Causal inference and causal discovery

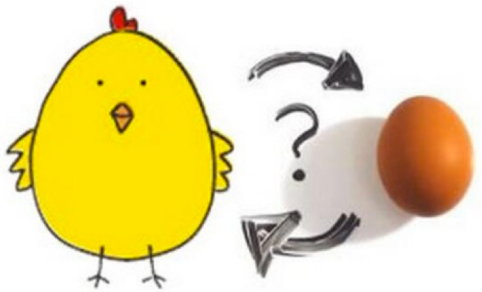
Shikui Tu

Department of Computer Science and Engineering, Shanghai Jiao Tong University

2018-04-12

什么叫因果？

- 如下说法是因果吗？
  - 善有善报，恶有恶报（佛法基本定律）
  - 人在做，天在看，举头三尺有神明，多行不义必自毙
  - 天网恢恢疏而不漏，不是不报时候未到
  - 满屋老鼠跑，必定有窟窿
  - 有风方起浪，无潮水自平
  - 不听老人言，吃亏在眼前
  - 种瓜得瓜，种豆得豆
  - 前因后果
  - 凡事预则立，不预则废



什么叫因果？

- 因果关系（西方哲学）
  - 亚里士多德的四因说：质料因、形式因、动力因、目的因
  - 托马斯·阿奎那排序：目的因>动力因>质料因>形式因。第一动力因是上帝。
  - 休谟：因果关系是是两个事件先后的恒常联结。
  - 马克思哲学原理：因果联系普遍存在，对立统一
- 因果关系（物理）
  - 原因先于效果（作为限制），可以预计信息不能比光速更快，否则，就可能在某个参照系（使用狭义相对论的洛伦兹变换）中观察者可以看到结果先于原因（即违反因果律的假定）。
  - 热力学第二定律确定了时间之箭
- 因果系统（工程）
  - 是指该系统的输出和内部状态取决于当前和以前的输入值。
- 因果关系（法律）
  - 根据法学理论，要认定被告对犯罪或侵权行为承担责任，必须证明存在法律上的因果关系。

Outline

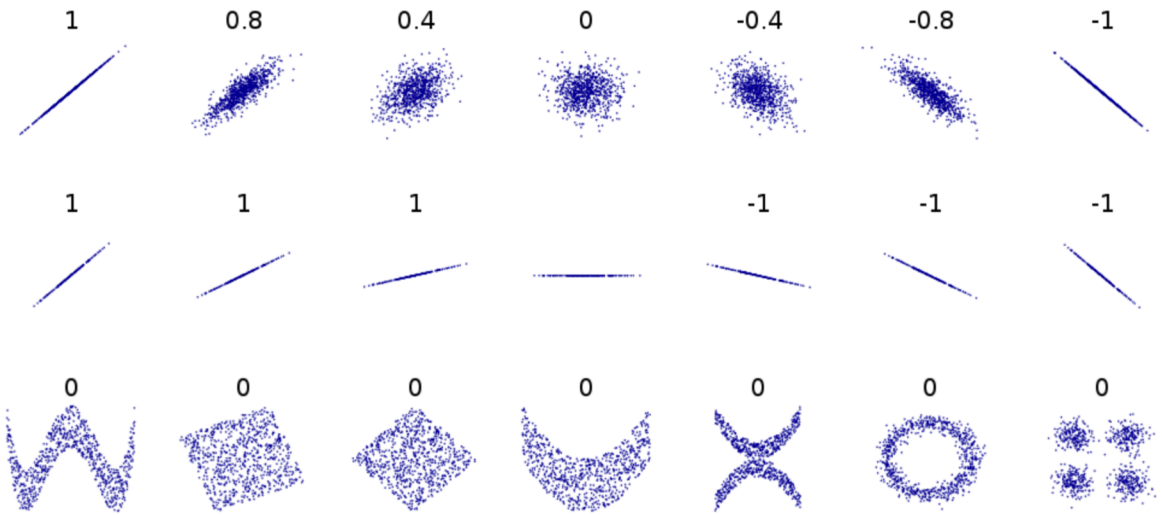
- **Association vs. Causation (Intervention)**
- Structural Causal Model
- Introduce a linear non-Gaussian model for causal discovery

What Is Association?

wiki

- In statistics, an **association** is any relationship between two measured quantities that renders them statistically dependent.

- The term "association" refers broadly to any such relationship, whereas the narrower term "correlation" refers to a linear relationship between two quantities.



Distinction between Associational and Causal Concepts

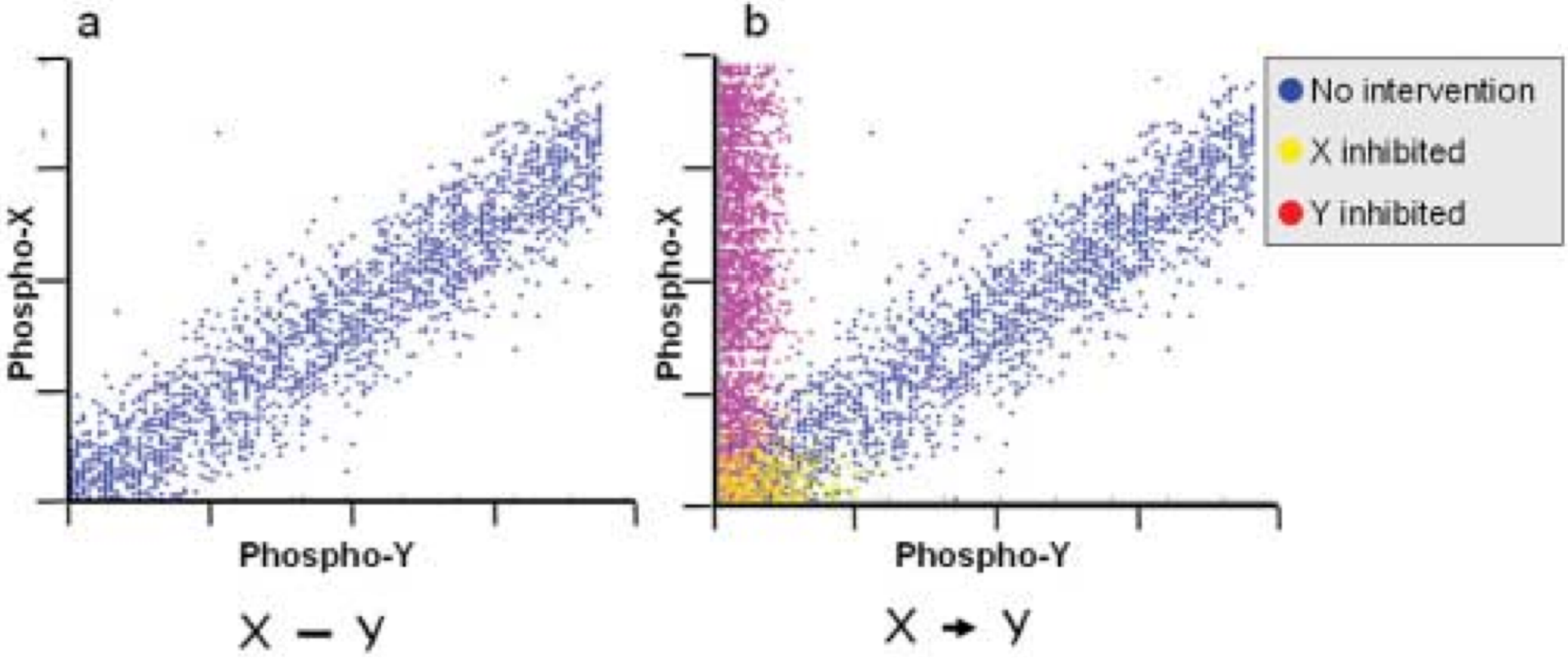
- Pearl ([2003], page 285) Pearl, J. (2003). Statistics and causal inference: A review. Test 12 281–318.
  - "An **associational** concept is any relationship that can be *defined in terms of a joint distribution of observed variables*, and a **causal** concept is any relationship that *cannot be defined from the distribution alone*. . . .
  - Every claim invoking **causal** concepts *must be traced to some premises* that invoke such concepts; it *cannot be inferred or derived from statistical associations alone*."

- About **Judea Pearl**
  - the development of Bayesian networks
  - the 2011 winner of the **ACM Turing Award**, "… for probabilistic and causal reasoning"

A Two-Variable Example

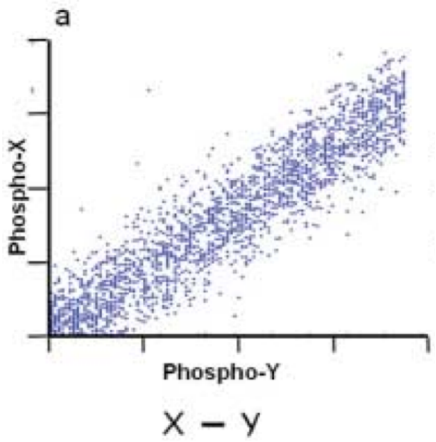
Correlation

Causation



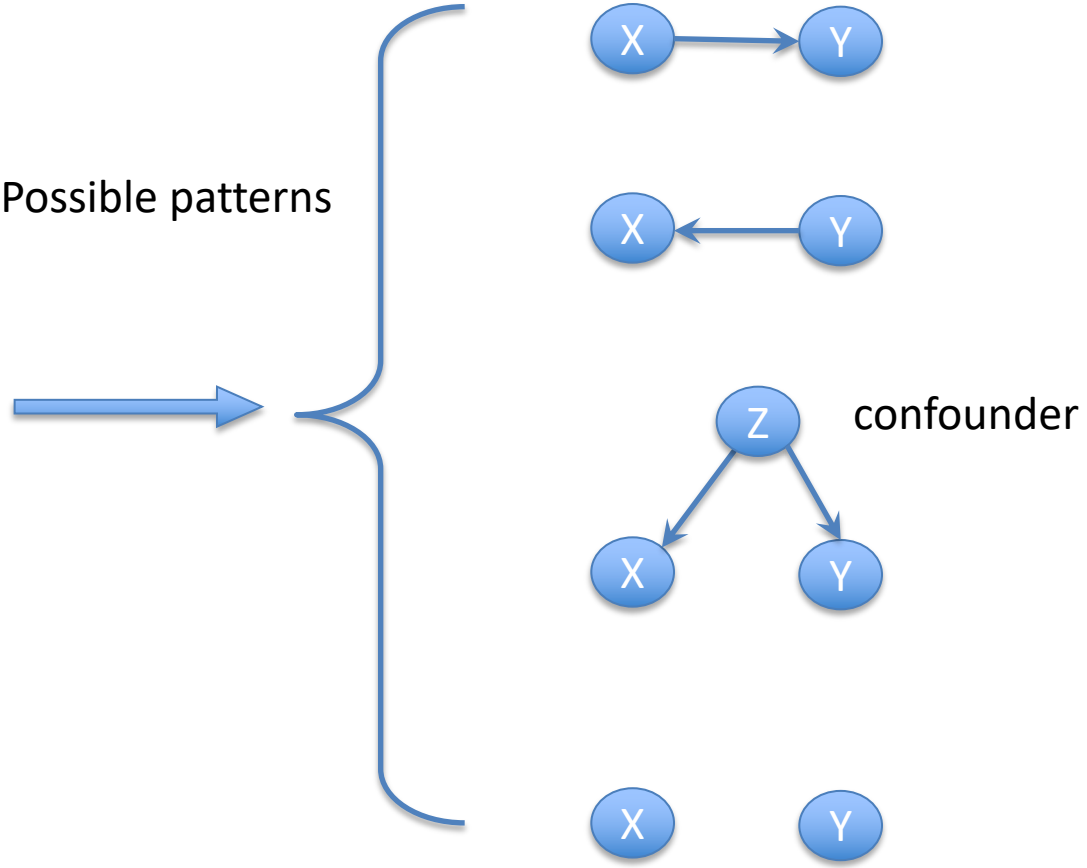
Sachs et. al., vol. 308, Science, 2005

Correlation Does Not Imply Causation



X and Y are correlated.

Possible patterns



No connection

Association vs. Causation

Within a probability distribution

- Correlation
- Regression
- Dependence
- Conditional independence
- Likelihood
- Collapsibility
- Propensity score
- Risk ratio
- Odds ratio
- Marginalization
- Conditionalization
- “Controlling for”
- …

Not distribution alone

- Randomization
- Influence
- Effect
- Confounding
- “holding constant”
- Disturbance
- Error terms
- Structural coefficients
- Spurious correlation
- Faithfulness/stability
- Instrumental variables
- Intervention
- Explanation, attribution
- …

Yule-Simpson's Paradox

-- Can statistics be used to study causality?

[Pearl 2000]

Table 1: Yule-Simpson's Paradox

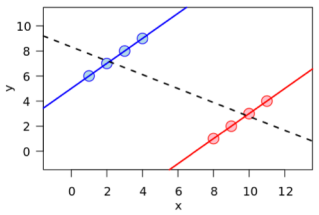
Population			
	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male			
	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female			
	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%

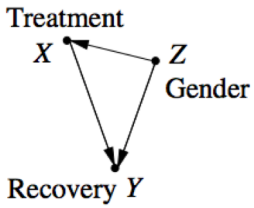
50% > 40%, positive

60% < 70%, negative

20% < 30%, negative

$$\frac{a}{b} < \frac{c}{d}, \frac{a'}{b'} < \frac{c'}{d'}, \frac{a+a'}{b+b'} > \frac{c+c'}{d+d'}$$





Paradox!

Paradox from Real Data

Science, New Series, Vol. 187, No. 4175 (Feb. 7, 1975), pp. 398-404

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

Table 1. Decisions on applications to Graduate Division for fall 1973, by sex of applicant—naive aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. $N = 12{,}763$, $\chi^2 = 110.8$, d.f. $= 1$, $P = 0$ (*18*).

Applicants	Outcome				Difference	
	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	Deny
Men	3738	4704	3460.7	4981.3	277.3	− 277.3
Women	1494	2827	1771.3	2549.7	− 277.3	277.3

8442 male, 44% admitted
4321 female, 35% admitted

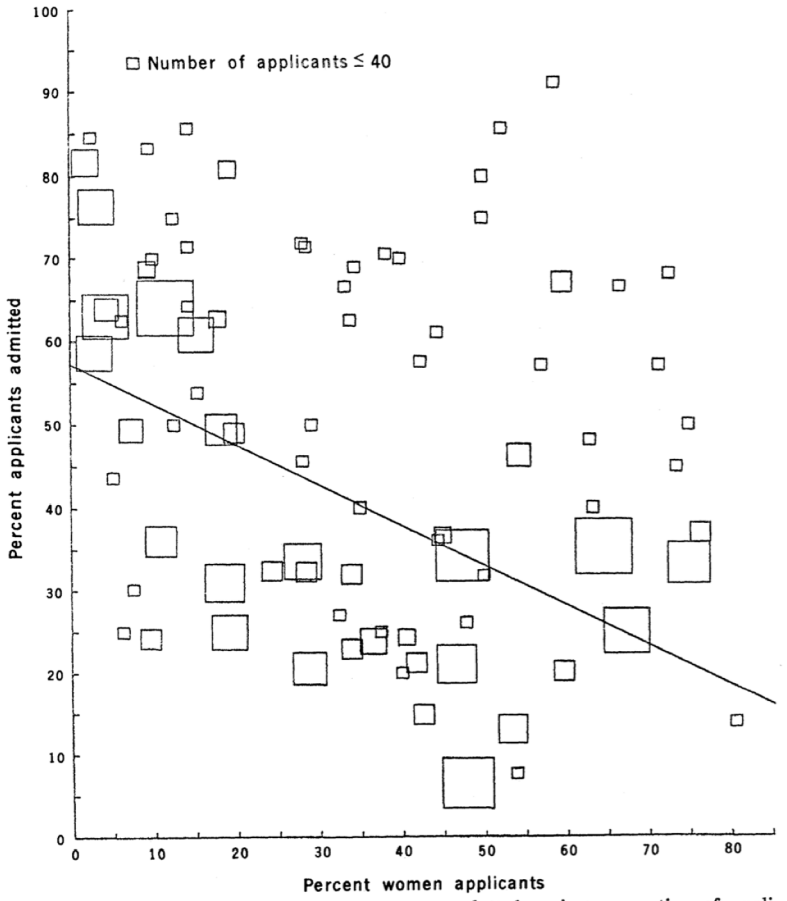


Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

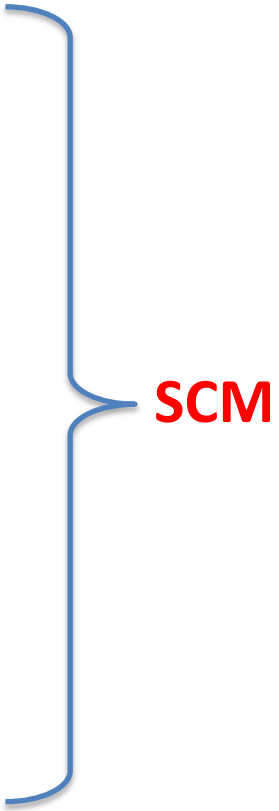
Outline

- Association vs. Causation (Intervention)
- **Structural Causal Model**
- Introduce a linear non-Gaussian model for causal discovery

Structural Causal Model (SCM)

a comprehensive **theory** of causation [Pearl 1995]

- **Potential-outcome framework** of Neyman (1923) and Rubin (1974)
- **Structural equation models** (SEM) used in economics and social science (Goldberger 1973, Duncan 1975)
- **Graphical models** developed for probabilistic reasoning and causal analysis (Pearl 1988, Lauritzen, 1996, Spirtes, Glymour, and Scheines, 2000, Pearl, 2000a).



POTENTIAL-OUTCOME FRAMEWORK

Neyman (1923) and Rubin (1974)

The Counterfactual (Potential Outcomes/Neyman-Rubin) Framework

Neyman (1923) and Rubin (1974)

- Accomplishments:
  - A precise definition of causal effects
  - A formal model of causality against which we can assess the adequacy of various estimators
- Approach:
  - Causal questions are “what if” questions.
  - Extend the logic of randomized experiments to observational data.

Example

- What is the causal effect of *attending catholic school* vs. *public school* on **high school graduation**?

Treatment T = { 0, 1 }
- *1: attending catholic school*
- *0: attending public school*

Each individual has two potential outcomes:
- Y_1: *potential outcome if attending catholic school*
- Y_0: *potential outcome if attending public school*

Individual causal effect (ICE) = $Y_1 - Y_0$

Outcome	Explanation
$Y_0 = Y_1 = 1$ ➔ ICE = 0	Kid would graduate from both catholic and public school. No effect.
$Y_0 = Y_1 = 0$ ➔ ICE = 0	Kid would neither graduate from catholic nor public school. No effect.
$Y_0 = 1, Y_1 = 0$ ➔ ICE = -1	Kid would not graduate from catholic but would graduate from public school. Negative effect.
$Y_0 = 0, Y_1 = 1$ ➔ ICE = 1	Kid would graduate from catholic but would not graduate from public school. Positive effect.

- Generally, either Y_1 or Y_0 can be observed, not both.
- The unobserved outcome is called the "counterfactual" outcome.

Fundamental Problem of Causal Inference

[Holland1986]

- One can never *directly observe* causal effects, because we can never observe both potential outcomes for any individual.

		Outcome	
		Treatment	Control
Assignment	Treatment	$(Y \mid T=1) = (Y_1 \mid T=1)$	???
	Control	???	$(Y \mid T=0) = (Y_0 \mid T=0)$

Missing data

Average causal effect (ACE) = $E[Y_1] - E[Y_0]$ where E[.] ranges over the entire population

population average of the individual level causal effects

Standard estimator S* = E[Y1 | T=1] – E[Y0 | T=0]

where E[.] ranges over the domain of the treatment and control groups, respectively

ACE measures causation ≠ S* measures association

Need assumptions (how the data were generated and collected) to fill in the missing values.

Solve the Problem by Randomized Experiments

Average causal effect (ACE) = $E[Y_1]$ – $E[Y_0]$

Standard estimator S* = E[Y1 | T=1] – E[Y0 | T=0]

if $E[Y_1|T=1] = E[Y_1|T=0] = E[Y_1]$
$E[Y_0|T=1] = E[Y_0|T=0] = E[Y_0]$

ACE = S*

This condition can be achieved by *random assignment* of individuals to the treatment and control group (in a randomized experiment).

What if random assignment is not possible ?

"Ignorability" $(Y_0, Y_1) \perp T$

S* is unbiased and consistent for ACE

"Conditional Ignorability" $(Y_0, Y_1) \perp T \mid X$

But how would we know whether ignorability holds?

STRUCTURAL EQUATION MODELS

by the geneticist Sewall Wright (1921)

Wright's Structural Equation Models

1921

How can one express mathematically the common understanding that symptoms do not cause diseases?

Linear equation:

$$y = \beta x + u_Y$$

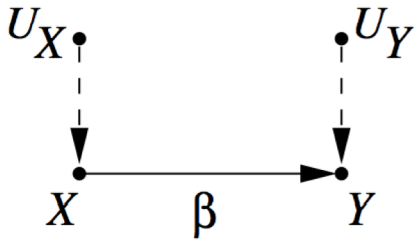
But if we rewrite it as:

$$x = (y - u_Y)/\beta$$

Then symptom influences disease!

Path Diagram:

$$x = u_X$$
$$y = \beta x + u_Y$$



$$Cov(X, Y) = \beta$$

- Encodes the <u>possible</u> existence of (direct) causal influence of X on Y
- Encode the absence of causal influence of Y on X

It is important to note that, in path diagrams, causal assumptions are encoded not in the links but, rather, in the missing links.

X: disease
Y: symptom

u_Y : all factors, other than the disease, that could possibly affect Y when X is held constant

β: path coefficient

U_X, U_Y: "exogenous", represent observed or unobserved background factors that the modeler decides to keep unexplained, that is, factors that influence but are not influenced by the other variables (called "endogenous") in the model.

DIRECTED ACYCLIC GRAPH

Judea Pearl and colleagues

Directed Acyclic Graph (DAG)

- DAGs originate from
  - Structural equation models (1930s+)
  - Social science path models (1960s+)
  - Bayesian networks (1980s+)

- Judea Pearl and colleagues synthesized and generalized these approaches to develop a powerful graphical syntax for causal inference.

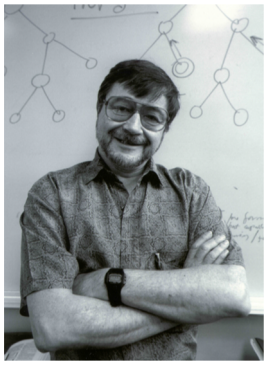
- Compatible with potential-outcome (Neyman-Rubin) framework

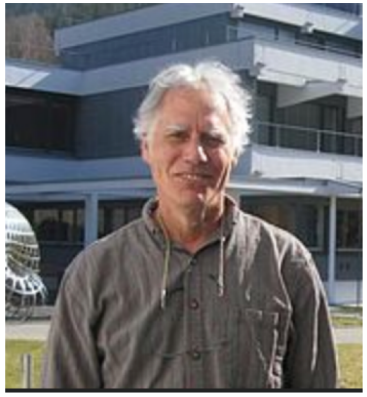
Same concepts, same theorems, different notations

Elwert, Felix. 2013. "Graphical Causal Models." pp. 245-273 in S.L. Morgan (ed.), Handbook of Causal Analysis for Social Research. New York: Sage Publications.

Some key names in the field

- Computer Science:
  Judea Pearl, Jin Tian, Thomas Verma
- Philosophy:
  Peter Spirtes, Clark Glymour, Richard Scheines
- Biostatistics:
  James Robins, Sander Greenland, Tyler VanderWeele, Miguel Hernan





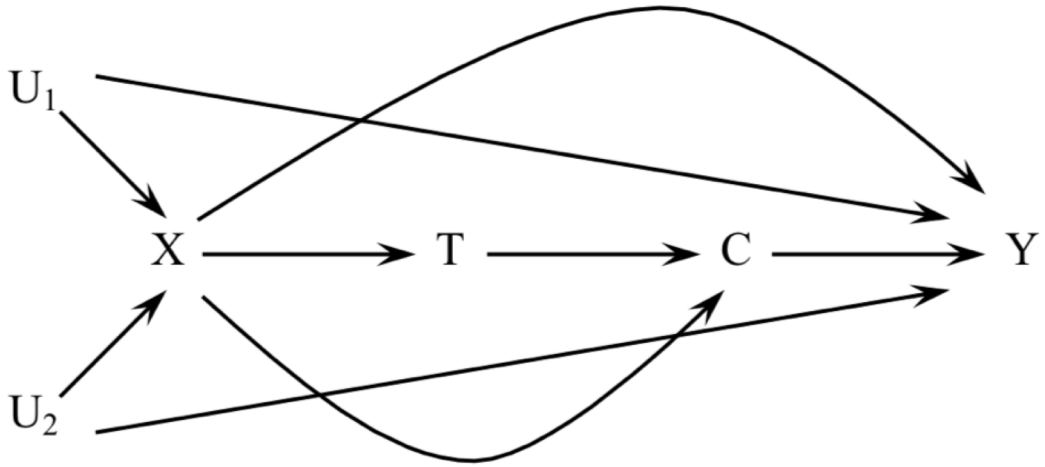
Identification: separating causal from non-causal associations

- Conditional ignorability
  - The total causal effect of T on Y is identifiable if

$$(Y_0, Y_1) \perp T \mid X$$

- "What control variables, X, must be included —and which variables mustn't be included— in the analysis to achieve identification? "

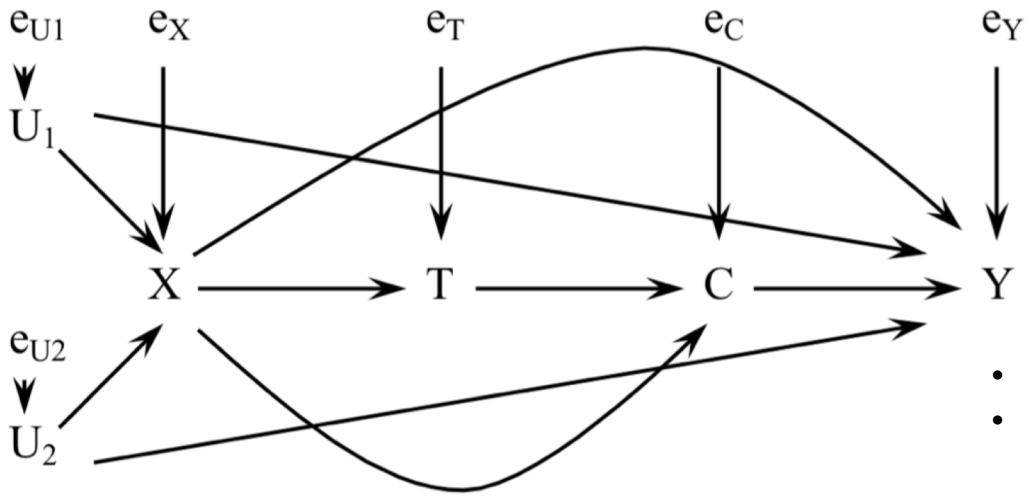
DAGs Encode Causal Knowledge



Arcs	Represent
Directed (presence)	Possible direct causal effects e.g., C may or may not directly cause Y
Missing (absence)	Sharp nulls of no-effect e.g., U_2 does not directly cause T

Only missing arcs encode causal assumptions, whereas directed arcs represent ignorance!

DAGs as SEM



usually omit the set of independent "error term" {e}

- Qualitative
- Population-level representation of data

f_v: some arbitrary deterministic function;
e_V: stochastic (independent) error

Each variable, V, is generated by $V = f_V(pa(V), e_V)$

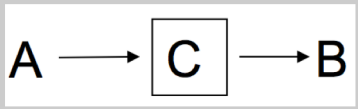
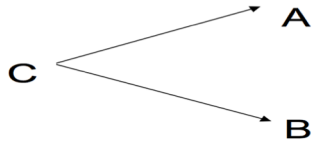
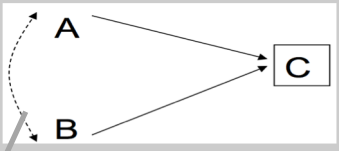
$$U_1 = f_{U1}(e_{U1})$$
$$U_2 = f_{U2}(e_{U2})$$
$$X = f_X(U_1, U_2, e_X)$$
$$T = f_T(X, e_T)$$
$$C = f_C(X, T, e_C)$$
$$Y = f_Y(U_1, U_2, X, C, e_Y)$$

Deriving Testable Implications of A DAG

Causal assumptions → (Encoded by) DAG → (Deduce) All associations in the system

Observations → testable

marginal and conditional dependences and independences

Bias in estimating the causal effect of A on B	Three elements	Sources of associations between A and B
Overcontrol: (conditional on C) intercepting the causal pathway		Direct and Indirect causation: $A \not\perp\!\!\!\perp B$ and $A \perp B \mid C$
Confounding bias: failure to condition on a common cause		Common cause confounding: $A \not\perp\!\!\!\perp B$ and $A \perp B \mid C$
Endogenous selection bias: mistaken conditioning on a common effect		Conditioning on a common effect ("collider"): Selection $A \perp B$ and $A \not\perp B \mid C$

Non-causal (spurious) association

[Elwert 2013]

d-separation

[Pearl 1988]

- A set S of nodes is said to block a path p if either
  1) p contains at least one arrow-emitting node that is in S, or
  2) p contains at least one collision node that is outside S and has no descendant in S.
- If S blocks all paths from X to Y , it is said to "d-separate X and Y," and then, X and Y are independent given S, written $X \perp Y | S$.

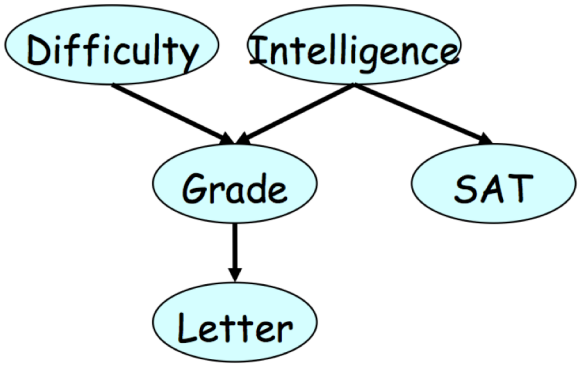
Active Trails

$d\text{-}sep_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$ ⟷ No active trails in G between X and Y given Z

- $X \rightarrow Y$
- $X \leftarrow Y$
- $X \rightarrow W \rightarrow Y$
- $X \leftarrow W \leftarrow Y$
- $X \leftarrow W \rightarrow Y$
- $X \rightarrow W \leftarrow Y$

blocking and the "flow" of association

$W \notin Z$	$W \in Z$
✓	✗
✓	✗
✓	✗
✗	✓



Daphne Koller

v-structure (W is a collider), e.g., dead battery --> **car won't start** <-- no gas

[Pearl (1988)]

Berkson's paradox (1946):

Given two independent events, if you only consider outcomes where at least one occurs, then they become negatively dependent.

Outline

- Association vs. Causation (Intervention)
- Structural Causal Model
- Introduce **a linear non-Gaussian model for causal discovery**

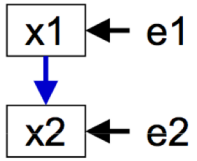
Linear Structural Equation Models

[Wright, 1921; Bollen, 1989]

- Use linear SEM to model the data generation process

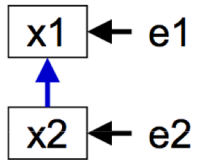
$x_1 = e_1$

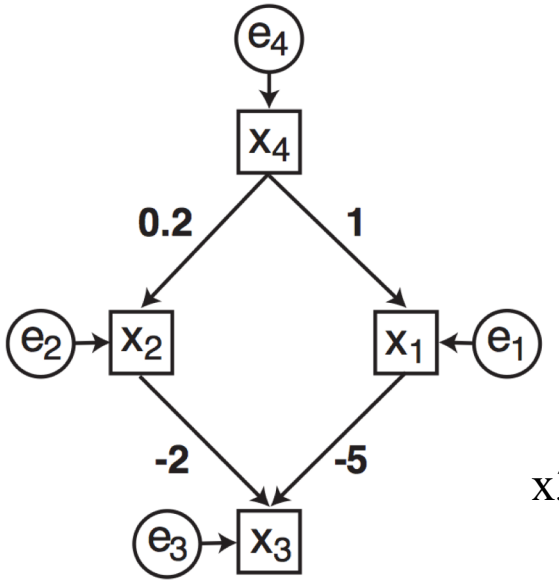
$x_2 = b_{21}x_1 + e_2$



$x_1 = b_{12}x_2 + e_1$

$x_2 = e_2$





x4 = e4

x2 = 0.2*x4 +e2

x1 = x4 +e1

x3 = −2*x2 −5*x1 +e3

p(x1,x2,x3,x4) = p(x4) p(x2|x4) p(x1|x4) p(x3|x2,x1)

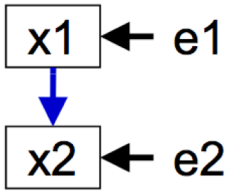
Shimizu et. al., JMLR 2006
Shimizu and Kawahara, UAI2010 Tutorial

Identify Which Model to Generate The Data

- Two models with Gaussian e_1 and e_2 :

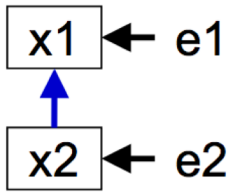
Model 1:

$$x_1 = e_1$$
$$x_2 = 0.8x_1 + e_2$$



Model 2:

$$x_1 = 0.8x_2 + e_1$$
$$x_2 = e_2$$



$$E(e_1) = E(e_2) = 0, \mathrm{var}(x_1) = \mathrm{var}(x_2) = 1$$

- Both introduce no conditional independence:

$$\mathrm{cov}(x_1, x_2) = 0.8 \neq 0$$

- Both induce the same Gaussian distribution:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \right)$$

Gaussian vs. Non-Gaussian

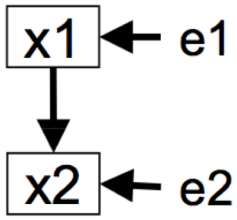
Gaussian

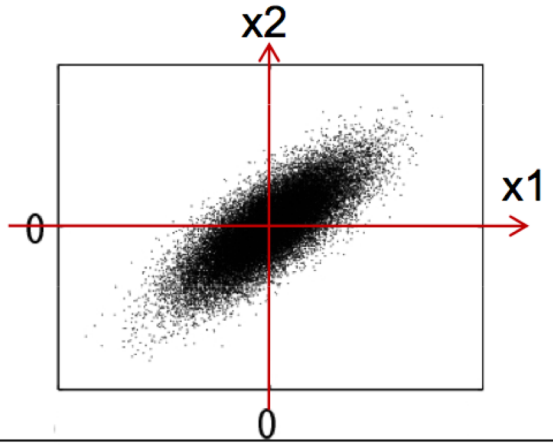
Non-Gaussian (uniform)

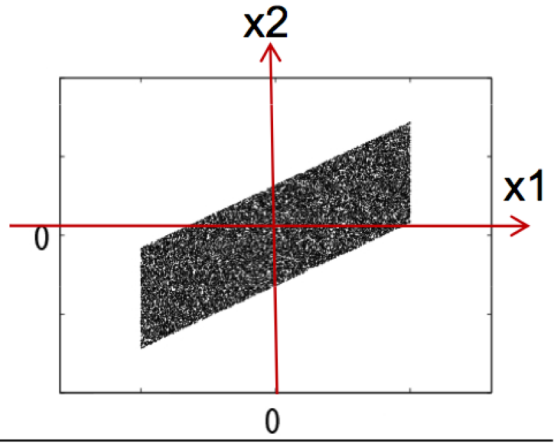
Model 1:

$x_1 = e_1$

$x_2 = 0.8x_1 + e_2$



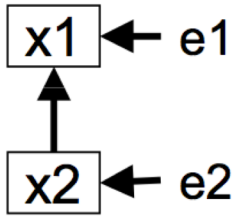


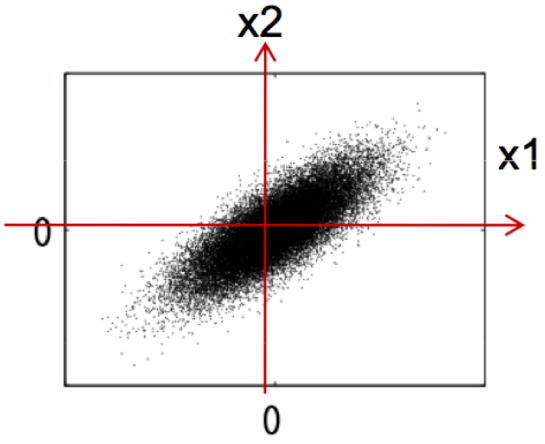


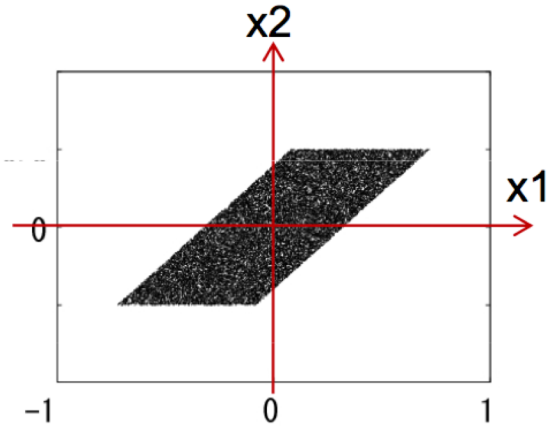
Model 2:

$x_1 = 0.8x_2 + e_1$

$x_2 = e_2$







$E(e_1) = E(e_2) = 0,$

$\mathrm{var}(x_1) = \mathrm{var}(x_2) = 1$

Linear Non-Gaussian Acyclic Model: LiNGAM

- Linear acyclic SEM

[Shimizu, Hyvarinen, Hoyer & Kerminen, JMLR 2006]

$$x_i = \sum_{j:\text{parents of } i} b_{ij} x_j + e_i \qquad \text{or} \qquad \mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

For example:

$$x_1 = 1.5x_3 + e_1$$
$$x_2 = -1.3x_1 + e_2$$
$$x_3 = e_3$$

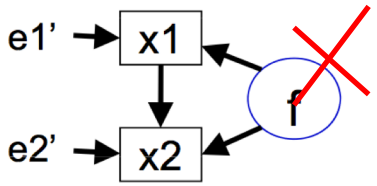
or

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & 1.5 \\ -1.3 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

Causal Markov condition holds:

$$p(\mathbf{x}) = \prod_{i=1}^{p} p(x_i \mid \text{parents of } x_i)$$

- Assumptions:
  - Directed acyclic graph (DAG): no directed cycles
  - External influences e_i are of non-zero variance, and are independent **non-Gaussian**
  - No latent confounders



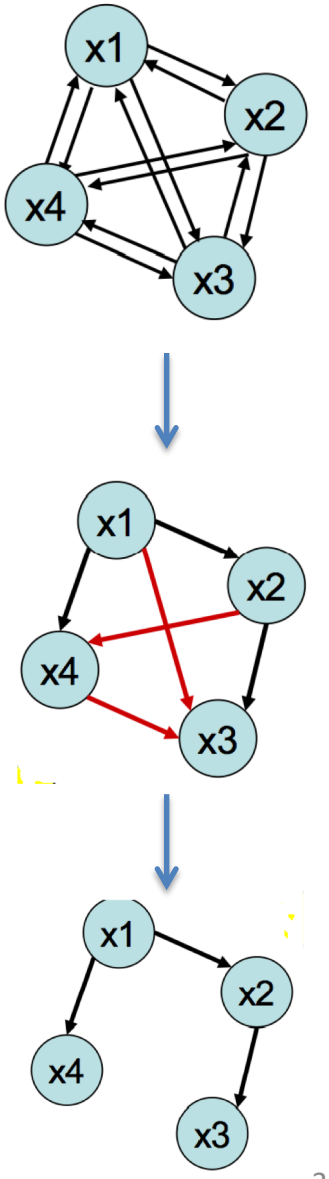
How *B* Is Estimated?

- Step 1: Estimate B by Independent Component Analysis (ICA) with post-processing

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \Leftrightarrow \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}$$
$$= \mathbf{A}\mathbf{e} = \mathbf{W}^{-1}\mathbf{e}$$

- Step 2: Find an order of the variables to get a DAG

- Step 3: Discard non-significant edges



[Shimizu, et. al. Jan 2008]

Performance of the algorithm

- Fast (ICA is fast)
- Possible local optimum problem (ICA is an iterative method)
- A good estimation needs >1000 sample size for >10 variables
- Not scale invariant

Applications

- Neuroinformatics
  – Brain connectivity analysis (Hyvarinen et al., JMLR, 2010)
- Bioinformatics
  – Gene network estimation (Sogawa et al., ICANN2010)
- Economics(Wan&Tan,2009; Moneta,Entner,Hoyer&Coad,2010)
- Genetics(Ozaki&Ando,2009)
- Environmental sciences(Niyogietal.,2010)
- Physics (Kawahara, Shimizu & Washio, 2010)
- Sociology (Kawahara, Bollen, Shimizu & Washio, 2010)

Thank you!