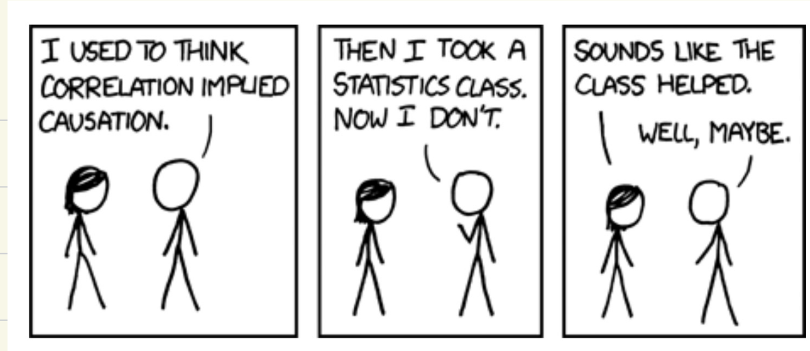
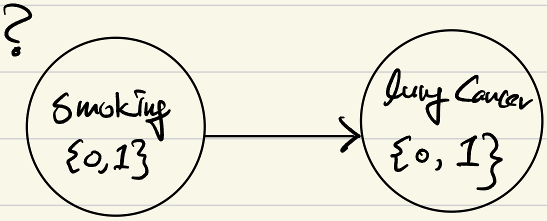
*Causal Structure Discovery.



- Does smoking cause lung cancer?

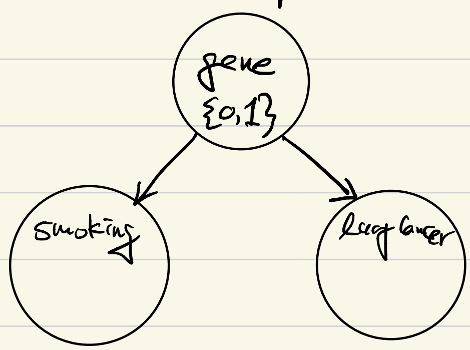


Observational Data

		lung Cancer Yes	lung Cancer No
Smoking	Yes	15%	85%
	No	6%	94%

Correlation does not imply causality.

Alternative Explanation



people with specific gene is likely to smoke AND get lung cancer.

→ those people would have gotten lung cancer even if they did not smoke.

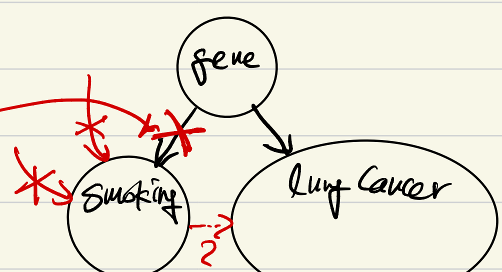
→ hence, smoking does not cause cancer.

Interventional Data.

Randomized trials.

random 50% of population set smoking = 1

random 50% of population set smoking = 0.



you can identify causality by intervention, but it can be unethical and/or expensive.

* All nodes are observed with Observational Data.

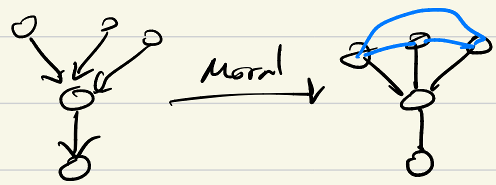
Recall: BN $G=(V,E)$ is a DAG with $P(x)=\prod_{i=1}^{n} P_i(x_i \mid x_{\pi_i})$

Def. Markov Equivalence Class (MEC)

$$G_1 \sim G_2 \overset{\text{iff}}{\Longleftrightarrow} I(G_1) = I(G_2)$$

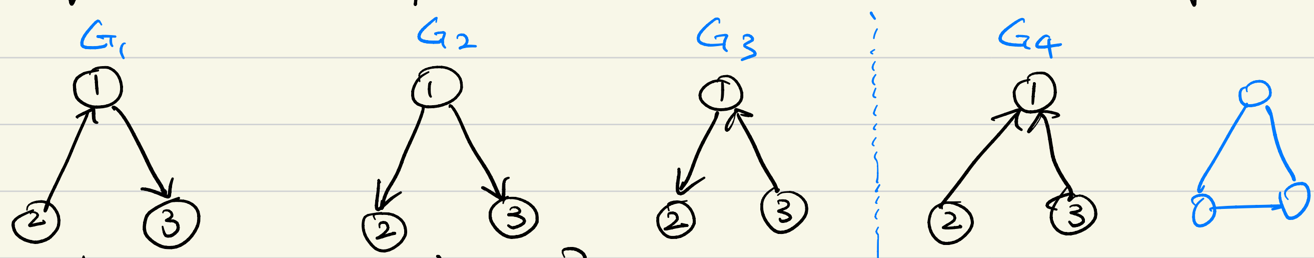
$\Updownarrow$ iff

[Skeleton is the same
[Moral Graph is the same.



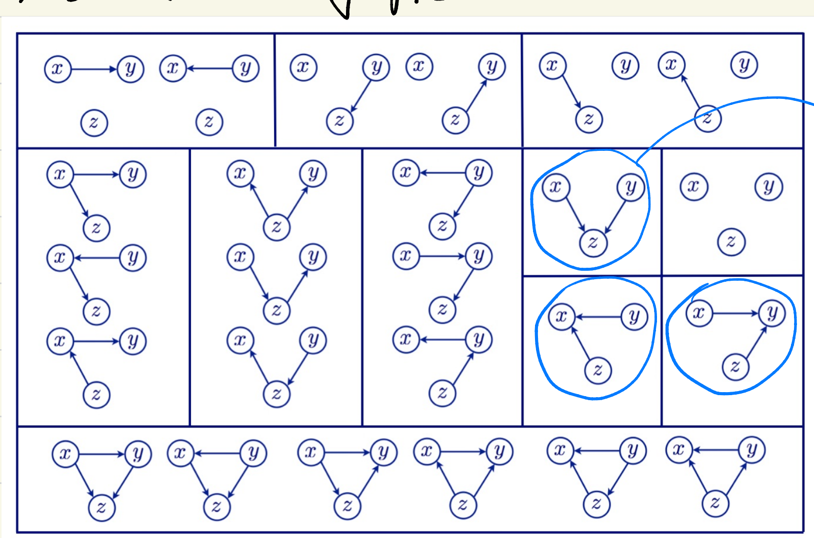
Claim: From observational data, we can only recover G up to its MEC.

proof: $G_1 \sim G_2$ implies $\forall P(x)$ that factorizes as G_1 also factorizes as G_2.



which ones are equivalent?

MEC on 3node graphs



- skeleton,
- Moral.

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

* Hence, G can only be partially identified

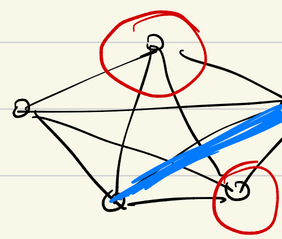
* to resolve the direction of edges within MEC, we need to use interventional data.

Constraint-based Algorithm [SGS-Algorithm]

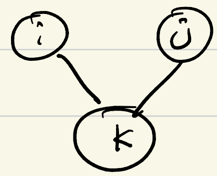
Sprites-Glymour-Scheines
2001.

Step 1. start with Complete Graph $G=(V,E)$. undirected graph

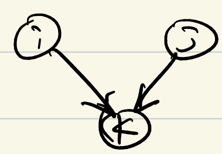
Step 2. Using observational data, for all $(i,j)\in V\times V$
remove (i,j) from E if $\exists S$ s.t. $X_i \perp\!\!\!\perp X_j \mid X_S$.

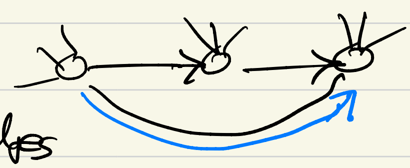


Step 3. for all triplets $(i,j,k)\in V\times V\times V$ s.t.



check if $X_i \perp\!\!\!\perp X_j \mid X_{rest} \setminus \{k\}$
If yes, direct edges as





Step 4. Orient remaining undirected edges
by consistency, recursively do this

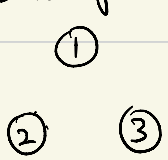
* How do we check $X_i \perp\!\!\!\perp X_j \mid X_S$?

$\underbrace{I(X_i;X_j\mid X_S)}_{\cong\ \hat{I}(X_i;X_j\mid X_S)} = 0$

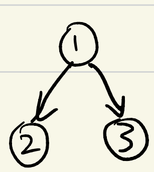
Q. When does SGS-Algorithm fail to recover MEC?

ex> $ICP(X_1,X_2,X_3) = \{X_1\perp\!\!\!\perp X_2\mid X_3,\ X_2\perp\!\!\!\perp X_3\mid X_1,\ X_1\perp\!\!\!\perp X_3\mid X_2\}$

SGS algorithm.

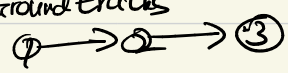


OUTPUT

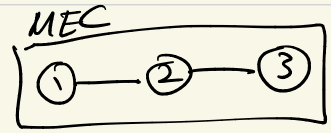


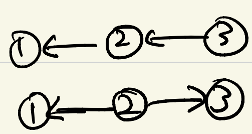
Possible Ground truth.

Ground truths



MEC





Recall. Global Markov Property.

If X_i and X_j are d-separated in G by S, then

$$X_i \perp\!\!\!\perp X_j \mid X_S.$$

Def. $P(x)$ is faithful w.r.t. G. if

$$X_i \perp\!\!\!\perp X_j \mid X_S \text{ for all } S \Rightarrow (i,j) \notin E.$$

Claim. If
- $X^{(1)}, \dots, X^{(N)}$ iid $P(x)$
- $P(x)$ is faithful to a graph G^*.
- all variables in G are observed.

Then SGS-algorithm is consistent, i.e.

$$\lim_{N\to\infty} \mathbb{P}(\underbrace{\hat{G}_{SGS} \not\approx G^*}_{\text{MEC}}) = 0$$

- $\underbrace{\text{Constraint-based}}_{\text{Cond independences as constraints.}}$ algorithm
  - requires lots of samples
  - faithfulness assumption.

* Score-based Algorithms

Recall: log-likelihood score of a DAG G. $x^{(1)} \cdots x^{(N)}$

$$\mathrm{SCORE}(G) = \underbrace{N \sum_{i=1}^{n} I_{\hat{p}}(X_i ; X_{\pi_i})}_{\text{depends on } G} - \underbrace{N \sum_{i=1}^{n} H_{\hat{p}}(X_i)}_{\text{Does not depend on } G}$$

G

without further restrictions on G, complete DAG always achieves max score.

Def. Bayesian Information Criteria (BIC) score.

$$\mathrm{SCORE}_{BIC}(G) \triangleq \underbrace{\mathrm{SCORE}(G)}_{\text{log-likelihood}} - \underbrace{\frac{\log N}{2} \cdot \dim(G)}_{\text{how complex model}}$$

how many variables required to describe $\hat{P}(x)$

Minimum Description Lengths (MDL) Principle.

$\prod_{i=1}^{n} \boxed{P_i(X_i | X_{\pi_i})}$

where $\dim(G) = \sum_{i=1}^{n} (|\mathcal{X}|-1) \cdot |\mathcal{X}|^{|\pi_i|}$

- first term log-likelihood scales as N. } $\to$ samples $\downarrow$ second Dominates.
  second term regularization scales as $\log N$. } samples $\uparrow$

$$\frac{N}{\log N} \cong \#\text{ variables in the model.}$$

* Properties :

① Score equivalent : $G_1 \underset{MEC}{\sim} G_2 \iff \mathrm{SCORE}_{BIC}(G_1) = \mathrm{SCORE}_{BIC}(G_2)$.

② Consistency : If G^* is a perfect map for $P(x)$.
Then as $N \to \infty$, G^* is the unique maximizer.

③ Decomposable : $\mathrm{SCORE}_{BIC} = \sum_{i=1}^{n} \widetilde{\mathrm{SCORE}}(X_i, X_{\pi_i})$.

$\Rightarrow$ Greedy Equivalence Search (GES).

Algorithm [Greedy Equivalence Search].

Initialize $G^{(1)} = (V, E=\phi)$

Phase I: $t=1 \cdots, T$ ← time until no more gain

add an edge that maximizes $SCORE_{BIC}(G^{(t+1)})$

phase II. $t=T+1, \cdots$

remove an edge that maximizes $SCORE_{BIC}(G^{(t+1)})$

Claim: As $N \to \infty$, GES correctly finds MEC under faithfulness.

*How long can T be?

* Permutation-based Greedy Search Algorithm

Idea:

Table 1: Equivalence Class Counts

n	Equivalence classes	Cl/ADG	Cl_1/Cl
1	1	1.00000	1.00000
2	2	0.66667	0.50000
3	11	0.44000	0.36364
4	185	0.34070	0.31892
5	8782	0.29992	0.29788
6	1067825	0.28238	0.28667
7	312510571	0.27443	0.28068
8	212133402500	0.27068	0.27754
9	326266056291213	0.26888	0.27590
10	1118902054495975141	0.26799	0.27507

(Gillispie & Perlman, 2001)

the # of MECs for n-node graph explodes.

⇓

we instead search over all permutations (and skeletons)

→ # MEC $\simeq 10^{18}$ vs. $10! = 3,628,800$

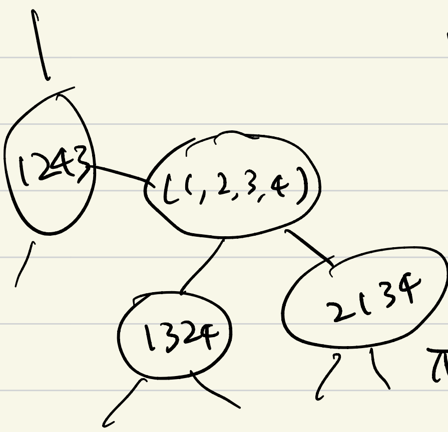
⇓

we apply Greedy Search.

Greedy Search for Sparsest Permutation [GSP] Algorithm.

Initialize: $\pi^{(1)}$ as arbitrary ordering.

Repeat : $t=1, \cdots$



for each permutation/ordering π in the neighborhood of $\pi^{(t)}$

construct a DAG G_π by

$$(\pi_i, \pi_j) \in E_\pi \iff X_{\pi_i} \not\perp\!\!\!\perp X_{\pi_j} \mid X_{\pi_1 \cdots \pi_{i-1}, \pi_{i+1} \cdots \pi_{j-1}}$$

Evaluate $SCORE_{BIC}(G_\pi)$

$\pi^{(t+1)} \leftarrow$ the best scoring candidate permutation.

- two permutations are neighboring if they differ only in two adjacent positions

  e.g. (2, 5, 3, 1, 4)

  (2, 3, 5, 1, 4)

- claim: GSP is consistent under strictly weaker condition than faithfulness.