

# Causal Models with Hidden Variables

Robin J. Evans

[www.stats.ox.ac.uk/~evans](http://www.stats.ox.ac.uk/~evans)

Department of Statistics, University of Oxford

Quantum Networks, Oxford

August 2017

# Correlation does not imply causation

Find a Job | M&S Wine | Feedback

## MailOnline

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

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

### How a short nap can raise blood pressure and cholesterol: Study finds people who take a siesta are more likely to have high blood pressure and high cholesterol

- Napping for more than 30 minutes at a time is associated with a higher risk of high blood pressure, according to a new study

They were much favoured by Margaret Thatcher, Albert Einstein and Benjamin Franklin. But while afternoon naps may revitalise tired brains, this is not always true, according to new research.

Wednesday, November 13, 2013



### 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<sup>a, b</sup>, Zhongliang Li<sup>a</sup>, Li Wu<sup>a</sup>, Zhongqiang Cao<sup>a</sup>, Yuan Liang<sup>a, c</sup>, Handong Yang<sup>d</sup>, Youjie Wang<sup>a, b</sup>, Tangchun Wu<sup>a</sup>

<sup>a</sup>Ministry of Education Key Laboratory of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, China

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

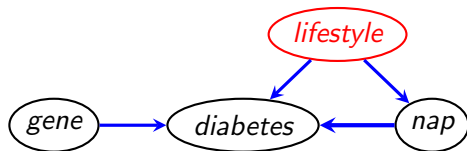
Abstract

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



# Distinguishing Between Causal Models

In order to compare the models, we need to understand in what ways causal models will differ, both:

- observationally;
- under interventions.


This question has been much studied in statistics and computer science: e.g. Robins (1986), Verma and Pearl (1990), Richardson and Spirtes (2002), Tian and Pearl (2002), Richardson et al. (2017).

It has also been of interest in the quantum literature: e.g. Bell (1964), Clauser et al. (1969), Fritz (2012), Chaves et al. (2014), Pienaar (2016).

# Outline

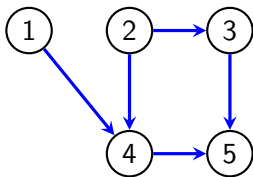
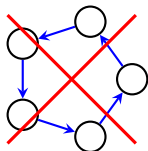
- 1 Introduction
- 2 DAG Models
- 3 Equality Constraints
- 4 Inequality Constraints
- 5 Testing, Fitting and Searching

# Directed Acyclic Graphs

vertices 

edges 

no directed cycles



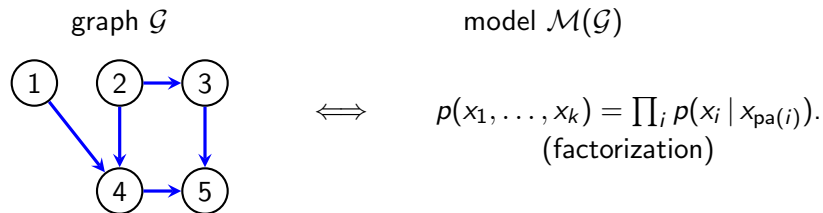
directed acyclic graph (DAG),  $\mathcal{G}$

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

If  $w \rightarrow \dots \rightarrow v$  then  $w$  is a **ancestor** of  $v$ .

An **ancestral set** contains all its own ancestors.

# DAG Models (aka Bayesian Networks)



So in example above:

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

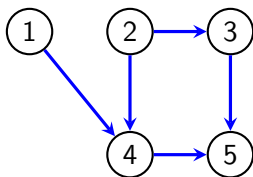
$\mathcal{M}(\mathcal{G})$  is also precisely those distributions such that:

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

so is defined by **polynomial constraints** (algebraic variety).

# 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  $X_4$ :

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

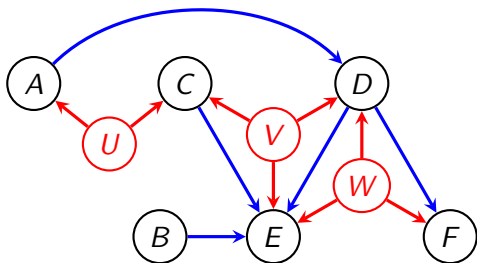
$$p(x_1, x_2, x_3, x_5 | \text{do}(x_4)) = p(x_1) \cdot p(x_2) \cdot p(x_3 | x_2) \cdot p(x_5 | x_3, x_4).$$

All other terms preserved.



# The Marginal Model

Can represent any causal model with hidden variables in following compact format; we call this an **mDAG** (Evans, 2016).

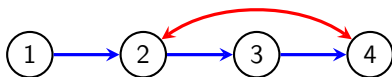


Only observed variables on graph  $\mathcal{G}$ ; latent variables represented by red hyper edges.

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

# Model Description

So we can associate an mDAG with a set of models.



But the definition of the marginal model is implicit:

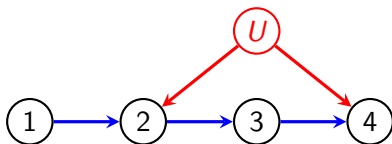
$$p(x_1, x_2, x_3, x_4) = \int p(u) p(x_1) p(x_2 | x_1, u) p(x_3 | x_2) p(x_4 | x_3, u) du$$

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.

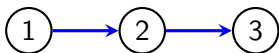
# Ancestral Sets



$$p(x_1, x_2, x_3, x_4)$$

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

# Ancestral Sets

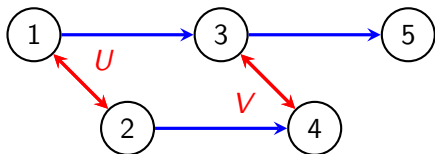


$$\begin{aligned} & p(x_1, x_2, x_3, x_4) \\ &= \int_u p(u) p(x_1) p(x_2 | x_1, u) p(x_3 | x_2) p(x_4 | x_3, u) du \\ &= \int_u p(u) p(x_1) p(x_2 | x_1, u) p(x_3 | x_2) \int_{x_4} p(x_4 | x_3, u) dx_4 du \\ &= \int_u p(u) p(x_1) p(x_2 | x_1, u) p(x_3 | x_2) du \\ &= p(x_1) p(x_3 | x_2) \int_u p(u) p(x_2 | x_1, u) du \\ &= p(x_1) p(x_3 | x_2) p(x_2 | x_1). \end{aligned}$$

Density has form corresponding to ancestral sub-graph.

# Factorization into Districts

A **district** is a maximal set connected by latent variables:

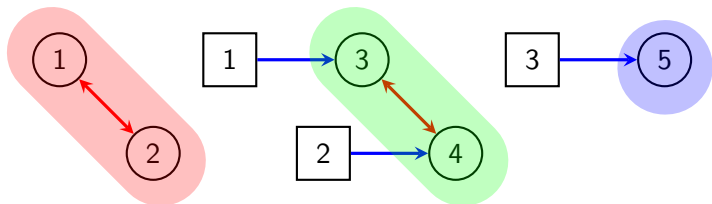


$$\begin{aligned} & \int_{u,v} p(u) p(x_1 | u) p(x_2 | u) p(v) p(x_3 | x_1, v) p(x_4 | x_2, v) p(x_5 | x_3) du dv \\ &= \int_u p(u) p(x_1 | u) p(x_2 | u) du \int_v p(v) p(x_3 | x_1, v) p(x_4 | x_2, v) dv p(x_5 | x_3) \\ &= q_{12}(x_1, x_2) \cdot q_{34}(x_3, x_4 | x_1, x_2) \cdot q_5(x_5 | x_3) \cdot \\ &= \prod_i q_{D_i}(x_{D_i} | x_{\text{pa}(D_i) \setminus D_i}) \end{aligned}$$

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

# Factorization into Districts

A **district** is a maximal set connected by latent variables:



$$\begin{aligned} & \int_{u,v} p(u) p(x_1 | u) p(x_2 | u) p(v) p(x_3 | x_1, v) p(x_4 | x_2, v) p(x_5 | x_3) du dv \\ &= \int_u p(u) p(x_1 | u) p(x_2 | u) du \int_v p(v) p(x_3 | x_1, v) p(x_4 | x_2, v) dv p(x_5 | x_3) \\ &= q_{12}(x_1, x_2) \cdot q_{34}(x_3, x_4 | x_1, x_2) \cdot q_5(x_5 | x_3) \cdot \\ &= \prod_i q_{D_i}(x_{D_i} | x_{\text{pa}(D_i) \setminus D_i}) \end{aligned}$$

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

# Nested Model

Using these two rules alternately leads to algorithm of Tian and Pearl (2002).

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

## 1. Ancestrality.

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

for each childless  $v \in V$ .

## 2. Factorization into districts.

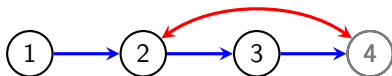
$$p(x_V | x_W) = \prod_D q_D(x_D | x_{\text{pa}(D) \setminus D})$$

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**  $\mathcal{N}(\mathcal{G})$ . (Richardson et al., 2017)

## Example



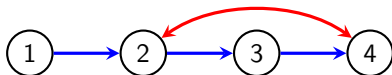
4 is childless, so if  $p \in \mathcal{N}(\mathcal{G})$ , then

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

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



# Example



Axiom 2:

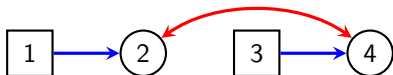
$$p(x_1, x_2, x_3, x_4) = q_1(x_2) \cdot q_3(x_3 | x_2) \cdot q_{24}(x_2, x_4 | x_1, x_3).$$

Can consider the district  $\{2, 4\}$  and factor  $q_{24} \dots$   
and then apply Axiom 1 to marginalize 2.

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

This places a non-trivial constraint on  $p$ .

## Example



Axiom 2:

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

Can consider the district  $\{2, 4\}$  and factor  $q_{24} \dots$   
and then apply Axiom 1 to marginalize 2.

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

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?

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).

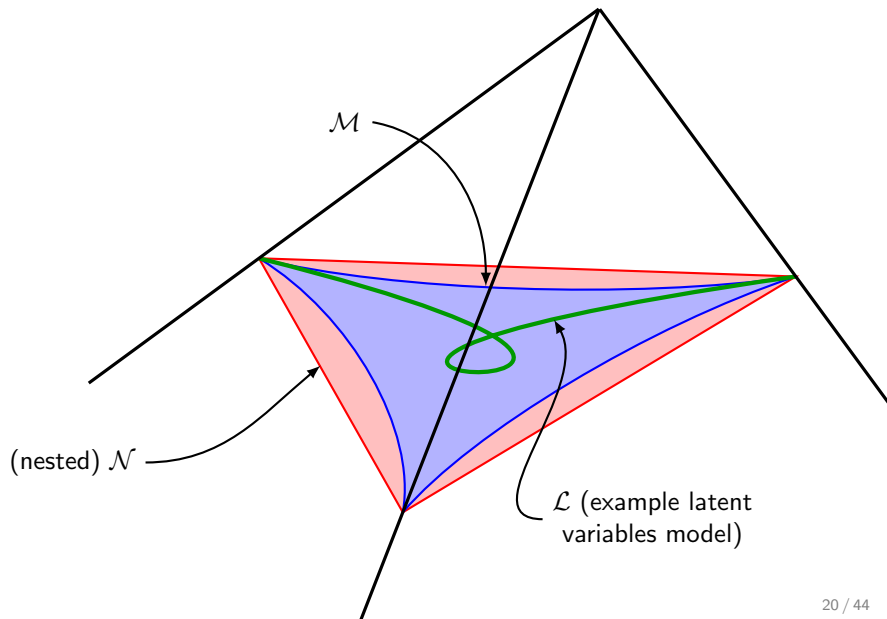
'Algebraically equivalent' = 'up to inequalities'.

Any 'gap'  $\mathcal{M}(\mathcal{G}) \subset \mathcal{N}(\mathcal{G})$  is due to inequality constraints (e.g. Bell/CHSH inequalities).

So in particular they have the same dimension.

The same result should apply to Bayesian networks with non-classical hidden variables.

# Getting the Picture



# Smoothness and Fitting

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

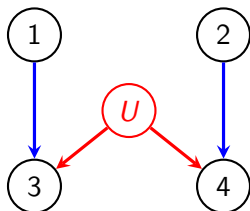
Theorem (Evans and Richardson, 2015)

Discrete nested models are curved exponential families (so the models are manifolds).

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

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

# The First Inequality



The nested model for this graph tells us that

$$X_1 \perp\!\!\!\perp X_2, X_4$$

$$X_2 \perp\!\!\!\perp X_1, X_3.$$

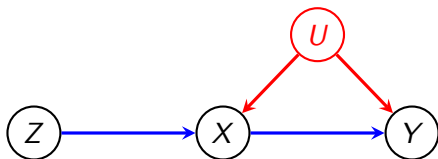
Letting (e.g.)  $p_{00|ij} = P(X_3 = 0, X_4 = 1 \mid X_1 = i, X_2 = j)$ , the CHSH inequalities are:

$$0 \leq p_{00|ij} + p_{11|i'j} + p_{11|ij'} - p_{11|i'j'} \leq 1$$

for all  $i, i', j, j' \in \{0, 1\}$ ,  $i + i' = 1$ ,  $j + j' = 1$ .

# The IV Model

Another important network is the **instrumental variables (IV)** model.

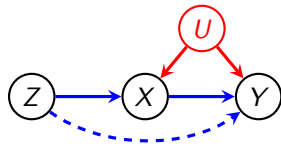


The DAG encodes a probability factorisation:

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

of which we can observe the  $Z, X, Y$  margin.

# Instrumental Inequalities



The assumption  $Z \not\leftrightarrow Y$  is important.

**Can we check it?**

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

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

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

Example:

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

No obvious graphical interpretation to Pearl's inequalities.



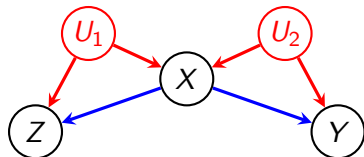
# The Problem

General inequalities seem much worse:

$$p(1, 2 | 2) + p(1, 1 | 3) + p(1, 2 | 1) + p(2, 2 | 2) + p(2, 1 | 1) \leq 2$$

where  $p(i, j | k) = P(X = i, Y = j | Z = k)$ ; (Bonet, 2001).

It's also not clear how to get inequalities for other graphs:

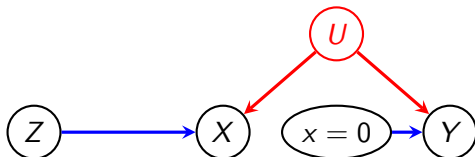


Pearl's proof does not obviously generalise.

Computational linear algebra only works for one latent variable.  
Also very computationally intensive.

Finding complete bounds in general is probably intractably hard.

# A Different Interpretation



Suppose we 'pretend' to  $Y$  that  $X = 0$ .

Use this to define a new (fictitious) distribution  $P^*$

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

**Can't observe  $P^*$  but:**

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

These constraints give precisely the IV inequality (Evans, 2012).

# Compatibility

Probabilities may not be compatible with independences.

Consider a partial probability table  $p(x = 0, y | z)$ :

$X = 0$	$Z = 0$	$Z = 1$
$Y = 0$	$2/3$	$0$
$Y = 1$	$0$	$2/3$

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 Generalisation

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

Theorem (Evans, 2012)

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

$$P(\mathbf{x}, \mathbf{y}, \mathbf{w} \mid \mathbf{z}), \quad \mathbf{x}, \mathbf{y}, \mathbf{z}.$$

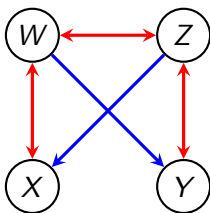
are **compatible with a distribution**  $P^*$ , in which  $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} [P^*]$ .

The simpler corollary:

Corollary

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

## Example



$X$  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  $w$ .

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

### Lemma

Testing compatibility of a probability distribution with a conditional independence is a convex optimisation problem.

## Limitations / Other Inequality Results

This method does not give Bell/CHSH type inequalities.

Fritz (2012) shows that pairwise latent variables do not induce saturated model. See also Evans (2016).



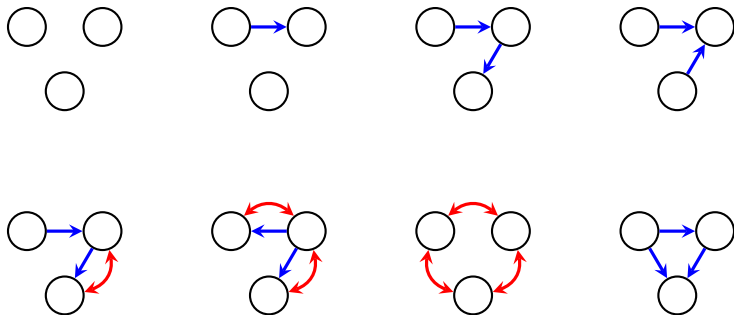
Chaves et al. (2014) derive entropic inequalities: completely non-parametric, but generally weaker. For the IV model:

$$I(Y : Z|X) + I(X : Z) \leq H(X).$$

# Equivalence on Three Variables

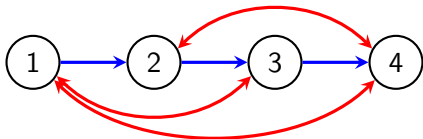
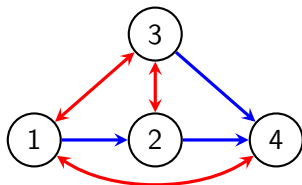
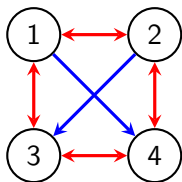
Markov equivalence (i.e. determining whether two models are observably the same) is hard.

Using Evans (2016) 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.

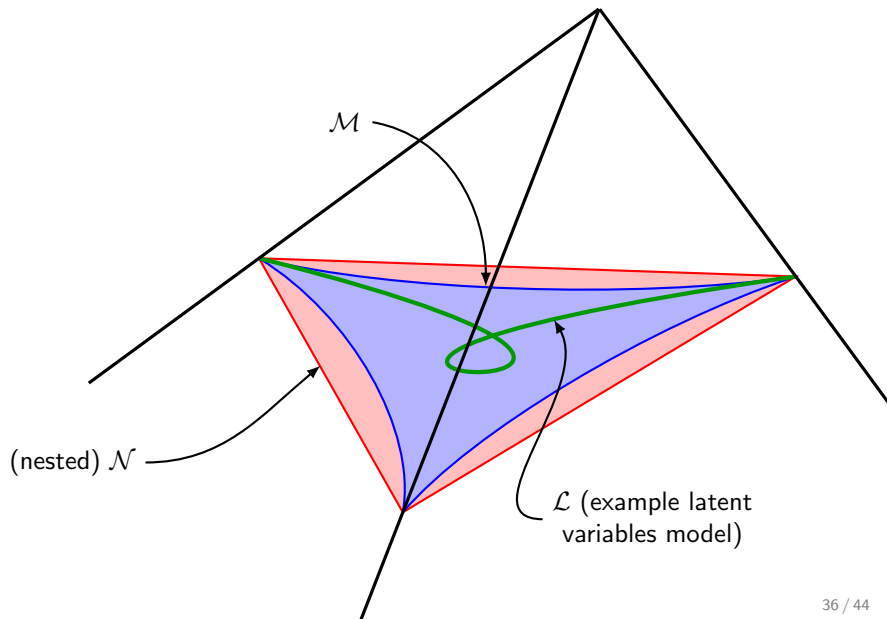
We can test constraints individually, but this is very inefficient.

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



## Example

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.

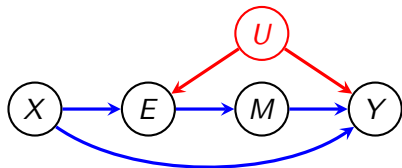
$X$  family income in 1957;

$E$  education level;

$M$  drafted into military;

$Y$  respondent income 1992;

$U$  unmeasured confounding.



# 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;

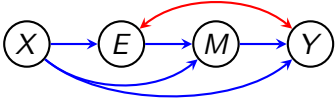
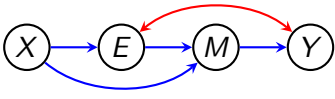
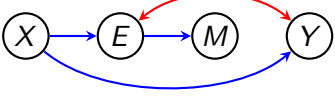
$E$  education post high school;

$M$  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

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:

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

## Some Extensions

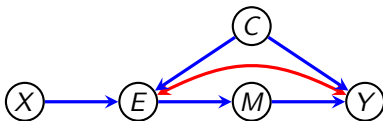
We know nested models are curved exponential families, so justifies classical statistical theory:

- likelihood ratio tests have asymptotic  $\chi^2$ -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;
- 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 I

Bell – On the Einstein-Podolsky-Rosen Paradox, *Physics*, 1964.

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

Chaves, . . . , Schölkopf – Inferring latent structures via information inequalities, *UAI*. 2014.

Clauser, Horne, Shimony, Holt – Proposed experiment to test local hidden-variable theories, *Phys. Rev. Lett.*, 1969.

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

Evans – Graphs for margins of Bayesian networks, *arXiv:1408.1809*, *Scand. J. Statist.*, 2016.

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.

Fritz – Beyond Bell's Theorem: Correlation Scenarios, *New J. Phys.*, 2012.

## References II

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

Pienaar – Which causal structures might support a quantum-classical gap?, *New J. Phys.*, 2017.

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

Richardson, Evans, Robins, and Shpitser – Nested Markov properties for ADMGs. *arXiv:1701.06686*, 2017.

Robins – A new approach to causal inference in mortality studies. . . , *Mathematical Modelling*, 1986.

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

Tian and Pearl – On the testable implications of causal models with hidden variables, *UAI*, 2002.

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

# Causal Coherence of mDAGs

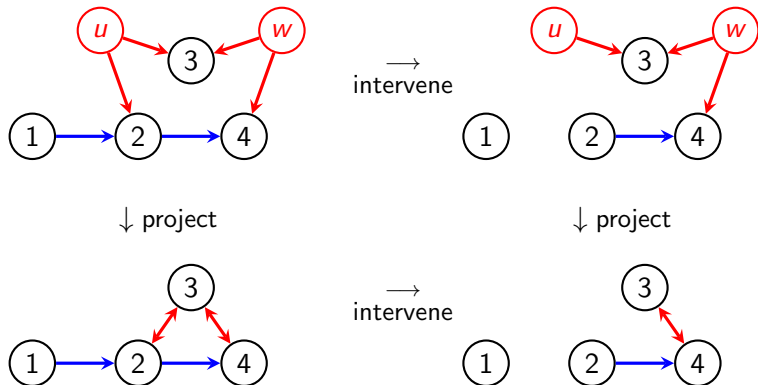
If  $P$  is represented by a DAG in a causally interpreted way, then intervening on some set of nodes  $C \subseteq V$  can be represented by deleting incoming edges to  $C$  in  $\mathcal{G}$ . Call that graph  $\mathcal{G}^{\bar{C}}$

Theorem (Evans, 2015)

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

# Causal Coherence

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

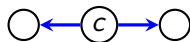


# d-Separation

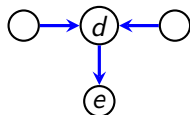
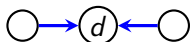
A **path** is a sequence of edges in the graph; vertices may not be repeated.

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  $C$ :

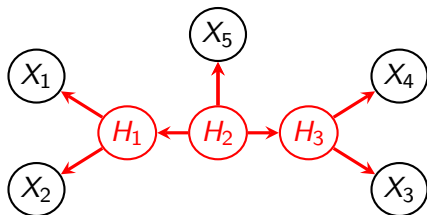


Two vertices  $v$  and  $w$  are **d-separated** given  $C \subseteq V \setminus \{v, w\}$  if **all** paths are blocked.

# 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)