{ for each } y, z, \qquad Y \perp\!\!\!\perp Z\,[p^*_\xi].$$

Does such distributions exist?

$$p^*_\xi(y \mid z) = p^*_\xi(y)$$

Derivation of Inequalities

For each $x = \xi$ we require p^*_ξ:

$$p_\xi(\xi, y \mid z) = p^*_\xi(\xi, y \mid z) \text{ for each } y, z, \qquad Y \perp\!\!\!\perp Z \, [p^*_\xi].$$

Does such distributions exist?

$$p^*_\xi(\xi, y \mid z) \le p^*_\xi(y \mid z) = p^*_\xi(y)$$

Derivation of Inequalities

For each $x = \xi$ we require p^*_ξ:

$$p_\xi(\xi, y \mid z) = p^*_\xi(\xi, y \mid z) \text{ for each } y, z, \qquad Y \perp\!\!\!\perp Z \, [p^*_\xi].$$

Does such distributions exist?

$$p(\xi, y \mid z) = p^*_\xi(\xi, y \mid z) \leq p^*_\xi(y \mid z) = p^*_\xi(y)$$

Derivation of Inequalities

For each $x = \xi$ we require p_ξ^*:

$$p_\xi(\xi, y \mid z) = p_\xi^*(\xi, y \mid z) \text{ for each } y, z, \qquad Y \perp\!\!\!\perp Z \, [p_\xi^*].$$

Does such distributions exist?

$$p(\xi, y \mid z) = p_\xi^*(\xi, y \mid z) \leq p_\xi^*(y \mid z) = p_\xi^*(y)$$

So clearly

$$\max_z p(\xi, y \mid z) \leq p_\xi^*(y)$$

Derivation of Inequalities

For each $x = \xi$ we require p^*_ξ:

$$p_\xi(\xi, y \mid z) = p^*_\xi(\xi, y \mid z) \text{ for each } y, z, \qquad\qquad Y \perp\!\!\!\perp Z\,[p^*_\xi].$$

Does such distributions exist?

$$p(\xi, y \mid z) = p^*_\xi(\xi, y \mid z) \leq p^*_\xi(y \mid z) = p^*_\xi(y)$$

So clearly

$$\max_z p(\xi, y \mid z) \leq p^*_\xi(y)$$

$$\sum_y \max_z p(\xi, y \mid z) \leq 1.$$

By varying ξ, the instrumental inequality follows.

Derivation of Inequalities

For each $x = \xi$ we require p^*_ξ:

$$p_\xi(\xi, y \mid z) = p^*_\xi(\xi, y \mid z) \text{ for each } y, z, \qquad\qquad Y \perp\!\!\!\perp Z\,[p^*_\xi].$$

Does such distributions exist?

$$p(\xi, y \mid z) = p^*_\xi(\xi, y \mid z) \leq p^*_\xi(y \mid z) = p^*_\xi(y)$$

So clearly

$$\max_z p(\xi, y \mid z) \leq p^*_\xi(y)$$

$$\sum_y \max_z p(\xi, y \mid z) \leq 1.$$

By varying ξ, the instrumental inequality follows.

We say that the probabilities $p(x, y \mid z)$ are **compatible** with $Y \perp\!\!\!\perp Z$.

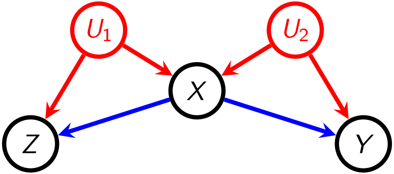
Generalizing

How does this help us with other graphs?

Generalizing

How does this help us with other graphs?

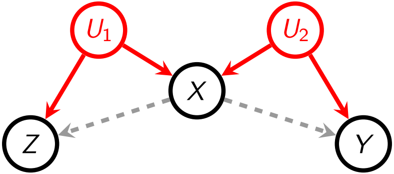
The argument works precisely because cutting edges led to a separation:



Generalizing

How does this help us with other graphs?

The argument works precisely because cutting edges led to a separation:

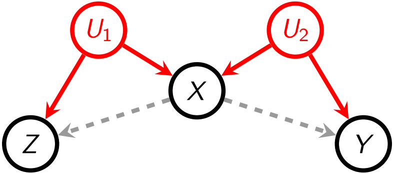


Z *is* d-separated from Y in the graph after cutting edges emanating from X.

Generalizing

How does this help us with other graphs?

The argument works precisely because cutting edges led to a separation:



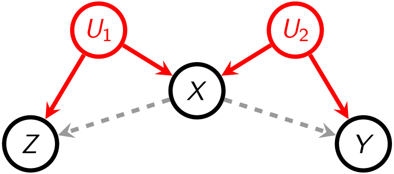
Z *is* d-separated from Y in the graph after cutting edges emanating from X.

So by the same argument, for fixed ξ, $p(\xi, y, z)$ must be compatible with a (fictitious) distribution p^*_ξ in which $Y \perp\!\!\!\perp Z$.

Generalizing

How does this help us with other graphs?

The argument works precisely because cutting edges led to a separation:



Z *is* d-separated from Y in the graph after cutting edges emanating from X.

So by the same argument, for fixed ξ, $p(\xi, y, z)$ must be compatible with a (fictitious) distribution p^*_ξ in which $Y \perp\!\!\!\perp Z$.

[Note for the IV model, the conditional distribution $p(\xi, y \mid z)$ had to be compatible.]

Compatibility

Probabilities may not be compatible with independences.

Compatibility

Probabilities may not be compatible with independences.

Consider a partial probability table $p(x = \xi, y, z)$:

	$Z = 0$	$Z = 1$
$Y = 0$	$1/3$	0
$Y = 1$	0	$1/3$

Compatibility

Probabilities may not be compatible with independences.

Consider a partial probability table $p(x = \xi, y, z)$:

	$Z = 0$	$Z = 1$
$Y = 0$	$1/3$	0
$Y = 1$	0	$1/3$

There is no way to construct a joint distribution over X, Y, Z with these probabilities such that Y and Z are independent.

Compatibility

Probabilities may not be compatible with independences.

Consider a partial probability table $p(x = \xi, y, z)$:

	$Z = 0$	$Z = 1$
$Y = 0$	$1/3$	0
$Y = 1$	0	$1/3$

There is no way to construct a joint distribution over X, Y, Z with these probabilities such that Y and Z are independent.

Most likely to happen if $p(x)$ is large for some value of x.

A Generalization

For a DAG $\mathcal{G}$ and set of variables $\boldsymbol{W}$, let $\mathcal{G}^{\underline{\boldsymbol{W}}}$ be the graph after removing edges pointing away from $\boldsymbol{W}$.

A Generalization

For a DAG $\mathcal{G}$ and set of variables $\boldsymbol{W}$, let $\mathcal{G}^{\underline{\boldsymbol{W}}}$ be the graph after removing edges pointing away from $\boldsymbol{W}$.

Theorem (Evans, 2012)

Let p be a discrete distribution in marginalized DAG model for $\mathcal{G}$.
Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}$ be sets of variables in $\mathcal{G}$.

A Generalization

For a DAG $\mathcal{G}$ and set of variables $\boldsymbol{W}$, let $\mathcal{G}^{\underline{\boldsymbol{W}}}$ be the graph after removing edges pointing away from $\boldsymbol{W}$.

Theorem (Evans, 2012)

Let p be a discrete distribution in marginalized DAG model for $\mathcal{G}$. Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}$ be sets of variables in $\mathcal{G}$.

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ in $\mathcal{G}^{\underline{\boldsymbol{W}}}$, then for each fixed $\{\boldsymbol{W} = \boldsymbol{\omega}\}$ the probabilities

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\omega} \mid \boldsymbol{z}), \qquad \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}.$$

are **compatible with a distribution** $p^*_{\boldsymbol{\omega}}$, in which $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \, [p^*_{\boldsymbol{\omega}}]$.

A Generalization

For a DAG $\mathcal{G}$ and set of variables $\boldsymbol{W}$, let $\mathcal{G}^{\underline{\boldsymbol{W}}}$ be the graph after removing edges pointing away from $\boldsymbol{W}$.

Theorem (Evans, 2012)

Let p be a discrete distribution in marginalized DAG model for $\mathcal{G}$.
Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}$ be sets of variables in $\mathcal{G}$.

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ in $\mathcal{G}^{\underline{\boldsymbol{W}}}$, then for each fixed $\{\boldsymbol{W} = \boldsymbol{\omega}\}$ the probabilities

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\omega} \mid \boldsymbol{z}), \qquad \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}.$$

are **compatible with a distribution** $p^*_{\boldsymbol{\omega}}$, in which $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}\,[p^*_{\boldsymbol{\omega}}]$.

If, in addition, $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ and $\boldsymbol{X}_2, \boldsymbol{Y}_2$ are not descendants of $\boldsymbol{W}$, then

$$p(\boldsymbol{x}_1, \boldsymbol{y}_1, \boldsymbol{\omega} \mid \boldsymbol{x}_2, \boldsymbol{y}_2, \boldsymbol{z}) \qquad \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{z}.$$

are **compatible with a distribution** $p^*_{\boldsymbol{\omega}}$, in which $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}\,[p^*_{\boldsymbol{\omega}}]$.

Missing Edges Give Constraints

This is nice because no previous derivation of inequalities was graphical: based on one of

- computational algebra (Bonet, 2001);
- algorithmic method (Kang and Tian, 2006);
- or convexity arguments (Pearl, 1995).

Missing Edges Give Constraints

This is nice because no previous derivation of inequalities was graphical: based on one of

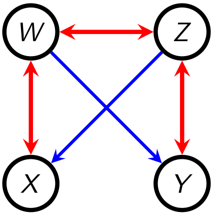
- computational algebra (Bonet, 2001);
- algorithmic method (Kang and Tian, 2006);
- or convexity arguments (Pearl, 1995).

Whereas...

Corollary

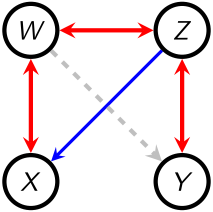
If X and Y are not joined by an edge, nor share a hidden common cause, then a constraint is always induced on the joint distribution.

Example 1



X and Y cannot be d-separated in this graph $\implies$ no independences.

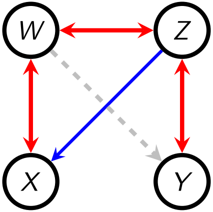
Example 1



X and Y cannot be d-separated in this graph $\implies$ no independences.

Remove edges emanating from W, see that now $X \perp\!\!\!\perp Y \mid Z$.
So $p(x, y, w \mid z)$ compatible with $X \perp\!\!\!\perp Y \mid Z$ for each w.

Example 1

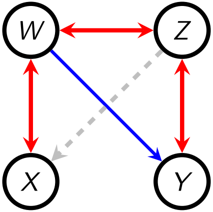


X and Y cannot be d-separated in this graph $\implies$ no independences.

Remove edges emanating from W, see that now $X \perp\!\!\!\perp Y \mid Z$.
So $p(x, y, w \mid z)$ compatible with $X \perp\!\!\!\perp Y \mid Z$ for each w.

In fact, Y not a descendant of Z, so $p(x, w \mid z, y)$ compatible.

Example 1



X and Y cannot be d-separated in this graph $\implies$ no independences.

Remove edges emanating from W, see that now $X \perp\!\!\!\perp Y \mid Z$.
So $p(x, y, w \mid z)$ compatible with $X \perp\!\!\!\perp Y \mid Z$ for each w.

In fact, Y not a descendant of Z, so $p(x, w \mid z, y)$ compatible.

By symmetry: $p(y, z \mid w, x)$ compatible with $X \perp\!\!\!\perp Y \mid W$ for each z.

Compatibility

Can we easily determine whether distributions are 'compatible' with independences?

Compatibility

Can we easily determine whether distributions are 'compatible' with independences?

Suppose we need $p(x, y, w \mid z)$ to be compatible with $X \perp\!\!\!\perp Y \mid Z \, [p^*]$.

Compatibility

Can we easily determine whether distributions are 'compatible' with independences?

Suppose we need $p(x, y, w \mid z)$ to be compatible with $X \perp\!\!\!\perp Y \mid Z \, [p^*]$.

In other words, for each z, w need a rank 1 matrix $B = (b_{xy})$ such that

$$b_{xy} \geq p(x, y, w \mid z) \qquad \text{and} \qquad \sum_{xy} b_{xy} \leq 1.$$

Compatibility

Can we easily determine whether distributions are 'compatible' with independences?

Suppose we need $p(x, y, w \mid z)$ to be compatible with $X \perp\!\!\!\perp Y \mid Z \, [p^*]$.

In other words, for each z, w need a rank 1 matrix $B = (b_{xy})$ such that

$$b_{xy} \geq p(x, y, w \mid z) \qquad \text{and} \qquad \sum_{xy} b_{xy} \leq 1.$$

Proposition

The existence of such a matrix is a **convex optimization problem**.

Compatibility

Can we easily determine whether distributions are 'compatible' with independences?

Suppose we need $p(x, y, w \mid z)$ to be compatible with $X \perp\!\!\!\perp Y \mid Z\,[p^*]$.

In other words, for each z, w need a rank 1 matrix $B = (b_{xy})$ such that

$$b_{xy} \geq p(x, y, w \mid z) \qquad \text{and} \qquad \sum_{xy} b_{xy} \leq 1.$$

Proposition

The existence of such a matrix is a **convex optimization problem**.

In general, Theorem 1 gives necessary but not sufficient conditions for p to be in the marginalized DAG model.

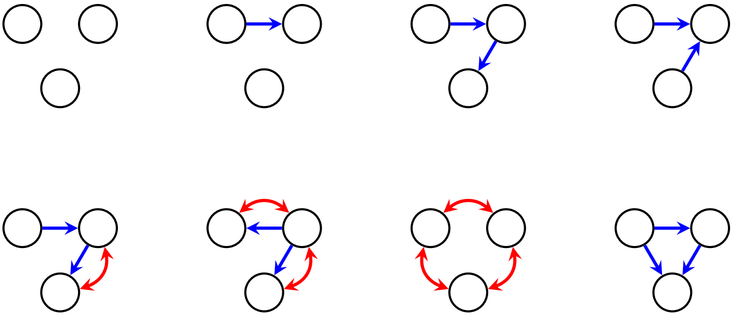
Equivalence on Three Variables

The previous method doesn't give all inequalities. This is generally an extremely hard problem, even in specific cases.

Equivalence on Three Variables

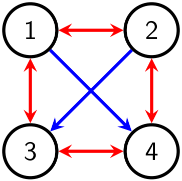
The previous method doesn't give all inequalities. This is generally an extremely hard problem, even in specific cases.

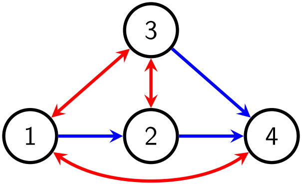
Even Markov equivalence is hard. Using Evans (2014), find 8 Markov equivalence classes on three variables.

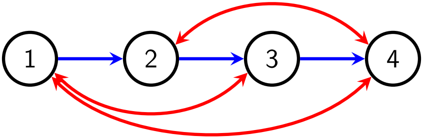


But Not on Four!

On four variables, it's still not clear whether or not the following models are saturated: (they are of full dimension in the discrete case)

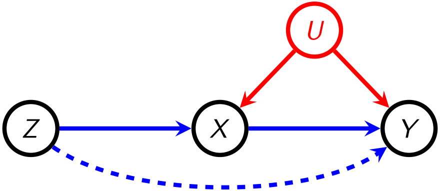






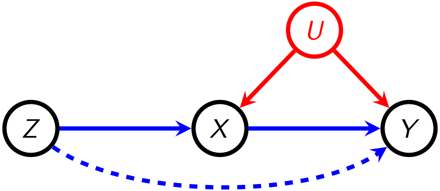
Outline

Causal Effects



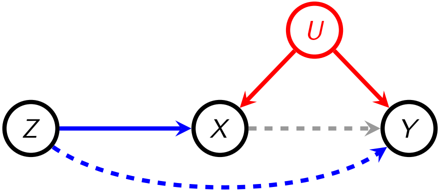
So far we've given inequalities which 'prove existence' for edges.

Causal Effects



So far we've given inequalities which 'prove existence' for edges.
Now we'd like to determine the strength of its causal effect.

Causal Effects

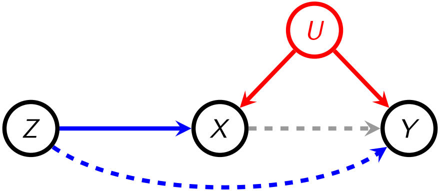


So far we've given inequalities which 'prove existence' for edges.
Now we'd like to determine the strength of its causal effect.

Construct p^* as before. Then

$$p(y \mid \mathrm{do}(x = \xi, z)) = p^*_\xi(y \mid z)$$

Causal Effects



So far we've given inequalities which 'prove existence' for edges.
Now we'd like to determine the strength of its causal effect.

Construct p^* as before. Then

$$\begin{aligned} p(y \mid \mathrm{do}(x = \xi, z)) &= p^*_\xi(y \mid z) \\ &= p(x, y \mid z) + \sum_{x' \neq \xi} p^*_\xi(x', y \mid z). \end{aligned}$$

Causal Bounds

This approach gives bounds on the interventional distributions (Evans, 2012) and, for example, the **average controlled direct effect**

$$\mathrm{ACDE}_{Z \to Y}(x) \equiv p(y = 1 \mid \mathrm{do}(x, z = 1)) - p(y = 1 \mid \mathrm{do}(x, z = 0)).$$

Causal Bounds

This approach gives bounds on the interventional distributions (Evans, 2012) and, for example, the **average controlled direct effect**

$$\mathrm{ACDE}_{Z \to Y}(x) \equiv p(y = 1 \mid \mathrm{do}(x, z = 1)) - p(y = 1 \mid \mathrm{do}(x, z = 0)).$$

Theorem

Let $X \to Y$, but otherwise d-separated in the graph $\mathcal{G}^{\underline{\boldsymbol{W}}}$. Then an upper-bound on $\mathrm{ACDE}_{X \to Y}(w)$ is given by maximizing

$$\frac{p(y = 1, x = 1, w) + \beta}{p(x = 1, w) + \beta} - \frac{p(y = 1, x = 0, w)}{p(x = 0, w) + 1 - p(w) - \beta}$$

over $0 \leq \beta \leq 1 - p(w)$.

Causal Bounds

This approach gives bounds on the interventional distributions (Evans, 2012) and, for example, the **average controlled direct effect**

$$\mathrm{ACDE}_{Z\to Y}(x) \equiv p(y=1 \mid \mathrm{do}(x, z=1)) - p(y=1 \mid \mathrm{do}(x, z=0)).$$

Theorem

Let $X \to Y$, but otherwise d-separated in the graph $\mathcal{G}^{\underline{\boldsymbol{W}}}$. Then an upper-bound on $\mathrm{ACDE}_{X\to Y}(w)$ is given by maximizing

$$\frac{p(y=1, x=1, w) + \beta}{p(x=1, w) + \beta} - \frac{p(y=1, x=0, w)}{p(x=0, w) + 1 - p(w) - \beta}$$

over $0 \leq \beta \leq 1 - p(w)$.

This is just a quadratic equation. There is an analogous lower-bound.

Bounds: Special Case

Theorem

Let $X \to Y$, but otherwise d-separated in the graph $\mathcal{G}^{\underline{\boldsymbol{W}}}$, and that X is not a descendant of any variable in $\boldsymbol{W}$. Then

$$p(y=0,\boldsymbol{\omega} \,|\, x=0) + p(y=1,\boldsymbol{\omega} \,|\, x=1) - 1 \\ \leq \mathrm{ACDE}(\boldsymbol{\omega}) \leq \\ 1 - p(y=0,\boldsymbol{\omega} \,|\, x=1) - p(y=1,\boldsymbol{\omega} \,|\, x=0).$$

For the IV model, this is the tight bound given by Cai et al (2008).

Bounds: Special Case

Theorem

Let $X \to Y$, but otherwise d-separated in the graph $\mathcal{G}^{\underline{\boldsymbol{W}}}$, and that X is not a descendant of any variable in $\boldsymbol{W}$. Then

$$
\begin{aligned}
&p(y=0,\boldsymbol{\omega} \mid x=0) + p(y=1,\boldsymbol{\omega} \mid x=1) - 1 \\
&\qquad \leq \mathrm{ACDE}(\boldsymbol{\omega}) \leq \\
&\qquad\qquad 1 - p(y=0,\boldsymbol{\omega} \mid x=1) - p(y=1,\boldsymbol{\omega} \mid x=0).
\end{aligned}
$$

For the IV model, this is the tight bound given by Cai et al (2008).

If bounds exclude zero then models violate Theorem 1 compatibility.

Outline

Summary

- (Causal) DAGs with latent variables induce non-parametric inequalities;

Summary

- (Causal) DAGs with latent variables induce non-parametric inequalities;
- some can be deduced as 'compatibility' of probabilities with conditional independences;

Summary

- (Causal) DAGs with latent variables induce non-parametric inequalities;
- some can be deduced as ‘compatibility’ of probabilities with conditional independences;
- there are other inequalities, including Bell’s inequality, see Evans (2014).

Summary

- (Causal) DAGs with latent variables induce non-parametric inequalities;
- some can be deduced as ‘compatibility’ of probabilities with conditional independences;
- there are other inequalities, including Bell’s inequality, see Evans (2014).

Some limitations:

Summary

- (Causal) DAGs with latent variables induce non-parametric inequalities;
- some can be deduced as 'compatibility' of probabilities with conditional independences;
- there are other inequalities, including Bell's inequality, see Evans (2014).

Some limitations:

- Complete inequality constraints seem very complicated (though some hope exists).

Summary

- (Causal) DAGs with latent variables induce non-parametric inequalities;
- some can be deduced as 'compatibility' of probabilities with conditional independences;
- there are other inequalities, including Bell's inequality, see Evans (2014).

Some limitations:

- Complete inequality constraints seem very complicated (though some hope exists).
- Performing inference for inequality constraints with finite samples is non-trivial.

Summary

- (Causal) DAGs with latent variables induce non-parametric inequalities;
- some can be deduced as 'compatibility' of probabilities with conditional independences;
- there are other inequalities, including Bell's inequality, see Evans (2014).

Some limitations:

- Complete inequality constraints seem very complicated (though some hope exists).
- Performing inference for inequality constraints with finite samples is non-trivial.
- Not obvious how to integrate inequalities into the previous parameterization.

Inequality References

Bonet – Instrumentality tests revisited, *UAI*, 2001.

Cai, Kuroki, Pearl and Tian – Bounds on direct effects in the presence of confounded intermediate variables, *Biometrics*, 64(3):695–701, 2008.

Evans – Graphical methods for inequality constraints in marginalized DAGs, *MLSP*, 2012.

Evans – Margins of discrete Bayesian networks, *arXiv:1501.02103*, 2015.

Kang and Tian – Inequality Constraints in Causal Models with Hidden Variables, *UAI*, 2006.

Pearl – 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

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

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

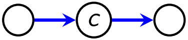
A **path** is a sequence of edges in the graph; vertices may not be repeated.

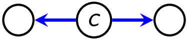
d-Separation

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:



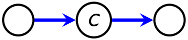


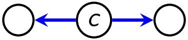
d-Separation

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

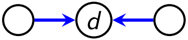
A path from v to w is **blocked** by $C \subseteq V \setminus \{v, w\}$ if either

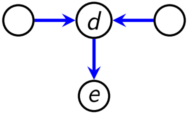
(i) any non-collider is in C:





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



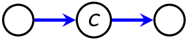


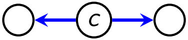
d-Separation

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

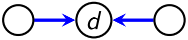
A path from v to w is **blocked** by $C \subseteq V \setminus \{v, w\}$ if either

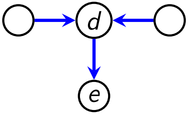
(i) any non-collider is in C:





(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.