Probabilistic Causal Models

A Short Introduction

Robin J. Evans
www.stat.washington.edu/∼rje42

ACMS Seminar, University of Washington

24th February 2011

Acknowledgements

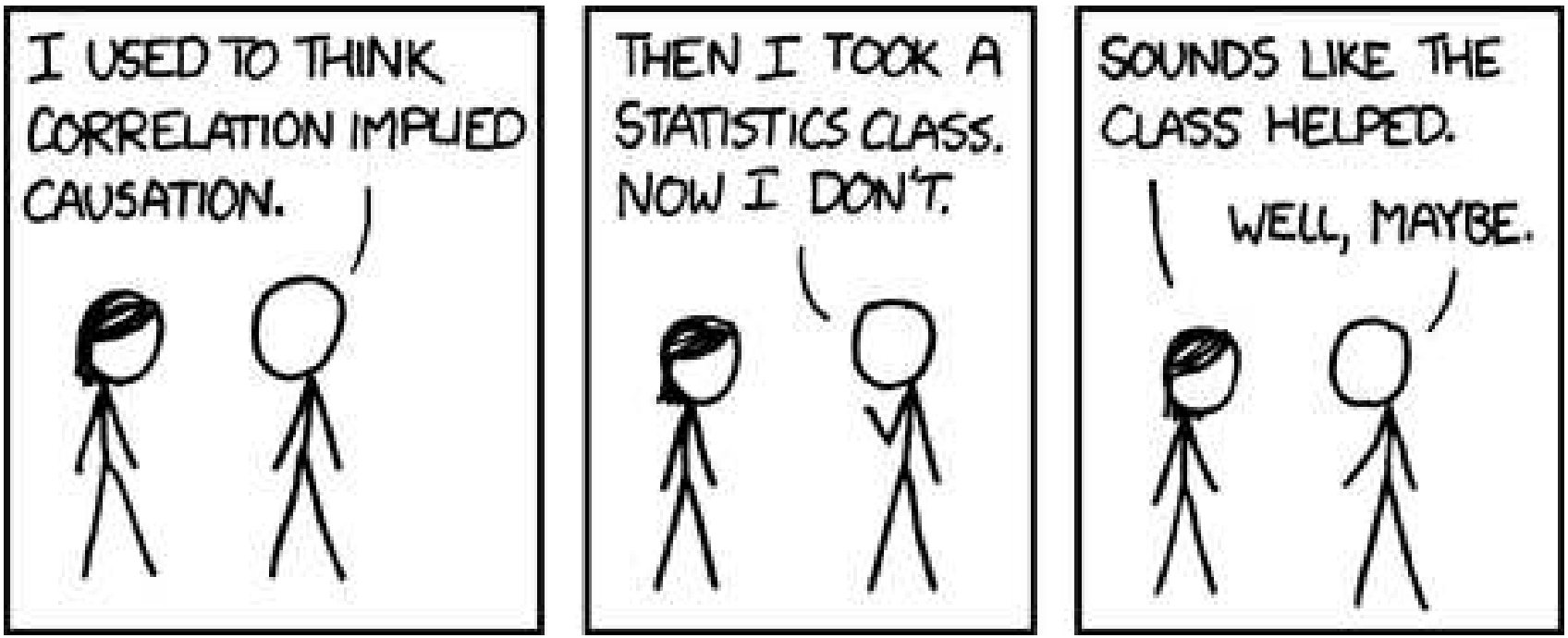
This work is joint with Thomas Richardson.

Thanks also to James Robins, Antonio Forcina and Tamás Rudas.

Outline

1. Introduction
2. Independence
3. ADMGs
4. Models and Parametrizations
5. Conclusions

The Problem of Inference



www.xkcd.com/552/

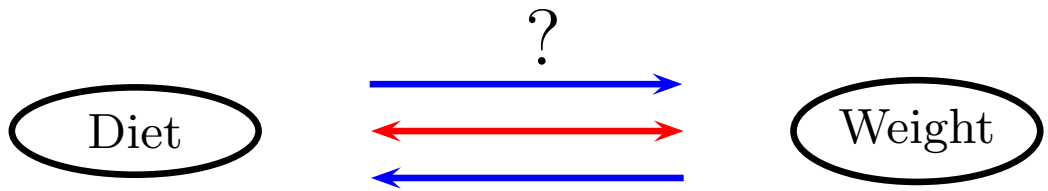
Causation vs Association

Some examples:

- smoking and lung cancer
- murder rates and ice cream sales
- use of antidepressants and sales of video games.

Randomized Controlled Trials

Scientists, and particularly medical researchers, want to answer causal questions all the time.



Solution: take 1,000 people, *randomly* divide into two groups. Assign one group to a healthy diet.

This breaks any possible confounding or effect of weight on diet. If we still measure any correlation, it must be causal[1].

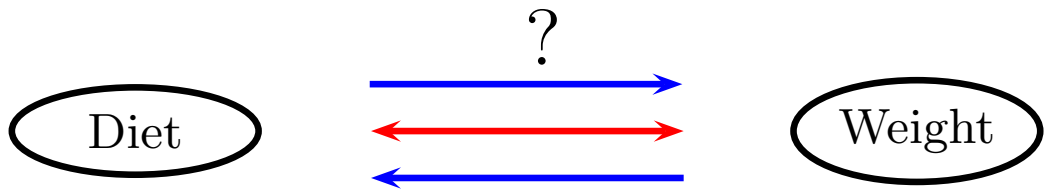
For this reason randomized controlled trials can be thought of as a 'gold standard' for the scientific method[2].

[1] Maybe.

[2] Or at least I think so.

Randomized Controlled Trials

Scientists, and particularly medical researchers, want to answer causal questions all the time.



Solution: take 1,000 people, *randomly* divide into two groups. Assign one group to a healthy diet.

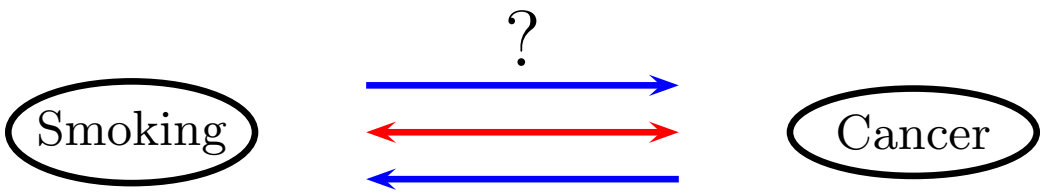
This breaks any possible confounding or effect of weight on diet. If we still measure any correlation, it must be causal[1].

For this reason randomized controlled trials can be thought of as a 'gold standard' for the scientific method[2].

[1] Maybe.

[2] Or at least I think so.

Gold Standards are Expensive...



In this case, forcing people to smoke is unethical.

In many examples, randomized trials are

(i) too expensive,

(ii) impractical,

(iii) unethical.

Thus we want to get the same information from observational data.

Technical Points

We will introduce three separate but related concepts:

Structural Equation Models (SEMs). This is our model of how causality 'works'.

Independence. SEMs impose independence constraints. This is how we will test different causal models.

Graphs. A coarser model than the SEM, but easier to deal with; they contain *only* independence constraints. They allow us to visualize the independences.

How does Causality work?

We need a model of what causality means in the real world. We use *structural equation models*: let $x_1, x_2, \ldots, x_k$ be our random variables.

Assume $$x_i = f_i(x_1, \ldots, x_{i-1}, u_i),$$

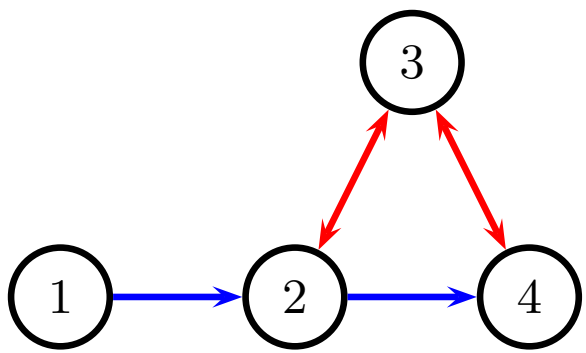
for some functions f_i and noise term u_i. Allow correlation between pairs of noise terms u_i and u_j.

'Nature' takes previous values and unobserved noise, and determines new values.

Example. Suppose that

$$x_1 = f_1(u_1) \qquad x_2 = f_2(x_1, u_2)$$
$$x_3 = f_3(u_3) \qquad x_4 = f_4(x_2, u_4)$$

with u_2, u_3 correlated and u_3, u_4 correlated.



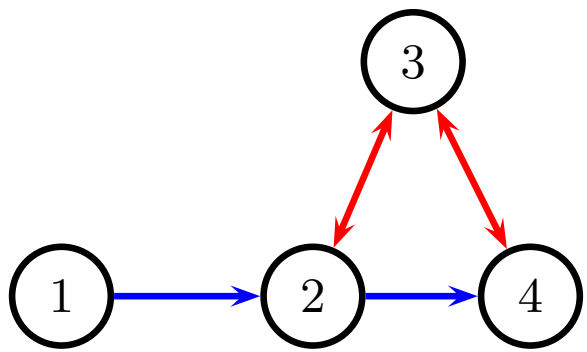
Interventions

This framework allows us to easily incorporate interventions, such as randomized experiments.

Suppose we wish to perform an experiment, by intervening on x_2:

$$x_1 = f_1(u_1) \qquad x_2 = f_2(x_1, u_2)$$
$$x_3 = f_3(u_3) \qquad x_4 = f_4(x_2, u_4)$$

with u_2, u_3 correlated and
u_3, u_4 correlated.



A randomized controlled trial is one kind of intervention.

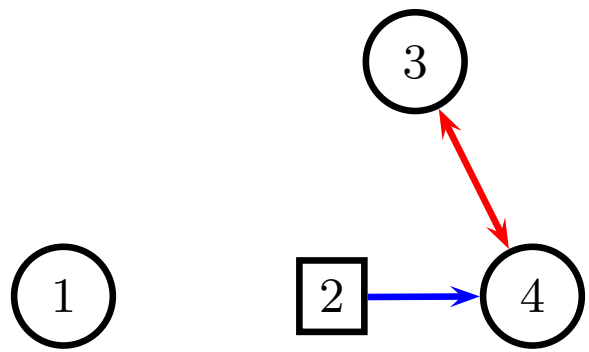
Interventions

This framework allows us to easily incorporate interventions, such as randomized experiments.

Suppose we wish to perform an experiment, by intervening on x_2:

$$x_1 = f_1(u_1) \qquad x_2 = x_2'$$
$$x_3 = f_3(u_3) \qquad x_4 = f_4(x_2, u_4)$$

with
u_3, u_4 correlated.



A randomized controlled trial is one kind of intervention.

Marginal and Conditional Independence

We say two random variables X and Y are *independent* if the chances of X taking particular values is unaffected by the value of Y:

$$P(X = x \,|\, Y = y) = P(X = x) \qquad \text{for all levels } y.$$

We write $X \perp\!\!\!\perp Y$. Example:

$$P(\text{raining} \,|\, \text{earthquake occurred}) = P(\text{raining}).$$

We say X and Y are *independent conditional upon* Z if X is unaffected by Y for each level z of Z:

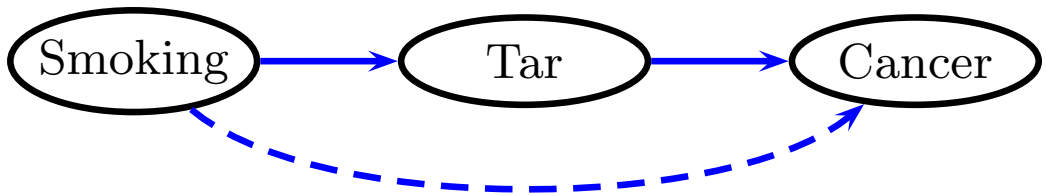
$$P(X = x \,|\, Y = y, Z = z) = P(X = x \,|\, Z = z) \qquad \text{for all levels } y, z.$$

We write $X \perp\!\!\!\perp Y \,|\, Z$. Example:

$$P(\text{ground wet} \,|\, \text{cloudy, rainfall}) = P(\text{ground wet} \,|\, \text{rainfall})$$

Using Independence

How are independence and causality related?



A causal question: does smoking cause cancer *other* than through putting tar in the lungs?

Suppose we only have observational data. How would it look if the dashed arrow were not present?

Then relationship between Smoking and Cancer entirely mediated by Tar; thus we would observe that

$$\text{Smoking} \perp\!\!\!\perp \text{Cancer} \,|\, \text{Tar}.$$

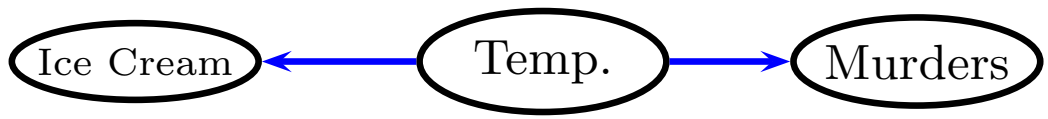
Using Independence

Is temperature the confounder for Ice Cream sales and Murder Rates?

Then we might expect that

$$\text{Ice Cream} \perp\!\!\!\perp \text{Murder} \mid \text{Temperature}.$$

Using Independence



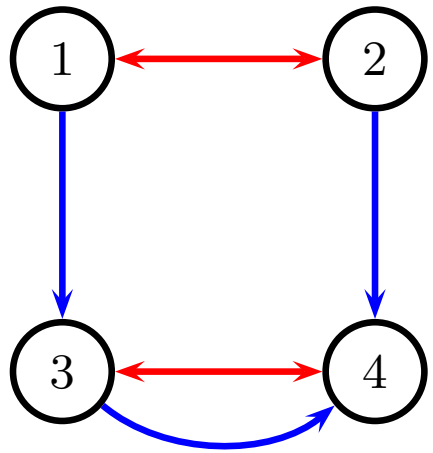
Is temperature the confounder for Ice Cream sales and Murder Rates?

Then we might expect that

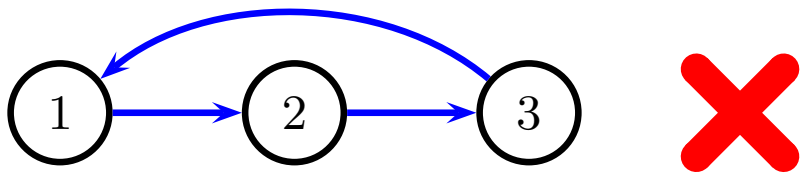
$$\text{Ice Cream} \perp\!\!\!\perp \text{Murder} \,|\, \text{Temperature}.$$

Acyclic Directed Mixed Graphs

We work with directed mixed graphs, which have 2 types of edges (directed and bidirected).



No directed cycles:



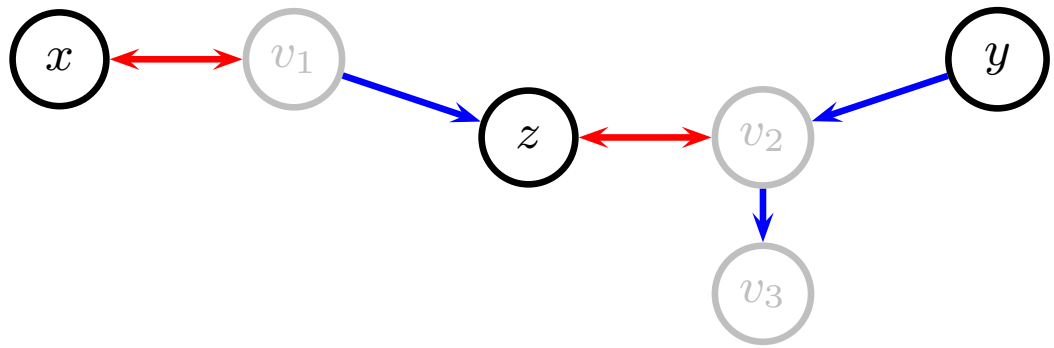
We call this class *Acyclic Directed Mixed Graphs* (ADMGs).

Without bidirected edges we get a *Directed Acyclic Graph* (DAG).

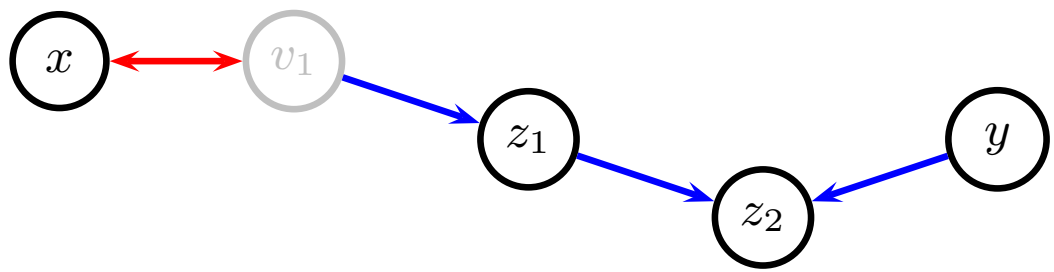
m-separation

Two vertices x and y are m-separated by a set Z if all paths from x to y are blocked by Z.

Either: at least one collider is not conditioned upon, and nor are any of its descendants:



Or: at least one non-collider is conditioned upon:



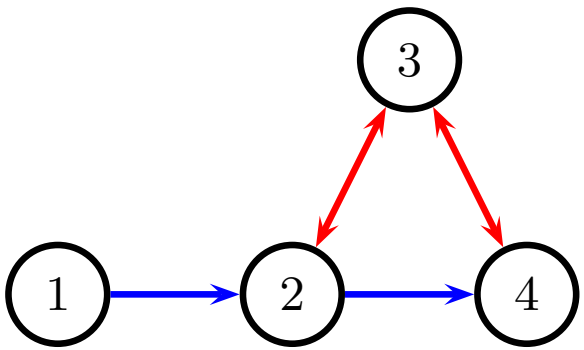
m-separation extends to sets X and Y if every $x \in X$ and $y \in Y$ are m-separated.

Global Markov Property

Let P be a distribution over the vertices of $\mathcal{G}$. The global Markov property (GMP) for ADMGs states that

$$X \text{ m-separated from } Y \text{ by } Z \quad \Longrightarrow \quad X \perp\!\!\!\perp Y \mid Z \; [P]$$

In fact, these independences hold in the structural equation model.

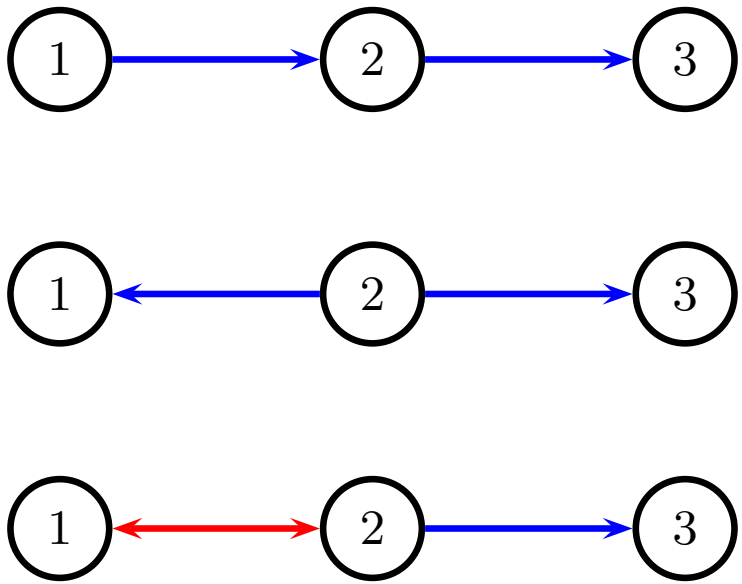


Here $1 \perp\!\!\!\perp 4 \mid 2$ and $1 \perp\!\!\!\perp 3$.

ADMGs represent a coarser model class than the corresponding structural equation models. In other words, SEMs may impose additional structure (see appendix).

Markov Equivalence

Two or more different ADMGs can represent the same conditional independences:



We cannot distinguish between these models observationally.

Set Up

We might be interested in some model

$$\mathcal{M}_{\mathcal{G}} = \{P \mid P \text{ obeys GMP for } \mathcal{G}\}.$$

For simplicity we assume the probability space is binary (two possible outcomes for each variable) and positive.

To fit a model like this to data, we need to characterize it explicitly. That is, need a smooth, bijective map

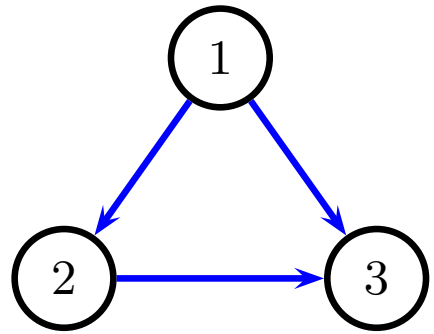
$$f : \mathcal{M}_{\mathcal{G}} \to A \subseteq \mathbb{R}^p$$

to some nice set A.

This is called a *parametrization*. There are many different choices, but p is fixed.

Saturated Model

Consider the model



This imposes no restrictions of any kind (*saturated*). The model is characterized by the 8 probabilities

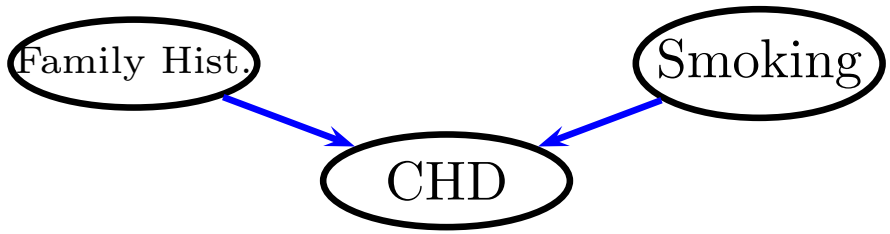
$$\begin{array}{cccc} P_{000} & P_{100} & P_{010} & P_{110} \\ P_{001} & P_{101} & P_{011} & P_{111}, \end{array}$$

where $P_{ijk} \equiv P(X_1 = i, X_2 = j, X_3 = k)$. We know they sum to 1, so picking any 7 of these is a parametrization of the model.

There are other parametrizations we could have used, but we always have 7 parameters.

Heart Disease Example

Imagine a simple system explaining coronary heart disease.



A priori, $P(\text{Family History}) = 0.4$ and $P(\text{Smoking}) = 0.25$, independently.
$P(\text{CHD} \mid \cdot)$ given by table:

		Smoking	
		0	1
FH	0	0.300	0.400
	1	0.350	0.456

The graph gives us Family History $\perp\!\!\!\perp$ Smoking.

These 6 probabilities are a parametrization of the model.

Model Selection

Goal: to select the correct ADMG from a data set.

A simple approach: fit a series of ADMG models to the data, see which one fits best according to some score.

This requires:

(i) a parametrization for the model (Richardson, 2009);

(ii) a fitting algorithm (Evans and Richardson, 2010).

The parametrization uses probabilities, as on the previous slide.

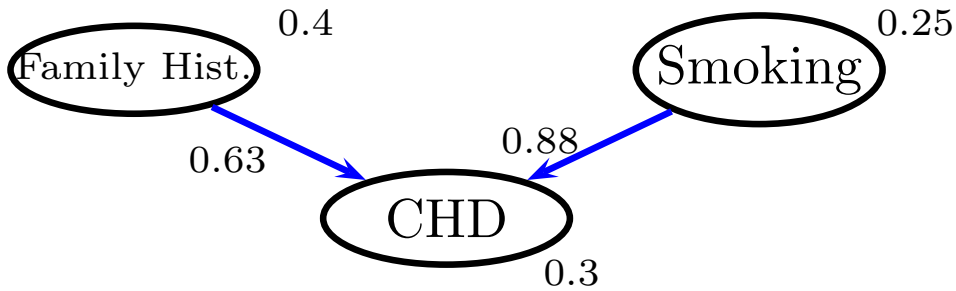
We use the *Bayesian information criterion* (BIC) to score models. This is

$$-2 \log L(P) + p \log n$$

where p is the number of parameters, n is sample size, L is likelihood. Models with more parameters are penalized.

Parametrizations

The parametrization of the heart disease model doesn't allow for no interactions between effects.



This doesn't impose any additional independences, so $\mathcal{G}$ is the smallest correct (graphical) model.

The parametrization in Richardson (2009) makes it hard to impose this extra constraint.

Alternative Parametrizations

In Richardson and Evans (2011) we provide a new parametrization which overcomes some of these difficulties. The DAG case was first given by Rudas et al. (2006).

In the heart disease example, the new parameters have the form

$$\log \frac{P(\text{Heart Disease} = 1 \,|\, \text{FH} = f, \text{S} = s)}{P(\text{Heart Disease} = 0 \,|\, \text{FH} = f, \text{S} = s)}.$$

We also get *variation independence* in some cases, which helps with interpretability as well as some applications.

Application

Maathuis et al. (2010) looked at gene expression data in yeast (5,361 genes). 234 observations from interventional experiments (single gene deletion) – taken to be 'truth'.

Observational data from 63 wild-type cultures. Aim to recover causal paths from just observational data.

The authors use only DAGs (not realistic, but simple and effective). Fit model and look at largest effects.

Of top 50 effects found, 2/3 were validated in interventional data. Good way of providing biologist with places to start looking!

We would like to extend this to ADMG models.

Summary

There is a close connection between graphical models and causal models.

Randomized experiments are the best way of making causal connections.

However, these are sometimes not feasible.

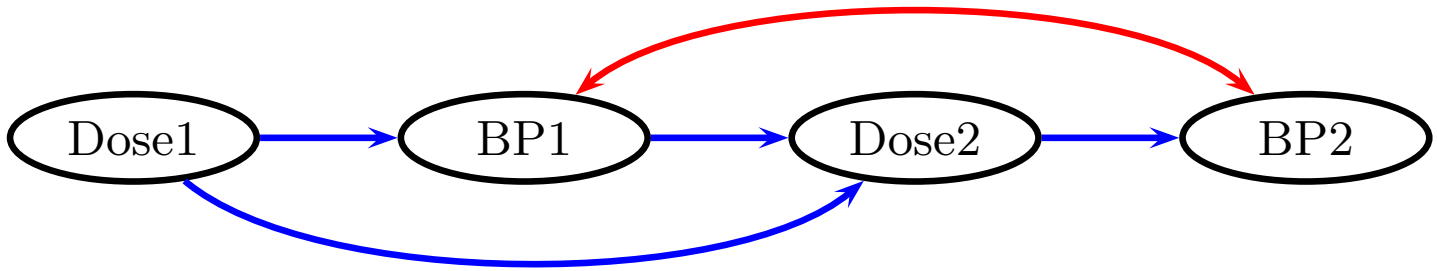
Sometimes a mixture of intuition, prior knowledge and observational data can go a long way.

Some parametrizations are better than others.

Thank you!

The Verma Constraint

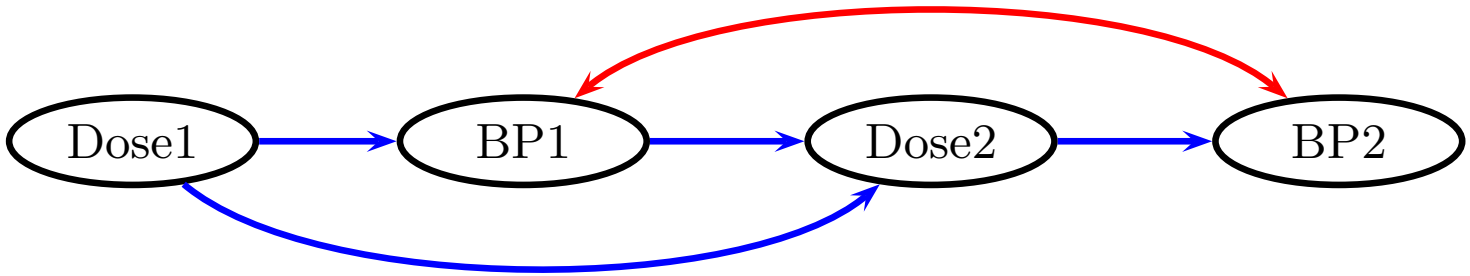
Consider a sequential clinical trial of a blood pressure drug:



No independences hold. However, there is no arrow between Dose1 and BP2. Can we test for the presence / absence of this arrow?

An Intervention

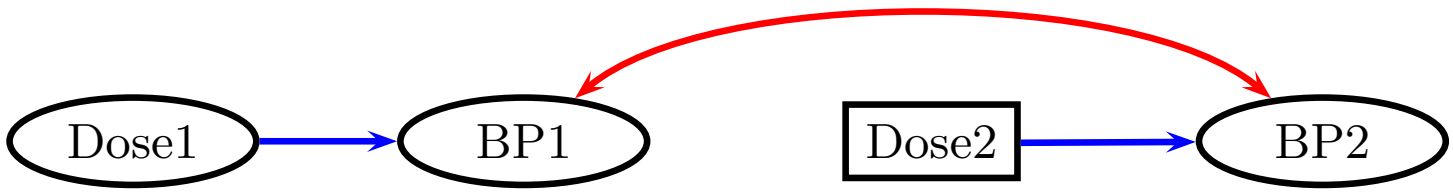
It is clear that if we could perform an experiment to randomize Dose2, we would see whether Dose1 and BP2 were independent.



However it turns out that we can see the effect of this intervention even without performing it!

An Intervention

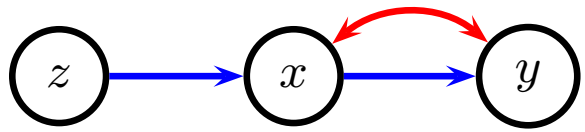
It is clear that if we could perform an experiment to randomize Dose2, we would see whether Dose1 and BP2 were independent.



However it turns out that we can see the effect of this intervention even without performing it!

Instrumental Variables

A well studied model is the instrumental variables model:



This model imposes no independence constraints (or Verma constraints), but there is no edge from z to y.

However, the SEM imposes certain inequality constraints such as:

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

These can be used to test the model (i.e. that there is no edge from z to y).