# Graphs for margins of Bayesian networks.

Robin J. Evans
www.stats.ox.ac.uk/∼evans

Department of Statistics, University of Oxford

ERCIM, Pisa
December 2014

# Correlation does not imply causation

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

**Mail**Online

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

sleepmedicine


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."

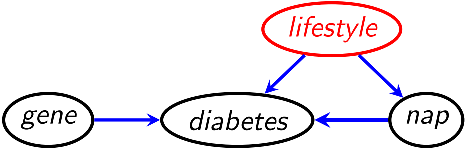

Abstract

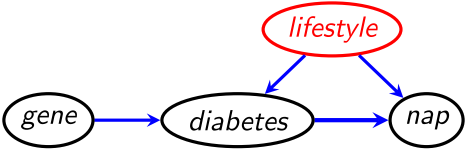
Objectives

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



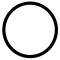




In order to compare the models, we need to understand in what ways causal models will differ, both:

- observationally;
- under interventions.

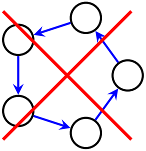
# Outline

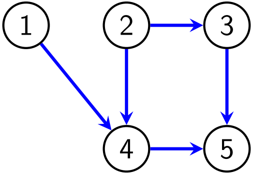
# Directed Acyclic Graphs

vertices 

edges 

no directed cycles






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

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

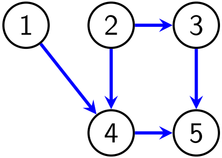
If $w \to \cdots \to v$ then $w$ is a **ancestor** of $v$

# DAG Models

vertex $\iff$ random variable


$X_a$

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




$$P : X_i \perp\!\!\!\perp X_{\mathrm{pre}(i)} \mid X_{\mathrm{pa}(i)} [P]$$

ordered local Markov property

So in example above:

$$X_2 \perp\!\!\!\perp X_1 \qquad X_3 \perp\!\!\!\perp X_1 \mid X_2$$

$$X_4 \perp\!\!\!\perp X_3 \mid X_1, X_2 \qquad X_5 \perp\!\!\!\perp X_1, X_2 \mid X_3, X_4$$

# Global Markov Property

$P$ satisfies the **global Markov property** if for all sets $A, B, C$,

$$A \text{ d-separated from } B \text{ by } C \implies X_A \perp\!\!\!\perp X_B \mid X_C \, [P].$$

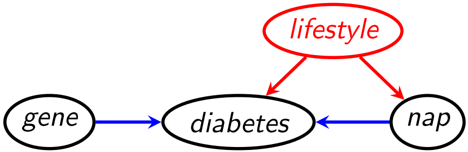
Theorem (Lauritzen et al, 1990)

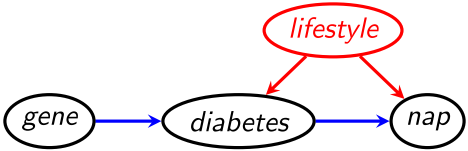
$P$ satisfies the global Markov property if and only if it satisfies the ordered local Markov property.

**Point:** the model can be defined in terms of 'paths of information'.

# Causal Interventions

If we interpret the DAG as representing structural assumptions, then if we intervene on a node, the graph of the resulting model is just locally altered:



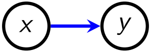

so if we force people to stop napping...

# Latent Projection

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

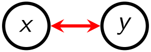
Whenever there is a path of the form


add



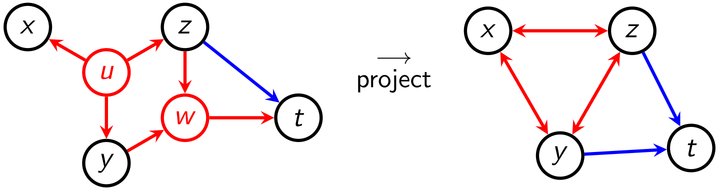

Whenever there is a path of the form


add




Then remove the latent variables $U$ from the graph.

# ADMGs



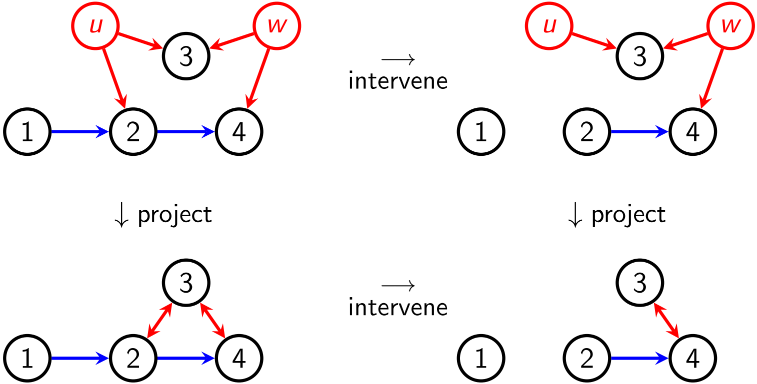

Latent projection leads to an **acyclic directed mixed graph** (ADMG) (equivalent to summary graph without undirected edges).

Can read off independences with d/m-separation. Like an ancestral graph, these are precisely observable independences from the original DAG. See Richardson (2003) for more details.

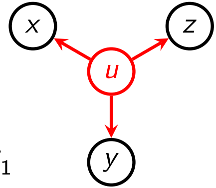
In addition, can see that projection preserves the causal structure; Verma and Pearl (1992).

# Causal Coherence

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

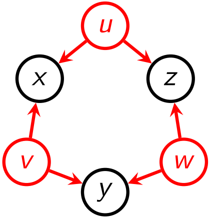


# ADMGs are not sufficient

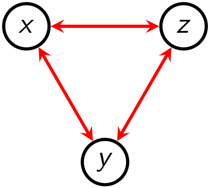



$\mathcal{G}_1$

$\mathcal{G}_2$




These have the same latent projection:




But the model over $(x, y, z)$ in $\mathcal{G}_2$ is not saturated. Still true if we dichotomize.

# The Problem

- Verma and Pearl's latent projection only uses paths, which are inherently 'pairwise';
- the paths are objects suited to conditional independence, but not all constraints on margins are conditional independences;
- ADMGs are not a sufficiently rich class of graphs to capture the different models one can obtain.

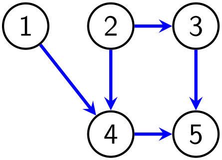
# Structural Equation Model View

There is another way to think about DAG models (e.g. Lauritzen et al, 1990).

$P \in \mathcal{M}(\mathcal{G})$ iff there exist functions $f_i$ and independent variables $E_i$ such that recursively setting

$$X_i = f_i(X_{\mathrm{pa}(i)}, E_i)$$

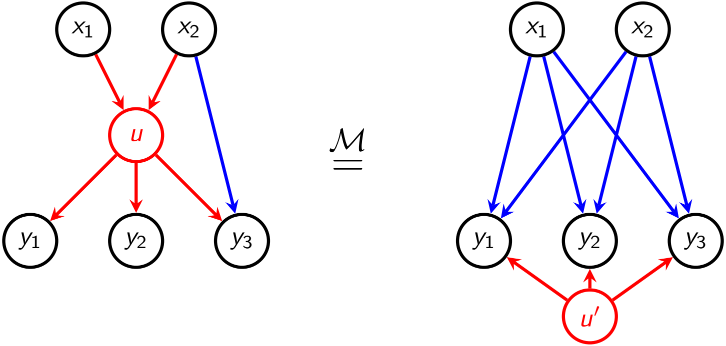
gives $X_V$ the distribution $P$.




$$\begin{aligned} X_1 &= f_1(E_1) \\ X_2 &= f_2(E_2) \\ X_3 &= f_3(X_2, E_3) \\ X_4 &= f_4(X_1, X_2, E_4) \\ X_5 &= f_5(X_3, X_4, E_5). \end{aligned}$$

# Simplifications

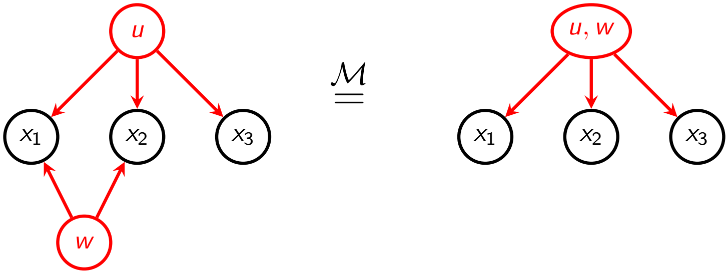
**Simplification 1.** WLOG latent vertices have no parents.




(Of course, this is not true if we assume a specific state-space: e.g. phylogenetic model)

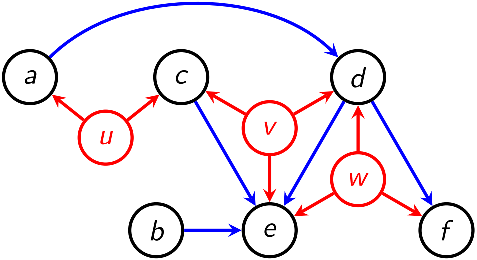
# Simplifications

**Simplification 2.** If $u, w$ are latent with $\mathrm{ch}_{\mathcal{G}}(w) \subseteq \mathrm{ch}_{\mathcal{G}}(u)$, then we don't need $w$.



# mDAGs

So we only need to consider models like this:



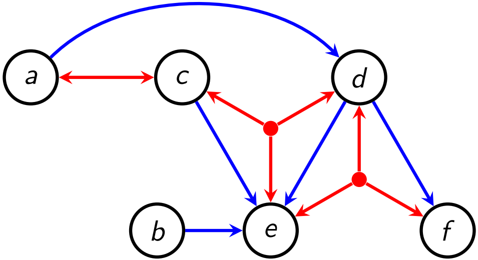

...which we represent with a hyper-graph called an **mDAG**.

Formally, an mDAG on $V$ is a DAG (in blue), together with some inclusion maximal collection subsets of size at least 2 (red).

Going backwards and replacing bidirected edges with latents gives us the **canonical DAG** $\bar{\mathcal{G}}$.

# mDAGs

So we only need to consider models like this:

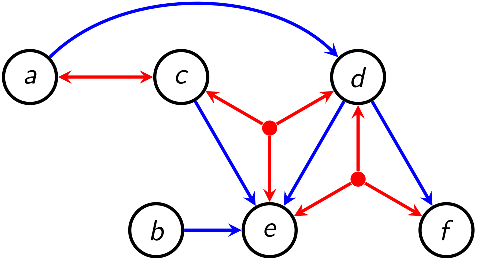



...which we represent with a hyper-graph called an **mDAG**.

Formally, an mDAG on $V$ is a DAG (in blue), together with some inclusion maximal collection subsets of size at least 2 (red).

Going backwards and replacing bidirected edges with latents gives us the **canonical DAG** $\bar{\mathcal{G}}$.

# Markov Properties




Given an mDAG $\mathcal{G}$ and distribution $P$, say $P$ is in the **complete model** for $\mathcal{G}$, or $P \in \mathcal{M}(\mathcal{G})$ if it is the margin of some distribution in the model for the canonical DAG $\mathcal{M}(\bar{\mathcal{G}})$.

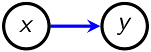
There are other (weaker) properties Shpitser et al. (2014).

# Latent Projection for mDAGs

For mDAG $\mathcal{G}$ and subset of vertices $O$, form latent projection $\mathfrak{p}(\mathcal{G}, O)$ by:

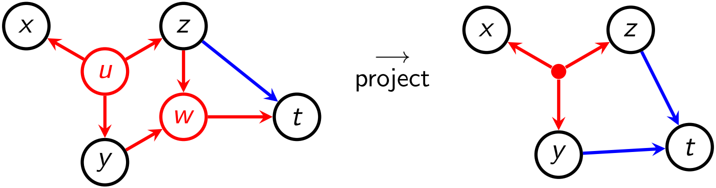
Whenever there is a path of the form


add




Whenever there is a maximal set $B = \{x_1, x_2, \ldots, x_k\}$ such that these variables share a hidden common cause, add hyper-edge $B$.

Then remove the latent variables $U$ from the graph.



# Results

The mDAG latent projection preserves the distinction between models.

**Theorem (Evans, 2014)**

If $\mathfrak{p}(\mathcal{G}, O) = \mathfrak{p}(\mathcal{G}', O)$ then the models induced by $\mathcal{M}(\mathcal{G})$ and $\mathcal{M}(\mathcal{G}')$ on the margin $O$ are the same.

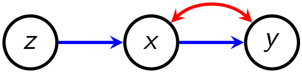
So the problem which arises with ADMGs never occurs for mDAGs.

**Theorem (Evans, 2014)**

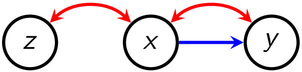
If $C \subseteq O$ then $\mathfrak{p}(\mathcal{G}_{\overline{C}}, O) = \mathfrak{p}(\mathcal{G}, O)_{\overline{C}}$; i.e. the projection respects causal interventions.

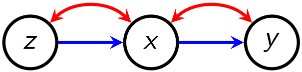
# Instrumental Variables

The Instrumental Variables model assumes causally exogenous variable $z$ affects the treatment $x$.



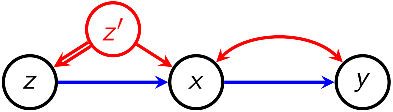

But it's well known that this is observationally indistinguishable from a hidden common cause for $x$ and $z$ (e.g. Didelez and Sheehan, 2007).



# Instrumental Variables

To see this, imagine $z$ is an exact copy of $z'$.




Doesn't really matter whether $x$ gets information from $z$ or $z'$.

Very hard to see this equivalence with conditional independence.

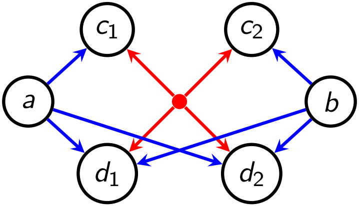
# Instrument Generalisation

Let $\mathcal{G}$ have bidirected edge $B = C \dot{\cup} D$ with:

(i) every $c \in C$ contained in no other bidirected edge;

(ii) $\text{pa}_{\mathcal{G}}(d) \supseteq \text{pa}_{\mathcal{G}}(C)$ for each $d \in D$.

Can 'split' $B$ into $C$ and $D$ and add edges $c \to d$ where necessary.

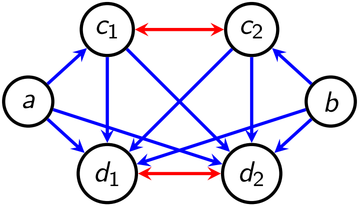


# Instrument Generalisation

Let $\mathcal{G}$ have bidirected edge $B = C \dot{\cup} D$ with:

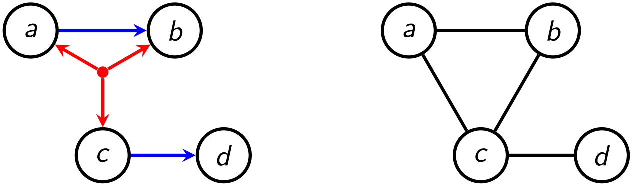
- (i) every $c \in C$ contained in no other bidirected edge;
- (ii) $\text{pa}_{\mathcal{G}}(d) \supseteq \text{pa}_{\mathcal{G}}(C)$ for each $d \in D$.

Can 'split' $B$ into $C$ and $D$ and add edges $c \to d$ where necessary.



# Skeletons

Define the **skeleton** of two mDAGs as the undirected graph with $v - w$ whenever $v$ and $w$ are contained in some edge together.



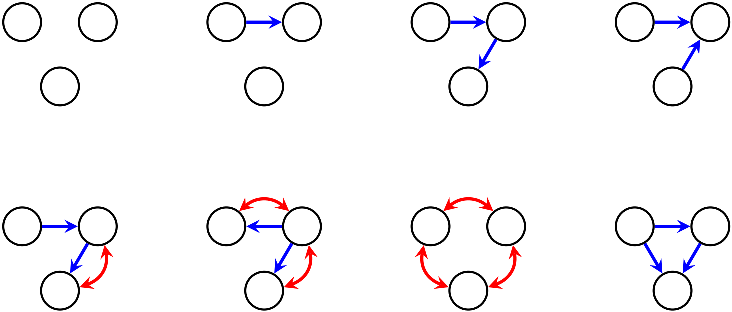

**Theorem**

mDAGs with different skeletons induce different models in general.

(Consequence of Theorem 4.2 of Evans, 2012)

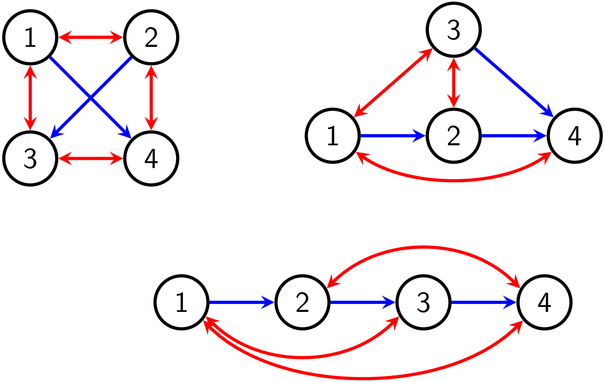
# Equivalence on Three Variables

Combining the previous results, there are 8 Markov equivalence classes on three variables.



# But Not on Four!

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



# Summary

We have seen that:

- graphs with ‘ordinary’ edges can give a causally coherent representation of marginal models;
- **but**: ordinary mixed graphs are not rich enough to represent all models;
- mDAGs provide the most general necessary framework for representing causal DAGs under marginalization;
- general Markov equivalence in this class is hard, but we’re getting there!

**Thank you!**

# Main References


Didelez and Sheehan. Mendelian Randomisation: Why Epidemiology needs a Formal Language for Causality, *SMMR*, 2007.

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

Evans. Margins of Bayesian networks. Draft, 2014.

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

Pearl. *Causality*. Cambridge, 2000.

Pearl and Verma. A statistical semantics for causation, *Stats and Computing*, 1992.

Richardson. Markov properties for ADMGs, *SJS*, 2003.

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

Shpitser et al. Introduction to nested Markov models. *Behaviormetrika*, 2014.

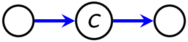
Verma. Invariant properties of causal models, *Tech Report R-134*, UCLA Cognitive Systems Laboratory, 1991.

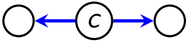
# d-Separation

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

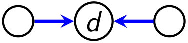
A path from $a$ to $b$ is **blocked** by $C \subseteq V \setminus \{a, b\}$ if either

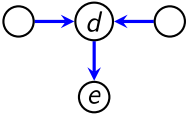
(i) any non-collider is in $C$:



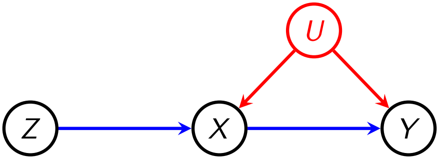

(ii) or any collider is not in $C$, nor has descendants in $C$:




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

# Inequality Results




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

$$\text{Let } p^*(x, y \mid z) \equiv \int p(u)\, p(x \mid z, u) \cdot p(y \mid x = 0, u)\, du$$

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

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

This 'compatibility' requirement turns out to place an inequality restriction on $p$:

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

# Inequality Results

Generalizing this argument, we find a rich theory of results on inequalities (Evans, 2012).

However these results are **not exhaustive**!
Finding **all** inequality constraints in marginal models is probably an NP hard problem.

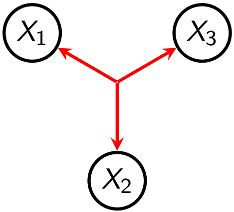
Additionally:

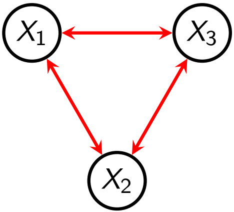
- fitting models with inequality constraints is not trivial;
- the usual asymptotic results do not necessarily apply.

Maybe the nested model is a good compromise!

# ADMGs are not sufficient

In general we need to distinguish between $\{1, 2, 3\}$ and $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$.




The model on the right is not saturated. Still true if we dichotomize.

# ADMGs are not sufficient

**Lemma**

Let $\mathcal{F}$, $\mathcal{G}$, $\mathcal{H}$ be mutually independent $\sigma$-algebrae (so that $\mathcal{F} \perp\!\!\!\perp \mathcal{G} \vee \mathcal{H}$ and so on), and let $X$, $Y$ and $Z$ be random variables such that

1. (i) $X$ is $\mathcal{F} \vee \mathcal{G}$-measureable;
2. (ii) $Y$ is $\mathcal{G} \vee \mathcal{H}$-measureable;
3. (iii) $Z$ is $\mathcal{F} \vee \mathcal{H}$-measureable.

Then $P(X = Y = Z) > 1 - \epsilon$ implies

$$\operatorname{Var} X < 3\epsilon.$$

# Causal Equivalence

The two mDAGs below are Markov equivalent, and lead to the same graph under any ordinary causal intervention.