Causal models and how to refute them.

Robin Evans
University of Oxford
www.stats.ox.ac.uk/∼evans

Statistics Seminar, University of York
26th November 2015

Acknowledgements

Thomas Richardson (U. of Washington)

James Robins (Harvard)

Ilya Shpitser (Johns Hopkins)

Correlation does not imply causation

Find a Job | M&S Wine | Feedback Wednesday, N

Mail Online

Home | News | U.S. | Sport | TV&Showbiz | Femail | Health | Science | Money | Video | Cof

Health Home | Health Directory | Health Boards | Diets | MyDish Recipe Finder

How a short nap can raise the risk of diabetes: Study finds people who have a siesta are more likely to have high blood pressure and high cholesterol

- **Napping for more than 30 minutes at a time can raise the risk of diabetes, according to a new study**
- **It can also increase likelihood of high blood pressure and high cholesterol**

By PAT HAGAN

PUBLISHED: 01:04, 21 September 2013 | **UPDATED:** 10:34, 21 September 2013

598 shares 102 View comments

They were much favoured by Margaret Thatcher, Albert Einstein and Winston Churchill.

But while afternoon naps may revitalise tired brains, they can also increase the risk of diabetes, according to new research.

Correlation does not imply causation

Find a Job | M&S Wine | Feedback   Wednesday, N

MailOnline

Home | News | U.S. | Sport | TV&Showbiz | Fem

Health Home | Health Directory | Health Boards | Diets | M

How a short nap can rais diabetes: Study finds pe siesta are more likely to pressure and high chole

- Napping for more than 30 minutes at a tim according to a new study
- It can also increase likelihood of high bloo

By PAT HAGAN

PUBLISHED: 01:04, 21 September 2013 | **UPDATED:** 10:34, 21 Septemb

598 shares

They were much favoured by Margaret Thatcher, Alber

But while afternoon naps may revitalise tired brains, the according to new research.

ELSEVIER

Sleep Medicine

Volume 14, Issue 10, October 2013, Pages 950–954

Original Article

Longer habitual afternoon napping is associated with a higher risk for impaired fasting plasma glucose and diabetes mellitus in older adults: results from the Dongfeng–Tongji cohort of retired workers

Weimin Fang[a, b], Zhongliang Li[a], Li Wu[a], Zhongqiang Cao[a], Yuan Liang[a, c], Handong Yang[d], Youjie Wang[a, b], Tangchun Wu[a]

[a] Ministry of Education Key Laboratory of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, China

[b] Department of Maternal and Child Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, China

[c] Department of Social Medicine, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, China

[d] Dongfeng General Hospital, Dongfeng Motor Corporation and Hubei University of Medicine, China

Abstract

Objectives

Afternoon napping is a common habit in China. We used data obtained from the Dongfeng–Tongji cohort to examine if duration of habitual afternoon napping was associated with risks for impaired fasting plasma glucose (IFG) and diabetes mellitus (DM) in a Chinese elderly population.

Methods

Correlation does not imply causation

Find a Job | M&S Wine | Feedback Wednesday, N

MailOnline

Home | News | U.S. | Sport | TV&Showbiz | Fem

Health Home | Health Directory | Health Boards | Diets | M

How a short nap can rais diabetes: Study finds pe siesta are more likely to pressure and high chole

- Napping for more than 30 minutes at a tim

598 shares

They were much favoured by Margaret Thatcher, Alber

But while afternoon naps may revitalise tired brains, the according to new research.

Sleep Medicine

Volume 14, Issue 10, October 2013, Pages 950–954

ELSEVIER

Original Article

Longer habitual afternoon napping is associated with a higher risk for impaired fasting plasma glucose and diabetes mellitus in older adults: results from the Dongfeng–Tongji cohort of retired workers

Weimin Fang[a, b], Zhongliang Li[a], Li Wu[a], Zhongqiang Cao[a], Yuan Liang[a, c], Handong Yang[d], Youjie Wang[a, b], Tangchun Wu[a]

[a] Ministry of Education Key Laboratory of Environment and Health, School of Public Health, Tongji Medical College, Huazhong

> "Dr Matthew Hobbs, head of research for Diabetes UK, said there was no proof that napping **actually caused** diabetes."

Abstract

Objectives

Afternoon napping is a common habit in China. We used data obtained from the Dongfeng–Tongji cohort to examine if duration of habitual afternoon napping was associated with risks for impaired fasting plasma glucose (IFG) and diabetes mellitus (DM) in a Chinese elderly population.

Methods

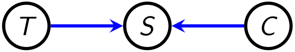
Distinguishing Between Causal Models

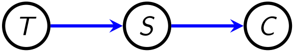
Causality is best inferred from experiments.
But doing experiments is hard (expensive, impractical, unethical...)

Distinguishing Between Causal Models

Causality is best inferred from experiments.
But doing experiments is hard (expensive, impractical, unethical...)

Collecting observational data is cheap.
Can we still tell what causes what from observational data?

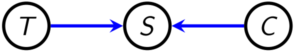


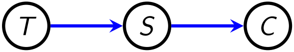


Distinguishing Between Causal Models

Causality is best inferred from experiments.
But doing experiments is hard (expensive, impractical, unethical...)

Collecting observational data is cheap.
Can we still tell what causes what from observational data?



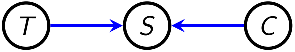


$$p(t, s, c) = p(t)p(c)p(s \mid t, c)$$
$$T \perp\!\!\!\perp C$$

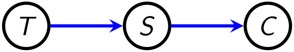
Distinguishing Between Causal Models

Causality is best inferred from experiments.
But doing experiments is hard (expensive, impractical, unethical...)

Collecting observational data is cheap.
Can we still tell what causes what from observational data?



$$p(t, s, c) = p(t)p(c)p(s \mid t, c)$$
$$T \perp\!\!\!\perp C$$

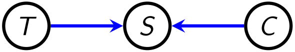


$$p(t, s, c) = p(t)p(s \mid t)p(c \mid s)$$
$$T \perp\!\!\!\perp C \mid S$$

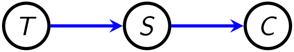
Distinguishing Between Causal Models

Causality is best inferred from experiments.
But doing experiments is hard (expensive, impractical, unethical...)

Collecting observational data is cheap.
Can we still tell what causes what from observational data?



$$p(t, s, c) = p(t)p(c)p(s \mid t, c)$$
$$T \perp\!\!\!\perp C$$



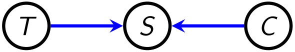
$$p(t, s, c) = p(t)p(s \mid t)p(c \mid s)$$
$$T \perp\!\!\!\perp C \mid S$$

Sometimes!

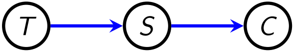
Distinguishing Between Causal Models

Causality is best inferred from experiments.
But doing experiments is hard (expensive, impractical, unethical...)

Collecting observational data is cheap.
Can we still tell what causes what from observational data?



$$p(t, s, c) = p(t)p(c)p(s \mid t, c)$$
$$T \perp\!\!\!\perp C$$



$$p(t, s, c) = p(t)p(s \mid t)p(c \mid s)$$
$$T \perp\!\!\!\perp C \mid S$$

Sometimes!

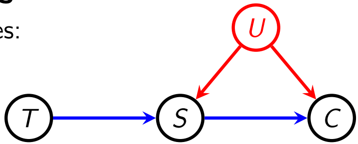
This is the basis of some causal search algorithms (e.g. PC, FCI).
Note: other methods (e.g. integer programming) are also used.

In order to do this well, we need to understand in what ways causal models will be **observationally** different.

When everything is observed this is (mathematically) easy.

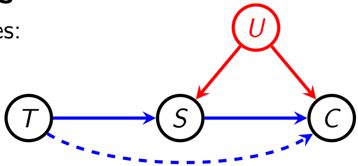
Hidden Variables

Instrumental variables:



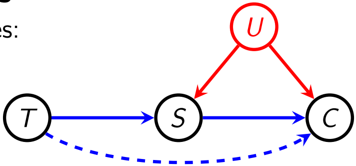
Hidden Variables

Instrumental variables:

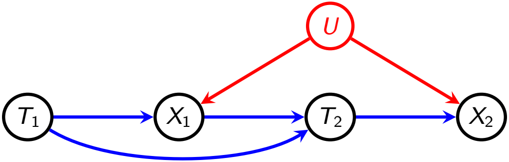


Hidden Variables

Instrumental variables:

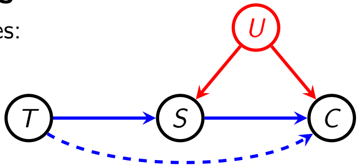


Dynamic treatment model / longitudinal exposure:

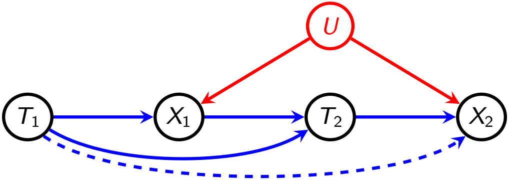


Hidden Variables

Instrumental variables:

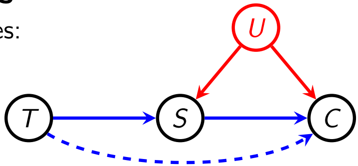


Dynamic treatment model / longitudinal exposure:

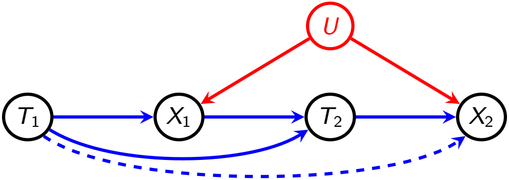


Hidden Variables

Instrumental variables:



Dynamic treatment model / longitudinal exposure:

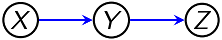


Principal aims:

- be able to test causal models;
- identify and bound causal effects;
- use constraints for model search.

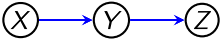
The Holy Grail: Structure Learning

Truth:

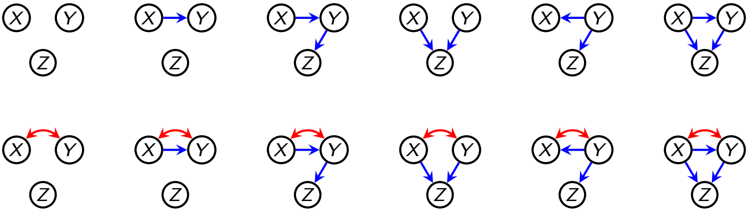


The Holy Grail: Structure Learning

Truth:

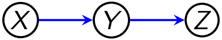


Given a distribution P (or rather data from P) and a set of possible causal models...

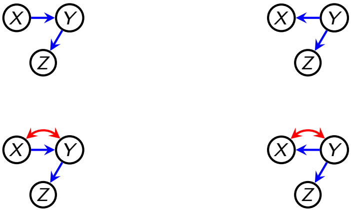


The Holy Grail: Structure Learning

Truth:



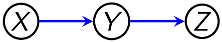
Given a distribution P (or rather data from P) and a set of possible causal models...



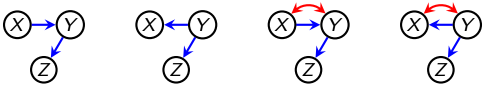
...return list of models which are compatible with data.

Experimental Design

Truth:

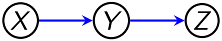


We could then identify an experiment to distinguish remaining models:

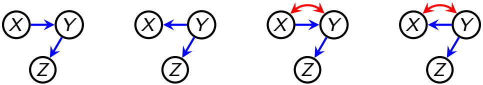


Experimental Design

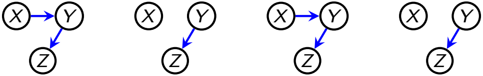
Truth:



We could then identify an experiment to distinguish remaining models:



intervene on X:

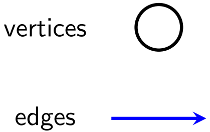


...return list of models which are compatible with data.

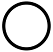
To do this we need to know what constraints the model places on the distribution (the focus of this talk).

Outline

Directed Acyclic Graphs

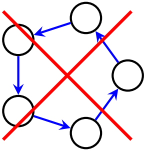


Directed Acyclic Graphs

vertices 

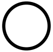
edges 

no directed cycles



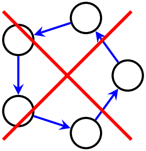
Directed Acyclic Graphs

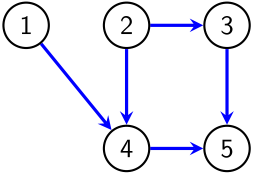
vertices



edges

no directed cycles

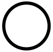




directed acyclic graph (DAG), $\mathcal{G}$

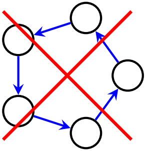
Directed Acyclic Graphs

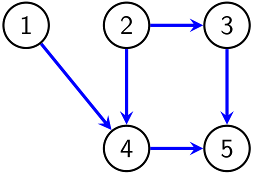
vertices



edges

no directed cycles





directed acyclic graph (DAG), $\mathcal{G}$

If $w \to v$ then w is a **parent** of v: $\mathrm{pa}_{\mathcal{G}}(4) = \{1, 2\}$.

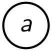
If $w \to \cdots \to v$ then w is a **ancestor** of v.
An **ancestral set** contains all its own ancestors.

DAG Models (aka Bayesian Networks)

vertex		random variable
a	$\iff$	X_a

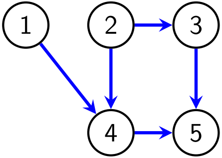
DAG Models (aka Bayesian Networks)

vertex $\iff$ random variable



X_a

graph $\mathcal{G}$ $\iff$ model $\mathcal{M}(\mathcal{G})$

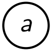


$$p(x_1, \ldots, x_k) = \prod_i p(x_i \mid x_{\mathrm{pa}(i)}).$$

(factorization)

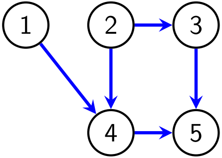
DAG Models (aka Bayesian Networks)

vertex $\iff$ random variable



X_a

graph $\mathcal{G}$ $\iff$ model $\mathcal{M}(\mathcal{G})$



$$p(x_1, \ldots, x_k) = \prod_i p(x_i \mid x_{\mathrm{pa}(i)}).$$

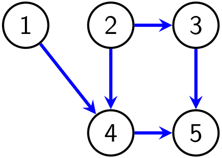
(factorization)

So in example above:

$$p(x_V) = p(x_1) \cdot p(x_2) \cdot p(x_3 \mid x_2) \cdot p(x_4 \mid x_1, x_2) \cdot p(x_5 \mid x_3, x_4)$$

DAG Models

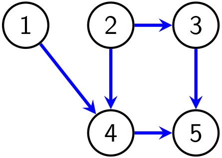
Can also define model as a list of conditional independences:



pick a topological ordering of the graph: $1, 2, 3, 4, 5$.

DAG Models

Can also define model as a list of conditional independences:



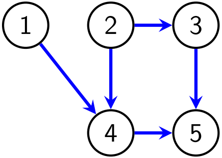
pick a topological ordering of the graph: $1, 2, 3, 4, 5$.

Can *always* factorize a joint distribution as:

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

DAG Models

Can also define model as a list of conditional independences:



pick a topological ordering of the graph: $1, 2, 3, 4, 5$.

Can *always* factorize a joint distribution as:

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

The model is the same as setting

$$p(x_i \mid x_1, x_2, \dots, x_{i-1}) = p(x_i \mid x_{\mathrm{pa}(i)}), \qquad \text{for each } i.$$

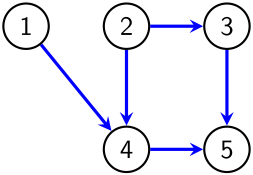
Thus $\mathcal{M}(\mathcal{G})$ is precisely distributions such that:

$$X_i \perp\!\!\!\perp X_{[i-1]\setminus \mathrm{pa}(i)} \mid X_{\mathrm{pa}(i)}, \qquad i \in V.$$

This is a constraint-based perspective.

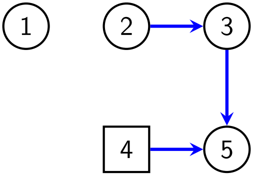
Causal Models

A DAG can also encode causal information:



Causal Models

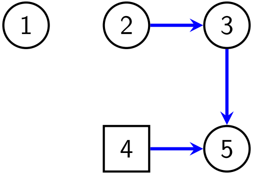
A DAG can also encode causal information:



If we intervene to experiment on X_4, just delete incoming edges.

Causal Models

A DAG can also encode causal information:



If we intervene to experiment on X_4, just delete incoming edges.

In distribution, just delete factor corresponding to X_4:

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1) \cdot p(x_2) \cdot p(x_3 \mid x_2) \cdot p(x_4 \mid x_1, x_2) \cdot p(x_5 \mid x_3, x_4).$$
$$p(x_1, x_2, x_3, x_5 \mid \text{do}(x_4)) = p(x_1) \cdot p(x_2) \cdot p(x_3 \mid x_2) \cdot p(x_5 \mid x_3, x_4).$$

All other terms preserved.

Outline

Marginalization

Very often causal models include random quantities that we cannot observe.

Wisconsin Longitudinal Study:

- over 10,000 Wisconsin high-school graduates from 1957;
- data on primary respondents collected in 1957, 1975, 1992, 2004.

Marginalization

Very often causal models include random quantities that we cannot observe.

Wisconsin Longitudinal Study:

- over 10,000 Wisconsin high-school graduates from 1957;
- data on primary respondents collected in 1957, 1975, 1992, 2004.

Suppose we want to know whether drafting has impact on future earnings, controlling for education/family background.

- X family income in 1957;
- E education level;
- M drafted into military;
- Y respondent income 1992;

Marginalization

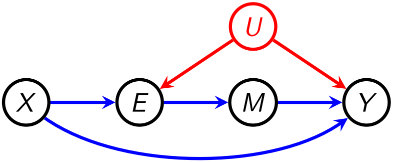
Very often causal models include random quantities that we cannot observe.

Wisconsin Longitudinal Study:

- over 10,000 Wisconsin high-school graduates from 1957;
- data on primary respondents collected in 1957, 1975, 1992, 2004.

Suppose we want to know whether drafting has impact on future earnings, controlling for education/family background.

X family income in 1957;
E education level;
M drafted into military;
Y respondent income 1992;
U unmeasured confounding.



Marginalization

Note we don't want to make assumptions about U; so this is **not** a latent variable model in the usual sense.

Model is defined (implicitly) by an integral:

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

No state-space is assumed for hidden variable (though uniform on $(0, 1)$ is sufficient).

Marginalization

Note we don't want to make assumptions about U; so this is **not** a latent variable model in the usual sense.

Model is defined (implicitly) by an integral:

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

No state-space is assumed for hidden variable (though uniform on $(0, 1)$ is sufficient).

But how can we

- characterize the model?
- test membership of the model?
- fit it to data?

Marginalization

Note we don't want to make assumptions about U; so this is **not** a latent variable model in the usual sense.

Model is defined (implicitly) by an integral:

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

No state-space is assumed for hidden variable (though uniform on $(0, 1)$ is sufficient).

But how can we

- characterize the model?
- test membership of the model?
- fit it to data?

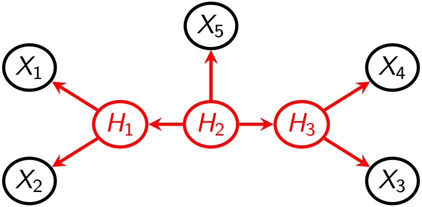
We aim to study the set of distributions constructed in this way.

Strategy: study constraints satisfied by these models.

Latent Variable Models

Traditional latent variable models would assume that the hidden variables are (e.g.) Gaussian, or discrete with some fixed number of states.

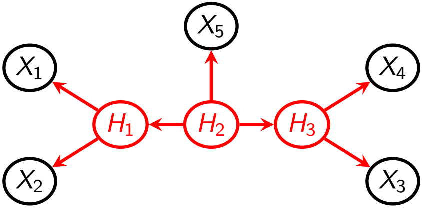
Advantages: can fit fairly easily (e.g. EM algorithm, Monte Carlo).



Latent Variable Models

Traditional latent variable models would assume that the hidden variables are (e.g.) Gaussian, or discrete with some fixed number of states.

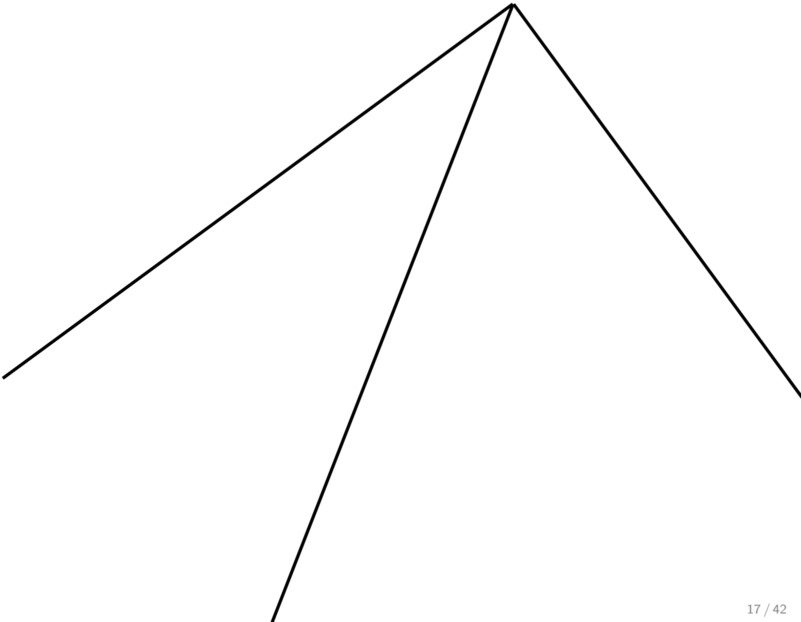
Advantages: can fit fairly easily (e.g. EM algorithm, Monte Carlo).



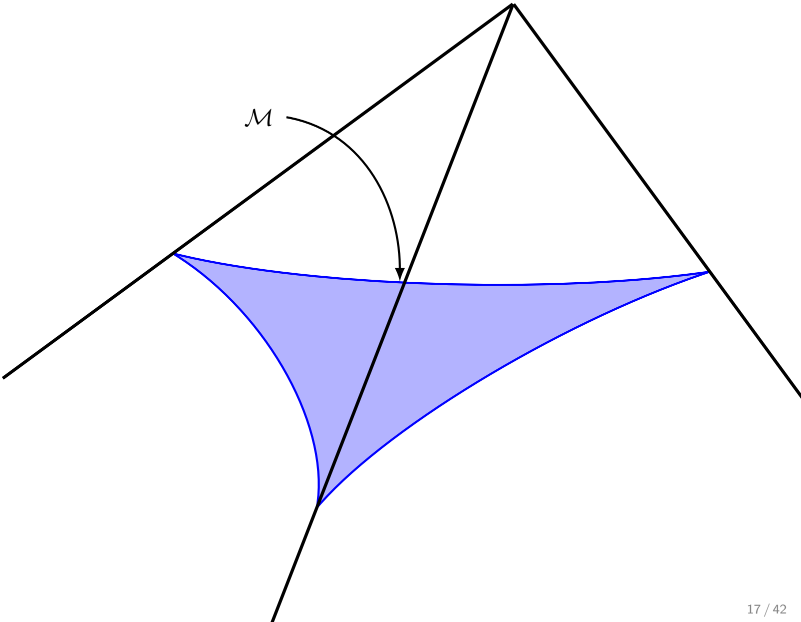
But:

- assumptions may be wrong!
- latent variables lead to singularities and nasty statistical properties (see e.g. Drton, Sturmfels and Sullivant, 2009)

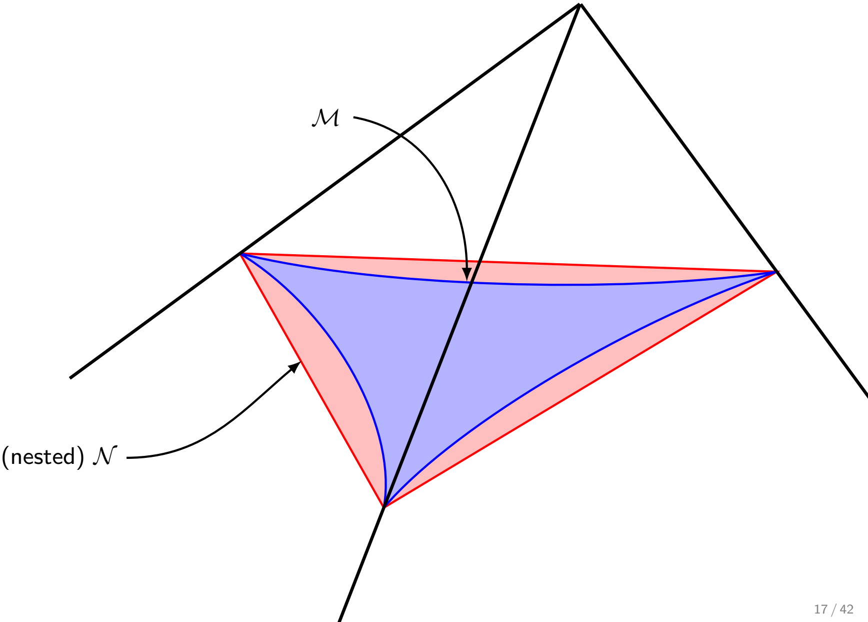
Getting the Picture



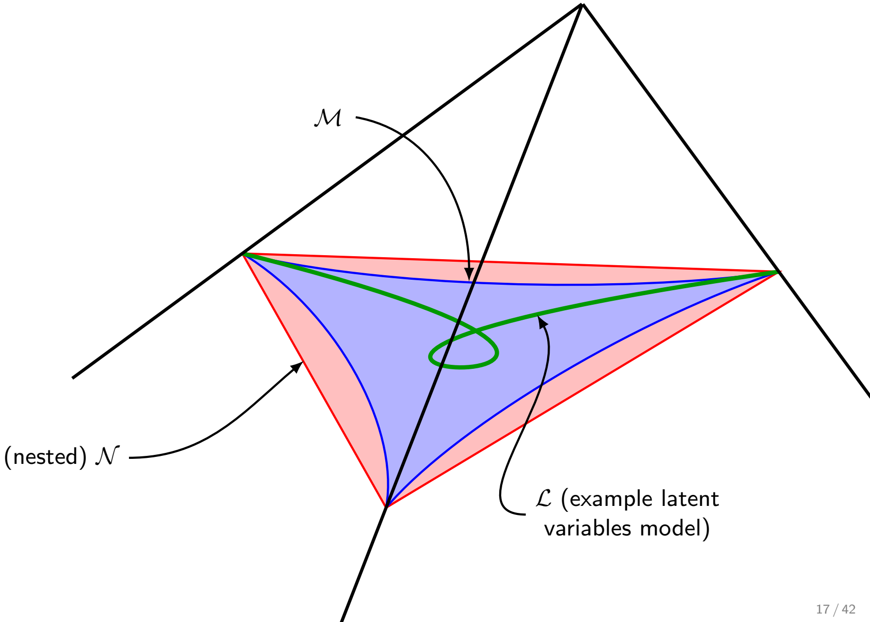
Getting the Picture



Getting the Picture



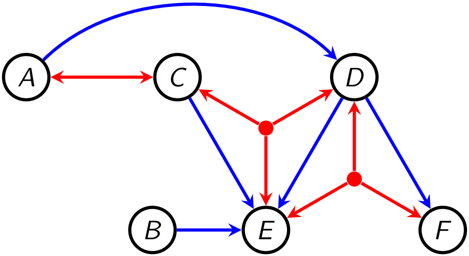
Getting the Picture



Outline

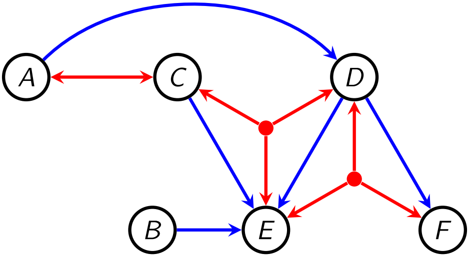
The Marginal Model

Can represent any causal model with hidden variables in following compact format; we call this an **mDAG** (Evans, 2015).



The Marginal Model

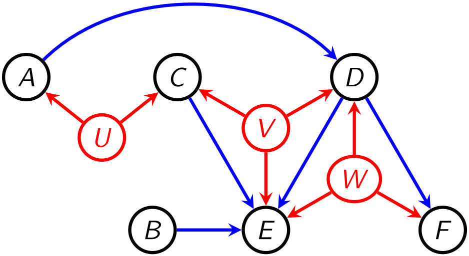
Can represent any causal model with hidden variables in following compact format; we call this an **mDAG** (Evans, 2015).



Only observed variables on graph $\mathcal{G}$; latent variables represented by red hyper edges.

The Marginal Model

Can represent any causal model with hidden variables in following compact format; we call this an **mDAG** (Evans, 2015).

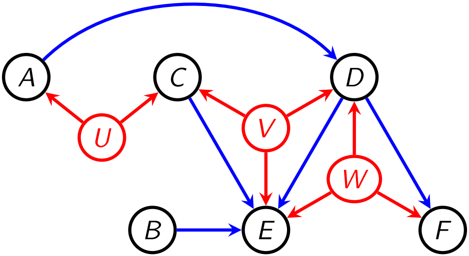


Only observed variables on graph $\mathcal{G}$; latent variables represented by red hyper edges.

Can put the latents back: call this the **canonical DAG** $\bar{\mathcal{G}}$.

The Marginal Model

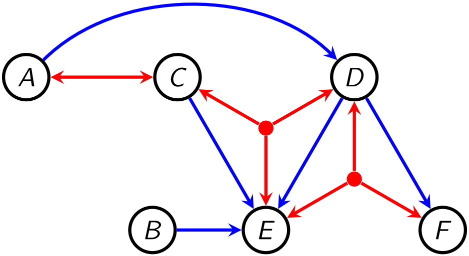
Can represent any causal model with hidden variables in following compact format; we call this an **mDAG** (Evans, 2015).



P satisfies the **marginal Markov property** for $\mathcal{G}$ if it is the margin of some distribution in $\mathcal{M}(\bar{\mathcal{G}})$.

The Marginal Model

Can represent any causal model with hidden variables in following compact format; we call this an **mDAG** (Evans, 2015).



P satisfies the **marginal Markov property** for $\mathcal{G}$ if it is the margin of some distribution in $\mathcal{M}(\bar{\mathcal{G}})$.

The **marginal model** is denoted $\mathcal{M}(\mathcal{G})$.

Model Description

We can write down a causal model, and collapse it to an mDAG, representing its margin.

But the definition of the marginal model is implicit:

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

Actually determining whether or not a distribution satisfies the marginal Markov property **is hard**.

Model Description

We can write down a causal model, and collapse it to an mDAG, representing its margin.

But the definition of the marginal model is implicit:

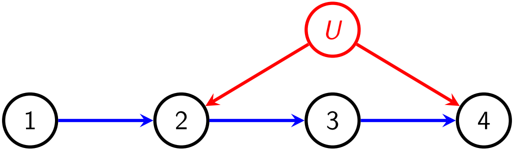
$$p(x_1, x_2, x_3, x_4) = \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(x_4 \mid x_3, u)\, du$$

Actually determining whether or not a distribution satisfies the marginal Markov property **is hard**.

Our strategy:

- derive some properties satisfied by the marginal model;
- define a new (larger) model that satisfies these properties;
- work with the larger model.

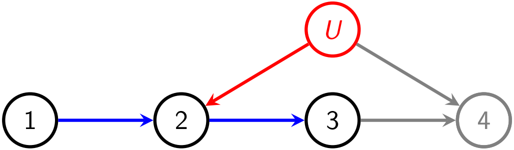
Ancestral Sets



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

.

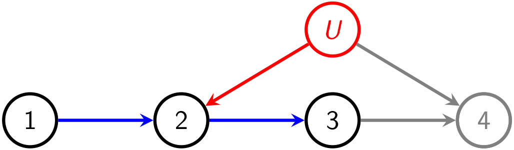
Ancestral Sets



$$p(x_1, x_2, x_3)$$
$$= \int_{x_4} \int_u p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(\boldsymbol{x_4} \mid x_3, u)\, du\, d\boldsymbol{x_4}$$

.

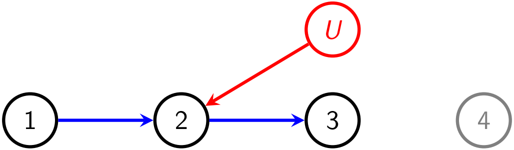
Ancestral Sets



$$p(x_1, x_2, x_3)$$
$$= \int_{\mathbf{x_4}} \int_u p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(\mathbf{x_4} \mid x_3, u)\, du\, d\mathbf{x_4}$$
$$= \int_u p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2) \int_{\mathbf{x_4}} p(\mathbf{x_4} \mid x_3, u)\, d\mathbf{x_4}\, du$$

.

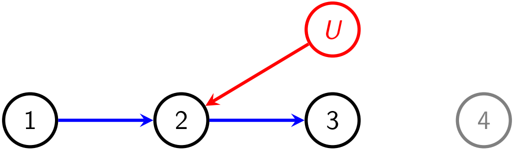
Ancestral Sets



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

.

Ancestral Sets



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

.

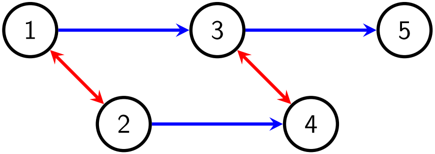
Ancestral Sets

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

Density has form corresponding to ancestral sub-graph.

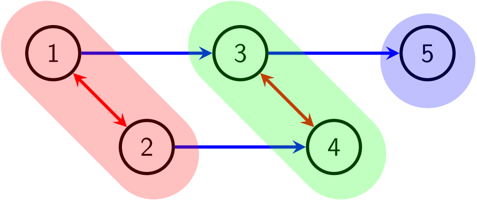
Factorization into Districts

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



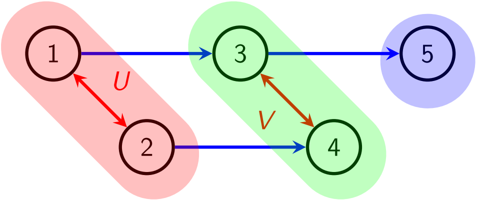
Factorization into Districts

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



Factorization into Districts

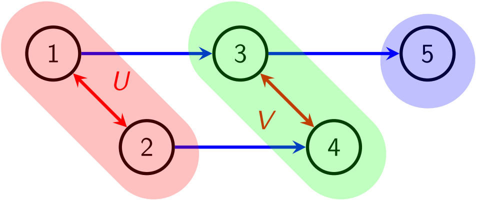
District is a maximal set connected by latent variables / bidirected edges:



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

Factorization into Districts

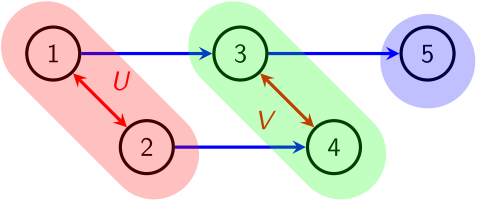
District is a maximal set connected by latent variables / bidirected edges:



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

Factorization into Districts

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

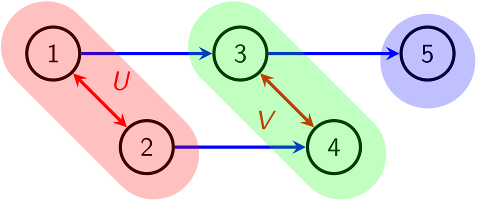


$$\int_{u,v} p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)\; du\; dv$$

$$= \int_u p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\, du \int_v p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\, dv\; p(x_5 \mid x_3)$$

Factorization into Districts

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



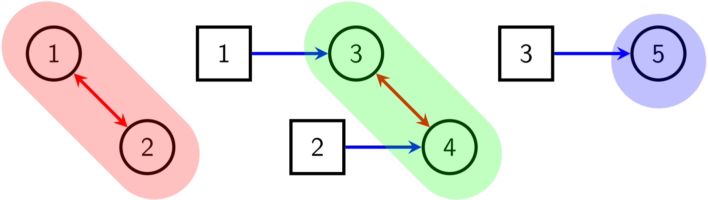
$$\int_{u,v} p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)\; du\; dv$$

$$= \int_u p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\, du \int_v p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\, dv\; p(x_5 \mid x_3)$$

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

Factorization into Districts

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



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

$$= \int_u p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\, du \int_v p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\, dv\; p(x_5 \mid x_3)$$

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

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

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

Nested Model

We use these two rules to define our model.

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

Nested Model

We use these two rules to define our model.

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

1. **Ancestrality.**

$$\int_{x_v} p(x_V \mid x_W)\, dx_v \in \mathcal{N}(\mathcal{G}_{-v})$$

   for each childless $v \in V$.

Nested Model

We use these two rules to define our model.

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

1. **Ancestrality.**
$$\int_{x_v} p(x_V \mid x_W)\, dx_v \in \mathcal{N}(\mathcal{G}_{-v})$$
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 $q_D \in \mathcal{N}(\mathcal{G}[D])$.

Nested Model

We use these two rules to define our model.

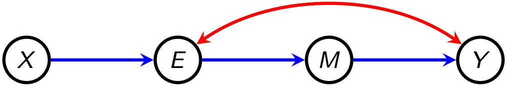
Say (conditional) probability distribution p **recursively factorizes** according to mDAG $\mathcal{G}$ and write $p \in \mathcal{N}(\mathcal{G})$ if:

1. **Ancestrality.**
$$\int_{x_v} p(x_V \mid x_W)\, dx_v \in \mathcal{N}(\mathcal{G}_{-v})$$
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** $\mathcal{N}(\mathcal{G})$. (Shpitser et al., 2014)

Example



Y childless,

Example

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

$$p(x, e, m) = p(x) \cdot p(e \mid x) \cdot p(m \mid e),$$

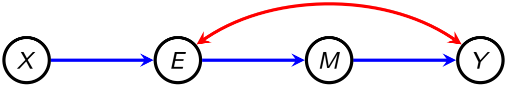
Example

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

$$p(x, e, m) = p(x) \cdot p(e \mid x) \cdot p(m \mid e),$$

and therefore $X \perp\!\!\!\perp M \mid E$.

Example



Axiom 2:

$$p(x, e, m, y) = q_X(x) \cdot q_M(m \mid e) \cdot q_{EY}(e, y \mid x, m).$$

Example

Axiom 2:

$$p(x, e, m, y) = q_X(x) \cdot q_M(m \mid e) \cdot q_{EY}(e, y \mid x, m).$$

Can consider the district $\{E, Y\}$ and factor q_{EY}...

Example

Axiom 2:

$$p(x, e, m, y) = q_X(x) \cdot q_M(m \mid e) \cdot q_{EY}(e, y \mid x, m).$$

Can consider the district $\{E, Y\}$ and factor q_{EY}...
and then apply Axiom 1 to marginalize E.

We see that $X \perp\!\!\!\perp M, Y \,[q_{EY}]$.

Example

Axiom 2:

$$p(x, e, m, y) = q_X(x) \cdot q_M(m \mid e) \cdot q_{EY}(e, y \mid x, m).$$

Can consider the district $\{E, Y\}$ and factor q_{EY}...
and then apply Axiom 1 to marginalize E.

We see that $X \perp\!\!\!\perp M, Y \, [q_{EY}]$.

This places a non-trivial constraint on p.

Completeness

We know

$$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}).$$

Could there be other constraints?

Completeness

We know

$$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}).$$

Could there be other constraints?
For discrete observed variables, we know not.

Theorem (Evans, 2015a)

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

Completeness

We know

$$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}).$$

Could there be other constraints?
For discrete observed variables, we know not.

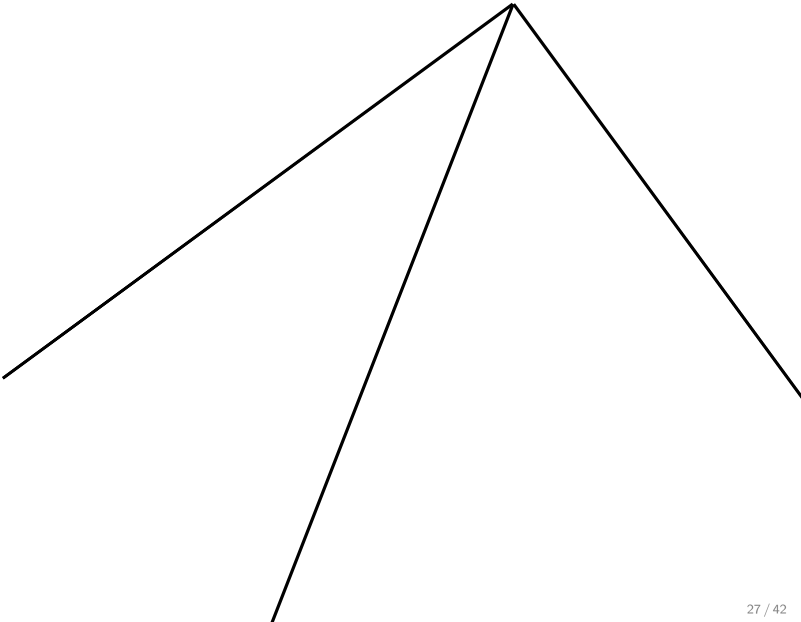
Theorem (Evans, 2015a)

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

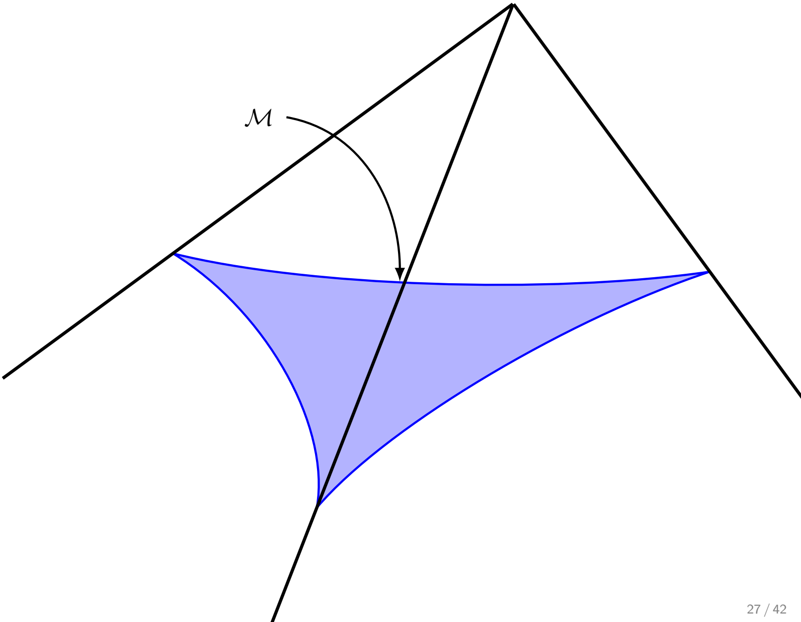
'Algebraically equivalent' $=$ 'up to inequalities'.
Any 'gap' $\mathcal{M}(\mathcal{G}) \subset \mathcal{N}(\mathcal{G})$ is due to inequality constraints.

So in particular they have the same dimension.

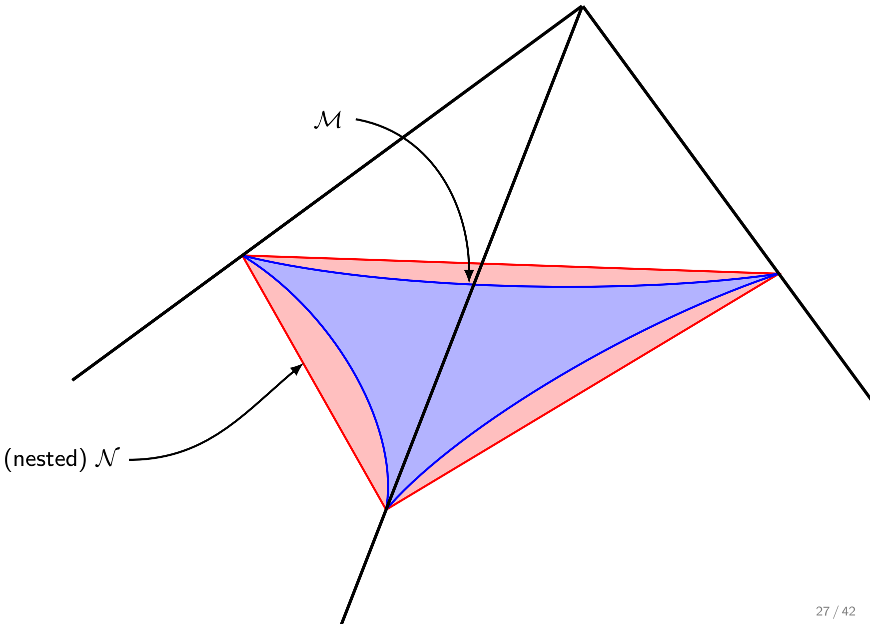
Getting the Picture



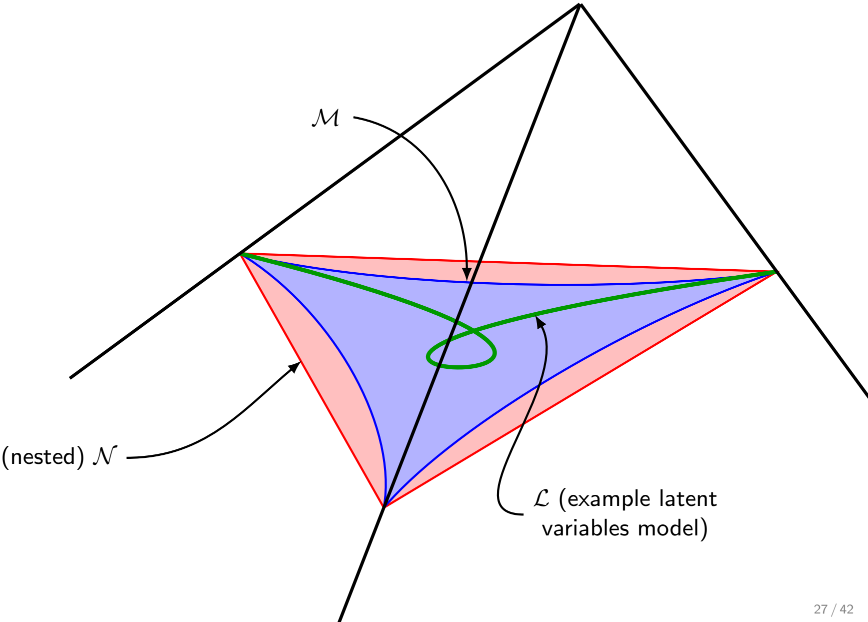
Getting the Picture



Getting the Picture



Getting the Picture



Main Result

Nested model is a good approximation to the marginal model: in the discrete case it can be explicitly parameterized and fitted.

Theorem (Evans and Richardson, 2015)

Discrete nested models are curved exponential families.

This has very nice statistical implications, including for the marginal model.

Main Result

Nested model is a good approximation to the marginal model: in the discrete case it can be explicitly parameterized and fitted.

Theorem (Evans and Richardson, 2015)

Discrete nested models are curved exponential families.

This has very nice statistical implications, including for the marginal model.

All parameters are of the form $p(\boldsymbol{X} \mid \mathrm{do}(\boldsymbol{Y}))$: easily interpretable.

Wisconsin Data Example

Take only male respondents who were either drafted or didn't enter military at all (before 1975).

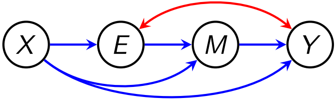
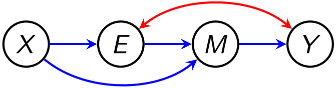
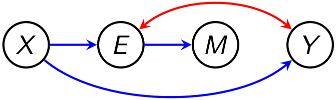
Continuous values dichotomised close to median.

Four binary indicators:

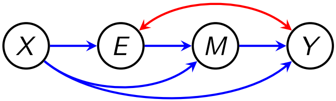
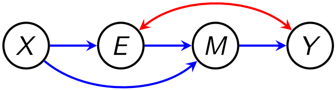
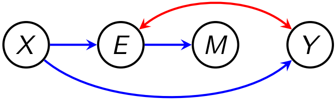
- X family income $>$\$5k in 1957;
- E education post high school;
- M drafted into military;
- Y respondent income $>$\$37k in 1992.

1,676 complete cases in 2^4 contingency table (minimum count 16).

Results

	model	deviance	d.f.
(a)		(saturated)	15
(b)		31.3	2
(c)		5.6	6

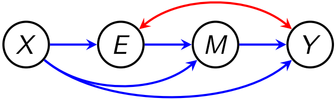
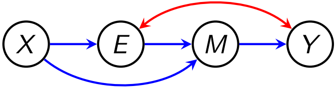
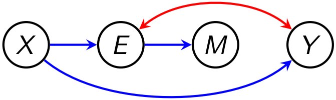
Results

	model	deviance	d.f.
(a)		(saturated)	15
(b)		31.3	2
(c)		5.6	6

No evidence that military service has any effect on income after controlling for education.

Removing any edges from (c) strongly rejected.

Results

	model	deviance	d.f.
(a)		(saturated)	15
(b)		31.3	2
(c)		5.6	6

No evidence that military service has any effect on income after controlling for education.

Removing any edges from (c) strongly rejected.

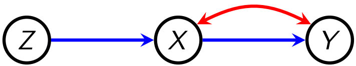
Also find strong residual income effect:

$$P(Y = 1 \mid \text{do}(X = 0)) = 0.36 \qquad P(Y = 1 \mid \text{do}(X = 1)) = 0.50.$$

Outline

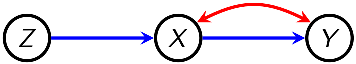
The IV Model

The **instrumental variables** model is represented by the mDAG below.



The IV Model

The **instrumental variables** model is represented by the mDAG below.

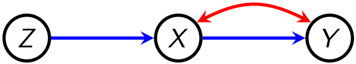


Assume all observed variables are discrete.

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

The IV Model

The **instrumental variables** model is represented by the mDAG below.



Assume all observed variables are discrete.

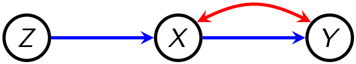
Nested Markov property gives saturated model, so true model of full dimension.

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

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

The IV Model

The **instrumental variables** model is represented by the mDAG below.



Assume all observed variables are discrete.

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

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

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

e.g.

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

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

Missing Edges Give Constraints

Proposition (Evans, 2012)

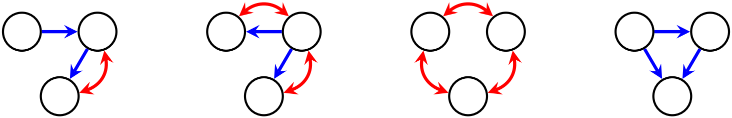
If X and Y are not joined by an edge in $\mathcal{G}$ there is always a constraint induced on a discrete joint distribution.

Outline

Equivalence on Three Variables

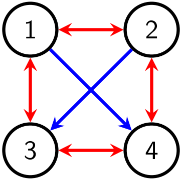
Markov equivalence (i.e. determining whether two models are observably the same) is hard.

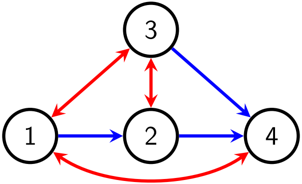
Using Evans (2015) there are 8 unlabelled marginal models on three variables.

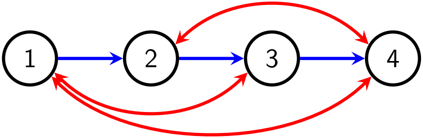


But Not on Four!

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)







Fitting Marginal Models

The ‘implicit’ nature of marginal models makes them hard to describe and to test.

We can test constraints individually, but this is very inefficient.

Fitting Marginal Models

The 'implicit' nature of marginal models makes them hard to describe and to test.

We can test constraints individually, but this is very inefficient.

On the other hand

- the nested model $\mathcal{N}(\mathcal{G})$ can be parameterized and fitted;
- latent variable models $\mathcal{L}(\mathcal{G})$ can be parameterized and fitted;
- $\mathcal{L}(\mathcal{G}) \subseteq \mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G})$.

Fitting Marginal Models

The ‘implicit’ nature of marginal models makes them hard to describe and to test.

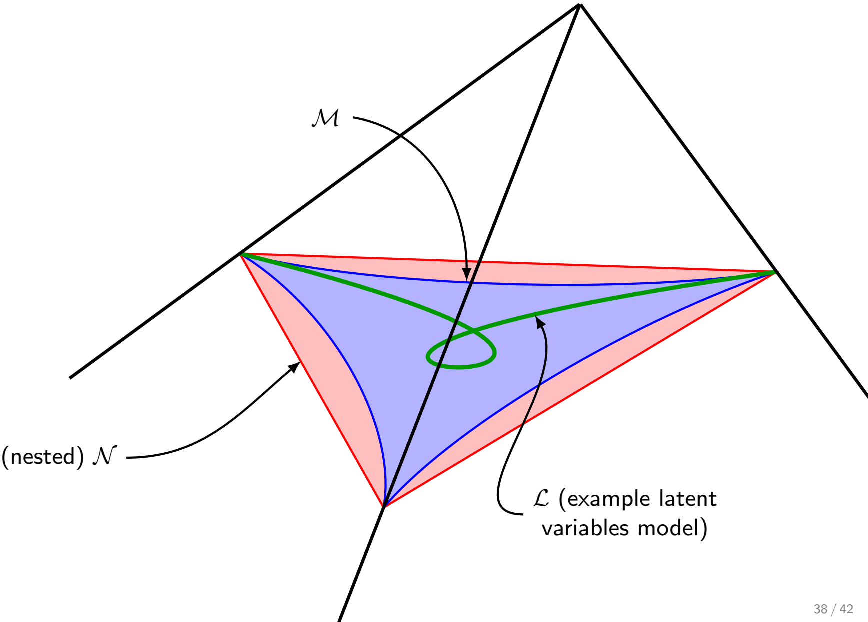
We can test constraints individually, but this is very inefficient.

On the other hand

- the nested model $\mathcal{N}(\mathcal{G})$ can be parameterized and fitted;
- latent variable models $\mathcal{L}(\mathcal{G})$ can be parameterized and fitted;
- $\mathcal{L}(\mathcal{G}) \subseteq \mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G})$.

So if we accept the latent variable model, or reject the nested model, same applies to marginal model.

That Picture Again



Some Extensions

We know nested models are curved exponential families, so justifies classical statistical theory:

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

Some Extensions

We know nested models are curved exponential families, so justifies classical statistical theory:

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

Since marginal models are the same dimension, they share these properties (except on their boundary).

Also, latent variable models **become** regular if state-space is large enough.

Some Extensions

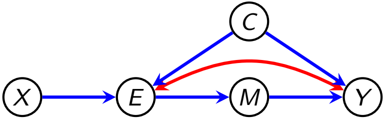
We know nested models are curved exponential families, so justifies classical statistical theory:

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

Since marginal models are the same dimension, they share these properties (except on their boundary).

Also, latent variable models **become** regular if state-space is large enough.

Can also include continuous covariates with outcome as multivariate response. e.g.:



Summary

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

Summary

- (Causal) DAGs with latent variables induce non-parametric constraints;
- can use these to define nested models;

Summary

- (Causal) DAGs with latent variables induce non-parametric constraints;
- can use these to define nested models;
- avoids some problems and assumptions of latent variable models: non-regularity, unidentifiability;

Summary

- (Causal) DAGs with latent variables induce non-parametric constraints;
- can use these to define nested models;
- avoids some problems and assumptions of latent variable models: non-regularity, unidentifiability;
- discrete parameterization and fitting algorithms available;

Summary

- (Causal) DAGs with latent variables induce non-parametric constraints;
- can use these to define nested models;
- avoids some problems and assumptions of latent variable models: non-regularity, unidentifiability;
- discrete parameterization and fitting algorithms available;
- solves some boundary issues (at expense of larger model class).

Summary

- (Causal) DAGs with latent variables induce non-parametric constraints;
- can use these to define nested models;
- avoids some problems and assumptions of latent variable models: non-regularity, unidentifiability;
- discrete parameterization and fitting algorithms available;
- solves some boundary issues (at expense of larger model class).

Some limitations:

Summary

- (Causal) DAGs with latent variables induce non-parametric constraints;
- can use these to define nested models;
- avoids some problems and assumptions of latent variable models: non-regularity, unidentifiability;
- discrete parameterization and fitting algorithms available;
- solves some boundary issues (at expense of larger model class).

Some limitations:

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

Summary

- (Causal) DAGs with latent variables induce non-parametric constraints;
- can use these to define nested models;
- avoids some problems and assumptions of latent variable models: non-regularity, unidentifiability;
- discrete parameterization and fitting algorithms available;
- solves some boundary issues (at expense of larger model class).

Some limitations:

- Complete inequality constraints seem very complicated (though some hope exists);
- nice rule for model equivalence not yet available for either nested or marginal models.

Thank you!

References

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

Evans – Graphs for margins of Bayesian networks, *arXiv:1408.1809*, *Scand. J. Statist.*, to appear, 2015.

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

Evans and Richardson – Smooth, identifiable supermodels of discrete DAG models with latent variables, *arXiv:1511.06813*, 2015.

Pearl – On the testability of causal models with latent and instrumental variables, *UAI*, 1995.

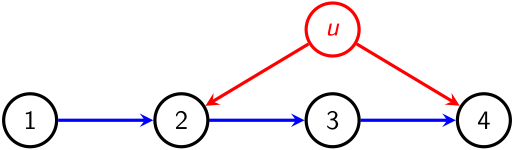
Richardson and Spirtes. Ancestral graph Markov models, *Ann. Stat.*, 2002.

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

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

Verma, Pearl. Equivalence and synthesis of causal models, *UAI*, 1990.

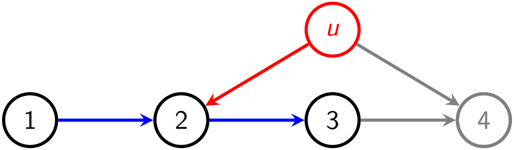
Ancestral Sets



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

.

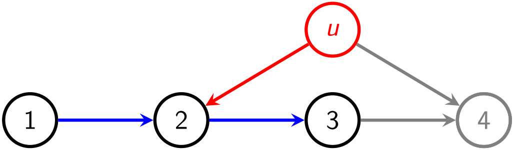
Ancestral Sets



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

.

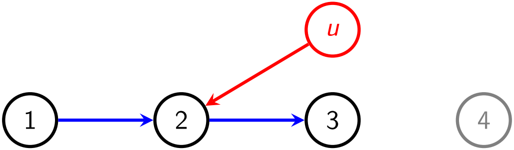
Ancestral Sets



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

.

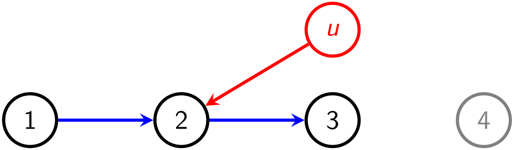
Ancestral Sets



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

.

Ancestral Sets



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

.

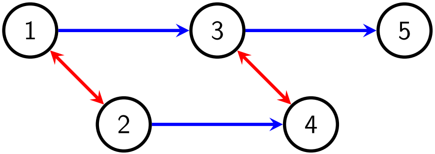
Ancestral Sets

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

Density has form corresponding to ancestral sub-graph.

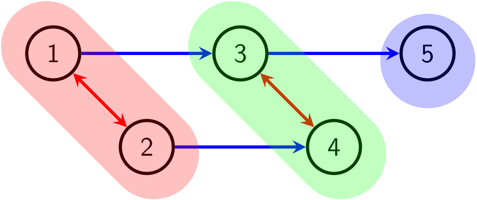
Factorization into Districts

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



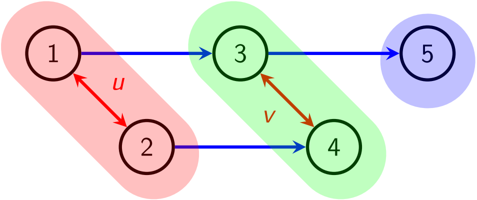
Factorization into Districts

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



Factorization into Districts

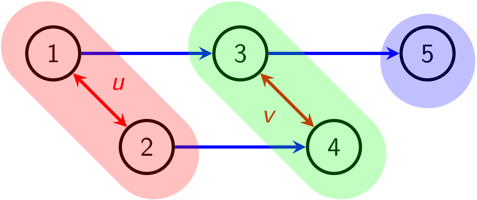
District is a maximal set connected by latent variables / bidirected edges:



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

Factorization into Districts

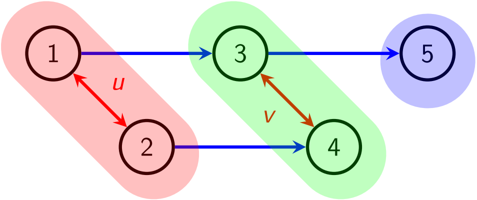
District is a maximal set connected by latent variables / bidirected edges:



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

Factorization into Districts

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

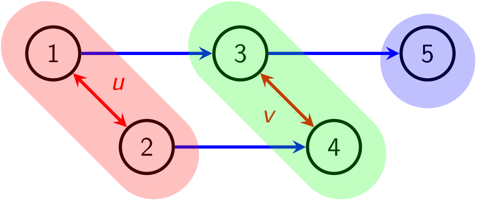


$$\sum_{u,v} p(u)\,p(x_1 \mid u)\,p(x_2 \mid u)\;\; p(v)\,p(x_3 \mid x_1, v)\,p(x_4 \mid x_2, v)\;\; p(x_5 \mid x_3)$$

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

Factorization into Districts

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



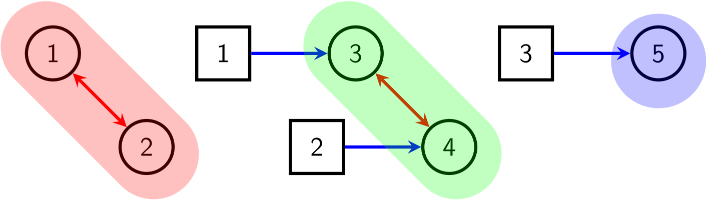
$$\sum_{u,v} p(u)\, p(x_1 \mid u)\, p(x_2 \mid u) \; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v) \; p(x_5 \mid x_3)$$

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

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

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 ($\mathcal{G}[D]$).

Nested Model

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:

Nested Model

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 $v \in V$.

Nested Model

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 $q_D \in \mathcal{N}(\mathcal{G}[D])$.

Nested Model

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

Causal Coherence of mDAGs

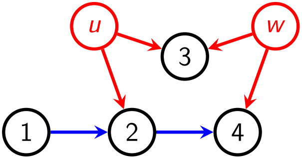
If P is represented be a DAG in a causally interpreted way, then intervening on some set of nodes $C \subseteq V$ can be represented by deleting incoming edges to C in $\mathcal{G}$. Call that graph $\mathcal{G}^{\overline{C}}$

Theorem (Evans, 2015)

If $C \subseteq O$ then $\mathfrak{p}(\mathcal{G}^{\overline{C}}, O) = \mathfrak{p}(\mathcal{G}, O)^{\overline{C}}$; i.e. the projection respects causal interventions.

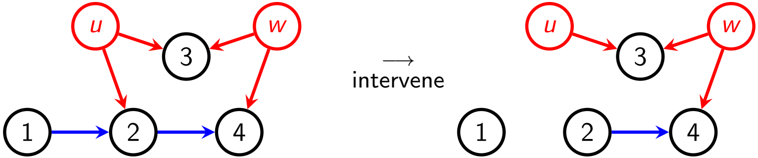
Causal Coherence

If we intervene on some observed variables, this 'breaks' their dependence upon their parents.



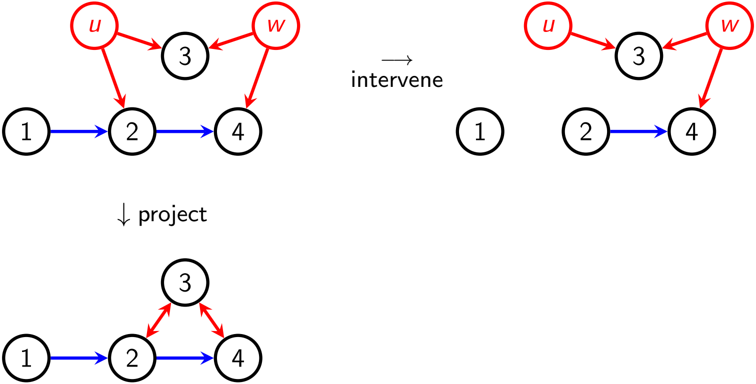
Causal Coherence

If we intervene on some observed variables, this ‘breaks’ their dependence upon their parents.



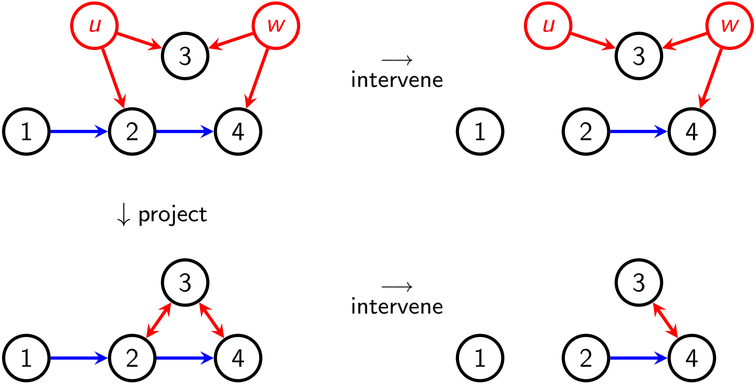
Causal Coherence

If we intervene on some observed variables, this ‘breaks’ their dependence upon their parents.



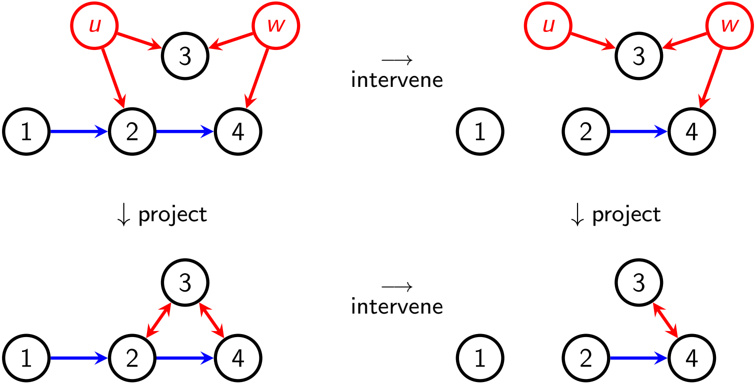
Causal Coherence

If we intervene on some observed variables, this 'breaks' their dependence upon their parents.

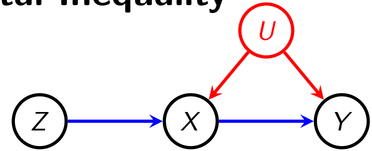


Causal Coherence

If we intervene on some observed variables, this ‘breaks’ their dependence upon their parents.

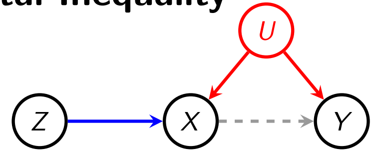


Proof of Instrumental Inequality



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

Proof of Instrumental Inequality



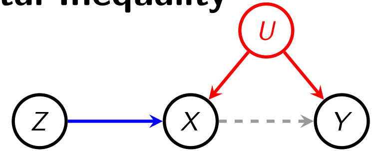
Have: $$p(x, y \mid z) = \int p(u)\, p(x \mid z, u)\, p(y \mid x, u)\, du.$$

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

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

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

Proof of Instrumental Inequality



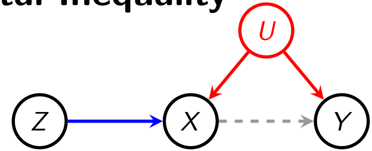
Have: $$p(x, y \mid z) = \int p(u)\, p(x \mid z, u)\, p(y \mid x, u)\, du.$$

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

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

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

Proof of Instrumental Inequality



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^*_ξ:

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

Now Y behaves as though $X = \xi$ 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:** $p(\xi, y \mid z) = p^*(\xi, y \mid z)$ for each z, y; and
- **Independence:** $Y \perp\!\!\!\perp Z$ under p^*.

Solution: A Different Proof

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 a distribution exist?

Solution: A Different Proof

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 a distribution exist?

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

Solution: A Different Proof

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 a distribution exist?

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

Solution: A Different Proof

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 a distribution exist?

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

Solution: A Different Proof

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 a distribution exist?

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

So clearly

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

Solution: A Different Proof

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 a distribution 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 maxing over ξ, the instrumental inequality follows.

Solution: A Different Proof

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 a distribution 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 maxing over ξ, the instrumental inequality follows.

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

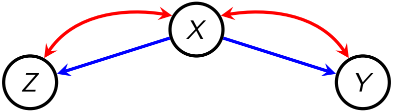
Generalizing

How does this help us with other graphs?

Generalizing

How does this help us with other graphs?

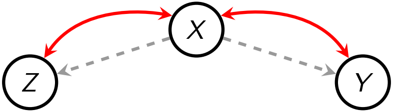
The argument works precisely because cutting edges led to an independence:



Generalizing

How does this help us with other graphs?

The argument works precisely because cutting edges led to an independence:

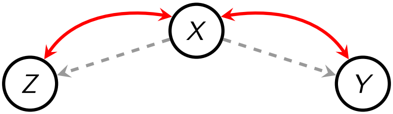


Z is independent of Y in the graph after cutting edges emanating from X.

Generalizing

How does this help us with other graphs?

The argument works precisely because cutting edges led to an independence:



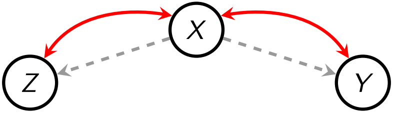
Z is independent of Y in the graph after cutting edges emanating from X.

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 an independence:



Z is independent of 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.]

d-Separation

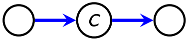
A **path** is a sequence of edges in the graph; vertices may not be repeated.

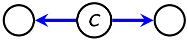
d-Separation

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:



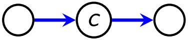


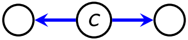
d-Separation

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

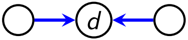
A path from v to w is **blocked** by $C \subseteq V \setminus \{v, w\}$ if either

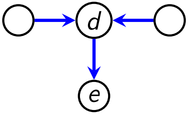
(i) any non-collider is in C:





(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.