Causal inference and causal discovery-part(2)

Shikui Tu

Department of Computer Science and Engineering, Shanghai Jiao Tong University

2018-04-19

Outline

- **Discover causal structure by conditional independence**
  - **PC algorithm**
  - **Markov equivalent class**
- Star causality structure
- A linear non-Gaussian model for causal discovery
- 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

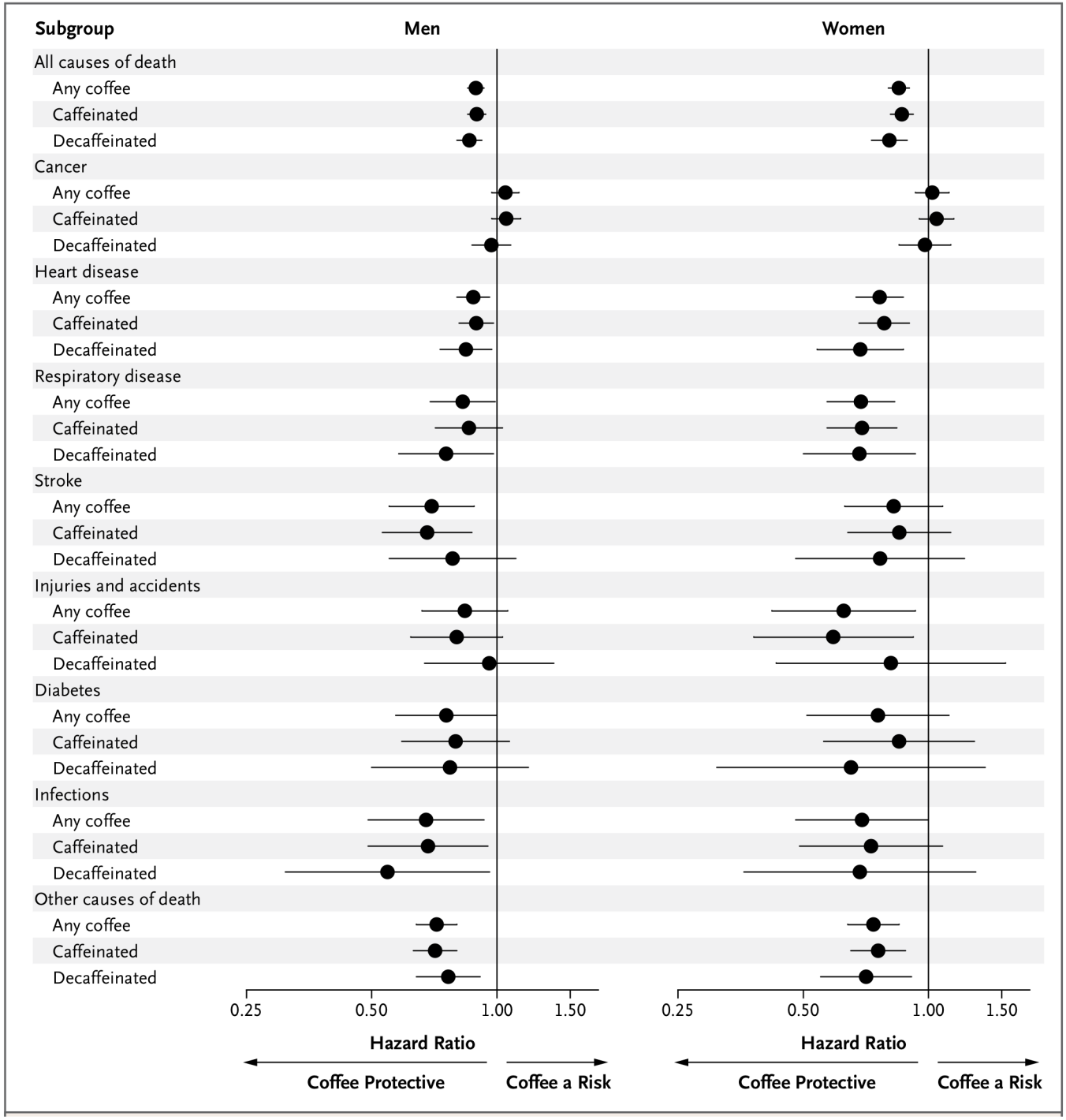
We examined the association of coffee drinking with subsequent total and cause-specific mortality among 229,119 men and 173,141 women in the National Institutes of Health–AARP Diet and Health Study who were 50 to 71 years of age at baseline. Participants with cancer, heart disease, and stroke were excluded.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

RESULTS

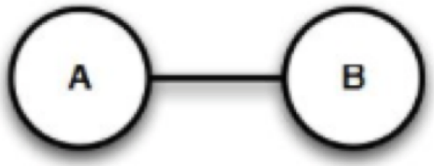
During 5,148,760 person-years of follow-up between 1995 and 2008, a total of 33,731 men and 18,784 women died. In age-adjusted models, the risk of death was increased among coffee drinkers. However, coffee drinkers were also more likely to smoke, and, after adjustment for tobacco-smoking status and other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

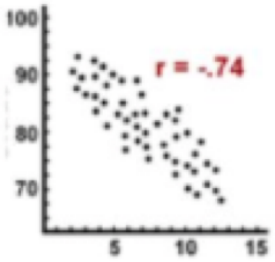
CONCLUSIONS

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

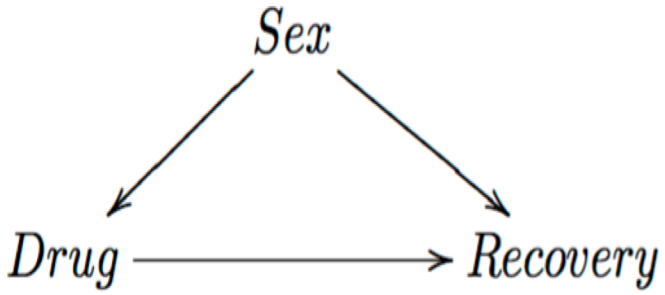
Association



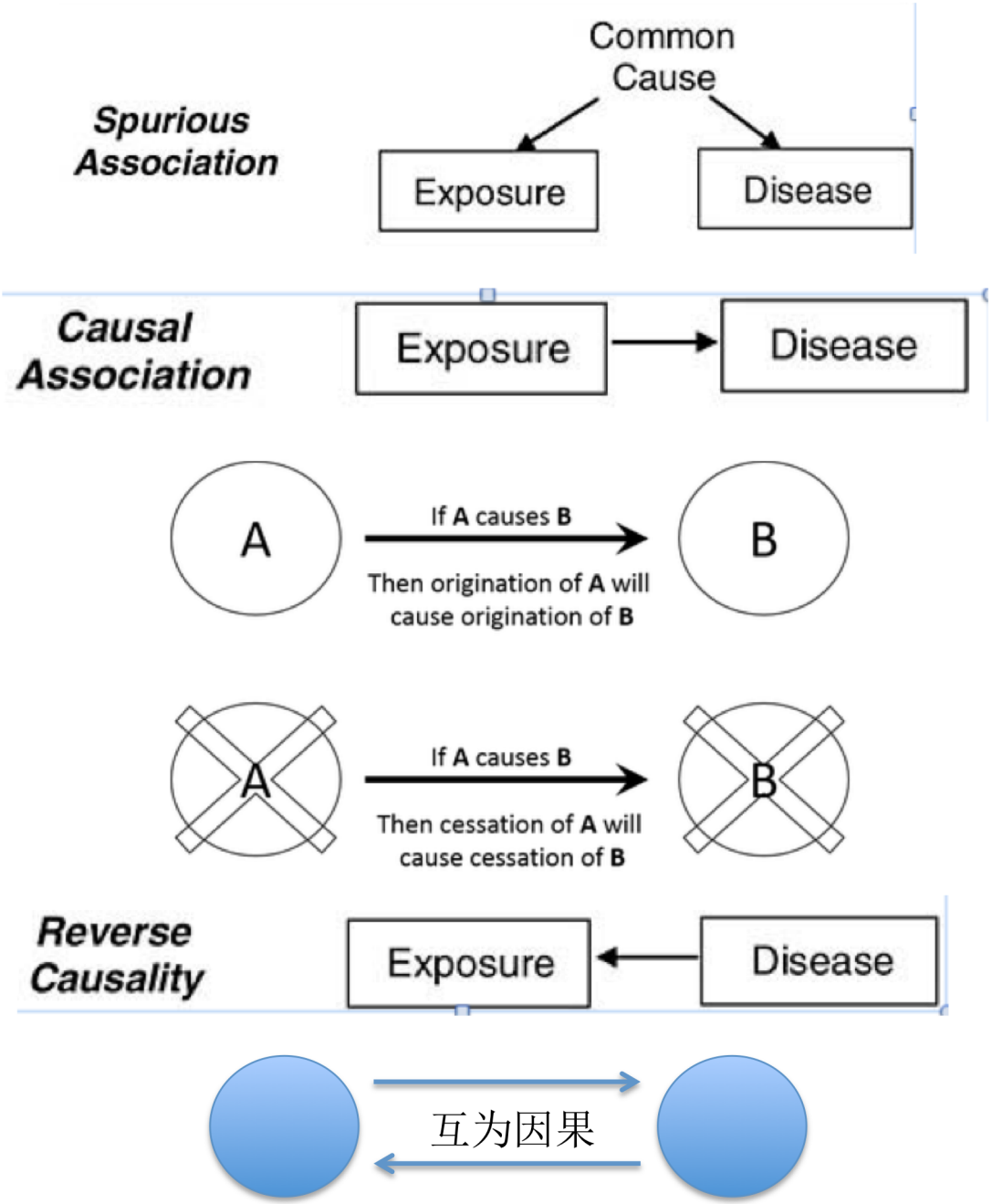


此悖论表明:
X 和Y 边缘上正相关，
但是,给定另外一个变量Z后，
在Z 的每一个水平上，X 和Y
可能负相关。

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%

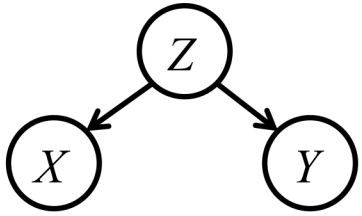

Spurious Association
Common Cause
Exposure
Disease
Causal Association
Exposure
Disease
A
If A causes B
Then origination of A will cause origination of B
B
A
If A causes B
Then cessation of A will cause cessation of B
B
Reverse Causality
Exposure
Disease
互为因果

Statistical Implications of Causality

(Reichenbach 1956)

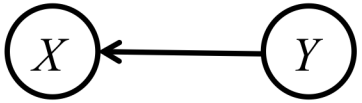
Reichenbach's
Common Cause Principle
links **causality** and **probability:**

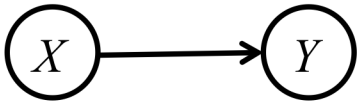
(i) if X and Y are statistically dependent, then there is a Z causally influencing both;



special cases:

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





Independence of random variables

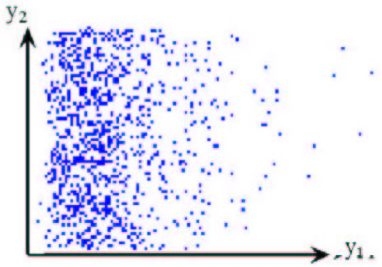
Two real-valued random variables X and Y are called *independent*,

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

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

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

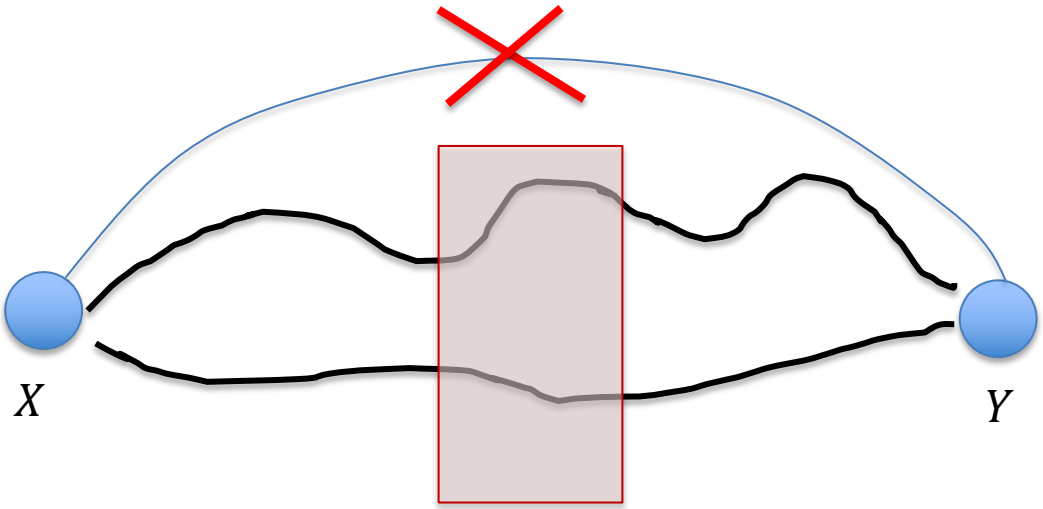
$$\int \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dp(x,y)$$

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

Discover causal structure by conditional independence

- Basic idea:

$X \perp Y \mid S$



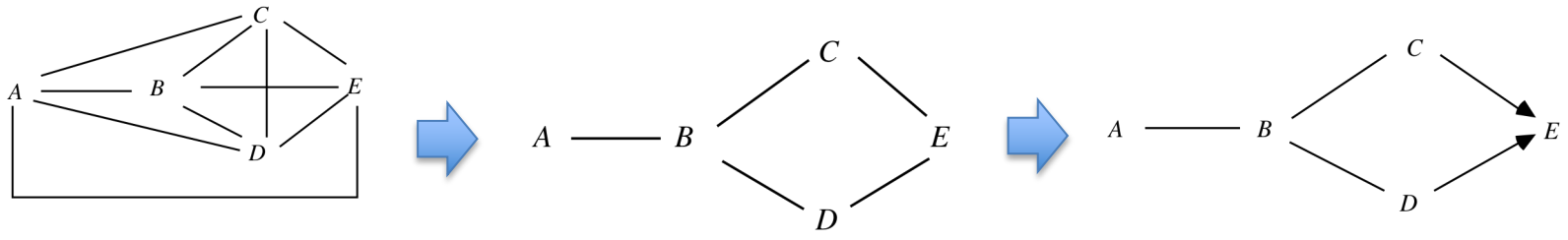
$S = \{Z_1, \ldots, Z_k\}$

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

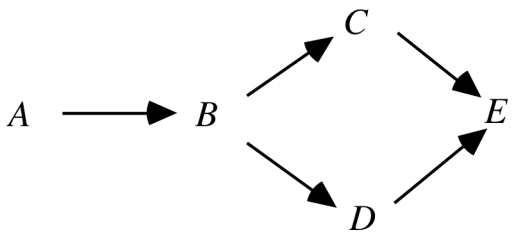
repeat

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);

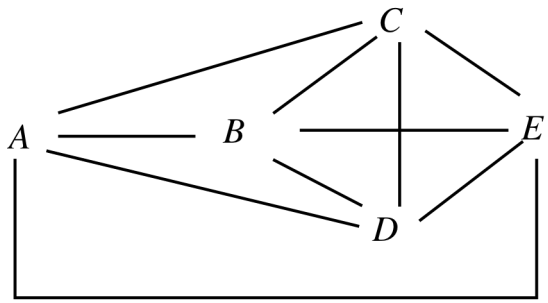
until all ordered pairs of adjacent variables X and Y such that **Adjacencies**(C,X)\{Y} has cardinality greater than or equal to n and all subsets **S** of **Adjacencies**(C,X)\{Y} of cardinality n have been tested for d-separation;

$n = n + 1$;

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

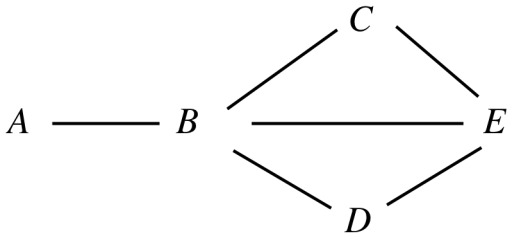
$n = 0$ No zero order independencies.

$n = 1$ First order independencies.

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

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

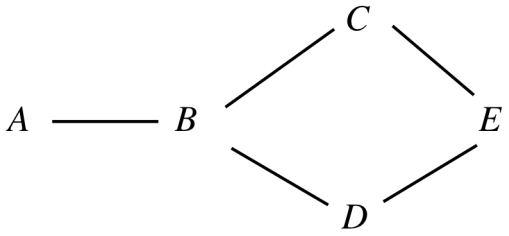
Resulting Adjacencies



$n = 2$ Second order independencies.

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

Resulting Adjacencies



Step C-D: Orient edges

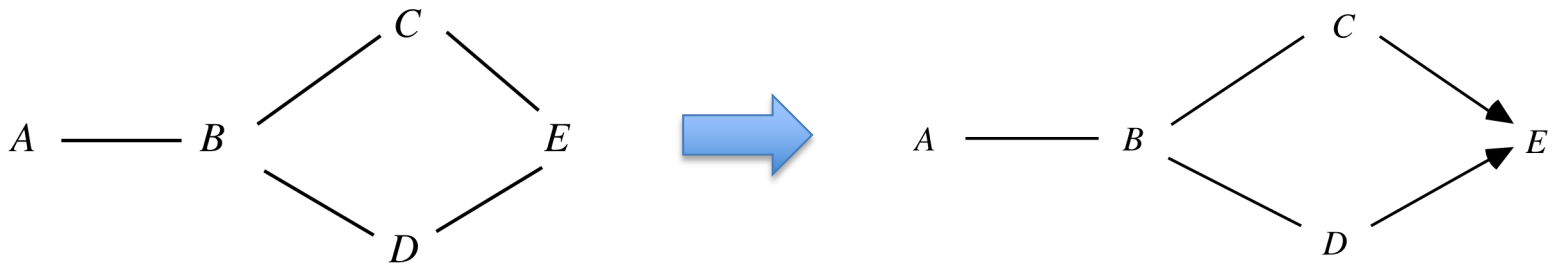
C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C, orient X - Y - Z as $X \rightarrow Y \leftarrow Z$ if and only if Y is not in **Sepset**(X,Z).

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.



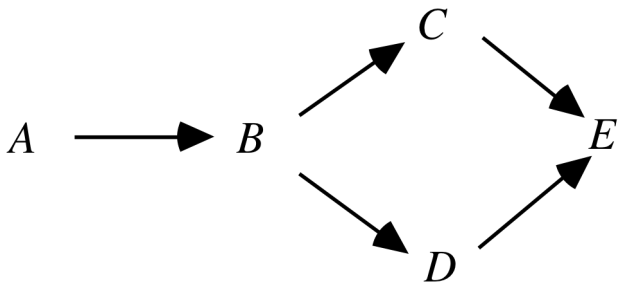
The triples of variables with only two adjacencies among them are:

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

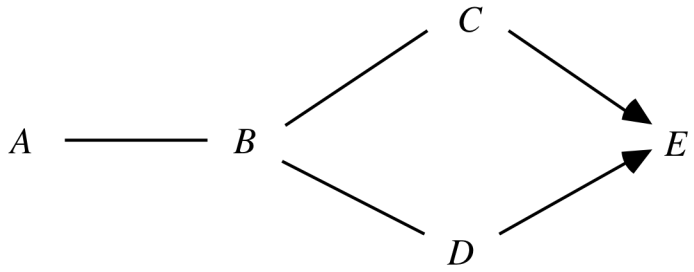
A - B -D;
B - C - E;
C - E - D

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

Results up to an indistinguishable class



True graph



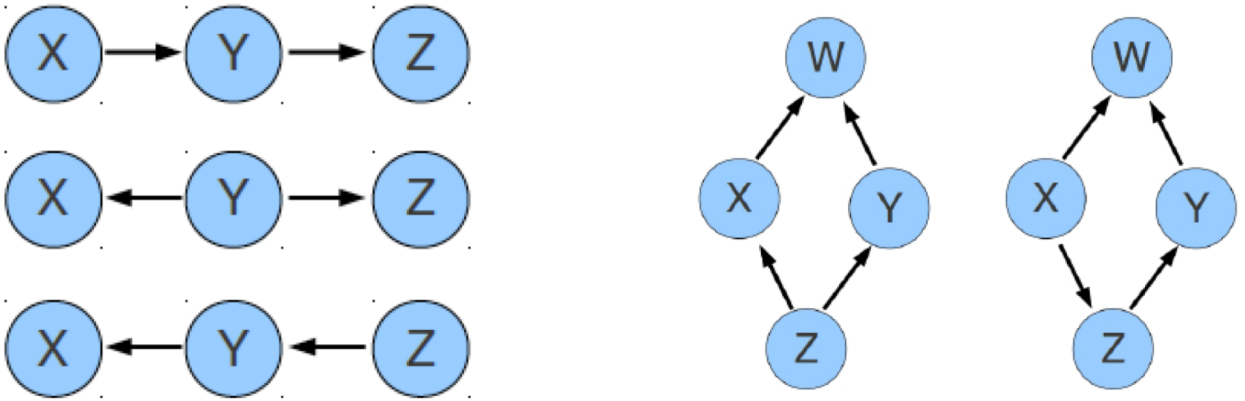
Result by PC algorithm

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

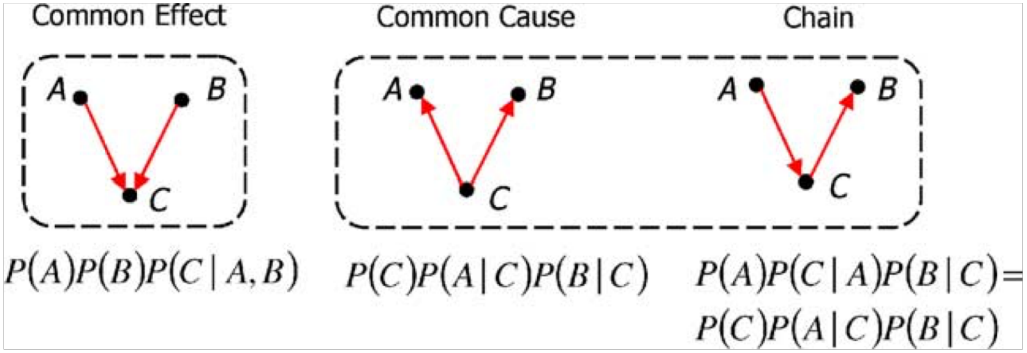
What is the complexity of the PC algorithm?

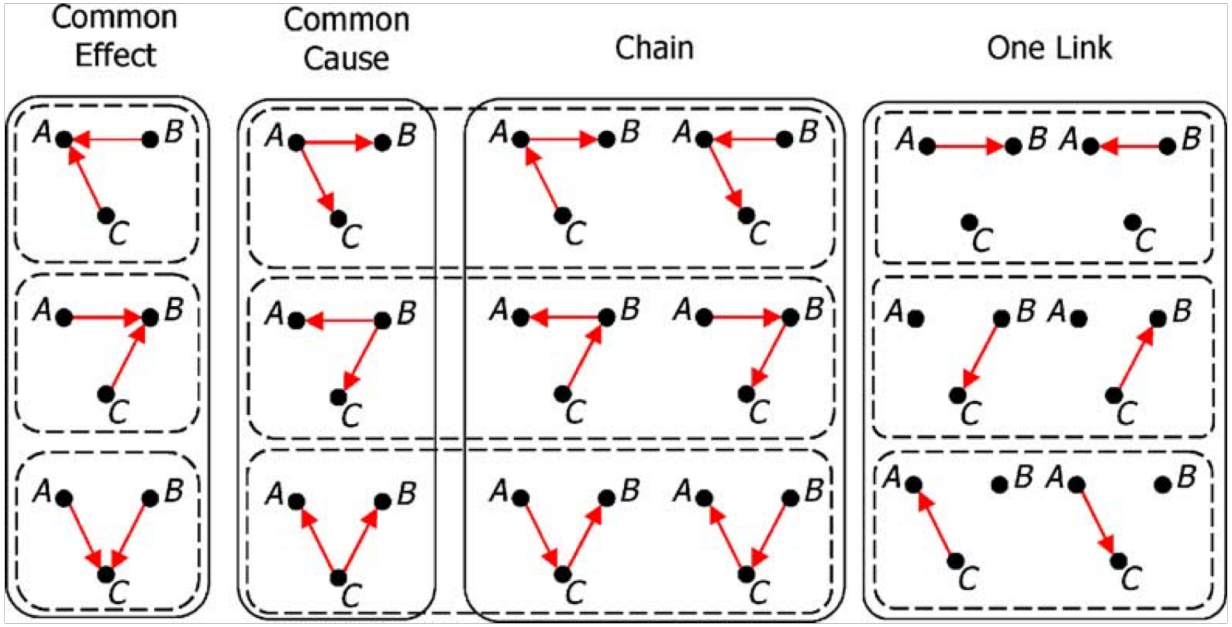
Markov equivalent class

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)

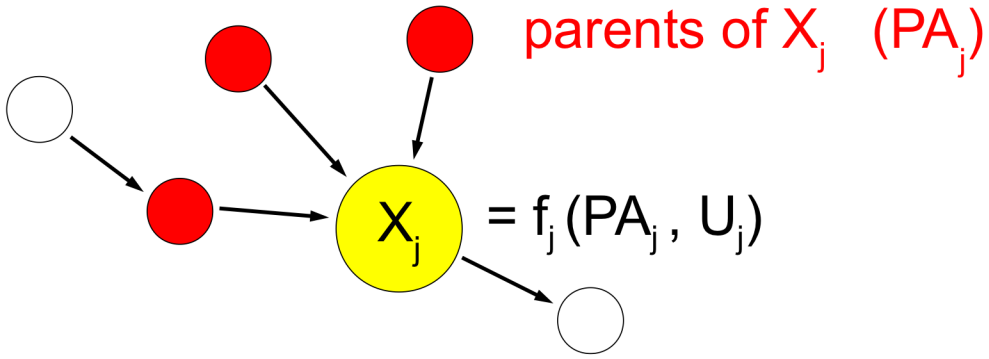




Solid lines group together networks of the same topological type.
Dashed lines delineate Markov equivalence classes.

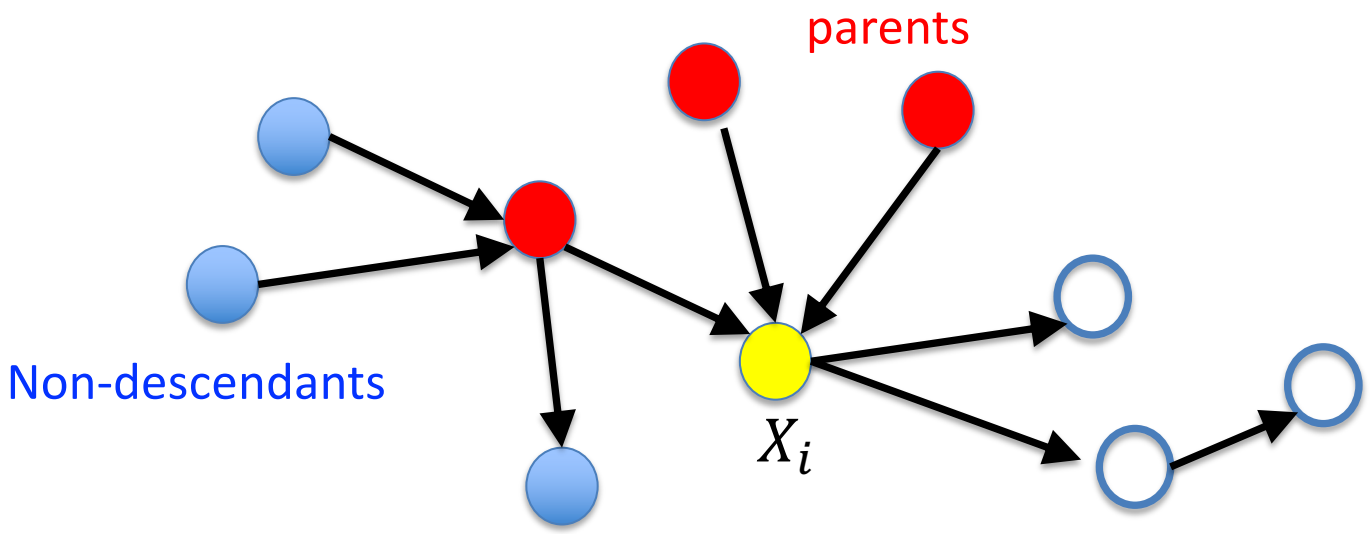
Functional Causal Model *(Pearl et al.)*

- 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

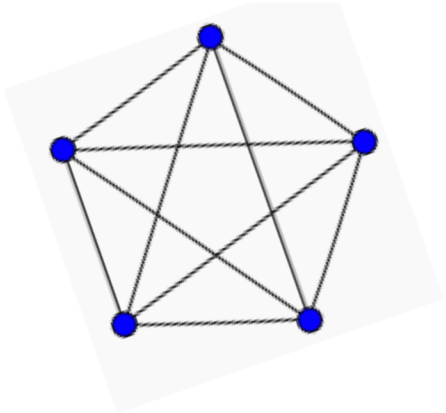
- Discover causal structure by conditional independence
  - PC algorithm
  - Markov equivalent class

- **Star causality structure**

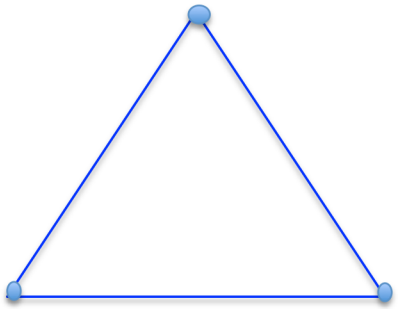
- A linear non-Gaussian model for causal discovery

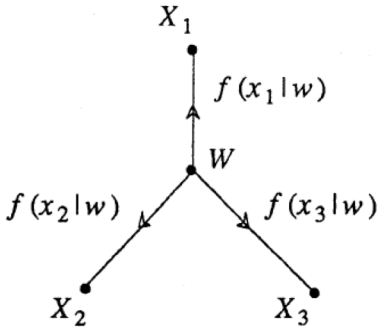
- Pearl's do-calculus

Gaussian networks



$$\Sigma = \begin{vmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{vmatrix} \quad \text{and} \quad \sigma_{ij} = \sigma_{ji}, \; i, j = 1, 2, 3$$





Xu L, Pearl J (1987) Structuring causal tree models with continuous variables. In: Proceedings of the 3rd annual conference on uncertainty in artificial intelligence. pp 170–179

STRUCTURING CAUSAL TREE MODELS WITH CONTINUOUS VARIABLES *

Lei Xu
Department of Automation
Tsinghua University
Beijing China

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
UCLA, Los Angeles, CA. 90024-1600

ABSTRACT

This paper considers the problem of invoking auxiliary, unobservable variables to facilitate the structuring of causal tree models for a given set of continuous variables. Paralleling the treatment of bi-valued variables in [Pearl 1986], we show that if a collection of coupled variables are governed by a joint normal distribution and a tree-structured representation exists, then both the topology and all internal relationships of the tree can be uncovered by observing pairwise dependencies among the observed variables (i.e., the leaves of the tree). Furthermore, the conditions for normally distributed variables are less restrictive than those governing bi-valued variables. The result extends the applications of causal tree models which were found useful in evidential reasoning tasks.

Causal structure determined by 2nd order statistics

Theorem 2:

1. A necessary and sufficient condition for three random variables with a joint normal distribution to be star-decomposable is that the correlation coefficients satisfy the inequalities:

$$\rho_{jk} \geq \rho_{ji}\rho_{ik} \qquad \rho_{ij}\,\rho_{ji}\,\rho_{ki} \geq 0 \tag{17}$$

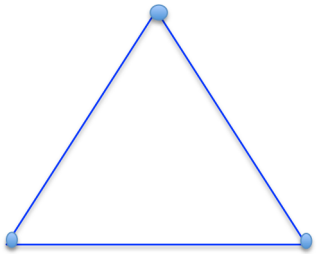
for all $i, j, k \in \{1, 2, 3\}$, and $i \neq j \neq k$.

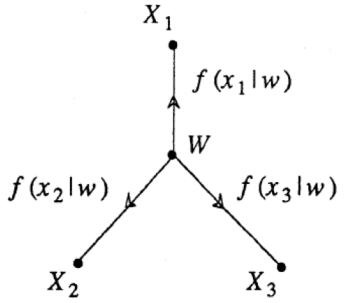
2. $f(x_i \mid w) \sim N(\mu_{i\mid w}, \sigma_{i\mid w})$, $i = 1,2,3$ are specified by the parameters

$$\sigma_{i\mid w} = \sigma_{ii}(1 - \rho_{iw}^2) = \sigma_{ii}(1 - \rho_{ji}\rho_{ik} / \rho_{jk})$$

$$\mu_{i\mid w} = \mu_i - \sigma_{wi}(w - \mu_w) / \sigma_{ww} = \mu_i - \rho_{wi}\sqrt{\frac{\sigma_{ij}}{\sigma_{ww}}}\,(w - \mu_w) \text{ and}$$

$f(w) \sim N(\mu_w, \sigma_w)$, where $\sigma_{ww} > 0$ and μ_w may be chosen arbitarily.





The proof of conditions for star-decomposability

Analogous with Section 3.2 in [Pearl, 1986], we can ask if $f(x_1, x_2, ..., x_n)$ can be represented as a marginal of an $n+1$ dimensional normal distribution of variables $\mathbf{x}_{n+1} = (w, \mathbf{x}_n^t)^t$ such that the x_i's are conditionally independent given w, i.e.

$$f(x_1, x_2, ..., x_n) = \int_{-\infty}^{+\infty} f_s(x_1, x_2, ..., x_n, w)\, dw \tag{2}$$

$$f_s(x_1, x_2, ..., x_n, w) = \prod_{i=1}^{n} f_s(x_i \mid w)\ f(w) \tag{3}$$

Where $f_s(x_i \mid w)$, $i = 1, ..., n$ relate each x_i to the central hidden variable w (see Figure 1).

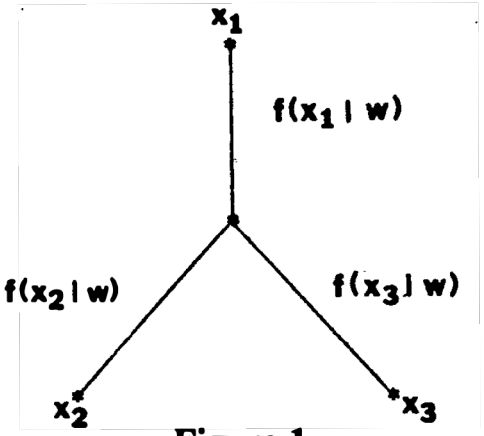


Figure 1

If the decomposition in (2) is possible, we name f_s a *star-distribution* and call f *star-decomposable*.

$$f(x_1, x_2, ..., x_n) = (2\pi)^{-\frac{n}{2}} (det \sum_n)^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2} (\mathbf{x}_n - \mu_n)^t \sum_n^{-1} (\mathbf{x}_n - \mu_n)\right]$$

Theorem 1: Let $\mathbf{x} = (\mathbf{x}_{(1)}^t, \mathbf{x}_{(2)}^t)^t$, $\mathbf{x}_{(1)}^t = (x_1 \cdots x_q)$, $\mathbf{x}_{(2)}^t = (x_{q+1} \cdots x_p)$
Let $\boldsymbol{\mu} = E\mathbf{x}$ be similarly partitioned as $\boldsymbol{\mu} = (\boldsymbol{\mu}_{(1)}, \boldsymbol{\mu}_{(2)})^t$ and let Σ be partitioned as

$$\Sigma = \begin{Bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{Bmatrix} \tag{5}$$

where Σ_{11} is the qxq upper left-hand corner submatrix of Σ. If $\mathbf{x}$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then

a. The vectors $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{x}_{(1)}$ are independently normally distributed with means $\boldsymbol{\mu}_{(1)}$, $\boldsymbol{\mu}_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\mu}_{(1)}$, and covariance matrices Σ_{11}, $\Sigma_{22-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, respectively.

b. The marginal distribution of $\mathbf{x}_{(1)}$ is q-variate normal with mean $\boldsymbol{\mu}_{(1)}$ and covariance matrix Σ_{11}.

c. The conditional distribution of $\mathbf{x}_{(2)}$ given $\mathbf{x}_{(1)}$ is normal with mean $\boldsymbol{\mu}_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)})$ and covariance matrix Σ_{22-1}.

The proof of the theorem is given in [Gigi, 1977, pp. 51-53] and it will be used in the next section.

Let $f(x_1, x_2, x_3)$ be a 3-dimensional joint normal distribution as in (1), for $n = 3$ with mean $\mu = (\mu_1\ \mu_2\ \mu_3)^t$ and covariance matrix

$$\Sigma = \begin{vmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{vmatrix} \quad \text{and} \quad \sigma_{ij} = \sigma_{ji},\ i, j = 1, 2, 3 \tag{6}$$

If $f(x_1, x_2, x_3)$ is star-decomposable, then it is a marginal of a 4-dimensional joint normal distribution $f_s(w, x_1, x_2, x_3)$ with mean $\mu_s = (\mu_w, \mu^t)^t$ and covariance matrix

$$\Sigma_s = \left|\begin{array}{c|ccc} \sigma_{ww} & \sigma_{1w} & \sigma_{2w} & \sigma_{3w} \\ \hline \sigma_{w1} & & & \\ \sigma_{w2} & & \Sigma & \\ \sigma_{w3} & & & \end{array}\right|$$

and

$$\sigma_{wi} = \sigma_{iw},\ i = 1, 2, 3$$

$$f(w, x_1, x_2, x_3) = f(x_1, x_2, x_3 \mid w)\, f(w) \tag{7}$$

$$f(x_1, x_2, x_3, \mid w) = f(x_1 \mid w) f(x_2 \mid w) f(x_3 \mid w) \tag{8}$$

Theorem 1 states that $f(x_1, x_2, x_3 \mid w)$, $f(w)$ and $f(x_i \mid w)$'s are also normal distributions, and the mean vector and covariance matrix of $f(x_1, x_2, x_3 \mid w)$ are given by

$$\mu_{1\cdot 2\cdot 3 \mid w} = \mu - \sigma_{ww}^{-1} \left[\sigma_{w_1}\ \sigma_{w_2}\ \sigma_{w_3}\right]^t (w - \mu_w) \tag{9}$$

$$\Sigma_{1\cdot 2\cdot 3 \mid w} = \Sigma - \sigma_{ww}^{-1} \left[\sigma_{w_1}\ \sigma_{w_2}\ \sigma_{w_3}\right]^t \left[\sigma_{w_1}\ \sigma_{w_2}\ \sigma_{w_3}\right] \tag{10}$$

Additionally, the conditional independence stated in (8) implies that $\sum_{1\cdot 2\cdot 3|w}$ must be a diagonal matrix, thus

$$\sigma_{ij} - \sigma_{wi}\sigma_{jw} / \sigma_{ww} = 0, \quad i \neq j \text{ and } i, j = 1, 2, 3. \tag{11}$$

and

$$\sigma_{ii} - \sigma_{iw}^2 / \sigma_{ww} > 0, \quad i = 1, 2, 3. \tag{12}$$

Using the correlation coefficients defined as

$$\rho_{ij} = \sigma_{ij} / (\sigma_{ii}\sigma_{ij})^{1/2} \tag{13}$$

(11) and (12) can be written as

$$\rho_{ij} = \rho_{iw}\rho_{jw}, \quad \text{for all } i, j \tag{14}$$

$$\rho_{iw}^2 \leq 1 \quad \text{for all } i. \tag{15}$$

Solving (14) for ρ_{iw}, we obtain

$$\rho_{1w} = (\rho_{12}\rho_{13} / \rho_{23})^{1/2} \quad \rho_{2w} = (\rho_{12}\rho_{23} / \rho_{13})^{1/2} \quad \rho_{3w} = (\rho_{13}\rho_{23} / \rho_{12})^{1/2} \tag{16}$$

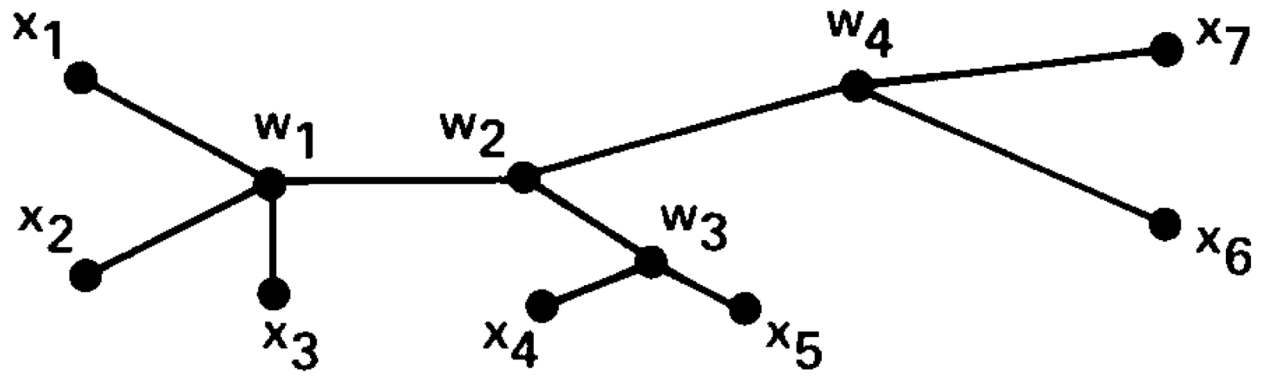
The requirement that the ρ_{iw}'s must be real numbers with magnitude not exceeding 1, yields the following two conditions for $f(x_1, x_2, x_3)$ to be star-decomposable:

a. $\rho_{12}, \rho_{13}, \rho_{23}$ are all positive, or two are negative and one is positive. In other words, the triplet (x_1, x_2, x_3) is positively correlated.

b. $\rho_{jk} \geq \rho_{ji}\rho_{ik}$ for all $i, j, k \in \{1,2,3\}$ and $i \neq j \neq k$

Summarizing the analysis above, we obtain Theorem 2.

Tree reconstruction, how?



We are now ready to confront the central problem of this section—given a tree-decomposable distribution $P(x_i, \ldots, x_n)$, can we uncover its underlying topology and the underlying tree-distribution $P_T(x_1, \ldots, x_n, w_1, \ldots, w_m)$?

Outline

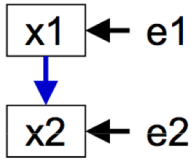
- Discover causal structure by conditional independence
  - PC algorithm
  - Markov equivalent class
- Star causality structure
- **A linear non-Gaussian model for causal discovery**
- Pearl's do-calculus

Linear Structural Equation Models

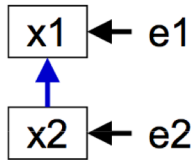
[Wright, 1921; Bollen, 1989]

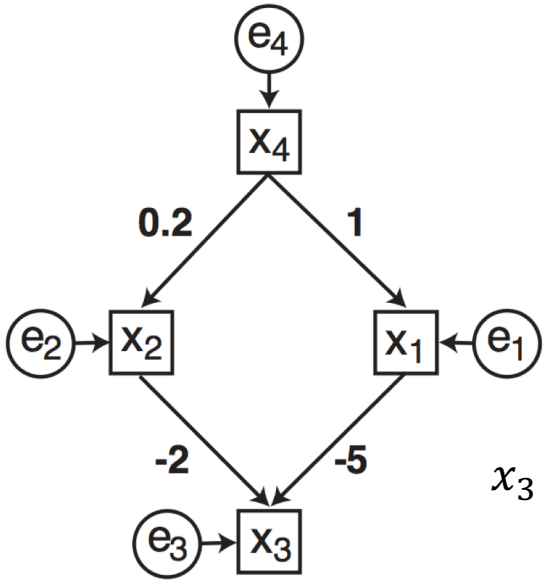
- Use linear SEM to model the data generation process

$$x_1 = e_1$$
$$x_2 = b_{21}x_1 + e_2$$



$$x_1 = b_{12}x_2 + e_1$$
$$x_2 = e_2$$





$$x_4 = e_4$$
$$x_2 = 0.2x_4 + e_2$$
$$x_1 = x_4 + e_1$$
$$x_3 = -2x_2 - 5x_1 + e_3$$

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

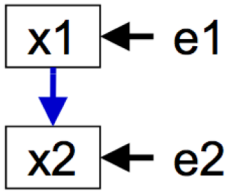
Shimizu et. al., JMLR 2006
Shimizu and Kawahara, UAI2010 Tutorial

Identify Which Model to Generate The Data

- Two models with Gaussian e_1 and e_2 :

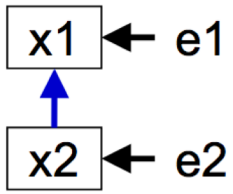
Model 1:

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



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

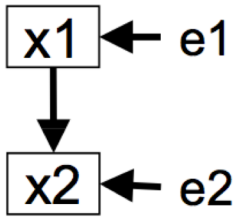
Gaussian

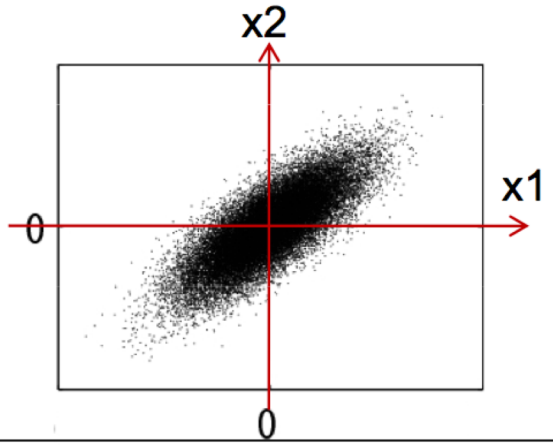
Non-Gaussian (uniform)

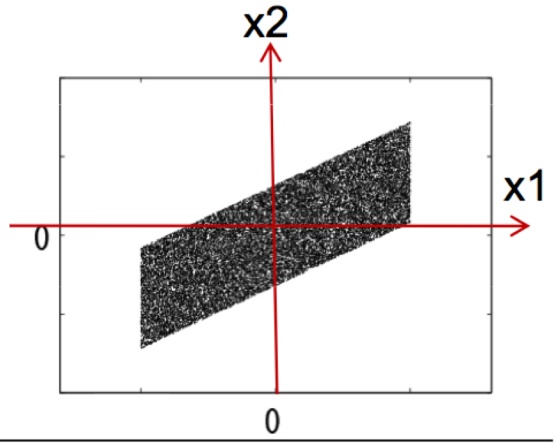
Model 1:

$x_1 = e_1$

$x_2 = 0.8x_1 + e_2$



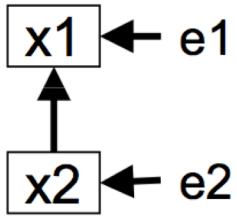


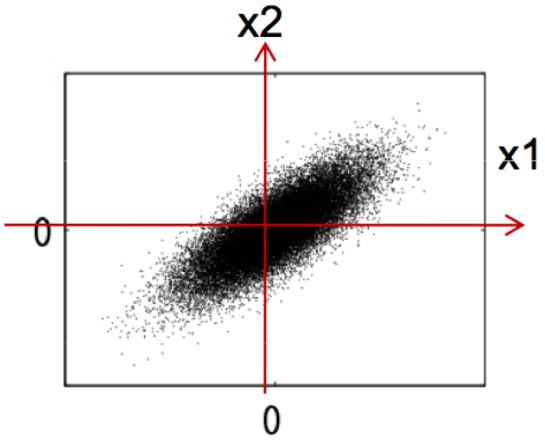


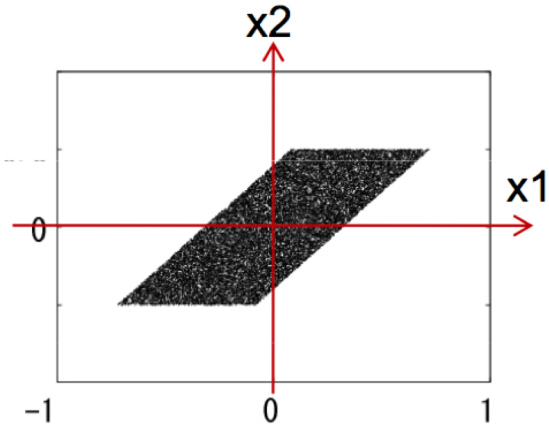
Model 2:

$x_1 = 0.8x_2 + e_1$

$x_2 = e_2$







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

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

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:

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

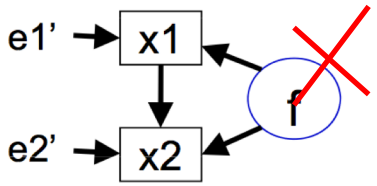
or

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

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



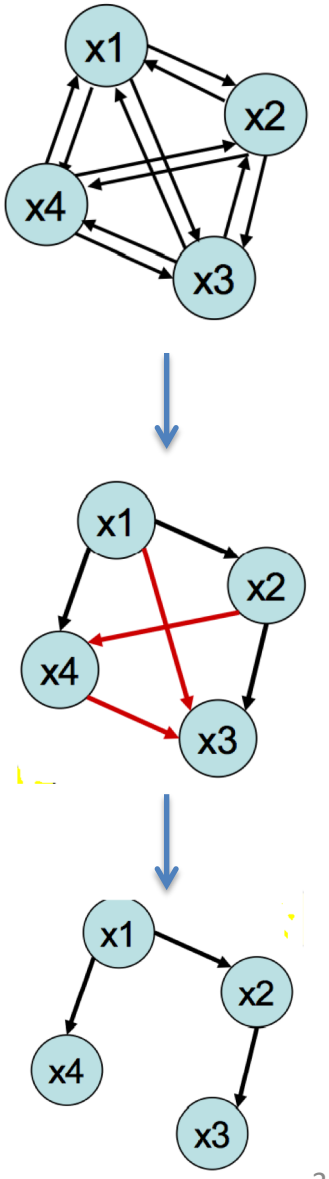
How *B* Is Estimated?

- 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)

Outline

- Discover causal structure by conditional independence
  - PC algorithm
  - Markov equivalent class
- Star causality structure
- A linear non-Gaussian model for causal discovery
- **Pearl's do-calculus**

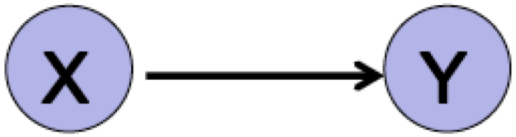
Pearl's do-calculus

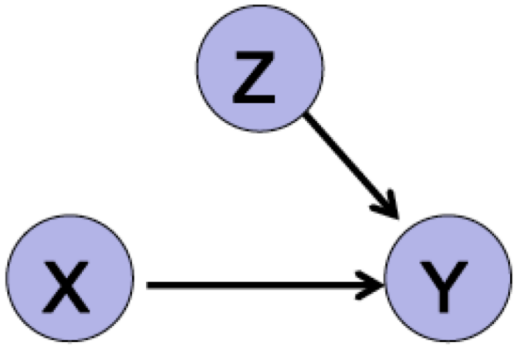
- Motivation: goal of causality is to infer the effect of interventions
- distribution of Y given that X is set to x: p(Y |do X = x) or p(Y |do x)

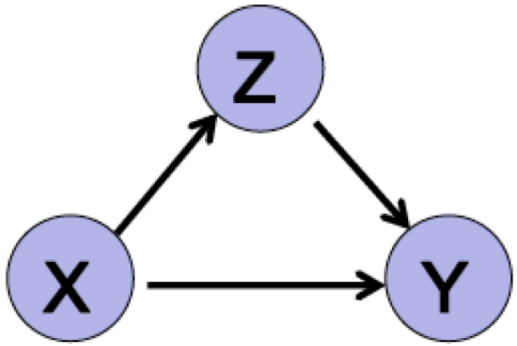
根据 do 操作，便可以定义因果作用，比如二值的变量 Z 对于 Y 的平均因果作用定义为

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

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

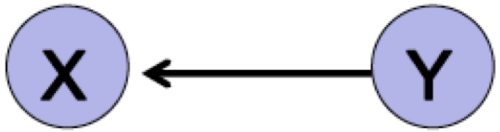




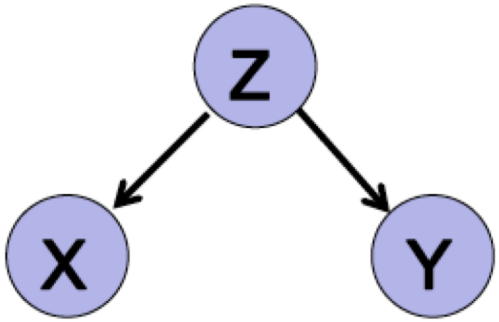


Examples for $p(.|do\,x) \neq p(.|x)$

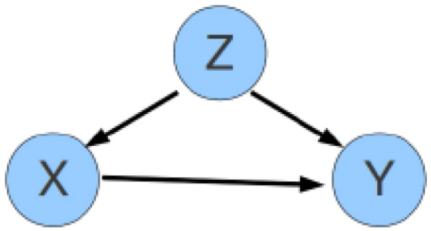
- $p(Y|do\,x) = P(Y) \neq P(Y|x)$



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



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!