Causal Bayesian networks

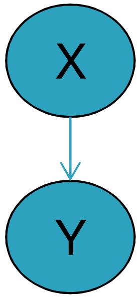
Peter Antal
antal@mit.bme.hu

Outline

- Can we represent exactly (in)dependencies by a BN?
- Can we interpret
  - edges as causal relations
    - with no hidden variables?
    - in the presence of hidden variables?
  - local models as autonomous mechanisms?
- Can we infer the effect of interventions?
- Optimal study design to infer the effect of interventions?

Motivation: from observational inference…

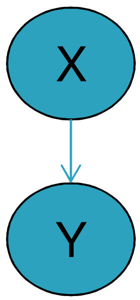
- In a Bayesian network, any query can be answered corresponding to passive observations: $p(Q=q|E=e)$.
  - What is the (conditional) probability of $Q=q$ given that $E=e$.
  - Note that Q can preceed temporally E, e.g. in diagnostic inference below or in smoothing in HMMs, but E is always passively observed.



- Specification: $p(X)$, $p(Y|X)$
- Joint distribution: $p(X,Y)$
- Inferences: $p(X)$, $p(Y)$, $p(Y|X)$, $p(X|Y)$

Motivation: to interventional inference…

- Perfect intervention: do(X=x) as set X to x.
- What is the relation of p(Q=q|E=e) and p(Q=q|do(E=e))?



- Specification: p(X), p(Y|X)
- Joint distribution: p(X,Y)
- Inferences:
  - $p(Y|X=x)=p(Y|do(X=x))$
  - $p(X|Y=y)\neq p(X|do(Y=y))$

- What is a formal knowledge representation of a causal model?
- What is the formal inference method?

Motivation: and to counterfactual inference

- Imagery observations and interventions:
  - We observed X=x, but imagine that x' would have been observed: denoted as X'=x'.
  - We set X=x, but imagine that x' would have been set: denoted as do(X'=x').
- What is the relation of
  - Observational p(Q=q|E=e, X=x')
  - Interventional p(Q=q|E=e, do(X=x'))
  - Counterfactual p(Q'=q'|Q=q, E=e, do(X=x), do(X'=x'))

- O: What is the probability that the patient recovers if he takes the drug *x'*.
- I:What is the probability that the patient recovers if we prescribe* the drug *x'*.
- C: Given that the patient did not recovered for the drug x, what would have been the probability that patient recovers if we had prescribed* the drug *x'*, instead of *x*.
- ~C (time-shifted): Given that patient did not recovered for the drug *x* and he has not respond well**, what is the probability that patient will recover if we change the prescribed* drug *x to x'*.

- *: Assume that the patient is fully compliant.
- **" expected to neither he will.

Challenges in a complex domain

The domain is defined by the joint distribution

$$P(X_1,\dots, X_n|\text{Structure},\text{parameters})$$

1. Representation of parameteres

   „small number of parameters”

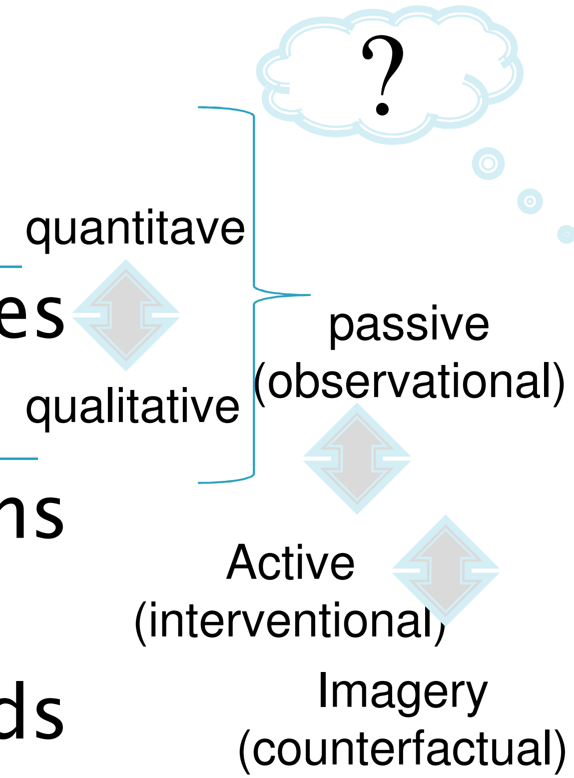
2. Representation of independencies

   „what is relevant for diagnosis”

3. Representation of causal relations

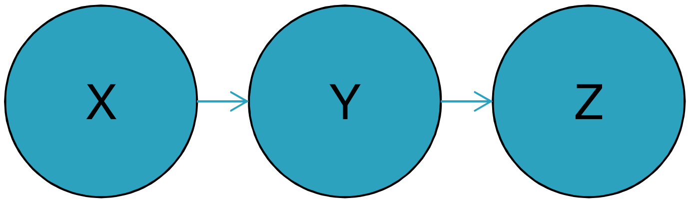
   „what is the effect of a treatment”

4. Representation of possible worlds



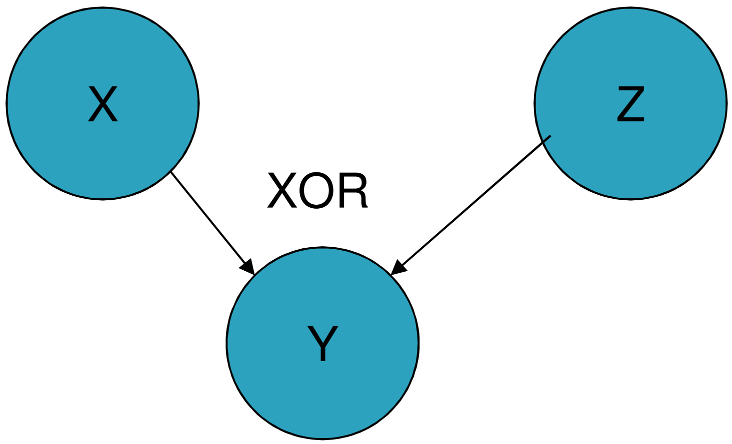
Parametrically encoded intransitivity of dependencies

- In the first order Markov chain below, despite the dependency of X-Y and Y-Z, X and Z can be independent (assuming non-binary Y).



Parametrically encoded pairwise in dependencies

- Pairwise independence does not imply multivariate independence!



Conditional independence

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z: $P(X;Y|z)=P(Y|z)\,P(X|z)$ for all z with $P(z)>0$.

(Almost) alternatively, $I_P(X;Y|Z)$ iff

$P(X|Z,Y)= P(X|Z)$ for all z,y with $P(z,y)>0$.

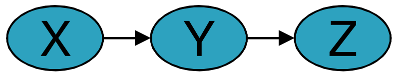
Other notations: $D_P(X;Y|Z)$ =def= $\neg\, I_P(X;Y|Z)$

Contextual independence: for not all z.

The independence model of a distribution

The independence map (model) M of a distribution P is the set of the valid independence triplets:

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$



If P(X,Y,Z) is a Markov chain, then
$M_P=\{D(X;Y), D(Y;Z), I(X;Z|Y)\}$
Normally/almost always: D(X;Z)
Exceptionally: I(X;Z)

The graphoid axioms

1. Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ iff\ I_p(\boldsymbol{Y};\boldsymbol{X}|\boldsymbol{Z})$$

2. Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z})$$

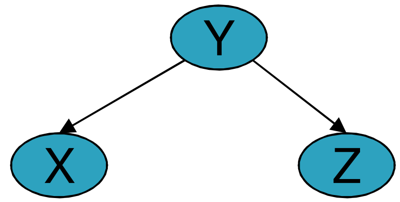
3. Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W})$$

4. Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y}) \Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

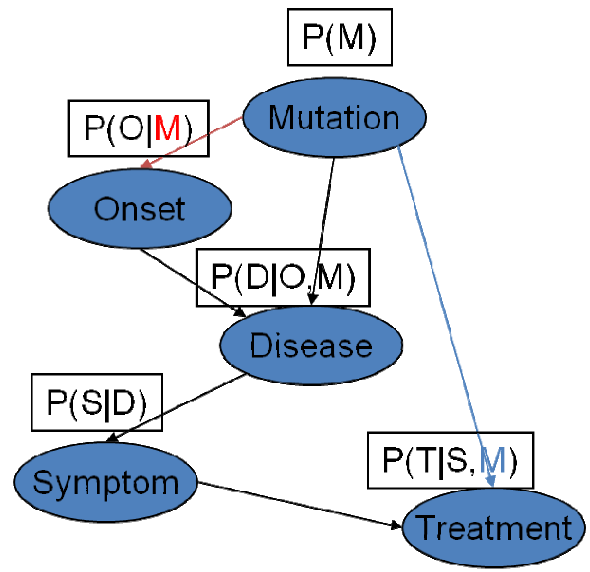
The independence map of a N-BN



If P(Y,X,Z) is a naive Bayesian network, then
M_P={D(X;Y), D(Y;Z), I(X;Z|Y)}
Normally/almost always: D(X;Z)
Exceptionally: I(X;Z)

Bayesian networks

- Directed acyclic graph (DAG)
  - nodes – random variables/domain entities
  - edges – direct probabilistic dependencies (edges– causal relations
- Local models – $P(X_i|Pa(X_i))$
- Three interpretations:



1. Causal model

3. Concise representation of joint distributions

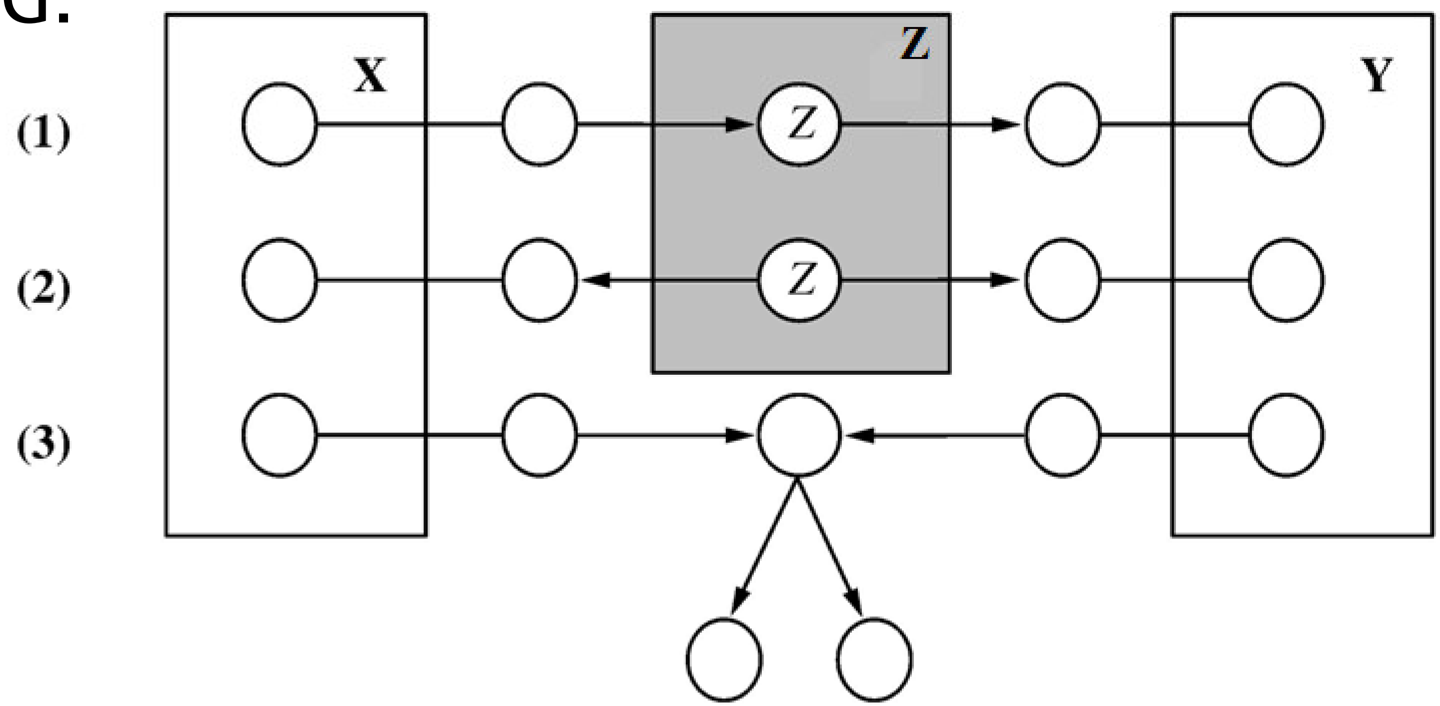
$$P(M,O,D,S,T) = P(M)P(O|M)P(D|O,M)P(S|D)P(T|S,M)$$

$$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots\}$$

2. Graphical representation of (in)dependencies

Inferring independencies from structure: d-separation

$I_G(X;Y|Z)$ denotes that X is d-separated (directed separated) from Y by Z in directed graph G.



d-separation and the global Markov condition

Definition 7 *A distribution $P(X_1, \ldots, X_n)$ obeys the* global Markov *condition w.r.t. DAG G, if*

$$\forall\, X, Y, Z \subseteq U \; (X \perp\!\!\!\perp Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y|Z)_P, \tag{9}$$

where $(X \perp\!\!\!\perp Y|Z)_G$ denotes that X and Y are d-separated *by Z, that is if every path p between a node in X and a node in Y is blocked by Z as follows*

1. *either path p contains a node n in Z with non-converging arrows (i.e. $\rightarrow n \rightarrow$ or $\leftarrow n \rightarrow$),*
2. *or path p contains a node n not in Z with converging arrows (i.e. $\rightarrow n \leftarrow$) and none of its descendants of n is in Z.*

Representation of independencies

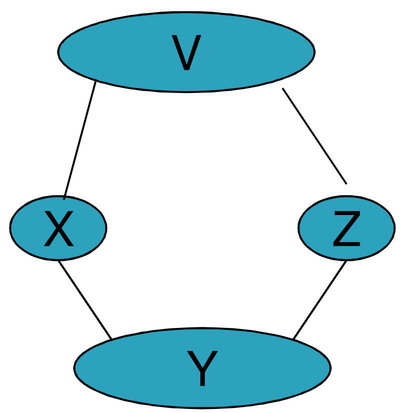
D-separation provides a sound and complete, computationally efficient algorithm to read off an (in)dependency model consisting the independencies that are valid in all distributions Markov relative to G, that is $\forall\ X, Y, Z \subseteq V$

$$(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow ((X \perp\!\!\!\perp Y|Z)_P \text{ in all P Markov relative to G}). \qquad (10)$$

For certain distributions exact representation is not possible by Bayesian networks, e.g.:

1. Intransitive Markov chain: X➔Y➔Z
2. Pure multivariate cause: {X,Z}➔Y
3. Diamond structure:

P(X,Y,Z,V) with M_P={D(X;Z), D(X;Y), D(V;X), D(V;Z), I(V;Y|{X,Z}), I(X;Z|{V,Y}).. }.



Markov conditions

Definition 4 *A distribution $P(X_1, \ldots, X_n)$ is Markov relative to DAG G or factorizes w.r.t G, if*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)), \qquad (6)$$

where $Pa(X_i)$ denotes the parents of X_i in G.

Definition 5 *A distribution $P(X_1, \ldots, X_n)$ obeys the ordered Markov condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_{\pi(i)} \perp\!\!\!\perp \{X_{\pi(1)}, \ldots X_{\pi(i-1)}\} / Pa(X_{\pi(i)}) | Pa(X_{\pi(i)}))_P, \qquad (7)$$

where $\pi()$ is some ancestral ordering w.r.t. G (i.e. compatible with arrows in G).

Definition 6 *A distribution $P(X_1, \ldots, X_n)$ obeys the local (or parental) Markov condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_i \perp\!\!\!\perp \mathrm{Nondescendants(X_i)} | \mathrm{Pa(X_i)})_{\mathrm{P}}, \qquad (8)$$

where $\mathrm{Nondescendants(X_i)}$ denotes the nondescendants of X_i in G.

Bayesian network definitions

Theorem 1 *Let $P(U)$ a probability distribution and **G** a **DAG**, then the conditions above (repeated below) are equivalent:*

F P *is Markov relative G or P factorizes w.r.t G,*

O P *obeys the ordered Markov condition w.r.t. G,*

L P *obeys the local Markov condition w.r.t. G,*

G P *obeys the global Markov condition w.r.t. G.*

Definition 8 *A directed acyclic graph **(DAG)** G is a Bayesian network of distribution $P(U)$ iff the variables are represented with nodes in G and (G, P) satisfies any of the conditions F, O, L, G such that G is minimal (i.e. no edge(s) can be omitted without violating a condition F, O, L, G).*

A practical definition

Definition 9 *A Bayesian network model M of domain with variables U consists of a structure G and parameters $\boldsymbol{\theta}$. The structure G is a* **DAG** *such that each node represents a variable and local probabilistic models $p(X_i|pa(X_i))$ are attached to each node w.r.t. the structure G, that is they describe the stochastic dependency of variable X_i on its parents $pa(X_i)$. As the conditionals are frequently from a certain parametric family, the conditional for X_i is parameterized by $\boldsymbol{\theta}_i$, and $\boldsymbol{\theta}$ denotes the overall parameterezation of the model.*

Association vs. Causation: Markov chain

Causal models:

$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ $\quad$ $X_1 \leftarrow X_2 \leftarrow X_3 \leftarrow X_4$

Markov chain

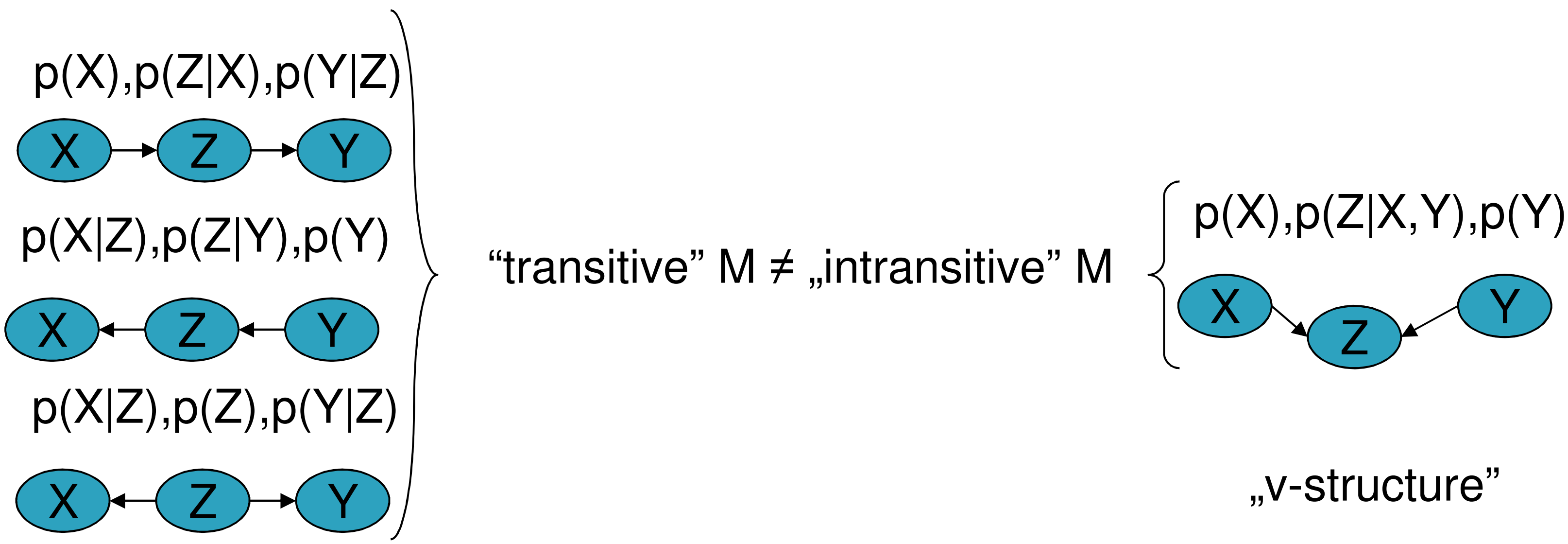
$P(X_1,\ldots)$

$M_P=\{I(X_{i+1};X_{i-1}|X_i)\}$

„first order Markov property”

Flow of time?

The building block of causality: v-structure (arrow of time)



M_P={D(X;Z), D(Z;Y), D(X,Y), I(X;Y|Z)}

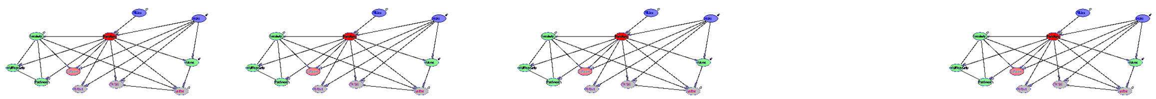
M_P={D(X;Z), D(Y;Z), I(X;Y), D(X;Y|Z) }

Often: present knowledge renders future states conditionally independent.
(confounding)

Ever(?): present knowledge renders past states conditionally independent.
(backward/atemporal confounding)

Observational equivalence of causal models

Causal models:



J.Pearl:
~„3D objects"

From passive observations:

$P(X_1,..., X_n)$

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),..., I_{P,K}(X_K;Y_K|Z_K)\}$

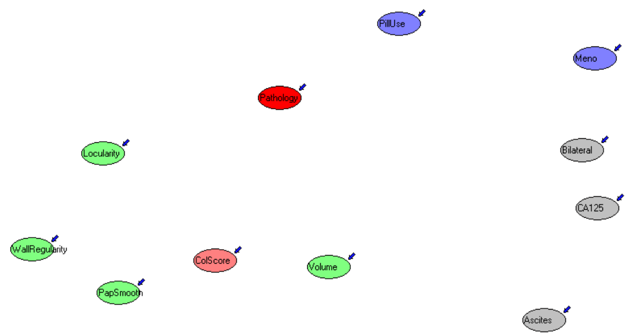
„2D projection"

Different causal models can have the same independence map!

Typically causal models cannot be identified from passive observations, they are ***observationally equivalent***.

Observational equivalence: total independence

„Causal" model:



One-to-one relation

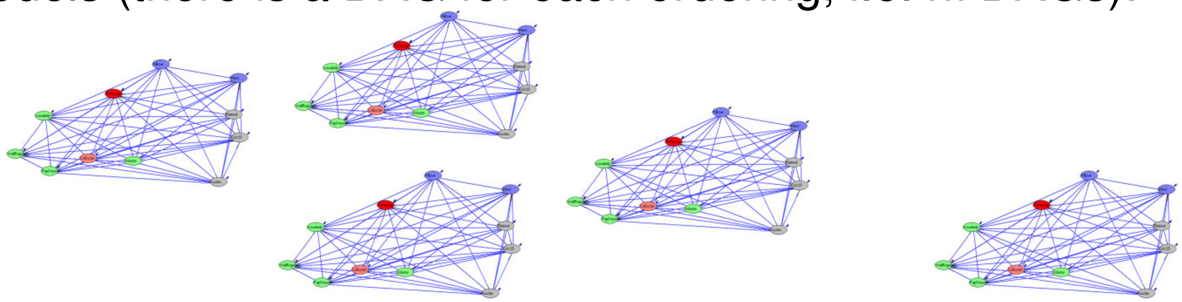
Dependency map:

$P(X_1,..., X_n)$

$M_P=\{I_{P,1}(X_1;X_2),...\}$

Observational equivalence: full dependence

„Causal” models (there is a DAG for each ordering, i.e. n! DAGs):



One-to-many relation

Dependency map:

$P(X_1,..., X_n)$

$M_P=\{D_{P,1}(X_1;X_2),...\}$

Observational equivalence of causal models

Definition 11 *Two DAGs G_1, G_2 are observationally equivalent, if they imply the same set of independence relations (i.e. $(X \perp\!\!\!\perp Y|Z)_{G_1}) \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_{G_2}$).*

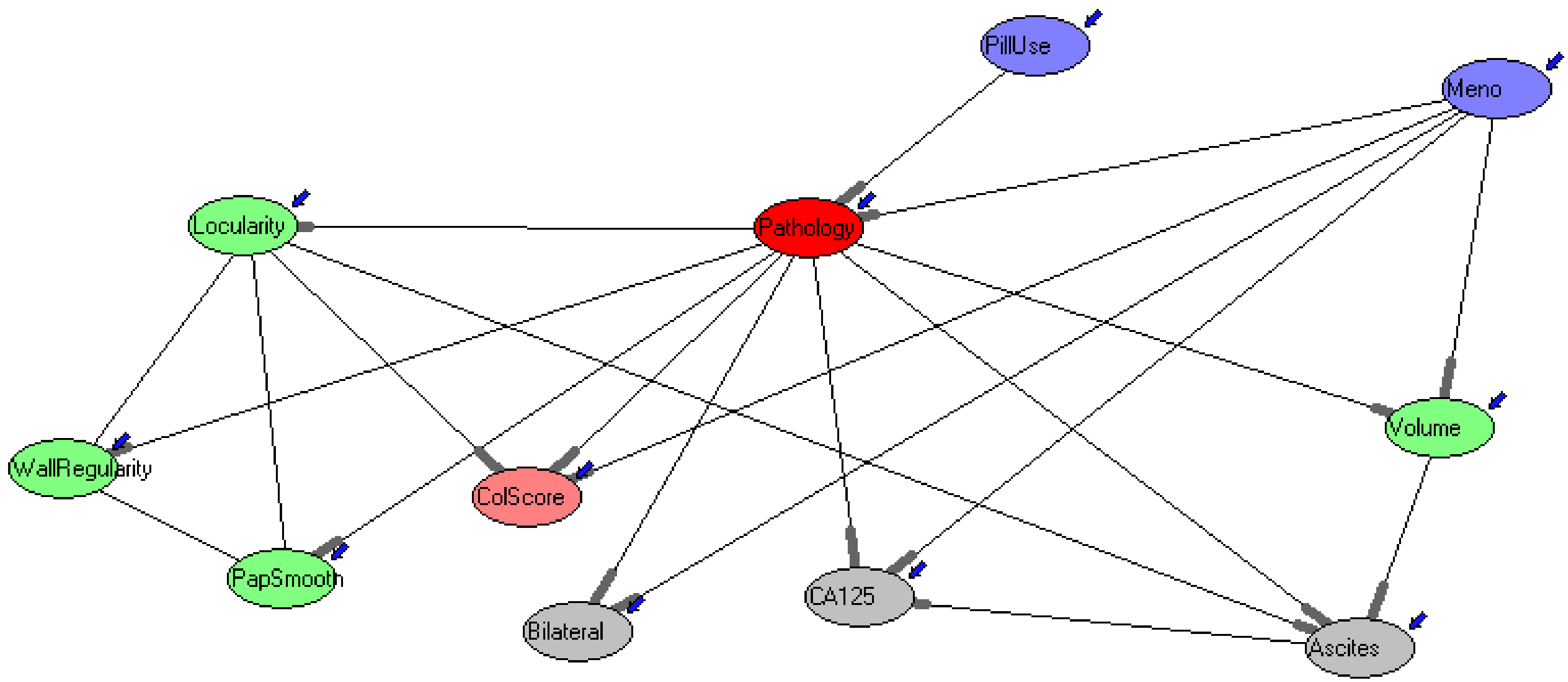
The implied equivalence classes may contain $n!$ number of DAGs (e.g. all the full networks representing no independencies) or just 1.

Theorem 2 *Two DAGs G_1, G_2 are observationally equivalent, iff they have the same skeleton (i.e. the same edges without directions) and the same set of v-structures (i.e. two converging arrows without an arrow between their tails).*

Definition 12 *The essential graph representing observationally equivalent DAGs is a partially oriented DAG (PDAG), that represents the identically oriented edges called compelled edges of the observationally equivalent DAGs (i.e. in the equivalence class), such a way that in the common skeleton only the compelled edges are directed (the others are undirected representing inconclusiveness).*

Compelled edges and PDAG

("can we interpret edges as causal relations?"→compelled edges)



The Causal Markov Condition

- A DAG is called a *causal structure* over a set of variables, if each node represents a variable and edges direct influences. A *causal model* is a causal structure extended with local probabilistic models.
- A causal structure *G* and distribution *P* satisfies the Causal Markov Condition, if P obeys the local Markov condition w.r.t. G.
- The distribution P is said to stable (or faithful), if there exists a DAG called *perfect map* exactly representing its (in)dependencies (i.e. $I_G(X;Y|Z) \Leftrightarrow I_P(X;Y|Z)\ \forall\ X,Y,Z \subseteq V$).

- CMC: sufficiency of G (there is no extra, acausal edge)
- Faithfulness/stability: necessity of G (there are no extra, parametric independency)

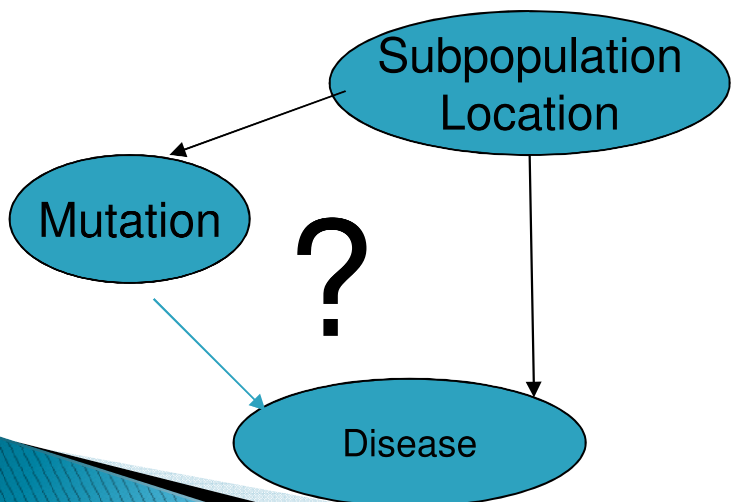
Interventional inference in causal Bayesian networks

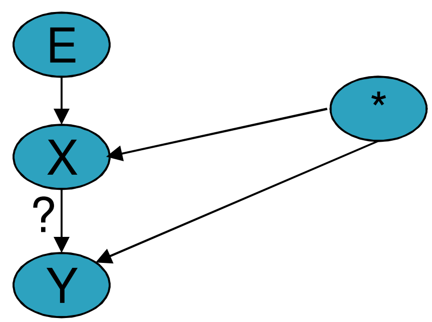
- (Passive, observational) inference
  - P(Query|Observations)
- **Interventionist inference**
  - **P(Query|Observations, Interventions)**
- Counterfactual inference
  - P(Query| Observations, Counterfactual conditionals)

Interventions and graph surgery

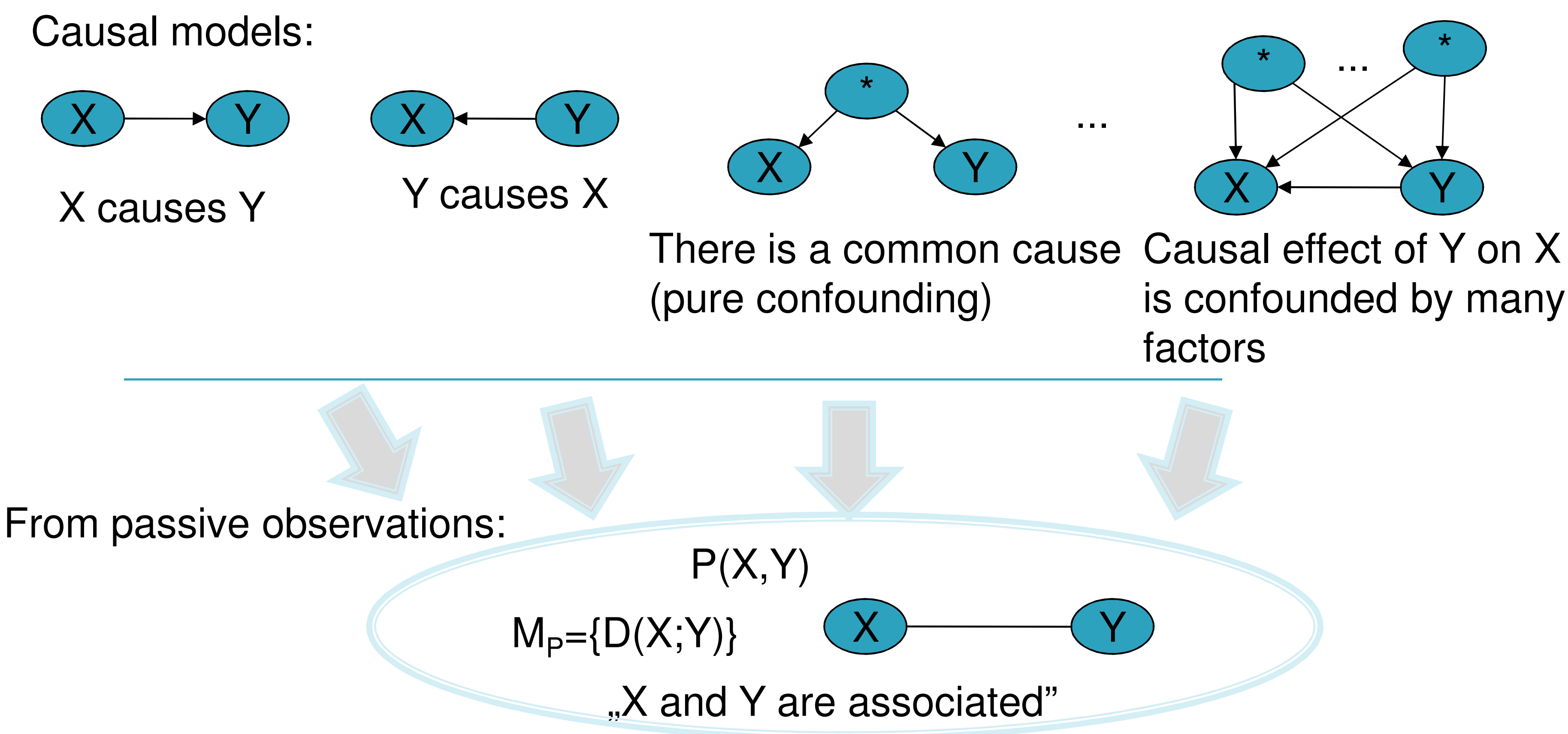
If G is a causal model, then compute p(Y|do(X=x)) by

1. deleting the incoming edges to X
2. setting X=x
3. performing standard Bayesian network inference.





Association vs. Causation



Reichenbach's Common Cause Principle:

a correlation between events *X* and *Y* indicates either that *X* causes *Y*, or that *Y* causes *X*, or that *X* and *Y* have a common cause.

Principles of causality

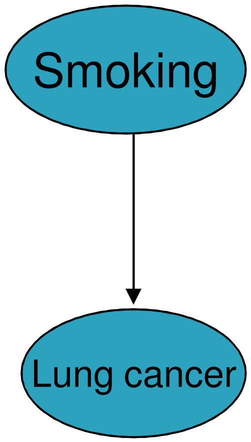
- strong association,
- X precedes temporally Y,
- plausible explanation without alternative explanations based on confounding,
- necessity (generally: if cause is removed, effect is decreased or actually: y would not have been occurred with that much probability if x had not been present),
- sufficiency (generally: if exposure to cause is increased, effect is increased or actually: y would have been occurred with larger probability if x had been present).

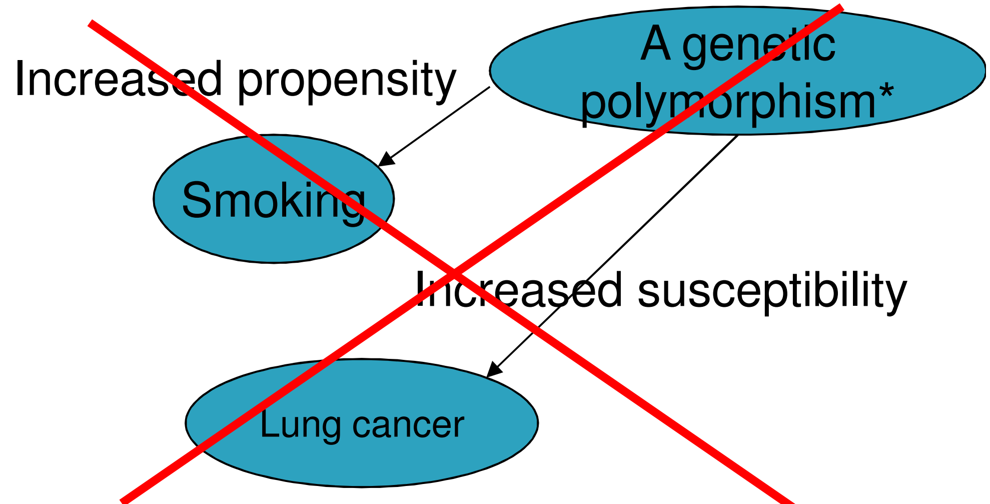
- The probabilistic definition of causation formalizes many, but for example not the counterfactual aspects.

Local Causal Discovery

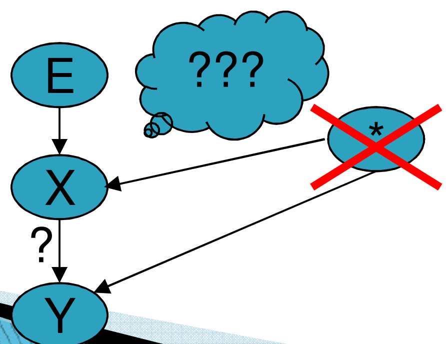
"can we interpret edges as causal relations in the presence of hidden variables?"

- Can we learn causal relations from observational data in presence of confounders???





- Automated, tabula rasa causal inference from (passive) observation is possible, i.e. hidden, confounding variables can be excluded



„Plato's two surprises:
1. Not all true theorems can be proved
2. Causal inference is possible from observations"

A deterministic concept of causation

- H.Simon
  - $X_i=f_i(X_1,..,X_{i-1})$ for i=1..n
  - In the linear case the sytem of equations indicates a natural causal ordering (flow of time?)

					X
				X	X
			X	X	X
		X	X	X	X
	....				

The probabilistic conceptualization is its generalization:

$$P(X_i,|X_1,..,X_{i-1}) \sim X_i=f_i(X_1,..,X_{i-1})$$

Summary

- Can we represent exactly (in)dependencies by a BN?
  - *almost always*
- Can we interpret
  - edges as causal relations
    - with no hidden variables?
      - ***compelled edges as a filter***
    - in the presence of hidden variables?
      - ***Sometimes, e.g. confounding can be excluded in certain cases***
    - in local models as autonomous mechanisms?
      - ***a priori knowledge, e.g. Causal Markov Assumption***
- Can we infer the effect of interventions in a causal model?
  - ***Graph surgery with standard inference in BNs***
- Optimal study design to infer the effect of interventions?
  - ***With no hidden variables: yes, in a non-Bayesian framework***
  - ***In the presence of hidden variables: open issue***
- Suggested reading
  - J. Pearl: Causal inference in statistics, 2009