Causal inference and causal discovery

Shikui Tu

Department of Computer Science and Engineering, Shanghai Jiao Tong University

2021-06-04

Outline

- **Discover causal structure by conditional independence**
  - **PC algorithm**
  - **Markov equivalent class**

- Pearl's do-calculus

The New England Journal of Medicine / Impact factor

72.406
2016

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Association of Coffee Drinking with Total and Cause-Specific Mortality

Neal D. Freedman, Ph.D., Yikyung Park, Sc.D., Christian C. Abnet, Ph.D., Albert R. Hollenbeck, Ph.D., and Rashmi Sinha, Ph.D.

BACKGROUND

Coffee is one of the most widely consumed beverages, but the association between coffee consumption and the risk of death remains unclear.

METHODS

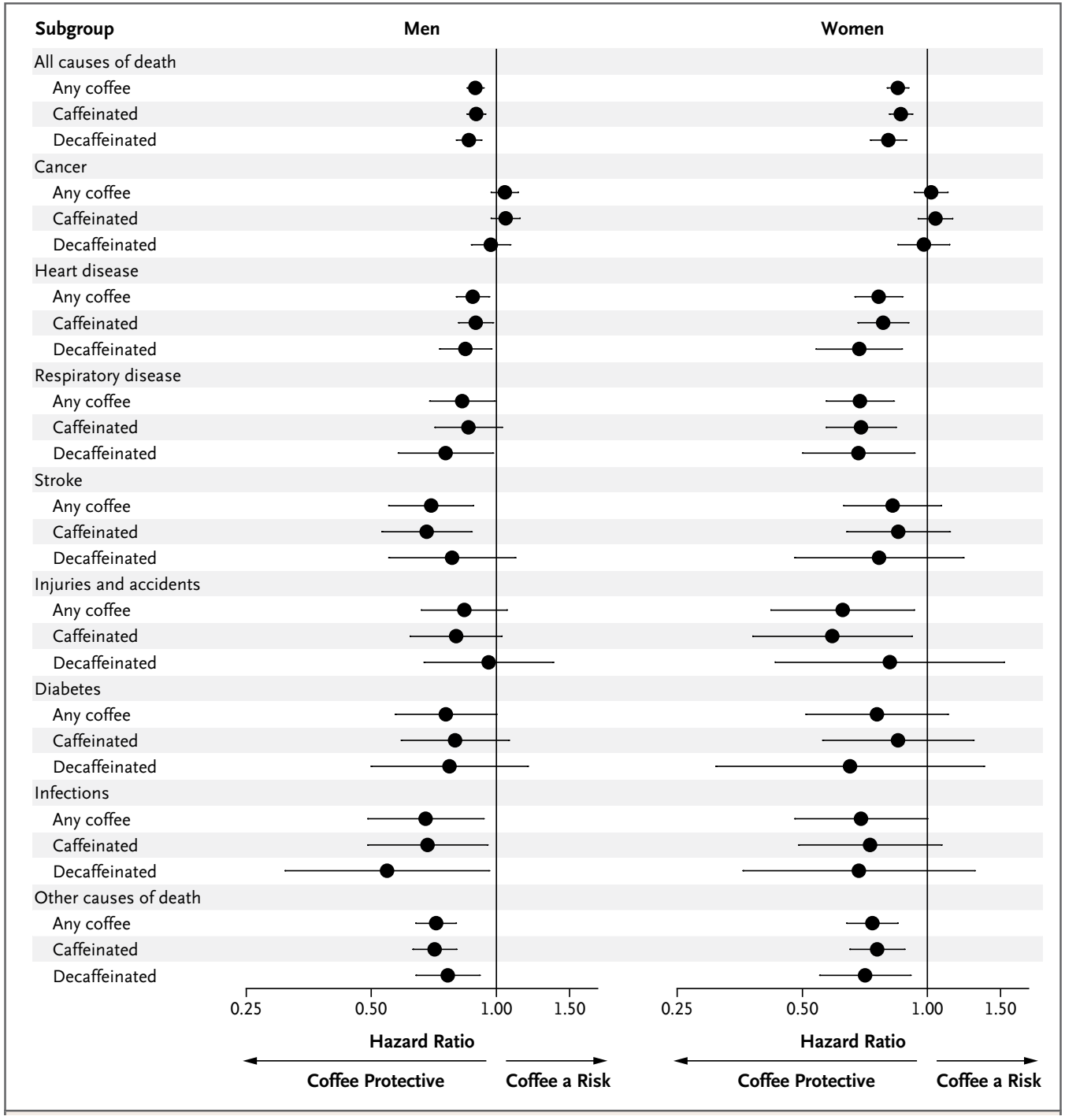
We examined the association of coffee drinking with subsequent total and cause-specific mortality among 229,119 men and 173,141 women in the National Institutes of Health–AARP Diet and Health Study who were 50 to 71 years of age at baseline. Participants with cancer, heart disease, and stroke were excluded. Coffee consumption was assessed once at baseline.

RESULTS

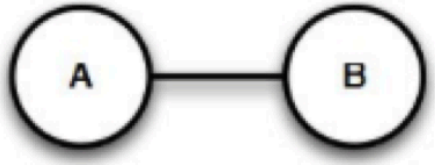
During 5,148,760 person-years of follow-up between 1995 and 2008, a total of 33,731 men and 18,784 women died. **In age-adjusted models, the risk of death was increased among coffee drinkers.** However, coffee drinkers were also more likely to **smoke**, and, after adjustment for tobacco-smoking status and other potential confounders, there was a significant **inverse association between coffee consumption and mortality**.

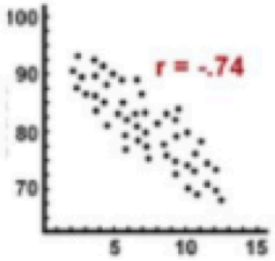
CONCLUSIONS

In this large prospective study, coffee consumption was inversely associated with total and cause-specific mortality. Whether this was a causal or associational finding cannot be determined from our data.


Subgroup
Men
Women
All causes of death
Any coffee
Caffeinated
Decaffeinated
Cancer
Any coffee
Caffeinated
Decaffeinated
Heart disease
Any coffee
Caffeinated
Decaffeinated
Respiratory disease
Any coffee
Caffeinated
Decaffeinated
Stroke
Any coffee
Caffeinated
Decaffeinated
Injuries and accidents
Any coffee
Caffeinated
Decaffeinated
Diabetes
Any coffee
Caffeinated
Decaffeinated
Infections
Any coffee
Caffeinated
Decaffeinated
Other causes of death
Any coffee
Caffeinated
Decaffeinated
0.25
0.50
1.00
1.50
Hazard Ratio
Coffee Protective
Coffee a Risk
0.25
0.50
1.00
1.50
Hazard Ratio
Coffee Protective
Coffee a Risk

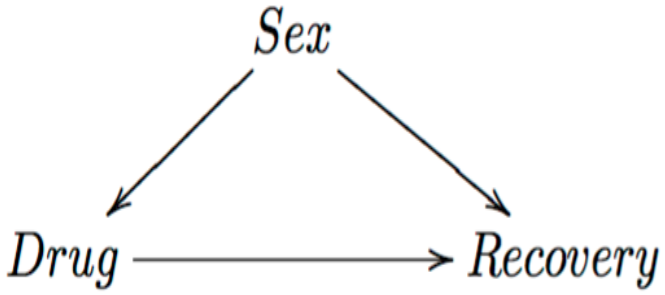
Association



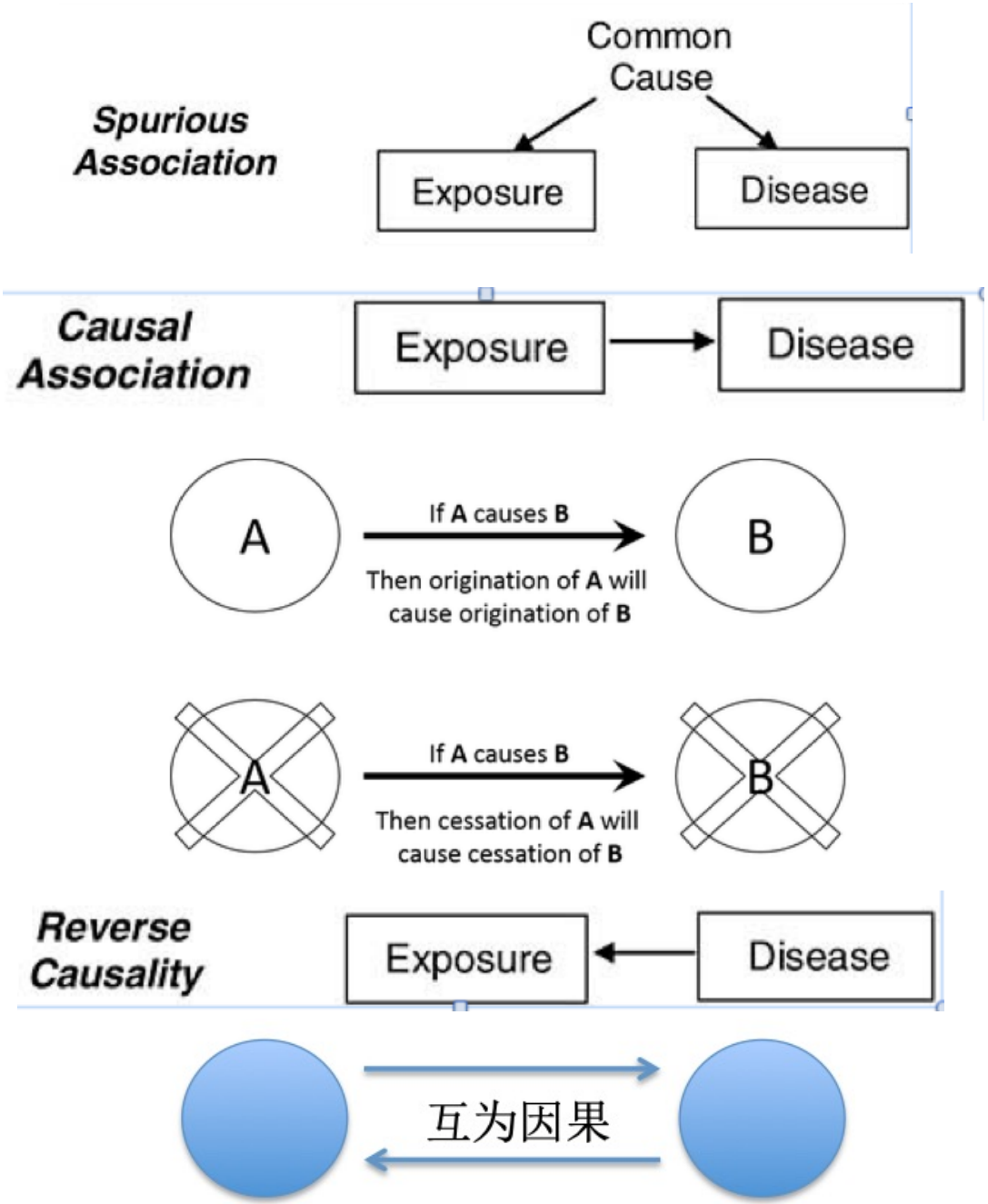


此悖论表明:
X 和Y 边缘上正相关，
但是,给定另外一个变量Z后，
在Z 的每一个水平上，X 和Y
可能负相关。

Yule-Simpson Paradox (Pearl，2000)



合并表	康复	未康复	康复率
吃药	20	20	50%
安慰剂	16	24	40%
男性	康复	未康复	康复率
吃药	18	12	60%
安慰剂	7	3	70%
女性	康复	未康复	康复率
吃药	2	8	20%
安慰剂	9	21	30%

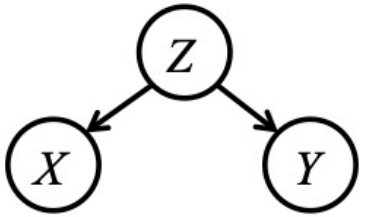

Spurious Association
Common Cause
Exposure
Disease
Causal Association
Exposure
Disease
A
If A causes B
Then origination of A will cause origination of B
B
A
If A causes B
Then cessation of A will cause cessation of B
B
Reverse Causality
Exposure
Disease
互为因果

Statistical Implications of Causality

(Reichenbach 1956)

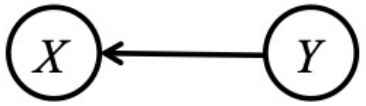
Reichenbach's
Common Cause Principle
links **causality** and **probability:**

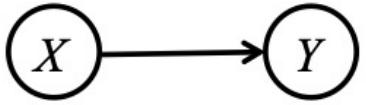
(i) if X and Y are statistically dependent, then there is a Z causally influencing both;



special cases:

(ii) Z screens X and Y from each other (given Z, the observables X and Y become independent)





Independence of random variables

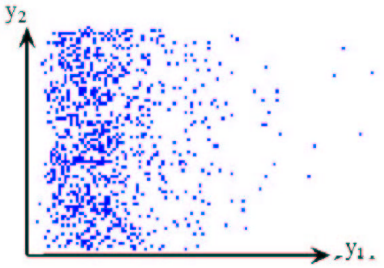
Two real-valued random variables X and Y are called *independent*,

$$X \perp\!\!\!\perp Y,$$

if for every $a, b \in \mathbb{R}$, the events $\{X \leq a\}$ and $\{Y \leq b\}$ are independent.

Equivalently, in terms of densities: for all x, y,

$$p(x, y) = p(x)p(y)$$



Note:
If $X \perp\!\!\!\perp Y$, then $E[XY] = E[X]E[Y]$, and $\text{cov}[X, Y] = E[XY] - E[X]E[Y] = 0$.

The converse is not true: $\text{cov}[X, Y] = 0 \not\Rightarrow X \perp\!\!\!\perp Y$.

However, we have, for large $\mathcal{F}$: $(\forall f, g \in \mathcal{F} : \text{cov}[f(X), g(Y)] = 0) \Rightarrow X \perp\!\!\!\perp Y$

Conditional Independence of random variables

Two real-valued random variables X and Y are called *conditionally independent* given Z,

$$(X \perp\!\!\!\perp Y)\,|Z \quad \text{or} \quad X \perp\!\!\!\perp Y\,|Z \quad \text{or} \quad (X \perp\!\!\!\perp Y\,|Z)_p$$

if

$$p(x, y|z) = p(x|z)p(y|z)$$

for all x, y, and for all z s.t. $p(z) > 0$.

Note: conditional independence neither implies nor is implied by independence.
I.e., there are X, Y, Z such that we have only independence or only conditional independence.

Conditional independence tests

- discrete case: contingency tables

		不吸烟	吸烟	合计
年龄< 40	呼吸正常	567	874	1441
	呼吸不正常	14	28	42
年龄40-59	呼吸正常	328	780	1108
	呼吸不正常	2	68	70

$$\left(\frac{p(x,y)}{p(x)p(y)} - 1\right)$$

C	A	B		
		B_1	$\cdots$	B_c
C_1	A_1	p_{111}	$\cdots$	p_{1c1}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_r	p_{r11}	$\cdots$	p_{rc1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_t	A_1	p_{11t}	$\cdots$	p_{1ct}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_r	p_{r1t}	$\cdots$	p_{rct}

- multivariate gaussian case: covariance matrix

$$\Sigma_s = \begin{array}{|c|cc|} \hline \sigma_{ww} & \sigma_{1w} & \sigma_{2w} \\ \hline \sigma_{w1} & & \\ \sigma_{w2} & & \Sigma \\ \hline \end{array} \quad ==\rangle\rangle \quad \sigma_{ij} - \sigma_{wi}\sigma_{jw} / \sigma_{ww} = 0, \quad i \neq j \text{ and } i, j = 1, 2,$$

$$\int \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dp(x,y)$$

- non-Gaussian continuous case: via reproducing kernel Hilbert spaces

Conditional Independence Testing using Generative Adversarial Networks

Alexis Bellot[1,2] **Mihaela van der Schaar**[1,2,3]
[1]University of Cambridge, [2]The Alan Turing Institute, [3]University of California Los Angeles
[abellot,mschaar]@turing.ac.uk

Abstract

We consider the hypothesis testing problem of detecting conditional dependence, with a focus on high-dimensional feature spaces. Our contribution is a new test statistic based on samples from a generative adversarial network designed to approximate directly a conditional distribution that encodes the null hypothesis, in a manner that maximizes power (the rate of true negatives). We show that such an approach requires only that density approximation be viable in order to ensure that we control type I error (the rate of false positives); in particular, no assumptions need to be made on the form of the distributions or feature dependencies. Using synthetic simulations with high-dimensional data we demonstrate significant gains in power over competing methods. In addition, we illustrate the use of our test to discover causal markers of disease in genetic data.

Kernel-based Conditional Independence Test and Application in Causal Discovery

Kun Zhang **Jonas Peters** **Dominik Janzing** **Bernhard Schölkopf**
Max Planck Institute for Intelligent Systems
Spemannstr. 38, 72076 Tübingen
Germany

Abstract

Conditional independence testing is an important problem, especially in Bayesian network learning and causal discovery. Due to the curse of dimensionality, testing for conditional independence of continuous variables is particularly challenging. We propose a Kernel-based Conditional Independence test (KCI-test), by constructing an appropriate test statistic and deriving its asymptotic distribution under the null hypothesis of conditional independence. The proposed method is computationally efficient and easy to implement. Experimental results show that it outperforms other methods, especially when the conditioning set is large or the sample size is not very large, in which case other methods encounter difficulties.

the continuous case – in particular, the variables are often assumed to have linear relations with additive Gaussian errors. In that case, $X \perp\!\!\!\perp Y|Z$ reduces to zero partial correlation or zero conditional correlation between X and Y given Z, which can be easily tested (for the links between partial correlation, conditional correlation, and CI, see Lawrance (1976)). However, nonlinearity and non-Gaussian noise are frequently encountered in practice, and hence this assumption can lead to incorrect conclusions.

Recently, practical methods have been proposed for testing CI for continuous variables without assuming a functional form between the variables as well as the data distributions, which is the case we are concerned with in this paper. To our knowledge, the existing methods fall into four categories. The first category is based on explicit estimation of the conditional densities or their variants. For example, Su and White (2008) define the test statistic as some distance be-

A Kernel Statistical Test of Independence

Arthur Gretton
MPI for Biological Cybernetics
Tübingen, Germany
arthur@tuebingen.mpg.de

Kenji Fukumizu
Inst. of Statistical Mathematics
Tokyo Japan
fukumizu@ism.ac.jp

Choon Hui Teo
NICTA, ANU
Canberra, Australia
choonhui.teo@gmail.com

Le Song
NICTA, ANU
and University of Sydney
lesong@it.usyd.edu.au

Bernhard Schölkopf
MPI for Biological Cybernetics
Tübingen, Germany
bs@tuebingen.mpg.de

Alexander J. Smola
NICTA, ANU
Canberra, Australia
alex.smola@gmail.com

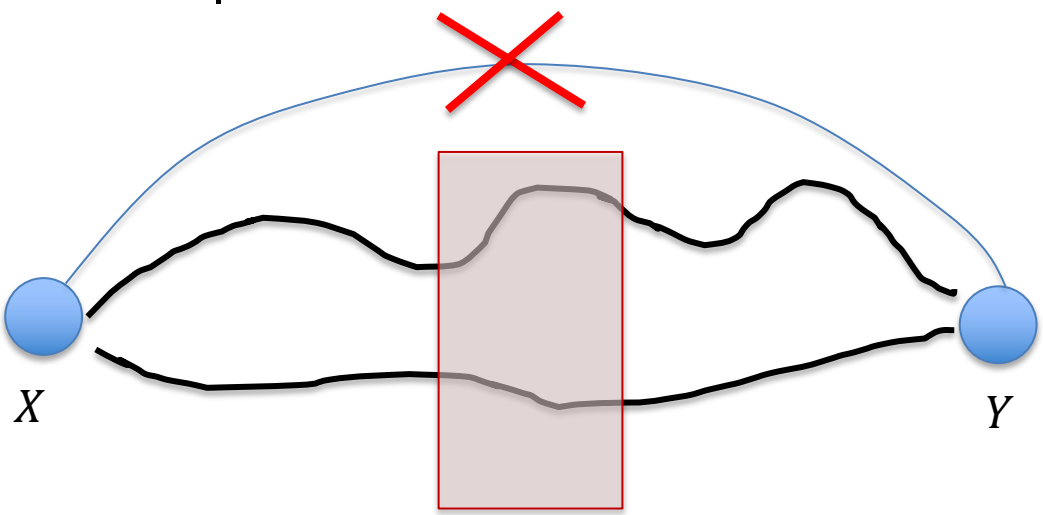
Abstract

Although kernel measures of independence have been widely applied in machine learning (notably in kernel ICA), there is as yet no method to determine whether they have detected statistically significant dependence. We provide a novel test of the independence hypothesis for one particular kernel independence measure, the Hilbert-Schmidt independence criterion (HSIC). The resulting test costs $\mathrm{O}(m^2)$, where m is the sample size. We demonstrate that this test outperforms established contingency table and functional correlation-based tests, and that this advantage is greater for multivariate data. Finally, we show the HSIC test also applies to text (and to structured data more generally), for which no other independence test presently exists.

Discover causal structure by conditional independence

- Basic idea:

$X \perp Y \mid S$



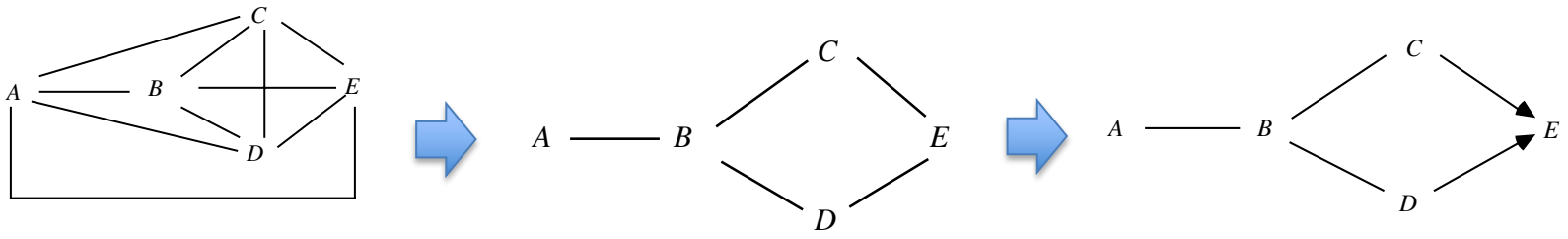
$S = \{Z_1, \ldots, Z_k\}$

X and Y are *d*-separated by S.

PC Algorithm

[Spirtes, Glymour, and Scheines 1991]

- Four steps.
  - A.) Form the complete undirected graph;
  - B.) Remove edges according to n-order conditional independence relations;
  - C.) Orient edges by v-structures
  - D.) Orient edges



Step B: Remove edges

B.)

$n = 0$.

repeat

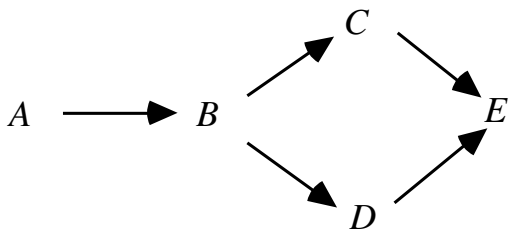
repeat

select an ordered pair of variables X and Y that are adjacent in C such that **Adjacencies**(C,X)\{Y} has cardinality greater than or equal to n, and a subset **S** of **Adjacencies**(C,X)\{Y} of cardinality n, and if X and Y are d-separated given **S** delete edge X - Y from C and record **S** in **Sepset**(X,Y) and **Sepset**(Y,X);

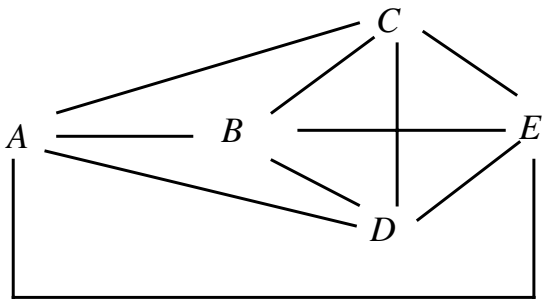
until all ordered pairs of adjacent variables X and Y such that **Adjacencies**(C,X)\{Y} has cardinality greater than or equal to n and all subsets **S** of **Adjacencies**(C,X)\{Y} of cardinality n have been tested for d-separation;

$n = n + 1$;

until for each ordered pair of adjacent vertices X, Y, **Adjacencies**(C,X)\{Y} is of cardinality less than n.



True Graph



Complete Undirected Graph

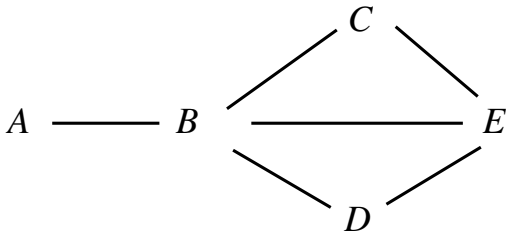
$n = 0$ No zero order independencies.

$n = 1$ First order independencies.

$A \perp C \mid B$ $A \perp D \mid B$

$A \perp E \mid B$ $C \perp D \mid B$

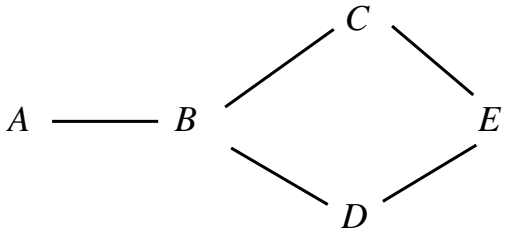
Resulting Adjacencies



$n = 2$ Second order independencies.

$B \perp E \mid \{C, D\}$

Resulting Adjacencies



Step C-D: Orient edges

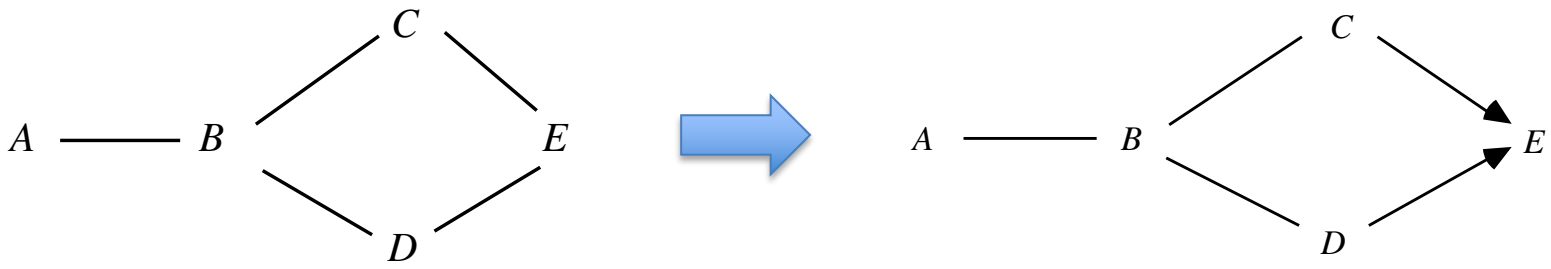
C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C, orient X - Y - Z as $X \rightarrow Y \leftarrow Z$ if and only if Y is not in **Sepset**(X,Z).

D.) repeat

If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient B - C as $B \rightarrow C$.

If there is a *directed* path from A to B, and an edge between A and B, then orient A - B as $A \rightarrow B$.

until no more edges can be oriented.



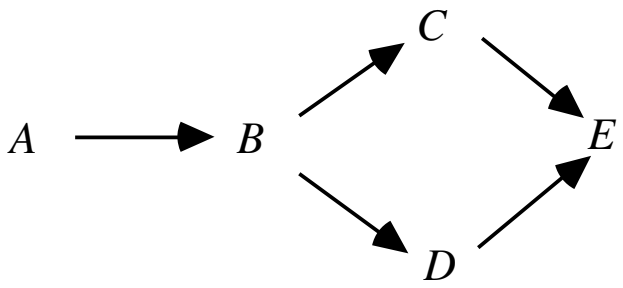
The triples of variables with only two adjacencies among them are:

A - B - C;
C - B - D;
B - D - E;

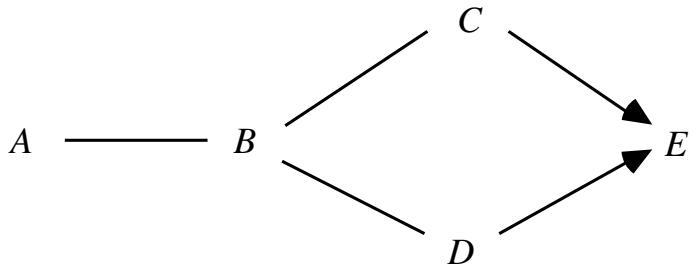
A - B -D;
B - C - E;
C - E - D

E is not in **Sepset**(C,D)
so C - E and E - D collide at E.

Results up to an indistinguishable class



True graph



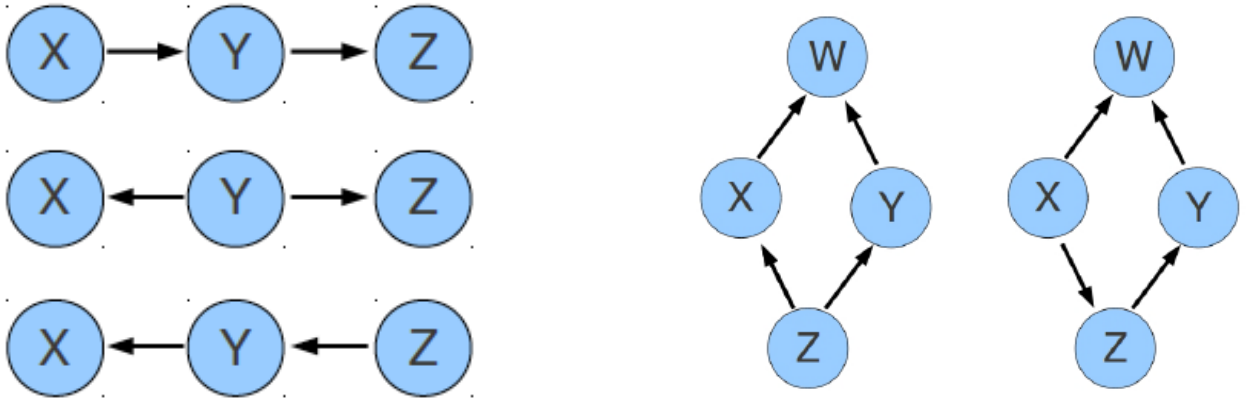
Result by PC algorithm

Every orientation of the undirected edges in the result by PC algorithm is permissible that does not include a collision at *B*.

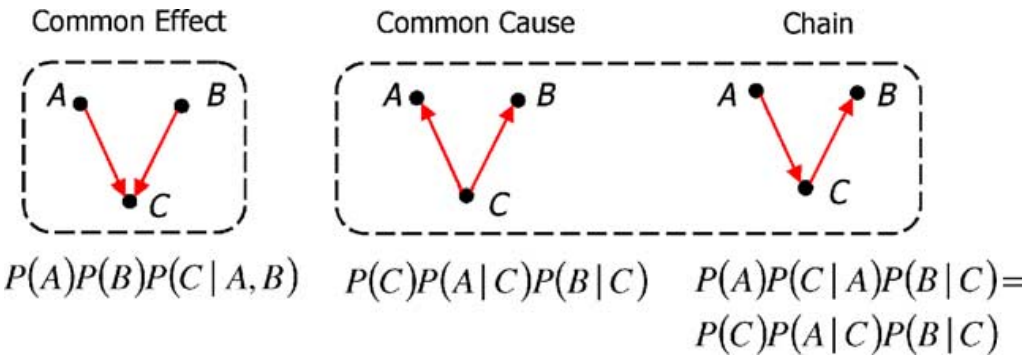
What is the complexity of the PC algorithm?

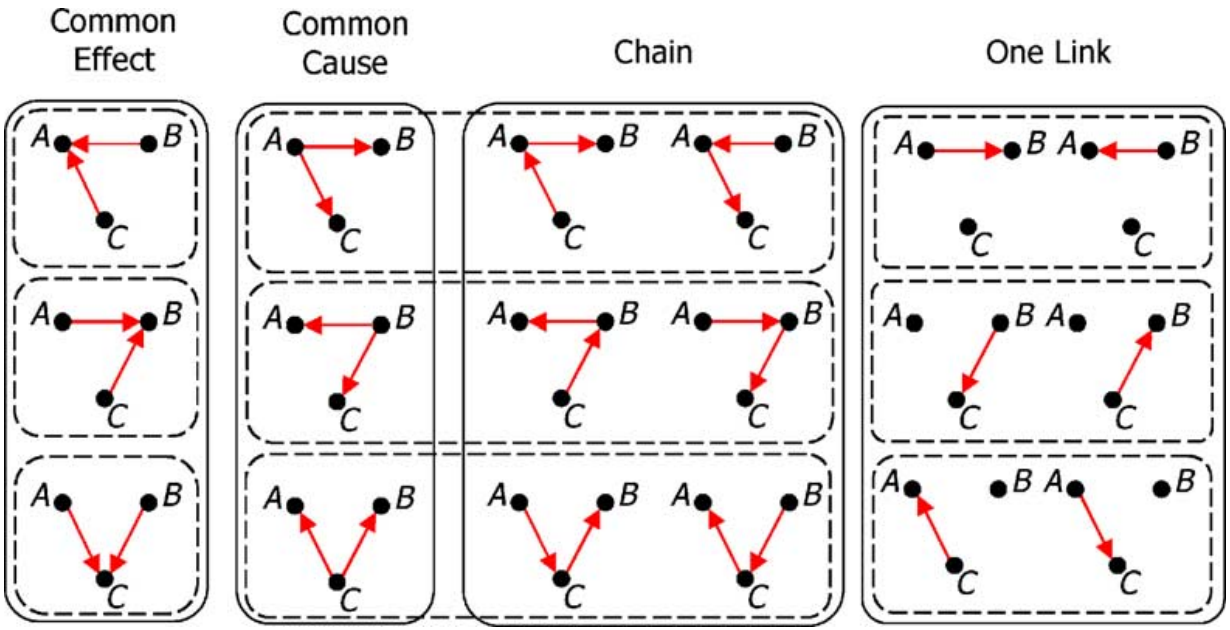
Markov equivalent class

Theorem (Verma and Pearl, 1990): two DAGs are Markov equivalent iff they have the same skeleton and the same v-structures.
skeleton: corresponding undirected graph
v-structure: substructure X ->Y <-Z with no edge between X and Z.



All three-node networks (1-2 arrows)

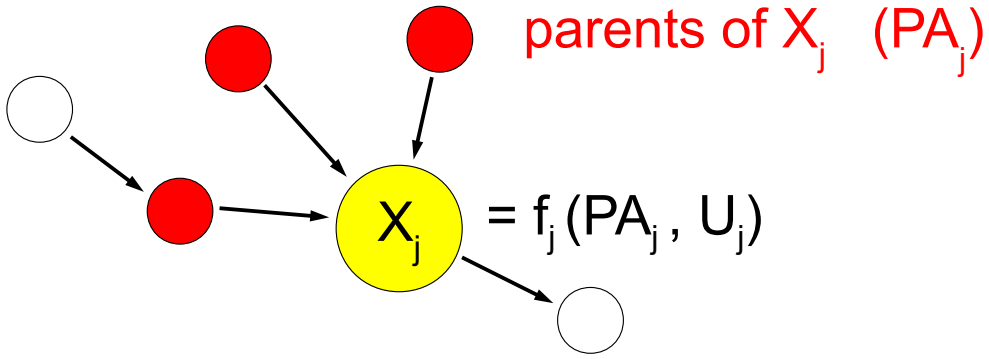




Solid lines group together networks of the same topological type.
Dashed lines delineate Markov equivalence classes.

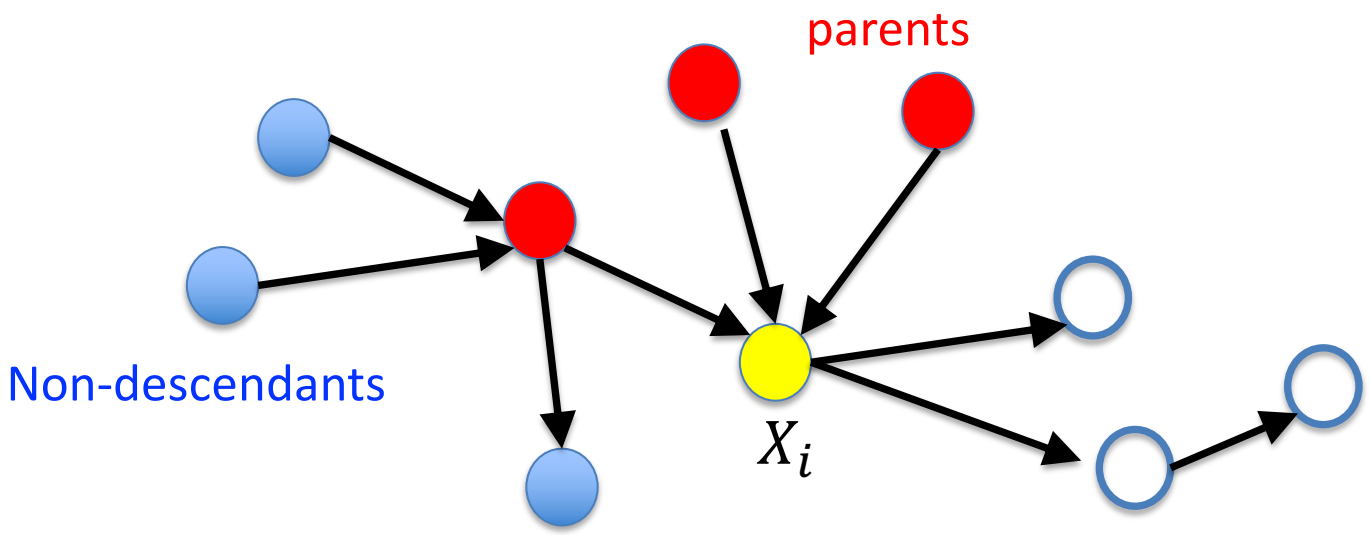
Functional Causal Model *(Pearl et al.)*

- Set of observables $X_1, \ldots, X_n$
- directed acyclic graph G with vertices $X_1, \ldots, X_n$
- Semantics: parents = direct causes
- $X_i = f_i(\text{ParentsOf}_i, \text{Noise}_i)$, with independent $\text{Noise}_1, \ldots, \text{Noise}_n$.
- "Noise" means "unexplained" (or "exogenous"), we use U_i
- Can add requirement that $f_1, \ldots, f_n, \text{Noise}_1, \ldots, \text{Noise}_n$ "independent" (cf. *Lemeire & Dirkx 2006, Janzing & Schölkopf 2010* — more below)



Theorem: the following are equivalent:

- Existence of a functional causal model
- Local Causal Markov condition: X_j *statistically independent* of non-descendants, given parents (i.e.: every information exchange with its non-descendants involves its parents)
- Global Causal Markov condition: d-separation (characterizes the set of independences implied by local Markov condition)
- Factorization $P(X_1, \ldots, X_n) = \prod_j P(X_j \mid \text{Parents}_j)$ (conditionals as causal mechanisms generating statistical dependence)



Outline

- Discover causal structure by conditional independence
  - PC algorithm
  - Markov equivalent class

- **Pearl's do-calculus**

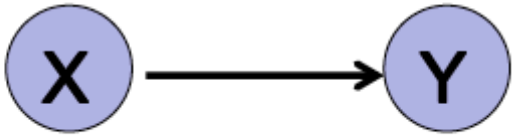
Pearl's do-calculus

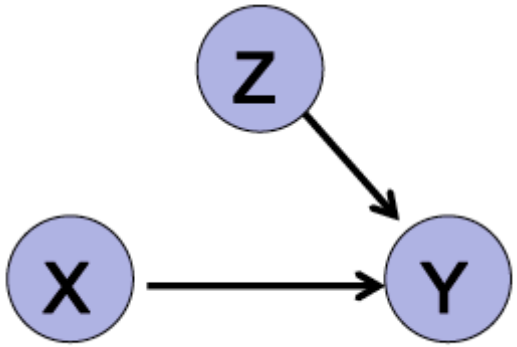
- Motivation: goal of causality is to infer the effect of interventions
- distribution of Y given that X is set to x: p(Y |do X = x) or p(Y |do x)

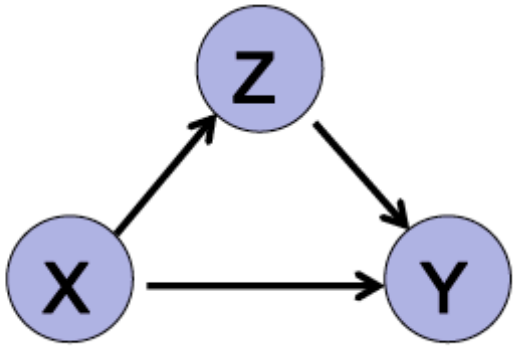
根据 do 操作，便可以定义因果作用，比如二值的变量 Z 对于 Y 的平均因果作用定义为

$$ACE(Z \to Y) = E(Y \mid do(Z) = 1) - E(Y \mid do(Z) = 0),$$

Examples for $p(.|do\,x) = p(.|x)$

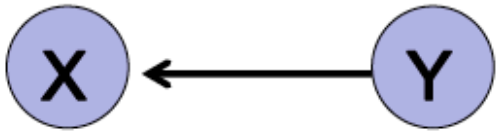




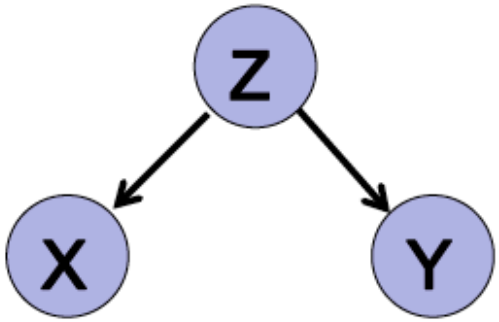


Examples for $p(.|do\,x) \neq p(.|x)$

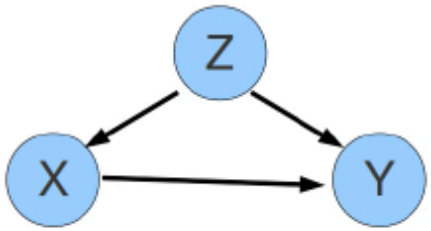
- $p(Y|do\,x) = P(Y) \neq P(Y|x)$



- $p(Y|do\,x) = P(Y) \neq P(Y|x)$



Example: controlling for confounding



$X \not\perp\!\!\!\perp Y$ partly due to the Z and partly due to $X \to Y$

- causal factorization

$$p(X, Y, Z) = p(Z)p(X|Z)p(Y|X, Z)$$

- replace $P(X|Z)$ with δ_{Xx}

$$p(Y, Z|do\,x) = p(Z)\ \delta_{Xx}\ p(Y|X, Z)$$

- marginalize

$$p(Y|do\,x) = \sum_z p(z)p(Y|x, z) \neq \sum_z p(z|x)p(Y|x, z) = p(Y|x)$$

Thank you!