

Causal Inference

Conditional Independences and Beyond

Dominik Janzing & Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Tübingen, Germany

<http://www.tuebingen.mpg.de/bs>



Roadmap

- informal motivation
- functional causal models
- causal graphical models;
d-separation, Markov conditions, faithfulness
- formalizing interventions
- causal inference...
 - using time order
 - using conditional independences
 - using restricted function classes
 - using “independence” of mechanisms
 - not using statistics



Dependence vs. Causation

Storks Deliver Babies ($p= 0.008$)

Robert Matthews

Article first published online: 25 DEC 2001

DOI: 10.1111/1467-9639.00013

Teaching Statistics Trust, 2000

Issue



Teaching Statistics
Volume 22, Issue 2, |
38, June 2000

Country	Area (km ²)	Storks (pairs)	Humans (10 ⁶)	Birth rate (10 ³ /yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	mailto:rajm@compuserve.com	
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries

Cargo cults— religious practices in pacific tribes around world war II, trying to obtain wealth (the "cargo") by building mock landing strips etc.

...the term cargo cult... is also idiomatically used (in the words of wikipedia) "to mean any group of people who imitate the superficial exterior of a process or system without having any understanding of the underlying substance".

(source: <http://philosophyisfashionable.blogspot.com/>)



MAX-PLANCK-GESellschaft



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.

N Engl J Med 2012; 366:1891-1904 | May 17, 2012

Abstract

Article

References

Citing Articles (1)

BACKGROUND

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

[Full Text of Background...](#)

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.

We present risk estimates separately for men and women. Multivariate models were adjusted for the following baseline factors: age; body-mass index (BMI); race or ethnic group; level of education; alcohol consumption; the number of cigarettes smoked per day, use or nonuse of pipes or cigars, and time of smoking cessation (<1 year, 1 to <5 years, 5 to <10 years, or ≥10 years before baseline); health status; presence or absence of diabetes; marital status; level of physical activity; total energy intake; consumption of fruits, vegetables, red meat, white meat, and saturated fat; and use of any vitamin supplement (yes vs. no). In addition, risk estimates for death from cancer were adjusted for history of cancer (other than nonmelanoma skin cancer) in a first-degree relative (yes vs. no). For women, status with respect to postmenopausal hormone therapy was also included in multivariate models. Less than 5% of the cohort lacked any single covariate; for each covariate, we

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. Adjusted hazard ratios for death among men who drank

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.

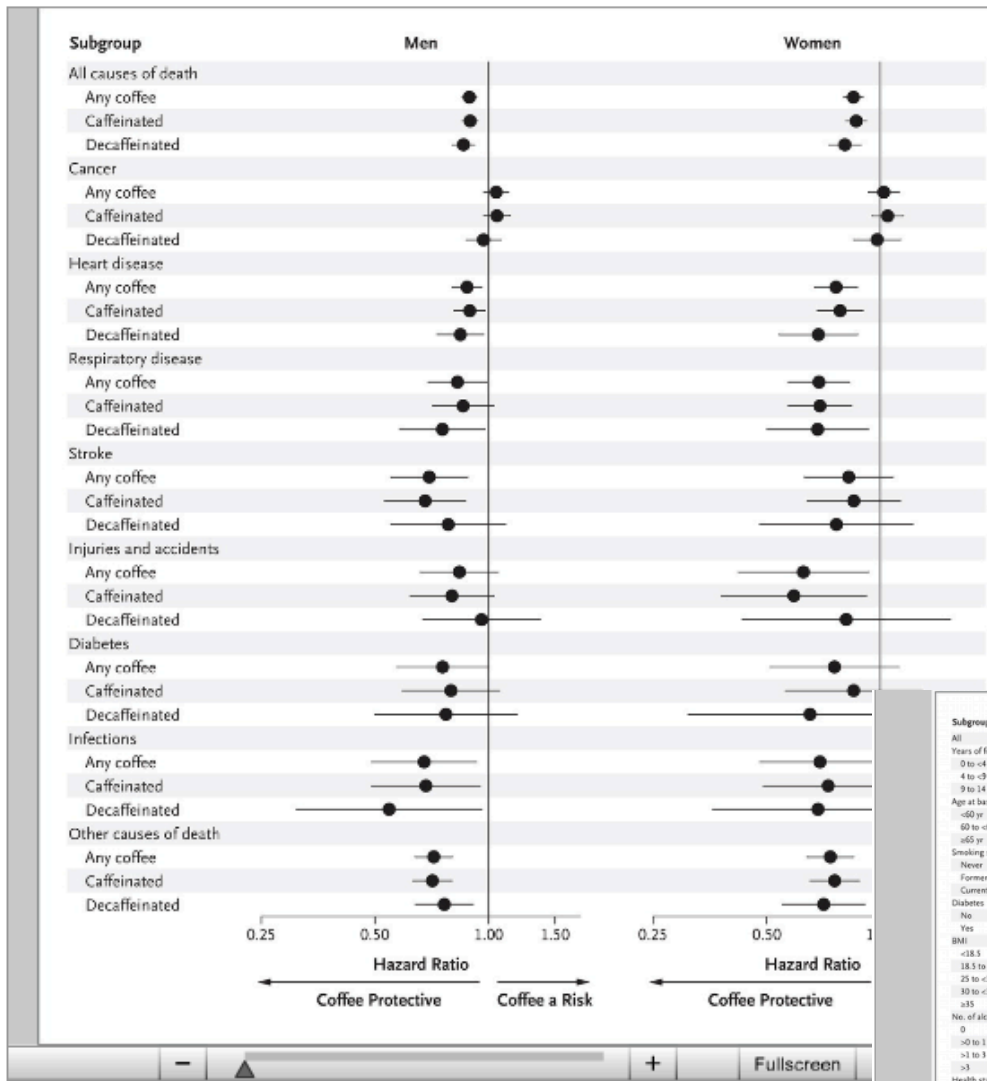


Figure 1. Subgroup Analysis of Associations between the Consumption of 4 or More Cups of Coffee per Day and Total and Cause-Specific Mortality.

Hazard ratios for death from all causes and from specific causes are for the comparison of men and women who drank 4 or more cups of coffee per day with those who did not drink coffee. Participants were classified as drinking caffeinated or decaffeinated coffee according to whether they reported drinking caffeinated or decaffeinated coffee more than half the time. Risk estimates for other categories of coffee consumption are shown in Tables 2 and 3 in the **Supplementary Appendix**. Risk estimates were adjusted for the following factors at baseline: age; body-mass index; race or ethnic group; level of education; alcohol consumption; the number of cigarettes smoked per day, use or nonuse of pipes or cigars, and time of smoking cessation (<1 year, 1 to <5 years, 5 to <10 years, or 10 years

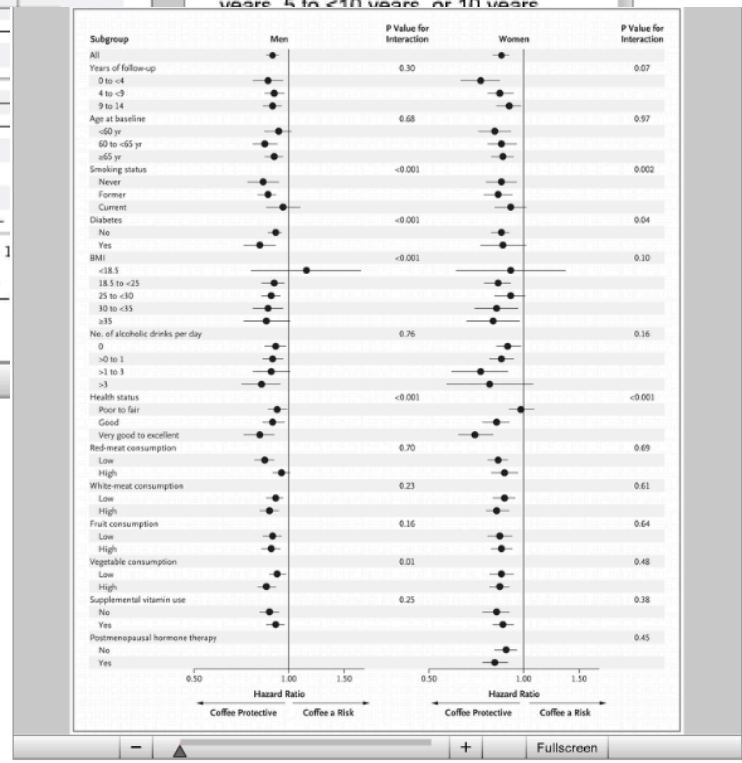


Figure 2. Subgroup Analysis of Associations between the Consumption of 4 or More Cups of Coffee per Day and Total Mortality.

Hazard ratios for death from any cause are for the comparison of men and women who drank 4 or more cups of coffee per day with those who did not drink coffee. The multivariate model was adjusted for the following factors at baseline: age; body-mass index (BMI); the weight in kilograms divided by the square of the height in meters; race or ethnic group; level of education; alcohol consumption; the number of cigarettes smoked per day, use or nonuse of pipes or cigars, and time of smoking cessation (<1 year, 1 to <5 years, 5 to <10 years, or 10 years before baseline); health status; diabetes (yes vs. no); marital status; physical activity; total energy intake; consumption of fruits, vegetables, red meat, white meat, and saturated fat; use or nonuse of vitamin supplements; and, in women, use or nonuse of postmenopausal hormone therapy. Risk estimates for other categories of coffee consumption are shown in Tables 4 and 5 in the **Supplementary Appendix**. High and low dietary-intake categories are split at the median. Horizontal lines represent 95% confidence intervals. P values for interactions were computed with the use of likelihood-ratio tests comparing Cox proportional-hazards models with and without cross-product terms for each level of baseline stratifying variables, with coffee consumption as an ordinal variable. P values for the years of follow-up were derived from testing the addition of a cross-product

12.12.2007

Deutsches Kinderkrebsregister untersucht Häufigkeit von Krebserkrankungen bei Kindern in der Nähe von Kernkraftwerken

Neue Studie veröffentlicht

Immer wieder wird der Verdacht geäußert, dass Kinder in der Nähe von Kernkraftwerken häufiger an Krebs erkranken. Eine frühere Studie des Kinderkrebsregisters mit Kindern unter 15 Jahren schien darauf hinzudeuten, dass speziell in den ersten Lebensjahren das Leukämie-Risiko in den betreffenden Gegenden erhöht war.

In diesen Tagen erscheinen zwei wissenschaftliche Veröffentlichungen über eine neue Studie des Deutschen Kinderkrebsregisters in Mainz. Das Ergebnis: In Deutschland findet man einen Zusammenhang zwischen der Nähe der Wohnung zu einem Kernkraftwerk und der Häufigkeit, mit der Kinder vor ihrem fünften Geburtstag an Krebs und besonders an Leukämie erkranken. Allerdings erlaubt die Studie keine Aussage darüber, wodurch sich die beobachtete Erhöhung der Anzahl von Kinderkrebsfällen in der Umgebung deutscher Kernkraftwerke erklären lässt. So kommt nach dem heutigen Wissensstand Strahlung, die von Kernkraftwerken im Normalbetrieb ausgeht, als Ursache für die beobachtete Risikoerhöhung nicht in Betracht. Denkbar wäre, dass bis jetzt noch unbekannte Faktoren beteiligt sind oder dass es sich doch um Zufall handelt.

✉ **Kontakt**

Dr. Peter Kaatsch
(Leiter des Deutschen
Kinderkrebsregisters)
Dt. Kinderkrebsregister am IMBEI
Tel +49 6131 17-3111

[E-Mail](#)

[Homepage](#)

Prof. Dr. Maria Blettner
(Direktorin des IMBEI)
Institut für Medizinische
Biometrie, Epidemiologie und
Informatik (IMBEI)
Tel +49 6131 17-3252

[E-Mail](#)

“Correlation does not tell us anything about causality”

- Better to talk of dependence than correlation
- Most statisticians would agree that causality does tell us something about dependence
- But dependence does tell us something causality too:



Statistical Implications of Causality

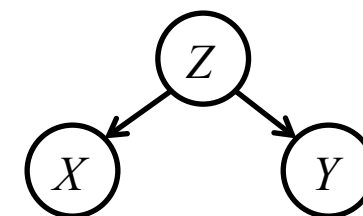
Reichenbach's

Common Cause Principle

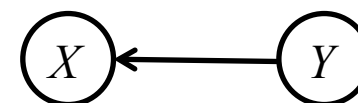
links **causality** and **probability**:

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

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



special cases:



Notation

- A, B event
- X, Y, Z random variable
- x value of a random variable
- \Pr probability measure
- P_X probability distribution of X
- p density
- p_X or $p(X)$ density of P_X
- $p(x)$ density of P_X evaluated at the point x
- always assume the existence of a joint density, w.r.t. a product measure

Independence

Two events A and B are called *independent* if

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B).$$

A_1, \dots, A_n are called *independent* if for every subset $S \subset \{1, \dots, n\}$ we have

$$\Pr\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \Pr(A_i).$$

Note: for $n \geq 3$, *pairwise independence* $\Pr(A_i \cap A_j) = \Pr(A_i) \cdot \Pr(A_j)$ for all i, j does not imply *independence*.

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. $p(z) > 0$.

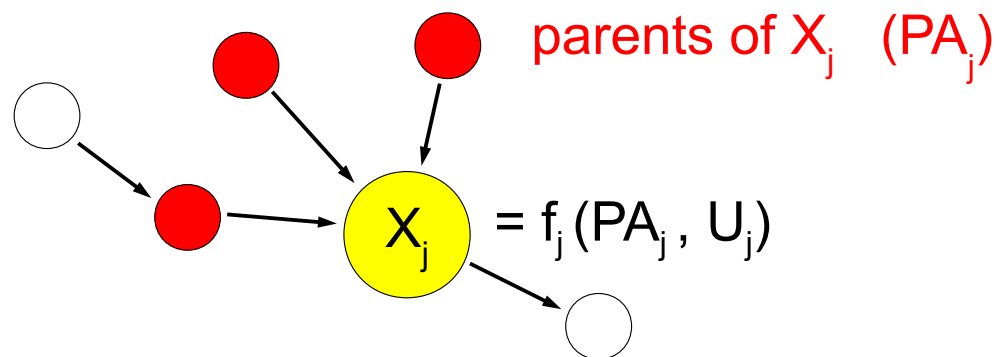
Note: conditional independence neither implies nor is implied by independence.

I.e., there are X, Y, Z such that we have only independence or only conditional independence.



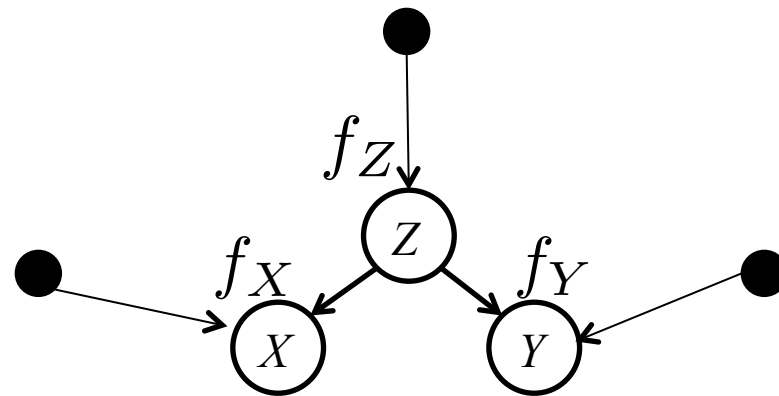
Functional Causal Model (*Pearl et al.*)

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



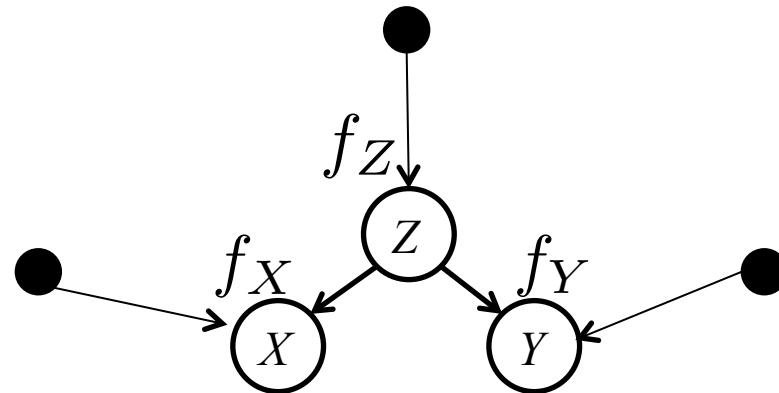
Functional Causal Model, ctd.

- this model can be shown to satisfy Reichenbach's principle:
 1. functions of independent variables are independent, hence dependence can only arise in two vertices that depend (partly) on the same noise term(s).
 2. if we condition on these noise terms, the variables become independent



Functional Causal Model, ctd.

- Independence of noises is a form of "causal sufficiency:" if the noises were dependent, then Reichenbach's principle would tell us the causal graph is incomplete
- Interventions are realized by replacing functions by values



- the model entails a joint distribution $p(X_1, \dots, X_n)$. **Questions:**
 - (1) What can we say about it?
 - (2) Can we recover G from p ?

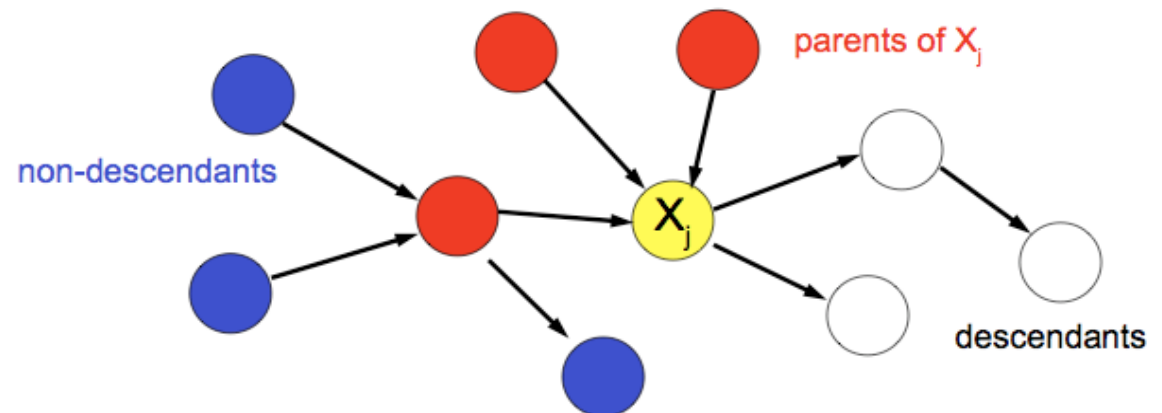
Functional Model and Markov conditions

(Lauritzen 1996, Pearl 2000)

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 $p(X_1, \dots, X_n) = \prod_j p(X_j \mid \text{Parents}_j)$ (conditionals as causal mechanisms generating statistical dependence)

(subject to technical conditions)



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 :

$$p(Y|do X = x) \quad \text{or} \quad p(Y|do x)$$

- can be computed from p and G

Computing $p(X_1, \dots, X_n | do x_i)$

from $p(X_1, \dots, X_n)$ and G

- Start with causal factorization

$$p(X_1, \dots, X_n) = \prod_{j=1}^n p(X_j | PA_j)$$

- Replace $p(X_i | PA_i)$ with $\delta_{X_i x_i}$

$$p(X_1, \dots, X_n | do x_i) := \prod_{j \neq i} p(X_j | PA_j) \delta_{X_i x_i}$$

Computing $p(X_k | do x_i)$

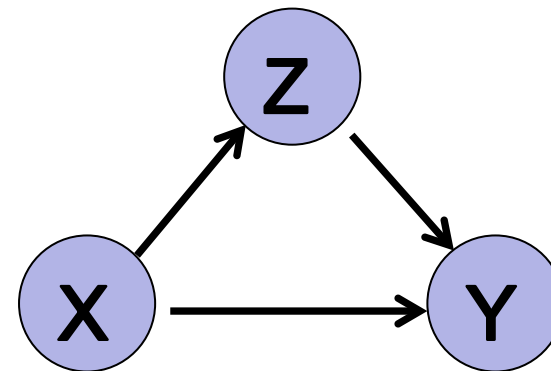
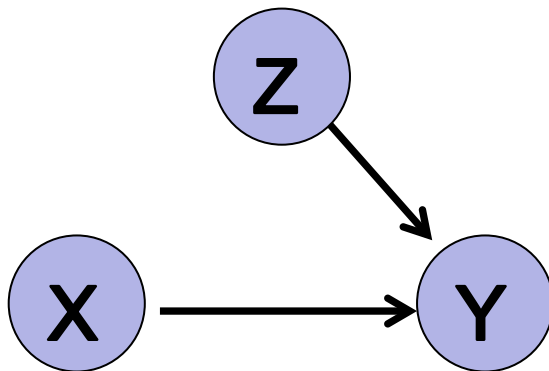
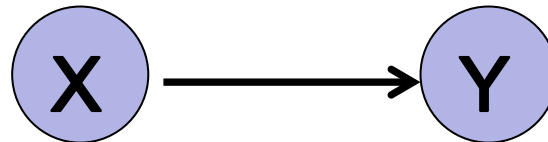
summation over x_i yields

$$p(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n | do x_i) = \prod_{j \neq i} p(X_j | PA_j(x_i)).$$

- distribution of X_j with $j \neq i$ is given by dropping $p(X_i | PA_i)$ and substituting x_i into PA_j to get $PA_j(x_i)$.
- obtain $p(X_k | do x_i)$ by marginalization

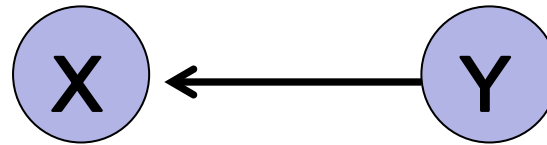


Examples for $p(.|do x) = p(.|x)$

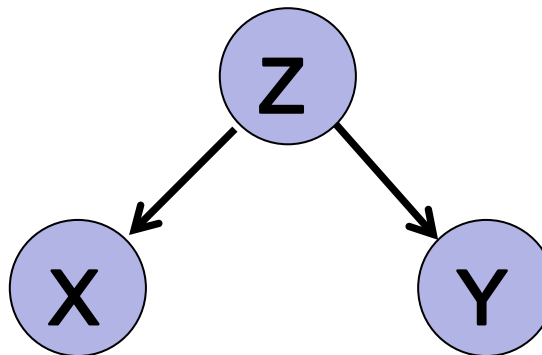


Examples for $p(.|do x) \neq p(.|x)$

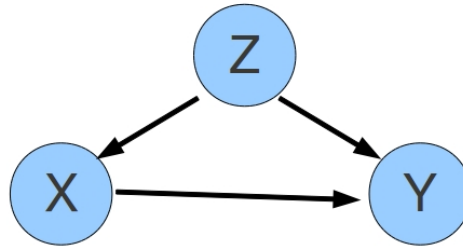
- $p(Y|do x) = P(Y) \neq P(Y|x)$



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



Example: controlling for confounding



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

- 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

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

Identifiability problem

e.g. Tian & Pearl (2002)

- given the causal DAG G and two nodes X_i, X_j
- which nodes need to be observed to compute $p(X_i | do x_j)$?



Inferring the DAG

- Key postulate: Causal Markov condition
- Essential mathematical concept: d-separation
(describes the conditional independences required by a causal DAG)



d-separation (Pearl 1988)

Path = sequence of pairwise distinct nodes where consecutive ones are adjacent

A path q is said to be **blocked** by the set Z if

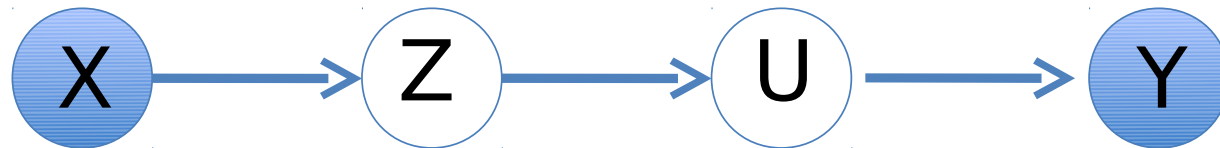
- q contains a *chain* $i \rightarrow m \rightarrow j$ or a *fork* $i \leftarrow m \rightarrow j$ such that the middle node is in Z , or
- q contains a *collider* $i \rightarrow m \leftarrow j$ such that the middle node is not in Z and such that no descendant of m is in Z .

Z is said to **d-separate** X and Y in the DAG G , formally

$$(X \perp\!\!\!\perp Y \mid Z)_G$$

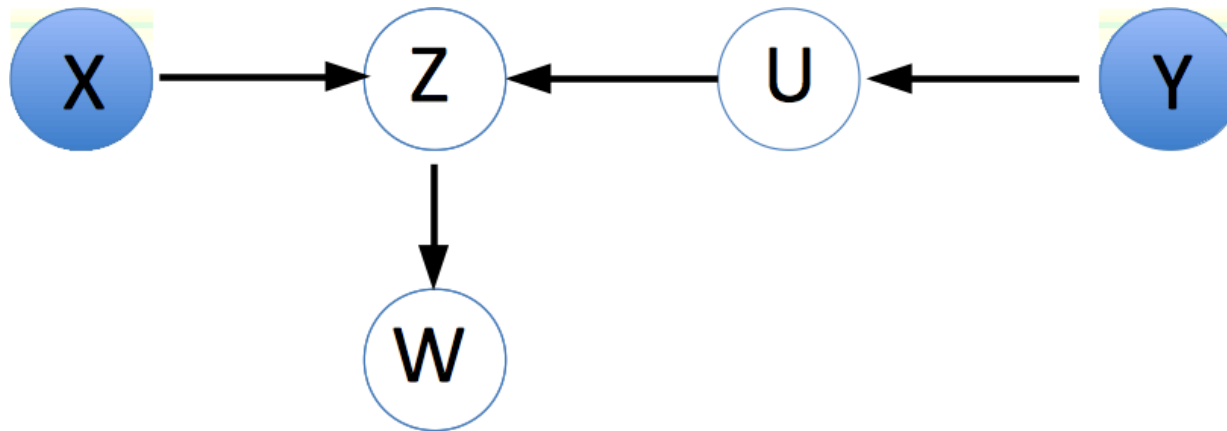
if Z blocks every path from a node in X to a node in Y .

Example (blocking of paths)



path from X to Y is blocked by conditioning on U or Z or both

Example (unblocking of paths)

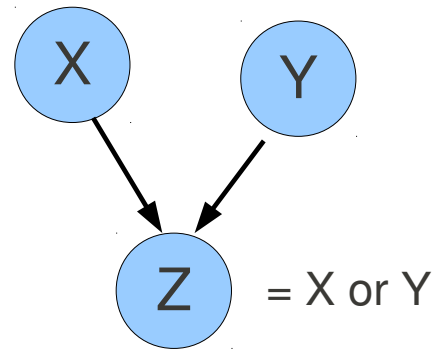


- path from X to Y is blocked by \emptyset
- unblocked by conditioning on Z or W or both

Unblocking by conditioning on common effects

Berkson's paradox (1946)

Example: X, Y, Z binary

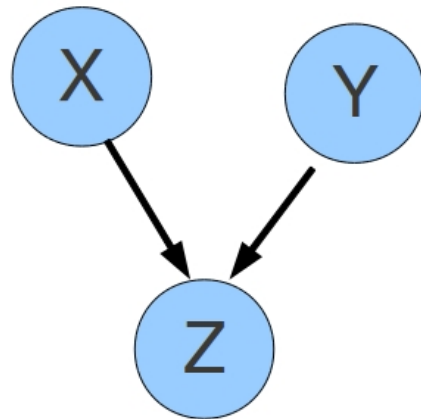


$$X \perp\!\!\!\perp Y \quad \text{but} \quad X \not\perp\!\!\!\perp Y | Z$$

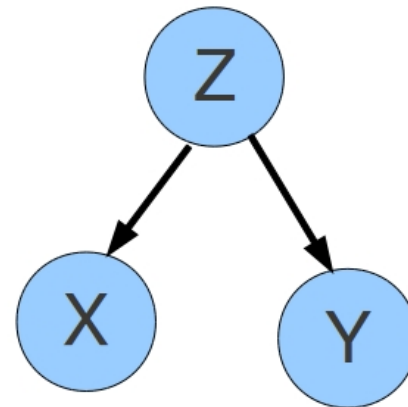
- assume: in paper submissions quality of theory is independent of quality of experiments
- papers get accepted for ICML if experiments or theory are strong
- for accepted ICML-papers, quality of theory and experiments are negatively correlated

Asymmetry under inverting arrows

(Reichenbach 1956)



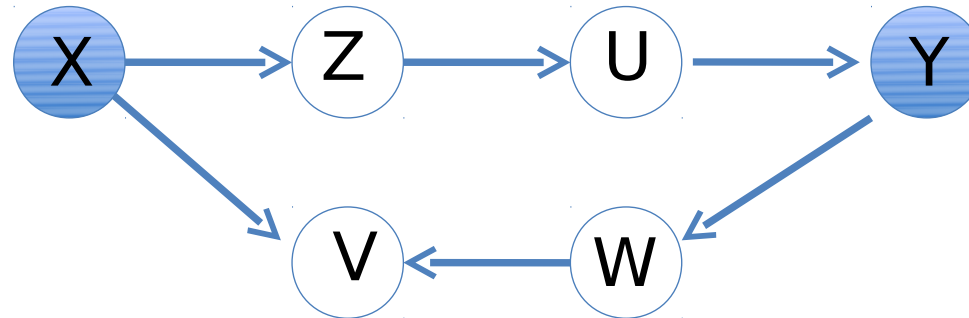
$$X \perp\!\!\!\perp Y$$
$$X \not\perp\!\!\!\perp Y \mid Z$$



$$X \not\perp\!\!\!\perp Y$$
$$X \perp\!\!\!\perp Y \mid Z$$



Examples (d-separation)



$$(X \perp\!\!\!\perp Y \mid ZW)_G$$

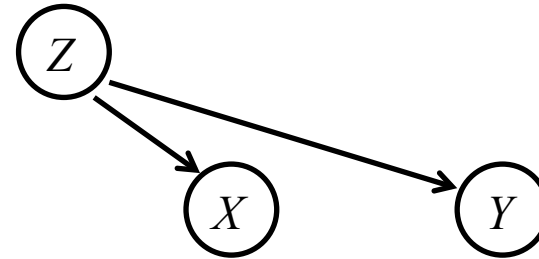
$$(X \perp\!\!\!\perp Y \mid ZUW)_G$$

$$(X \perp\!\!\!\perp Y \mid VZUW)_G$$

$$(X \not\perp\!\!\!\perp Y \mid VZU)_G$$

Causal inference for time-ordered variables

assume $X \not\perp\!\!\!\perp Y$ and X earlier. Then $X \leftarrow Y$ excluded, but still two options:



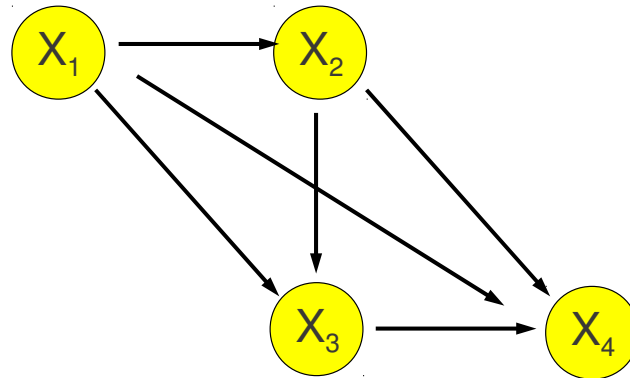
Example (Fukumizu 2007): barometer falls before it rains, but it does not cause the rain

Conclusion: time order makes causal problem (slightly?) easier but does not solve it

Causal inference for time-ordered variables

assume X_1, \dots, X_n are time-ordered and **causally sufficient**

- start with complete DAG



- remove as many parents as possible:

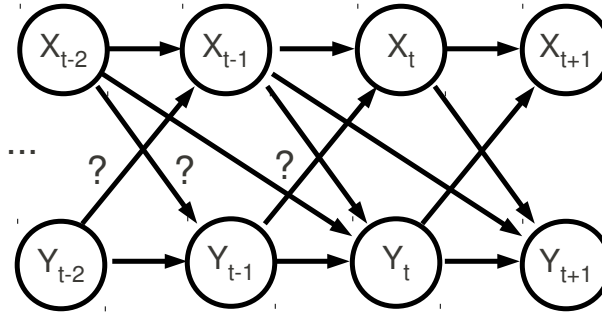
$p \in PA_j$ can be removed if

$$X_j \perp\!\!\!\perp p \mid PA_j \setminus p$$

(going from potential arrows to true arrows “only” requires statistical testing)

Time series and Granger causality

Does X cause Y and/or Y cause X ?



exclude instantaneous effects and common causes

- if

$$Y_t \not\perp\!\!\!\perp X_{[t-1, -\infty)} \mid Y_{[t-1, -\infty)}$$

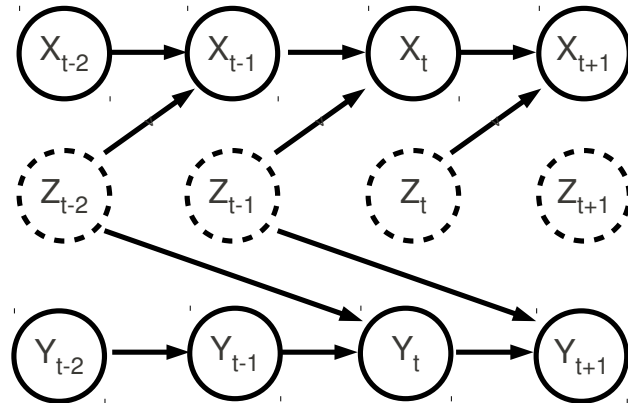
there must be arrows from X to Y (otherwise d-separation)

- Granger (1969): the past of X helps when predicting Y_t from its past
- strength of causal influence often measured by transfer entropy

$$I(Y_t; X_{[t-1, -\infty)} \mid Y_{[t-1, -\infty)})$$

Confounded Granger

Hidden common cause Z relates X and Y



due to different time delays we have

$$Y_t \not\perp\!\!\!\perp X_{[t-1, -\infty)} \mid Y_{[t-1, -\infty)}$$

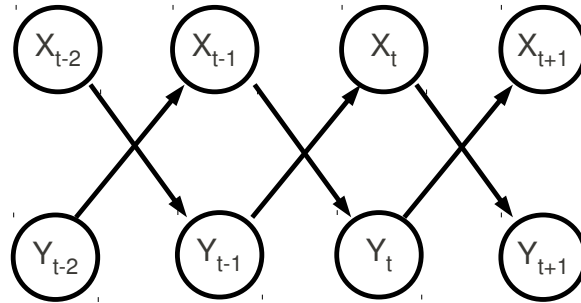
but

$$X_t \perp\!\!\!\perp Y_{[t-1, -\infty)} \mid X_{[t-1, -\infty)}$$

Granger infers $X \rightarrow Y$

Why transfer entropy does not quantify causal strength (Ay & Polani, 2008)

deterministic mutual influence between X and Y



- although the influence is strong

$$I(Y_t; X_{[t-1, -\infty)} | Y_{[t-1, -\infty)}) = 0,$$

because the past of Y already determines its future

- quantitatively still wrong for non-deterministic relation
- recent preprint on definitions of causal strength: Janzing, Balduzzi, Grosse-Wentrup, Schölkopf 2012

Inferring the causal DAG without time information

- Setting: given observed n -tuples drawn from $p(X_1, \dots, X_n)$, infer G
- Key postulates: Causal Markov condition and causal faithfulness

Causal faithfulness

Spirtes, Glymour, Scheines



p is called faithful relative to G if only those independences hold true that are implied by the Markov condition, i.e.,

$$(X \perp\!\!\!\perp Y | Z)_G \iff (X \perp\!\!\!\perp Y | Z)_p$$

Recall: Markov condition reads

$$(X \perp\!\!\!\perp Y | Z)_G \implies (X \perp\!\!\!\perp Y | Z)_p$$



Examples of unfaithful distributions (1)

Cancellation of direct and indirect influence in linear models

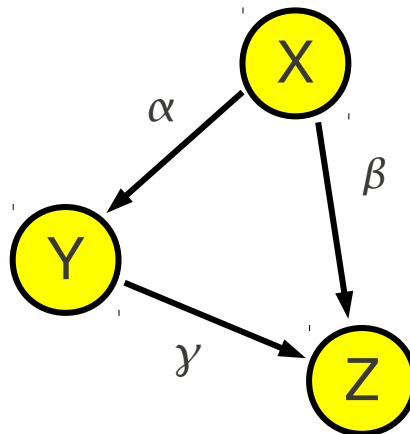
$$X = U_X$$

$$Y = \alpha X + U_Y$$

$$Z = \beta X + \gamma Y + U_Z$$

with independent noise terms U_X, U_Y, U_Z

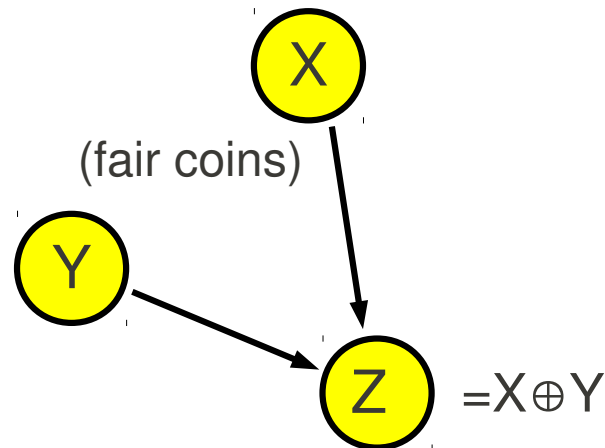
$$\beta + \alpha\gamma = 0 \quad \Rightarrow \quad X \perp\!\!\!\perp Z$$



Examples of unfaithful distributions (2)

binary causes with XOR as effect

- for $p(X), p(Y)$ uniform: $X \perp\!\!\!\perp Z, Y \perp\!\!\!\perp Z$.
i.e., unfaithful (since X, Z and Y, Z are connected in the graph).
- for $p(X), p(Y)$ non-uniform: $X \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z$.
i.e., faithful



unfaithfulness considered unlikely because it only occurs for non-generic parameter values

Conditional-independence based causal inference

Spirtes, Glymour, Scheines and Pearl

Causal Markov condition + Causal faithfulness:

- accept only those DAGs G as causal hypotheses for which

$$(X \perp\!\!\!\perp Y | Z)_G \iff (X \perp\!\!\!\perp Y | Z)_p .$$

- identifies causal DAG up to Markov equivalence class
(DAGs that imply the same conditional independences)



Markov equivalence class

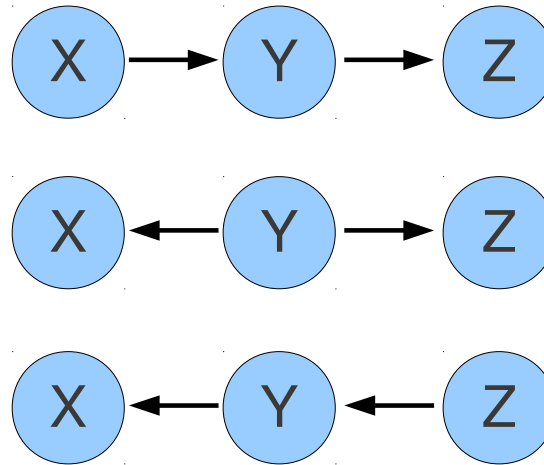
Theorem (Verma and Pearl, 1990): two DAGs are Markov equivalent iff they have the same skeleton and the same v -structures.

skeleton: corresponding undirected graph

v -structure: substructure $X \rightarrow Y \leftarrow Z$ with no edge between X and Z



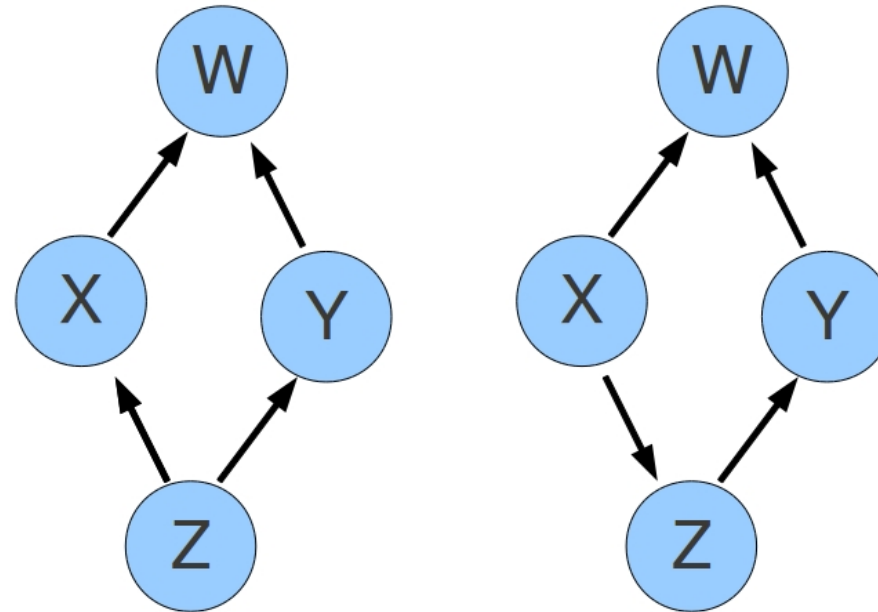
Markov equivalent DAGs



same skeleton, no v -structure

$$X \perp\!\!\!\perp Z \mid Y$$

Markov equivalent DAGs



same skeleton, same v-structure at W

Algorithmic construction of causal hypotheses

IC algorithm by Verma & Pearl (1990) to reconstruct DAG from p

idea:

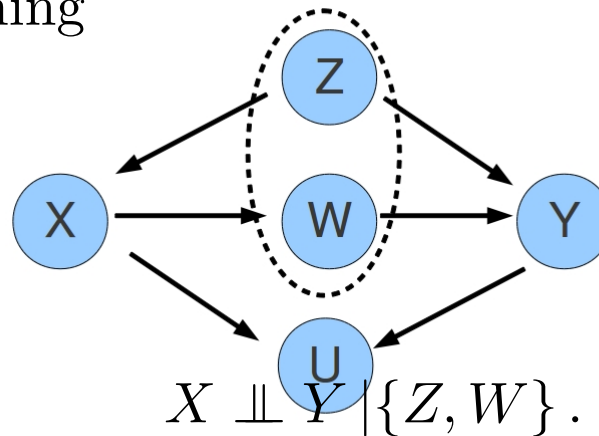
1. Construct skeleton
2. Find v-structures
3. direct further edges that follow from
 - graph is acyclic
 - all v-structures have been found in 2)

Construct skeleton

Theorem: X and Y are linked by an edge iff there is no set S_{XY} that d-separates them,

$$(X \perp\!\!\!\perp Y \mid S_{XY})_G .$$

Explanation: dependence that is due to indirect links can be screened off by conditioning



Faithfulness implies: edge $X - Y$ exists iff there is a set $S_{X,Y}$ such that

$$(X \perp\!\!\!\perp Y \mid S_{XY})_p .$$

S_{XY} is called a Sepset for X, Y

Efficient construction of skeleton

PC algorithm by Spirtes & Glymour (1991)

iteration over size of Sepset

1. remove all edges $X - Y$ with $X \perp\!\!\!\perp Y$
2. remove all edges $X - Y$ for which there is a neighbor $Z \neq Y$ of X with $X \perp\!\!\!\perp Y \mid Z$
3. remove all edges $X - Y$ for which there are two neighbors $Z_1, Z_2 \neq Y$ of X with $X \perp\!\!\!\perp Y \mid Z_1, Z_2$
4. ...

Advantages

- many edges can be removed already for small sets
- testing all sets S_{XY} containing the adjacencies of X is sufficient
- depending on sparseness, algorithm only requires independence tests with small conditioning tests
- polynomial for graphs of bounded degree

Find v-structures

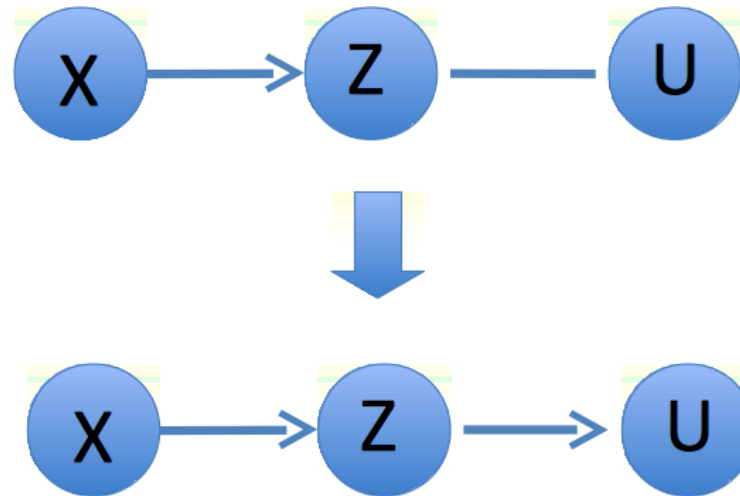
- given $X - Y - Z$ with X and Y non-adjacent
- given S_{XY} with $X \perp\!\!\!\perp Y \mid S_{XY}$

a priori, there are 4 possible orientations:

$$\left. \begin{array}{l} X \rightarrow Z \rightarrow Y \\ X \leftarrow Z \rightarrow Y \\ X \leftarrow Z \leftarrow Y \end{array} \right\} Z \in S_{XY}$$
$$X \rightarrow Z \leftarrow Y \quad Z \notin S_{XY}$$

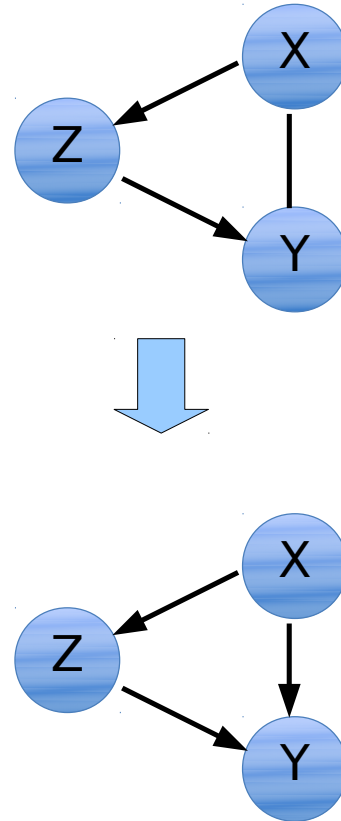
Orientation rule: create v-structure if $Z \notin S_{XY}$

Direct further edges (Rule 1)



(otherwise we get a new v-structure)

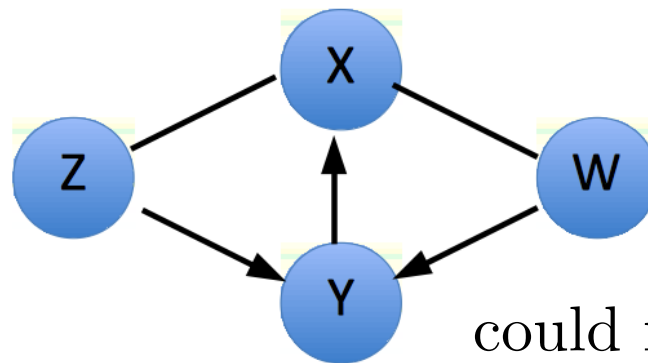
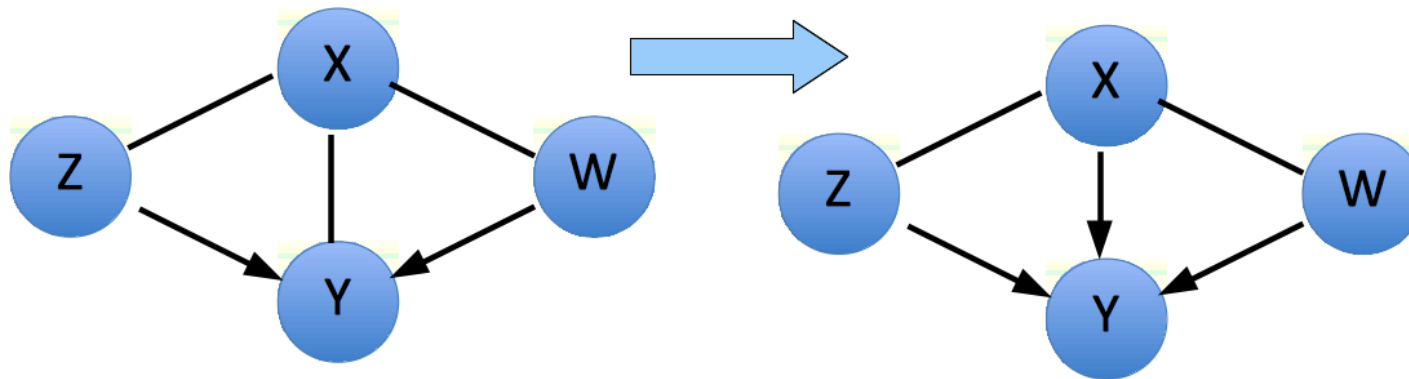
Direct further edges (Rule 2)



(otherwise one gets a cycle)

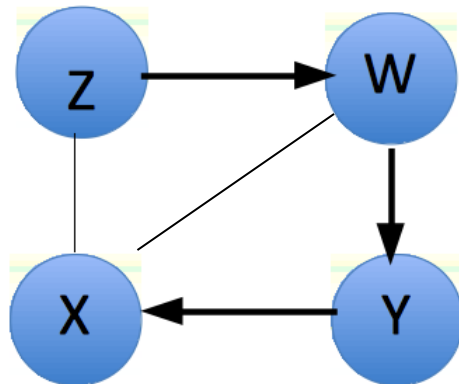
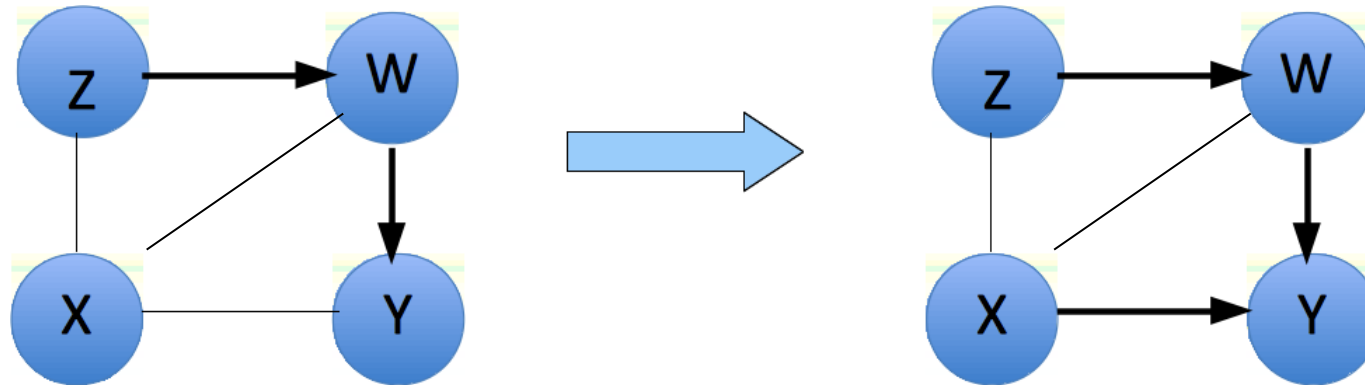


Direct further edges (Rule 3)



could not be completed
without creating a new v-structure

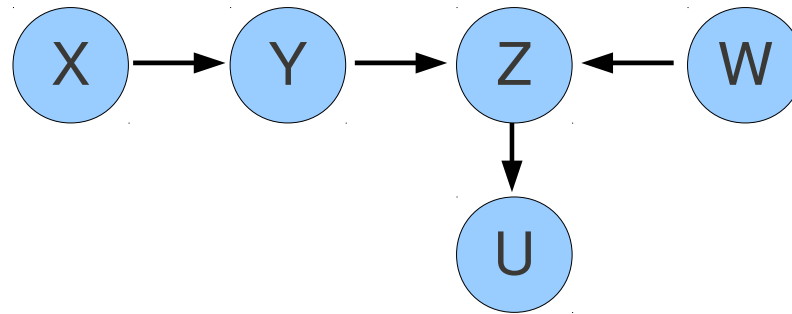
Direct further edges (Rule 4)



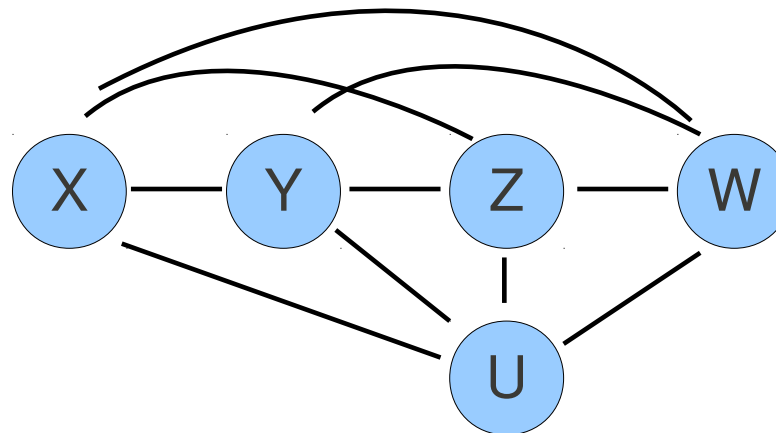
could not be completed
without creating a cycle
or a new v-structure

Examples

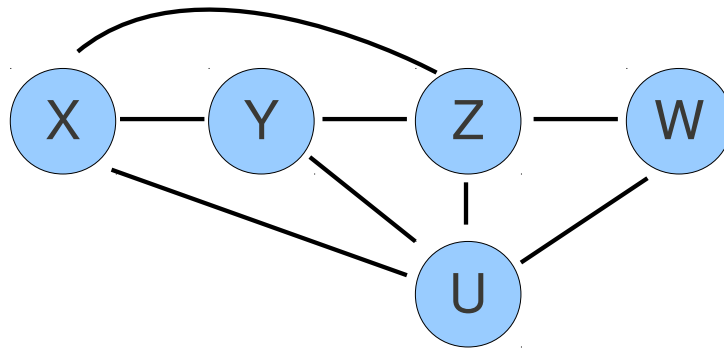
(taken from Spirtes et al, 2010)
true DAG



start with fully connected undirected graph

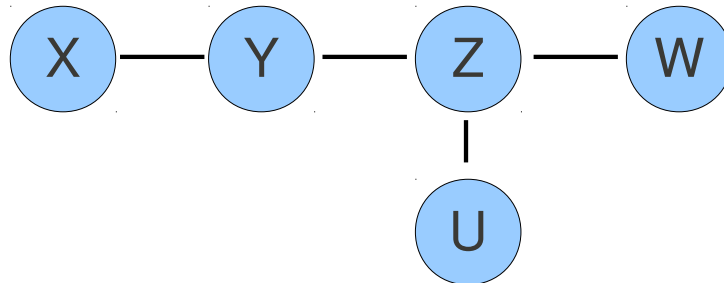


remove all edges X_Y with $X \perp\!\!\!\perp Y \mid \emptyset$



$$X \perp\!\!\!\perp W \quad Y \perp\!\!\!\perp W$$

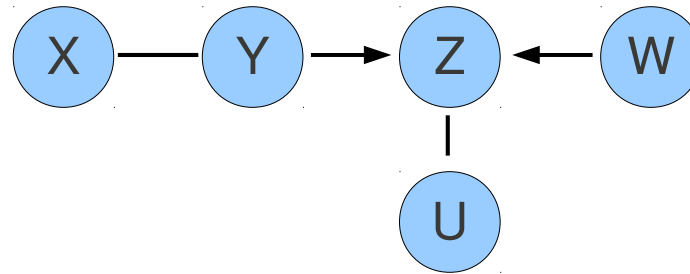
remove all edges having Sepset of size 1



$$X \perp\!\!\!\perp Z \mid Y \quad X \perp\!\!\!\perp U \mid Y \quad Y \perp\!\!\!\perp U \mid Z \quad W \perp\!\!\!\perp U \mid Z$$

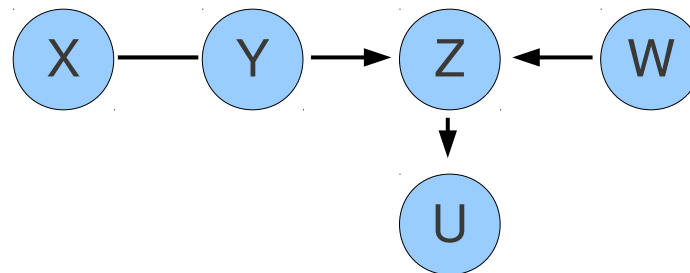


find v-structure



$$Z \notin S_{YW}$$

orient further edges (no further v-structure)



edge $X - Y$ remains undirected

Conditional independence tests

- **discrete case:** contingency tables
- **multi-variate gaussian case:**
covariance matrix

non-Gaussian continuous case: challenging, recent progress via reproducing kernel Hilbert spaces (Fukumizu...Zhang...)

Improvements

- CPC (conservative PC) by Ramsey, Zhang, Spirtes (1995) uses weaker form of faithfulness
- FCI (fast causal inference) by Spirtes, Glymour, Scheines (1993) and Spirtes, Meek, Richardson (1999) infers causal links in the presence of latent common causes
- for implementations of the algorithms see homepage of the TETRAD project at Carnegie Mellon University Pittsburgh

Bayesian approach e.g. Cooper, Heckerman, Meek (1997)

idea:

- The conditionals $p(X_j|PA_j)$ are free parameters in the factorization

$$p(X_1, \dots, X_n) = \prod_{j=1}^n p(X_j|PA_j)$$

- define prior over possible DAGs
- define priors on the parameter space of each DAG
- compute posterior probabilities of DAGs

implicit preference of faithful DAGs

Note: whether Markov equivalent DAGs obtain the same posterior probability depends on the prior

Large scale evaluation of PC-related approach

Maathuis, Colombo, Kalisch & Bühlmann (2007)

Given

- **Observational data:** expression profiles of 5,361 genes of yeast (wild type)
- **Interventional data:** expression profiles of 5,361 genes for interventions on 234 genes

Evaluation:

- use observational data to select the genes that are most influenced by the interventions
(new method: compute lower bound on the effect over all equivalent DAGs)
- compare with those selected from interventional data

success rates clearly significant: e.g. 33 true positive instead of 5



Saccharomyces 57 full-genome yeast deletion data), together with regression profiles of gene deletions (observed under the same conditions and data cleaning procedures), the intervention profiles (measured from 234 single-gene deletions) and the observed expression measurements for 63 genes for 63

experimental data as the target set for identifying the total set of deleted genes on the yeast (i.e., $234 \times 5,360$ genes) (Methods). We used the top 10% of these genes as our target set for IDA could identify the largest predicted set of 50, 250, 1,000

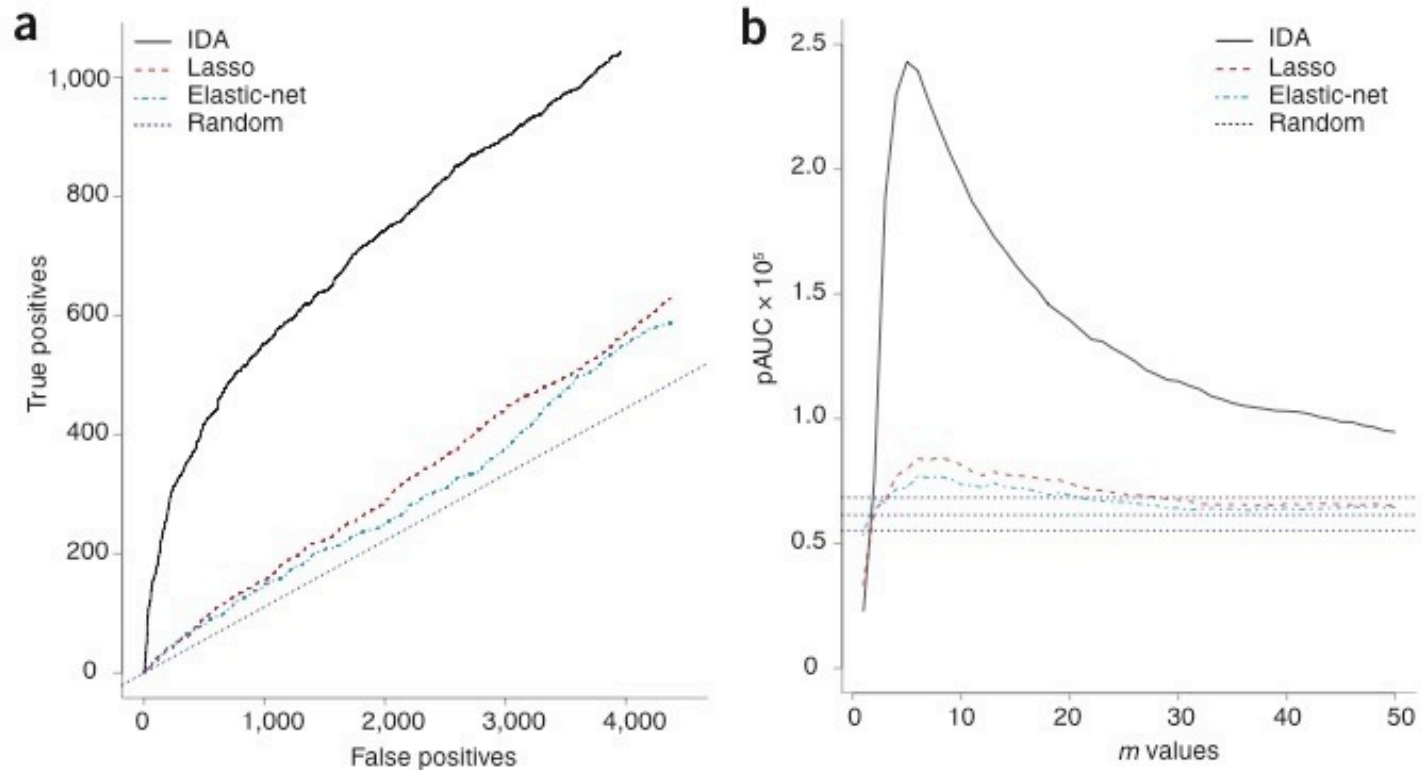


Figure 1 | Predicting causal effects from observational data (data are from ref. 3). (a) The number of true positives versus the number of false positives are plotted for the indicated methods, for the top 5,000 predicted effects from the observational data. The target set is the top 10% of the effects as computed from the interventional data. (b) The partial area under the receiver operating characteristic curve (pAUC) is plotted versus m values, when the target set is the top m percentage of the effects as computed from the interventional data. The pAUC was computed up to the false-positive rate determined by the top 5,000 effects from IDA for $m = 10$. The three horizontal lines for random guessing correspond to the 2.5th, 50th and 97.5th percentiles of a simulated distribution based on random orderings of effects.



INTERVAL?



Kernel Independence Testing (*Gretton et al., 2007*)

k bounded p.d. kernel; P Borel probability measure

Define the *kernel mean map*

$$\mu : P \mapsto \mathbf{E}_{x \sim P}[k(x, \cdot)].$$

Theorem: If k is universal, μ is injective.

Discussion: a measure can be represented as an element of the RKHS associated with k without loss of information.

Let's represent $p(X, Y)$ and $p(X)p(Y)$ — they will only map to the same element if they are equal, i.e., if $X \perp\!\!\!\perp Y$.

Kernel Independence Testing: HSIC

Corollary: $x \perp\!\!\!\perp y \iff \Delta := \|\mu(p_{xy}) - \mu(p_x \times p_y)\| = 0.$

- For $k((x, y), (x', y')) = k_x(x, x')k_y(y, y')$:
 $\Delta^2 =$ HS-norm of cross-covariance operator between the two RKHSes (HSIC, Gretton et al., 2005)
- empirical estimator $\frac{1}{n^2} \text{tr}[K_x K_y]$ (ignoring centering)
- Why does this characterize independence: $x \perp\!\!\!\perp y$ iff

$$\sup_{f, g \in \text{RKHS unit balls}} \text{cov}(f(x), g(y)) = 0$$

(cf. Kernel ICA, *Bach & Jordan, 2002*)

Hilbert-Schmidt Normalized Independence Criterion

(*Fukumizu et al., 2007*)

- normalize out variance of X and Y to get HSNIC; can be shown to equal the *mean squared contingency*

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

independent of the (characteristic/universal) kernel

- can be shown to be upper bounded by the mutual information,

$$\text{HSNIC}(X, Y) \leq \text{MI}(X, Y) = \int \log \left(\frac{p(x, y)}{p(x)p(y)} \right) dp(x, y)$$



Approximating the null distribution

- to construct a test, need to compute the *null distribution* of our test statistic (HSIC): how is the empirical HSIC distributed if $X \perp\!\!\!\perp Y$?
- can use a (complicated) asymptotic expression for HSIC (Gretton et al., 2008), but there's an easy practical method to generate samples consistent with the null hypothesis (independence), and the original marginals $p(X), p(Y)$:
- given a permutation σ , turn $(x_1, y_1), \dots, (x_n, y_n)$ into $(x_1, y_{\sigma(1)}), \dots, (x_n, y_{\sigma(n)})$
- the case of conditional independence is harder: given $(x_1, y_1, z_1), \dots, (x_n, y_n, z_n)$, need to generate samples consistent with $X \perp\!\!\!\perp Y | Z$, and original $p(X|Z), p(Y|Z)$.
- if z only takes few values, can permute within groups having the same value of z (Fukumizu et al., 2007)
- general case is an open problem, but see e.g. Zhang et al., UAI 2011

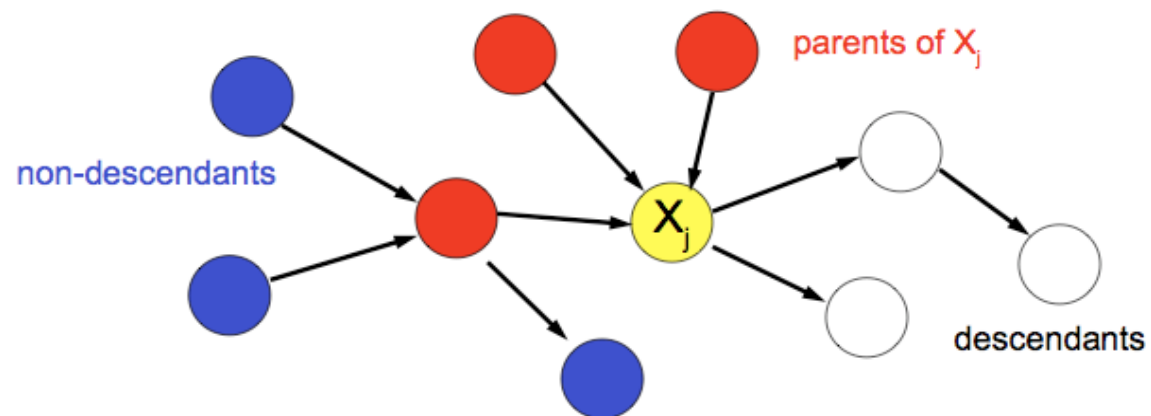


Equivalence of Markov conditions

Theorem: the following are equivalent:

- Existence of a functional causal model
- Local Causal Markov condition: X_j statistically independent of non-descendants, given parents
- Global Causal Markov condition: d-separation
- Factorization $p(X_1, \dots, X_n) = \prod_j p(X_j \mid PA_j)$

(subject to technical conditions)



Local Markov \Rightarrow factorization (Lauritzen 1996)

- proof by induction. Note the factorization is trivial for $n = 1$.
- assume that local Markov for $n - 1$ nodes implies

$$p(x_1, \dots, x_{n-1}) = \prod_{j=1}^{n-1} p(x_j | pa_j).$$

- By local Markov, $X_n \perp\!\!\!\perp ND_n | PA_n$. Assume X_n is a terminal node, i.e., it has no descendants, then $ND_n = \{X_1, \dots, X_{n-1}\}$. Thus

$$X_n \perp\!\!\!\perp \{X_1, \dots, X_{n-1}\} | PA_n$$

and hence the general decomposition

$$p(x_1, \dots, x_n) = p(x_n | x_1, \dots, x_{n-1}) p(x_1, \dots, x_{n-1}).$$

becomes

$$p(x_1, \dots, x_n) = p(x_n | pa_n) p(x_1, \dots, x_{n-1}).$$

- By induction,

$$p(x_1, \dots, x_n) = \prod_{j=1}^n p(x_j | pa_j).$$

Factorization \Rightarrow global Markov

(Lauritzen 1996)

Need to prove $(X \perp\!\!\!\perp Y | Z)_G \Rightarrow (X \perp\!\!\!\perp Y | Z)_p$.

Assume $(X \perp\!\!\!\perp Y | Z)_G$

- define the smallest subgraph G' containing X, Y, Z and all their ancestors
- consider moral graph G'^m (undirected graph containing the edges of G' and links between all parents)
- use results that relate factorization of probabilities with separation in undirected graphs

Global Markov \Rightarrow local Markov

Know that if Z d-separates X, Y , then $X \perp\!\!\!\perp Y | Z$.

Need to show that $X_j \perp\!\!\!\perp ND_j | PA_j$.

Simply need to show that the parents PA_j d-separate X_j from its non-descendants ND_j :

All paths connecting X_j and ND_j include a $P \in PA_j$, but never as a collider

$$\cdot \rightarrow P \leftarrow X_j$$

Hence all paths are chains

$$\cdot \rightarrow P \rightarrow X_j$$

or forks

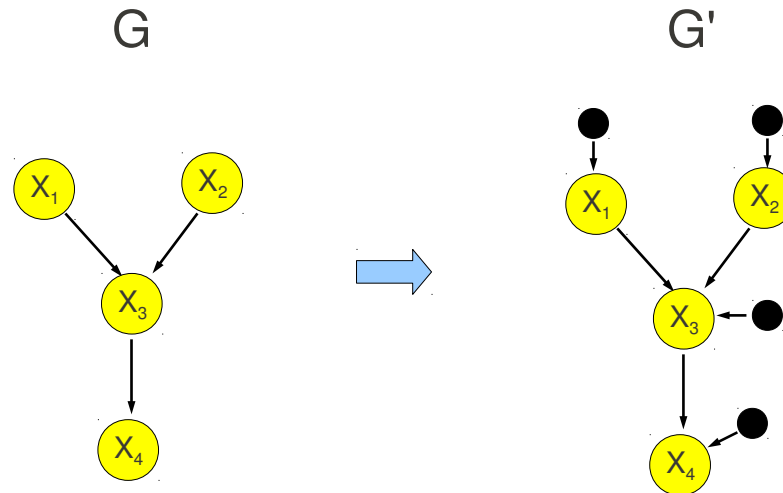
$$\cdot \leftarrow P \rightarrow X_j$$

Therefore, the parents block every path between X_j and ND_j .



functional model \Rightarrow local Markov condition

(Pearl 2000)



- augmented DAG G' contains unobserved noise
- local Markov-condition holds for G' :
 - (i): the unexplained noise terms U_j are jointly independent, and thus (unconditionally) independent of their non-descendants
 - (ii): for the X_j , we have

$$X_j \perp\!\!\!\perp ND'_j \mid PA'_j$$

because X_j is a (deterministic) function of PA'_j .

- local Markov in G' implies global Markov in G'
- global Markov in G' implies local Markov in G (proof as last slide)

factorization \Rightarrow functional model

Generate each $p(X_j|PA_j)$ in

$$p(X_1, \dots, X_n) = \prod_{j=1}^n p(X_j|PA_j)$$

by a deterministic function and a noise variable.

- Idea: “encode” $X_j|pa_j$ (for all values pa_j) into the noise U_j , and pick out the right one depending on pa_j .
- formally, U_j is a map satisfying

$$pa_j \mapsto X_j|pa_j$$

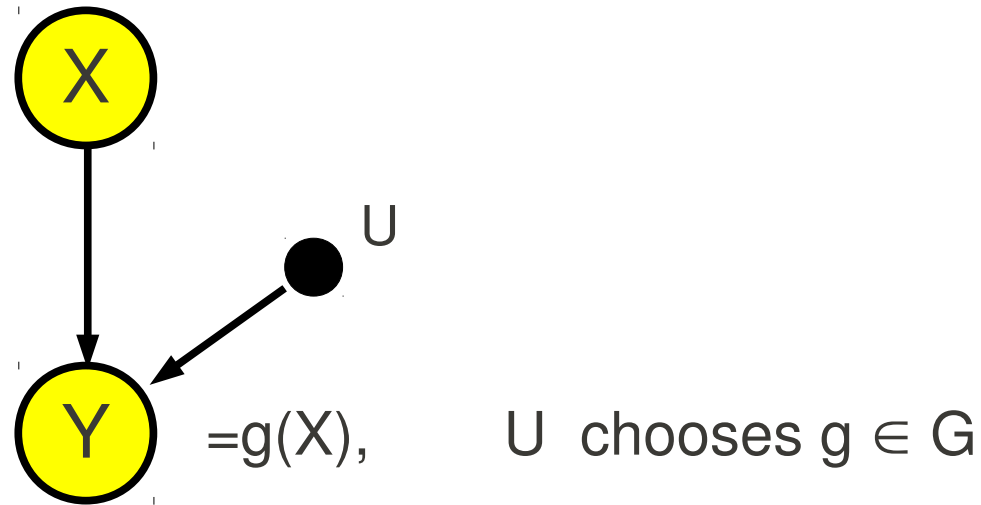
- define structural equation

$$f_j(pa_j, U_j) := U_j(pa_j) = X_j|pa_j$$

- special case: if PA_j only takes d values, U_j is an d -dimensional random vector, and the structural equation picks out a component of the vector
- Note that we could use another joint distribution of the components of U_j , with the same marginals. The causal graphical model only depends on the marginals, but the structural equation model contains more information.

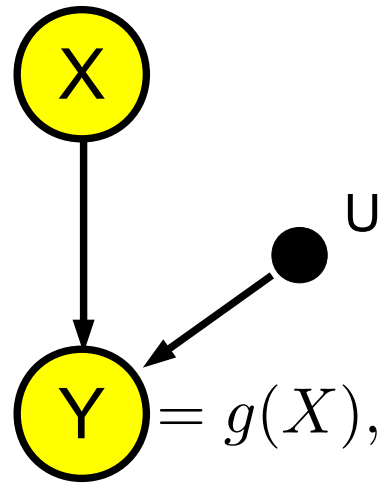


different point of view



- G denotes set of deterministic mechanisms
- U randomly chooses a mechanism

Example: X, Y binary



the same $p(X, Y)$ can be induced by different distributions on G :

- model 1 (no causal link from X to Y)

$$P(g = 0) = 1/2, \quad P(g = 1) = 1/2$$

- model 2 (random switching between ID and NOT)

$$P(g = ID) = 1/2, \quad P(g = NOT) = 1/2$$

both induce the uniform distribution for Y , *independent* of X

Restricting the Functional Model

- general functional model

$$X_i = f_i(\text{Parents}_i, \text{Noise}_i)$$

Note: if Noise_i can take N different values, then it could switch randomly between mechanisms $g_i^1(\text{Parents}_i), \dots, g_i^N(\text{Parents}_i)$

- additive noise model

$$X_i = f_i(\text{Parents}_i) + \text{Noise}_i$$

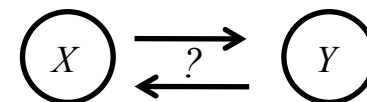
(Noise_i jointly independent)



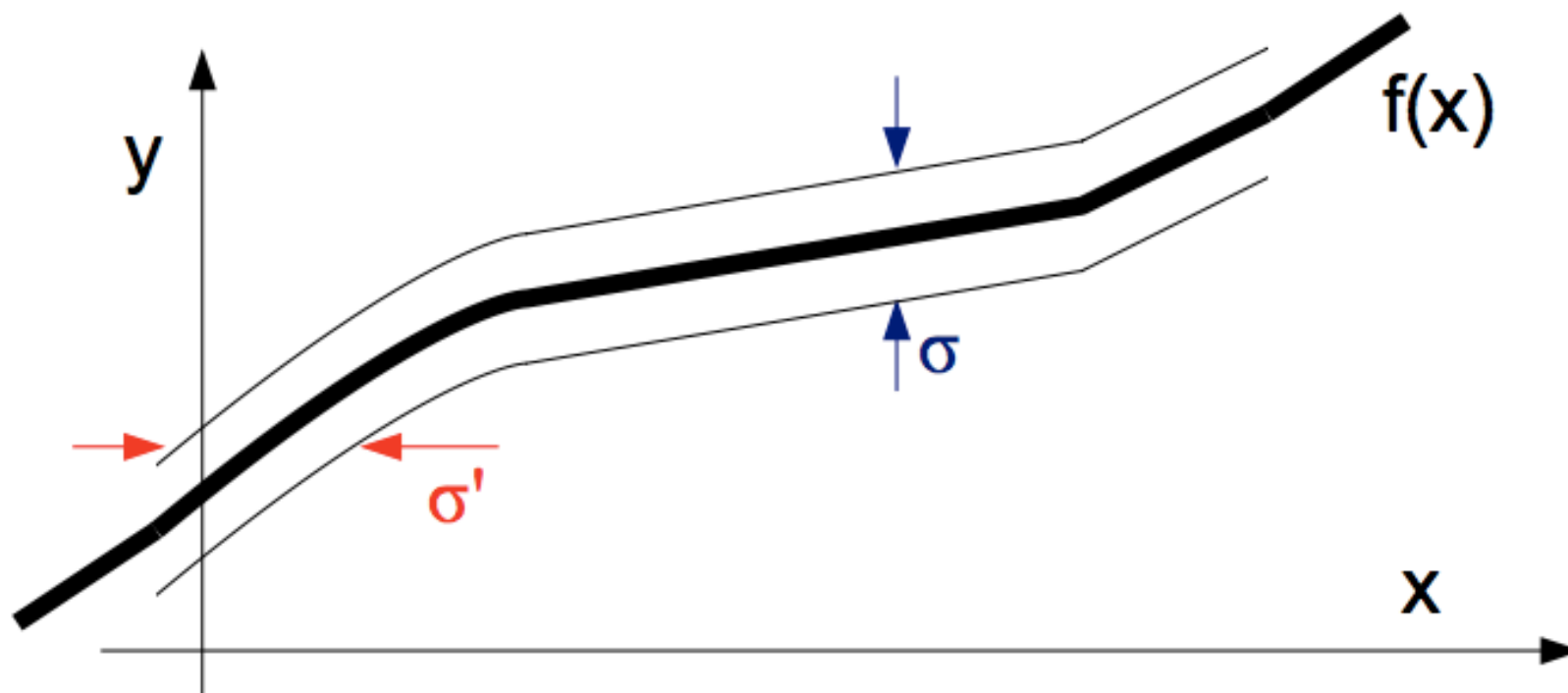
Causal Inference with Additive Noise, 2-Variable Case

Forward model:

$$y := f(x) + n, \text{ with } x \perp\!\!\!\perp n$$



Identifiability: when is there a backward model of the same form?



Hoyer et al.: Nonlinear causal discovery with additive noise models. *NIPS* 21, 2009

Peters et al.: Detecting the Direction of Causal Time Series. *ICML* 2009

Dominik Janzing & Bernhard Schölkopf, June 26, 2012



Identifiability Result *(Hoyer, Janzing, Mooij, Peters, Schölkopf, 2008)*

Theorem 1 *Let the joint probability density of x and y be given by*

$$p(x, y) = p_n(y - f(x))p_x(x), \quad (1)$$

where p_n, p_x are positive probability densities on \mathbb{R} . If there is a backward model

$$p(x, y) = p_{\tilde{n}}(x - g(y))p_y(y), \quad (2)$$

then, denoting $\nu := \log p_n$ and $\xi := \log p_x$ and assuming sufficient differentiability, the triple (f, p_x, p_n) must satisfy the following differential equation for all x, y with $\nu''(y - f(x))f'(x) \neq 0$:

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'}, \quad (3)$$

where we have skipped the arguments $y - f(x)$, x , and x for ν , ξ , and f and their derivatives, respectively. Moreover, if for a fixed pair (f, ν) there exists $y \in \mathbb{R}$ such that $\nu''(y - f(x))f'(x) \neq 0$ for all but a countable set of points $x \in \mathbb{R}$, the set of all p_x for which p has a backward model is contained in a 3-dimensional affine space.

Corollary 1 *Assume that $\nu''' = \xi''' = 0$ everywhere. If a backward model exists, then f is linear.*



Idea of the proof

If $p(x, y)$ admits an additive noise model

$$Y = f(X) + E$$

we have

$$p(x, y) = q(x)r(y - f(x)).$$

It then satisfies the differential equation

$$\frac{\partial}{\partial x} \left(\frac{\partial^2 \log p(x, y) / \partial x^2}{\partial^2 \log p(x, y) / \partial x \partial y} \right) = 0.$$

If it also holds with exchanging x and y , only specific cases remain.

Causal Inference Method

Prefer the causal direction that can better be fit with an additive noise model.

Implementation:

- Compute a function f as non-linear regression of X on Y
- Compute the residual

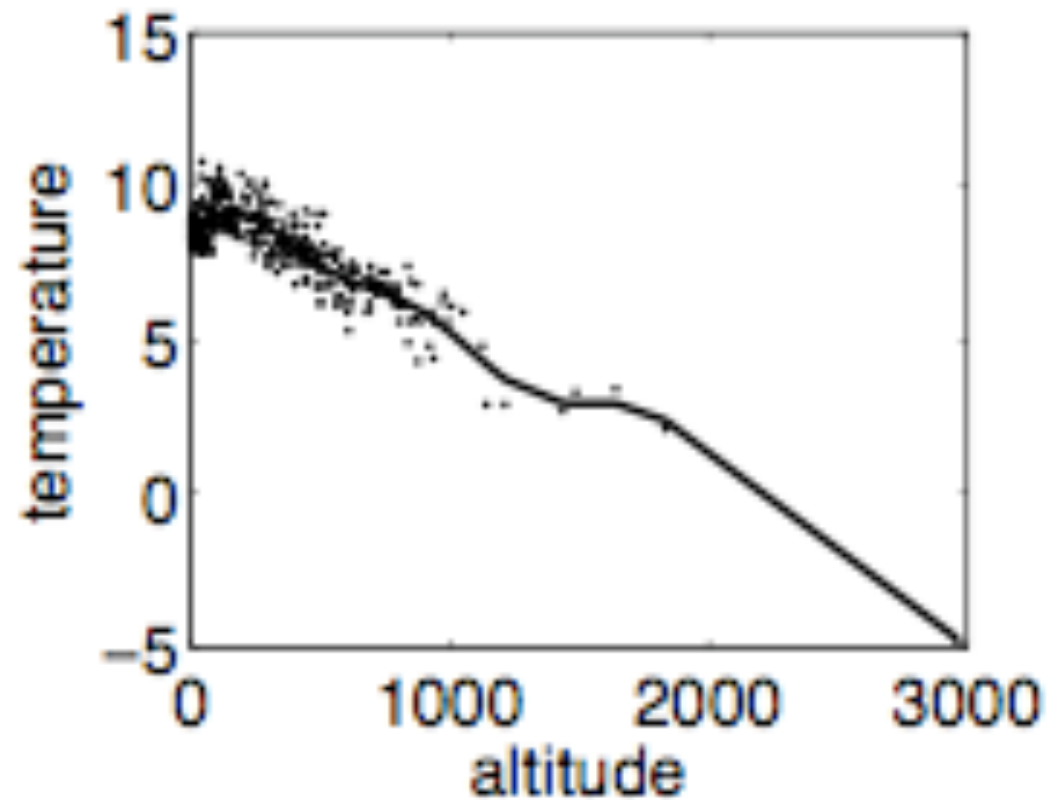
$$E := Y - f(X)$$

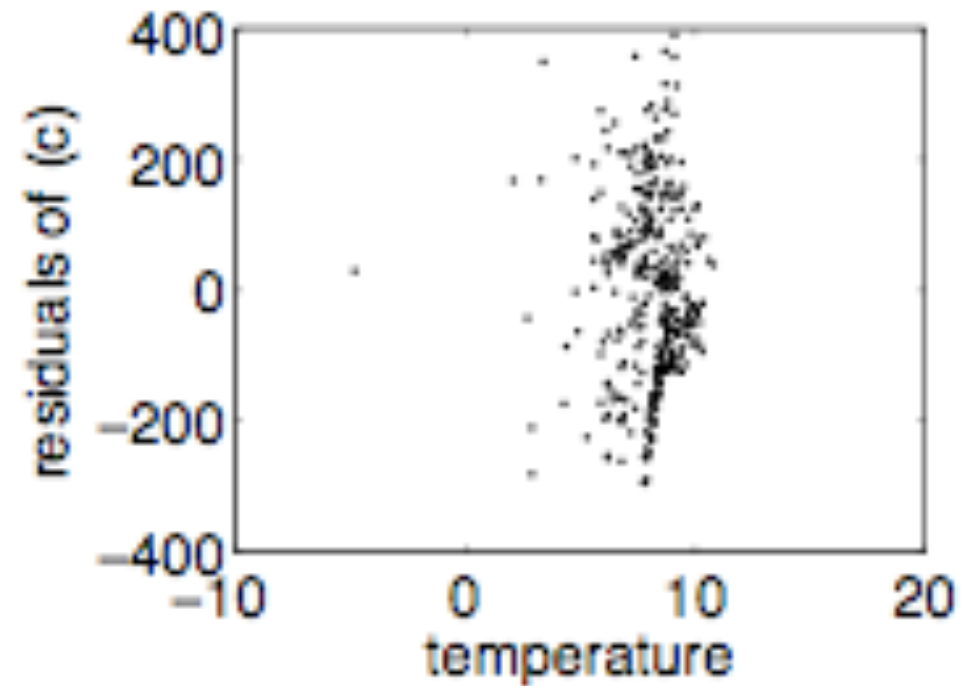
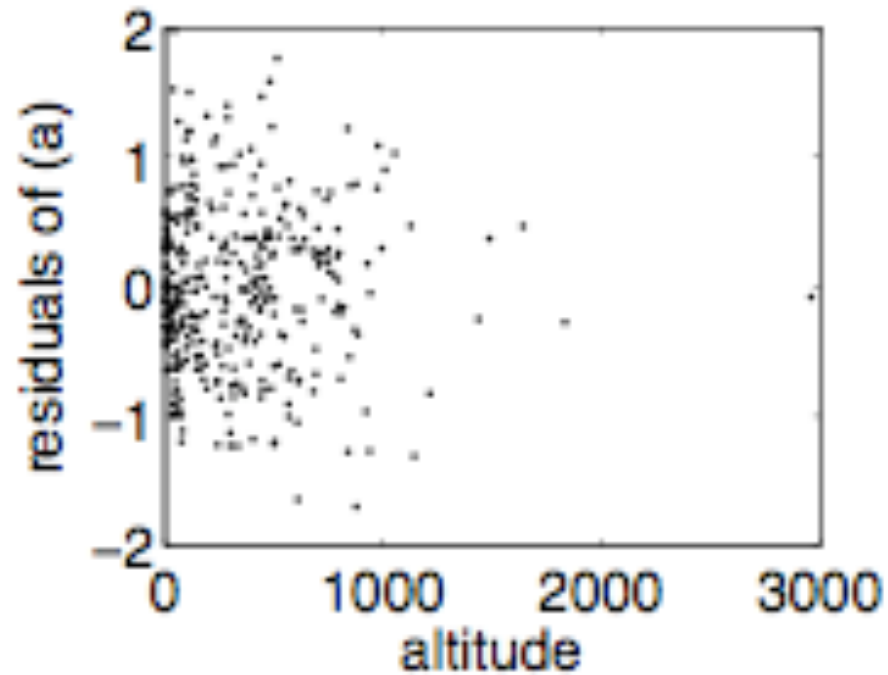
- check whether E and X are statistically independent (uncorrelated is not enough)



Experiments

Relation between altitude (cause) and average temperature (effect) of places in Germany





Our independence tests detect strong dependence.
Hence the method prefers the correct direction

altitude \rightarrow temperature



- Generalization to post-nonlinear additive noise models: Zhang & Hyvärinen: *On the Identifiability of the Post-Nonlinear Causal Model*, UAI 2009
- Generalization to graphs with more than two vertices: Peters, Mooij, Janzing, Schölkopf: *Identifiability of Causal Graphs using Functional Models*, UAI 2011
- Generalization to two-vertex-graphs with loops: Mooij, Janzing, Heskes, Schölkopf: *Causal discovery with Cyclic additive noise models*, NIPS 2011



Independence-based Regression *(Mooij et al., 2009)*

- Problem: many regression methods assume a particular noise distribution; if this is incorrect, the residuals may become dependent
- Solution: minimize dependence of residuals rather than maximizing likelihood of data in regression objective
- Use HSIC as a dependence measure

Mooij, Janzing, Peters, Schölkopf: Regression by dependence minimization and its application to causal inference. ICML 2009.

Yamada & Sugiyama: Dependence Minimizing Regression with Model Selection for Non-Linear Causal Inference under Non-Gaussian Noise. AAAI 2010.

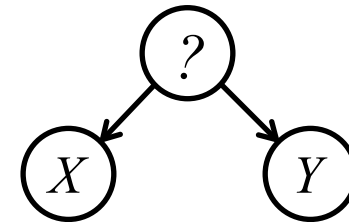
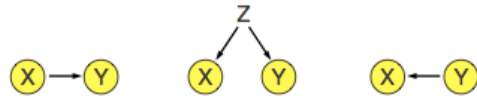


Detection of Confounders

Given $p(X, Y)$, infer whether

▶ $X \rightarrow Y$

▶ $Y \rightarrow X$



▶ $X \leftarrow T \rightarrow Y$ for some (possibly) unobserved variable T

- Confounded additive noise (CAN) models

$$X = f_X(T) + U_X$$

$$Y = f_Y(T) + U_Y$$

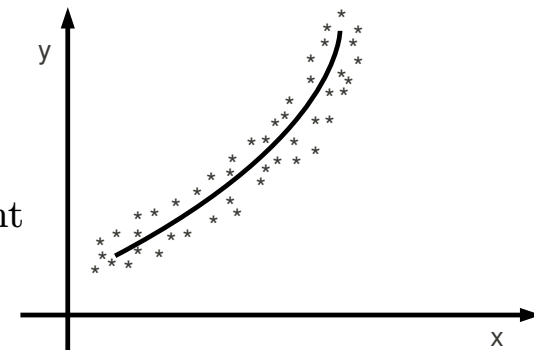
with functions f_X, f_Y and U_X, U_Y, T jointly independent

Note: includes the case

$$Y = f(X) + U$$

by setting $f_X = id$ and $U_X = 0$.

- Estimate $(f_X(T), f_Y(T))$ using dimensionality reduction
- If U_X or U_Y is close to zero, output 'no confounder'
- Identifiability result for small noise



Janzing, Peters, Mooij, Schölkopf: Identifying latent confounders using additive noise models.

UAI 2009



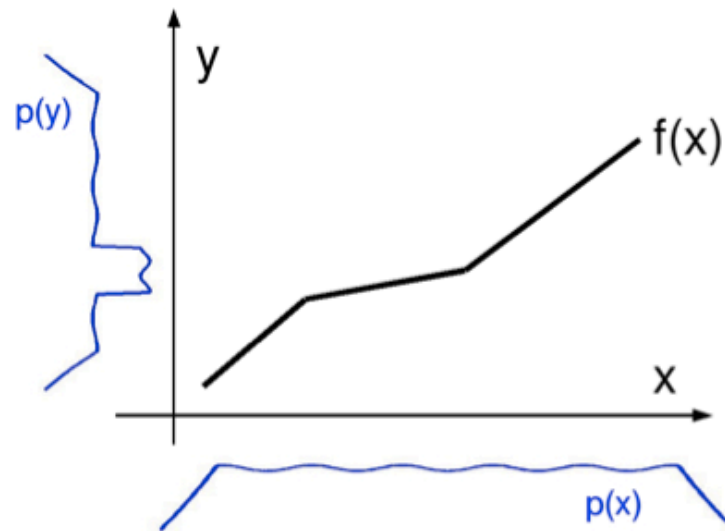
However, employing properties of the noise is not the only way of inferring cause and effect.

What about the noiseless case?



Inferring deterministic causal relations

- Idea: If $X \rightarrow Y$ then f and the density p_X are chosen “independently” by nature
- Hence, peaks of the density p_X do not correlate with the slope of f .
- Then, peaks of p_Y necessarily correlate with the slope of f^{-1} .



Daniusis, Janzing, Mooij, Zscheischler, Steudel, Zhang, Schölkopf: Inferring deterministic causal relations, UAI 2010



Causal independence implies anticausal dependence

Assume that f is a monotonously increasing bijection of $[0, 1]$.

View p_x and $\log f'$ as RVs on the prob. space $[0, 1]$ w. Lebesgue measure.

Postulate (independence of mechanism and input):

$$\text{Cov}(\log f', p_x) = 0$$

Note: this is equivalent to

$$\int_0^1 \log f'(x) p(x) dx = \int_0^1 \log f'(x) dx,$$

since

$$\text{Cov}(\log f', p_x) = E[\log f' \cdot p_x] - E[\log f'] E[p_x] = E[\log f' \cdot p_x] - E[\log f'].$$

Proposition:

$$\text{Cov}(\log f^{-1'}, p_y) \geq 0$$

with equality iff $f = Id$.



u_x, u_y uniform densities for x, y

v_x, v_y densities for x, y induced by transforming u_y, u_x via f^{-1} and f

Equivalent formulations of the postulate:

Additivity of Entropy:

$$S(p_y) - S(p_x) = S(v_y) - S(u_x)$$

Orthogonality (information geometric):

$$D(p_x \parallel v_x) = D(p_x \parallel u_x) + D(u_x \parallel v_x)$$

which can be rewritten as

$$D(p_y \parallel u_y) = D(p_x \parallel u_x) + D(v_y \parallel u_y)$$

Interpretation:

irregularity of p_y = irregularity of p_x + irregularity introduced by f

Slope-Based Estimator

Slope-based IGCI: infer $X \rightarrow Y$ whenever

$$\int_0^1 \log |f'(x)| P(x) dx < \int_0^1 \log |g'(y)| P(y) dx.$$

We introduce the following estimator:

$$\hat{C}_{X \rightarrow Y} := \int \log |f'(x)| P(x) dx \approx \frac{1}{m-1} \sum_{i=1}^{m-1} \log \left| \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \right|$$

where the x_i values are ordered.

► infer $X \rightarrow Y$ whenever

$$\hat{C}_{X \rightarrow Y} < \hat{C}_{Y \rightarrow X}.$$

80 Cause-Effect Pairs

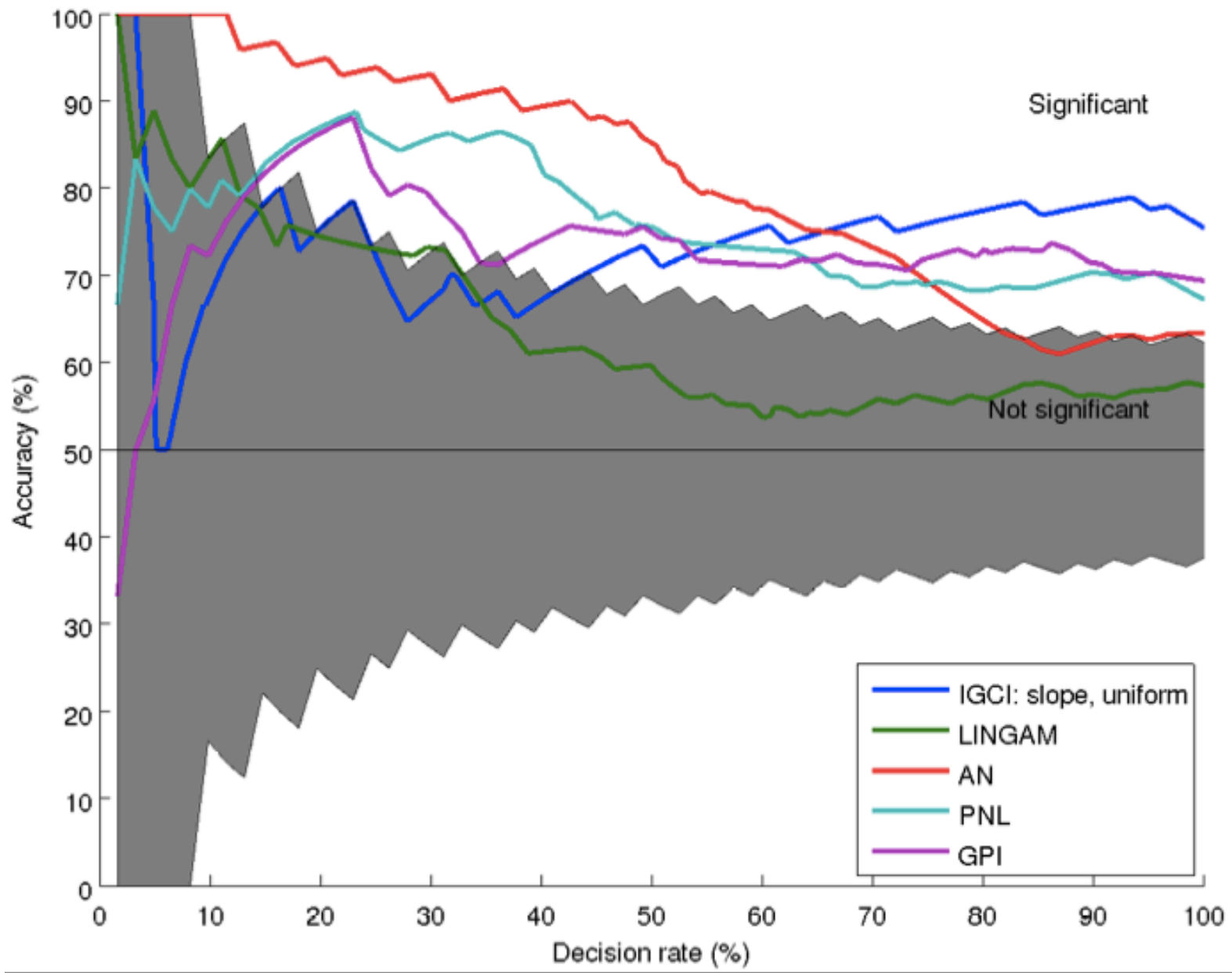


80 Cause-Effect Pairs – Examples

	var 1	var 2	dataset	ground truth
pair0001	Altitude	Temperature	DWD	→
pair0005	Age (Rings)	Length	Abalone	→
pair0012	Age	Wage per hour	census income	→
pair0025	cement	compressive strength	concrete_data	→
pair0033	daily alcohol consumption	mcv mean corpuscular volume	liver disorders	→
pair0040	Age	diastolic blood pressure	pima indian	→
pair0042	day	temperature	B. Janzing	→
pair0047	#cars/24h	specific days	traffic	←
pair0064	drinking water access	infant mortality rate	UNdata	→
pair0068	bytes sent	open http connections	P. Daniusis	←
pair0069	inside room temperature	outside temperature	J. M. Mooij	←
pair0070	parameter	sex	Bülthoff	→
pair0072	sunspot area	global mean temperature	sunspot data	→
pair0074	GNI per capita	life expectancy at birth	UNdata	→
pair0078	PPFD (Photosynth. Photon Flux)	NEP (Net Ecosystem Productivity)	Moffat A. M.	→

<http://webdav.tuebingen.mpg.de/cause-effect/>





IGCI:
Deterministic
Method

LINGAM:
Shimizu et al.,
2006

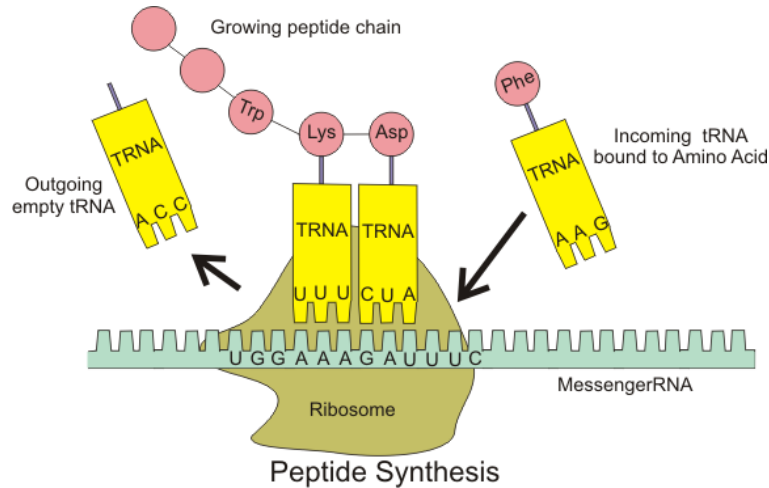
AN:
Additive Noise
Model (nonlinear)

PNL:
AN with post-
nonlinearity

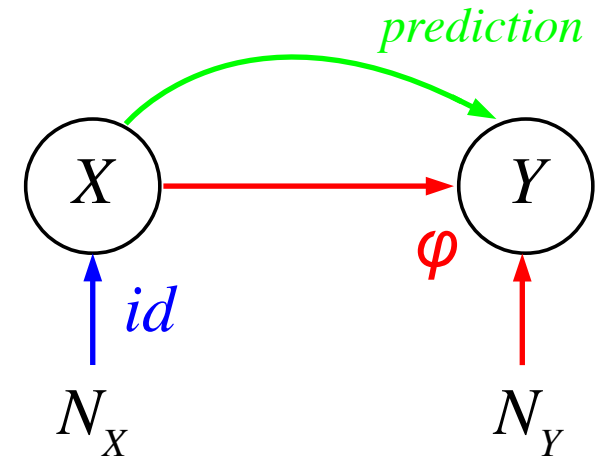
GPI:
Mooij et al.,
2010

Causal Learning and Anticausal Learning

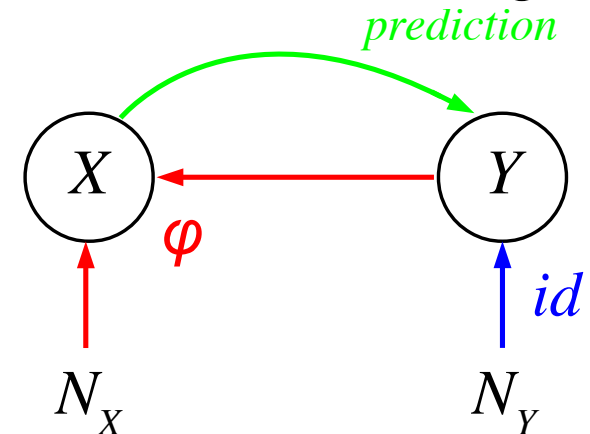
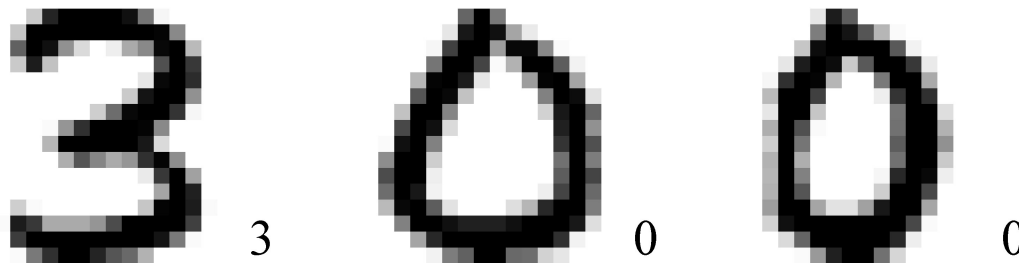
- example 1: predict gene from mRNA sequence



Source: http://commons.wikimedia.org/wiki/File:Peptide_syn.png



- example 2: predict class membership from handwritten digit



Dominik Janzing & Bernhard Schölkopf, June 26, 2012

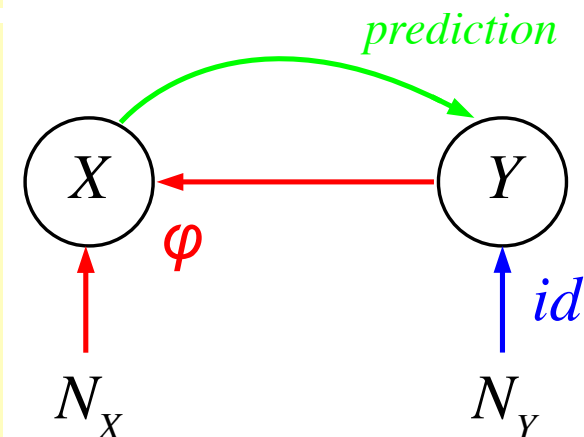
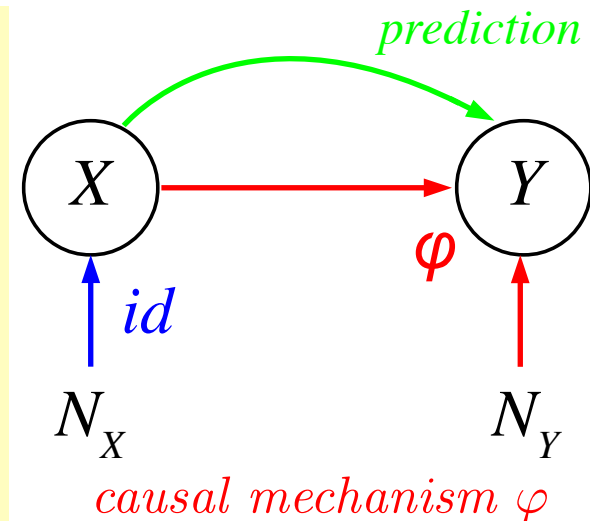
Covariate Shift and Semi-Supervised Learning

Assumption: $p(\text{cause})$ and mechanism $p(\text{effect}|\text{cause})$ are independently chosen by nature;

Goal: learn $X \mapsto Y$, i.e., estimate (properties of) $p(Y|X)$

- covariate shift (i.e., $p(X)$ changes): mechanism $p(Y|X)$ is unaffected by assumption
- semi-supervised learning: impossible, since $p(X)$ contains no information about $p(Y|X)$
- transfer learning (N_X, N_Y change, φ not): could be done by additive noise model with conditionally independent noise

- covariate shift (i.e., $p(X)$ changes): need to decide if change is due to mechanism $p(X|Y)$ or cause distribution $p(Y)$ (nontrivial)
- semi-supervised learning: possible, since $p(X)$ contains information about $p(Y|X)$ — e.g., cluster assumption.



High-dimensional variables

Let X and Y be n and m dimensional variables (possibly Gaussian), infer whether the causal model reads:

$$Y = AX + U \quad \text{or} \quad X = \tilde{A}Y + \tilde{U}$$

Idea:

- $P(X)$ described by covariance matrix Σ_X
- $P(Y|X)$ described by structure matrix A
- hence Σ_X and A are “independent” if $X \rightarrow Y$
- therefore the trace formula holds:

$$\frac{1}{n} \text{tr}(A \Sigma_X A^T) \approx \frac{1}{n} \text{tr}(\Sigma_X) \frac{1}{n} \text{tr}(A A^T)$$

- uses $\frac{1}{d} \text{tr}(CD) \approx \frac{1}{d} \text{tr}(C) \frac{1}{d} \text{tr}(D)$, for two d -dimensional matrices whose entries are drawn independently



Deterministic case

$$Y = AX \quad \text{or} \quad X = A^{-1}Y$$

Theorem (Janzing, Hoyer, Schölkopf 2010): if trace formula

$$\frac{1}{n} \text{tr}(A \Sigma_X A^T) = \frac{1}{n} \text{tr}(\Sigma_X) \frac{1}{n} \text{tr}(A A^T)$$

for $X \rightarrow Y$ holds exactly, then it is violated for $Y \rightarrow X$:

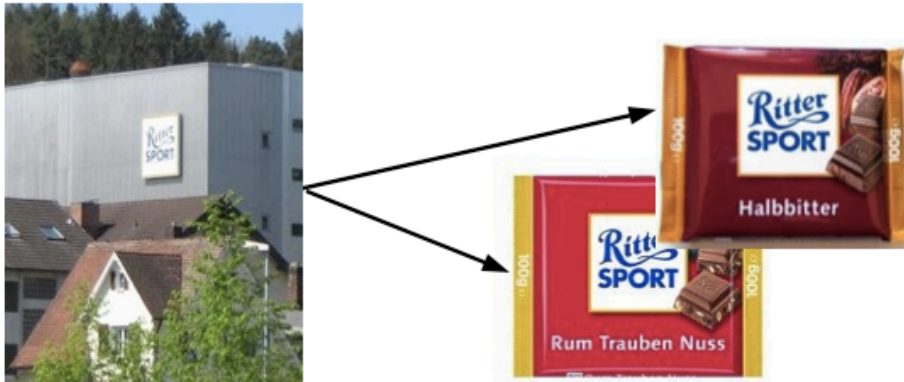
$$\frac{1}{n} \text{tr}(A^{-1} \Sigma_Y A^{-T}) < \frac{1}{n} \text{tr}(\Sigma_Y) \frac{1}{n} \text{tr}(A^{-1} A^{-T})$$

Inference rule: Prefer the direction for which violation of trace formula is smaller (works also for the nondeterministic case)



Causal Inference for Individual Objects *(Janzing & Schölkopf, 2010)*

Similarities between single objects also indicate causal relations:



However, if similarities are too simple there need not be a common cause:



Causal Markov Conditions

- Recall the **(Local) Causal Markov condition**:
An observable is statistically independent of its non-descendants, given parents
- Reformulation:
Given all direct causes of an observable, its non-effects provide no additional *statistical* information on it



Causal Markov Conditions

- Generalization:
Given all direct causes of an observable, its non-effects provide no additional *statistical* information on it
- **Algorithmic Causal Markov Condition:**
Given all direct causes of an object, its non-effects provide no additional *algorithmic* information on it



Kolmogorov complexity

(Kolmogorov 1965, Chaitin 1966, Solomonoff 1964)

of a binary string x

- $K(x) :=$ length of the shortest program with output x (on a Turing machine)
- interpretation: number of bits required to describe the rule that generates x
- equality " $=$ " is always understood up to string-independent additive constants
- $K(x)$ is uncomputable
- probability-free definition of information content



Conditional Kolmogorov complexity

- $K(y | x)$: length of the shortest program that generates y from the shortest description of the input x .
- number of bits required for describing y if the shortest description of x is given
- note: x can be generated from its shortest description but not vice versa because there is no algorithmic way to find the shortest compression



Algorithmic mutual information *(Chaitin, Gacs)*

Information of x about y

- $I(x : y) := K(x) + K(y) - K(x, y)$
 $= K(x) - K(x | y) = K(y) - K(y | x)$
- Interpretation: number of bits saved when compressing x, y jointly rather than independently
- Algorithmic independence $x \perp\!\!\!\perp y : \iff I(x : y) = 0$

Conditional algorithmic mutual information

Information that x has on y (and vice versa) when z is given

- $I(x : y | z) := K(x | z) + K(y | z) - K(x, y | z)$
- Analogy to statistical mutual information:

$$I(X : Y | Z) = S(X | Z) + S(Y | Z) - S(X, Y | Z)$$

- Conditional algor. independence $x \perp\!\!\!\perp y | z : \iff I(x : y | z) = 0$



Algorithmic mutual information: example

$$I(\text{★} : \text{★}) = K(\text{★})$$

Postulate: Local Algorithmic Markov Condition

Let x_1, \dots, x_n be observations (formalized as strings). Given its direct causes pa_j , every x_j is conditionally algorithmically independent of its non-effects nd_j

$$x_j \perp\!\!\!\perp nd_j \mid pa_j$$



Equivalence of Algorithmic Markov Conditions

For n strings x_1, \dots, x_n the following conditions are equivalent

- Local Markov condition

$$I(x_j : nd_j | pa_j) = 0$$

- Global Markov condition:

If R d-separates S and T then $I(S : T | R) = 0$

- Recursion formula for joint complexity

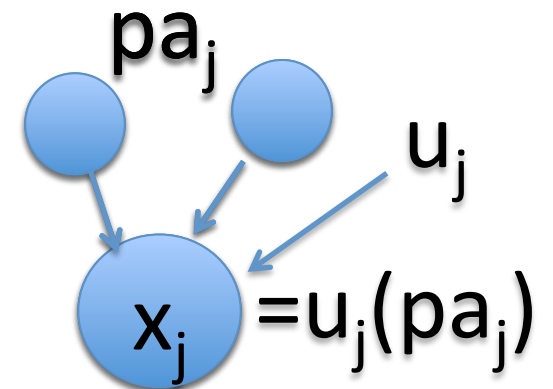
$$K(x_1, \dots, x_n) = \sum_{j=1}^n K(x_j | pa_j)$$

Janzing & Schölkopf, IEEE Trans. Information Theory, 2010



Algorithmic model of causality

- for every node x_j there exists a program u_j that computes x_j from its parents pa_j



- all u_j are jointly independent
- the program u_j represents the causal mechanism that generates the effect from its causes
- u_j are the analog of the unobserved noise terms in the statistical functional model

Theorem: this model implies the algorithmic Markov condition

“Independent” = algorithmically independent?

Postulate (Janzing & Schölkopf, 2010, inspired by Lemeire & Dirkx, 2006):
The causal conditionals $p(X_j|PA_j)$ are algorithmically independent

- special case: $p(X)$ and $p(Y|X)$ are alg. independent for $X \rightarrow Y$
- can be used as justification for novel inference rules (e.g., for additive noise models: Steudel & Janzing 2010)
- excludes many, but not all violations of faithfulness (Lemeire & Janzing, 2012)



Generalized independences *Steudel, Janzing, Schölkopf (2010)*

Given n objects $\mathcal{O} := \{x_1, \dots, x_n\}$

Observation: if a function $R : 2^{\mathcal{O}} \rightarrow \mathbb{R}_0^+$ is submodular, i.e.,

$$R(S) + R(T) \geq R(S \cup T) + R(S \cap T) \quad \forall S, T \subset \mathcal{O}$$

then

$$I(A; B | C) := R(A \cup C) + R(B \cup C) - R(A \cup B \cup C) - R(C) \geq 0$$

for all disjoint sets $A, B, C \subset \mathcal{O}$

Interpretation: I measures conditional dependence

(replace R with Shannon entropy to obtain usual mutual information)

Generalized Markov condition

Theorem: the following conditions are equivalent for a DAG G

- local Markov condition

$$x_j \perp\!\!\!\perp nd_j \mid pa_j$$

- global Markov condition: d-separation implies independence

- sum rule

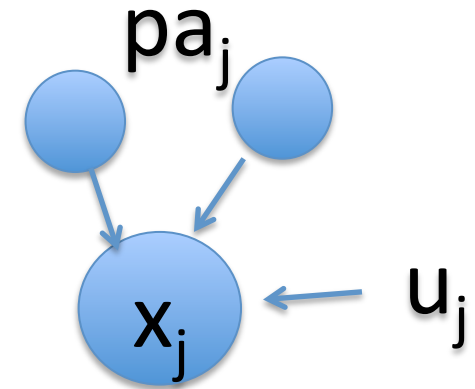
$$R(x_1, \dots, x_n) = \sum_{j=1}^n R(x_j \mid pa_j)$$

–but can we postulate that the conditions hold w.r.t. to the true DAG?

Generalized functional model

Theorem:

- assume there are unobserved objects u_1, \dots, u_n



- assume

$$R(x_j, pa_j, u_j) = R(pa_j, u_j)$$

(x_j contains only information that is already contained in its parents + noise object)

then x_1, \dots, x_n satisfy the Markov conditions

\Rightarrow causal Markov condition is justified provided that mechanisms fit to information measure

Generalized PC

PC algorithm also works with generalized conditional independence

Examples:

1. $R :=$ number of different words in a text
2. $R :=$ compression length (e.g. Lempel Ziv is approximately submodular)
3. $R :=$ logarithm of period length of a periodic function

example 2 yielded reasonable results on simple real texts (different versions of a paper abstract)

Summary

- conventional causal inference algorithms use conditional statistical dependences
- more recent approaches also use other properties of the joint distribution
- non-statistical dependences also tell us something about causal directions



Selection within Markov equivalence classes

different approaches

- some “independence” condition between $p(X_j|PA_j)$
Information-geometric method, Trace Method
- restricting conditionals/functional models to subsets
Additive-noise models, post-nonlinear model
- define priors on $p(X_j|PA_j)$ that can yield different posteriors for equivalent DAGs
Gaussian process based prior by Mooij, Stegle, Janzing, Schölkopf (2010)



- ?

Thank you for your attention



Jonas Peters,



Kun Zhang,



Joris Mooij,



Oliver Stegle,



Eleni Sgouritsa,



Jakob Zscheischler