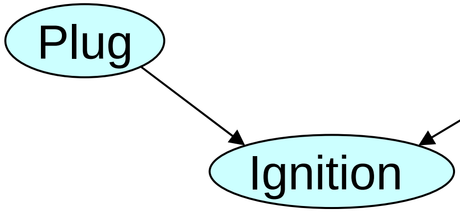
Structure Learning

How to give a network?

- Prior knowledge

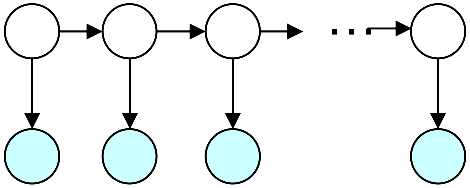
A graphical model may given by the prior knowledge on the problem.

e.g.1) Diagnosis system



The problem is to estimate the probabilities (parameters).

e.g.2) HMM



- Structure learning

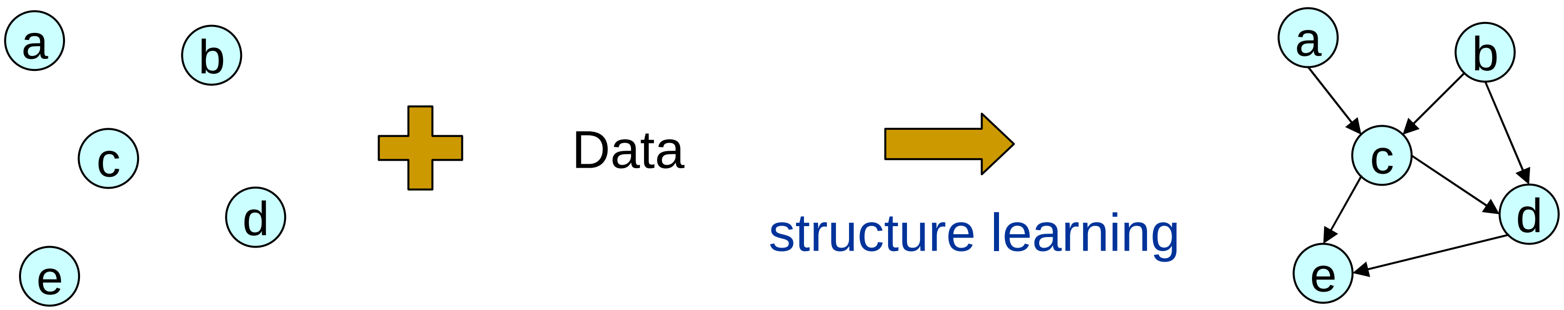
If it is difficult to assume an appropriate model, the graph structure must be learned from given data.

Structure Learning

Variables: $X_1, ..., X_m$

Data: $(X_1^{(1)}, ..., X_m^{(1)}), ..., (X_1^{(N)}, ..., X_m^{(N)})$

Output of structure learning = a directed / undirected graph associated with the probability of $(X_1, ..., X_m)$.



Difficulty: the number of possible directed graphs = 3^m

The search space is very large.

Learning of Directed Graph

- Constraint-based method
  - Determine the conditional independence of the underlying probability by statistical tests.
  - Many statistical tests are required.
  - Often referred to as causal learning.

- Score-based method
  - Use a global score to match a graph and data.
  - Optimization in huge search space.
  - Able to use informative prior on graphs.
  - Usually, discrete variables are assumed.
  - Often referred to as Bayesian structure learning

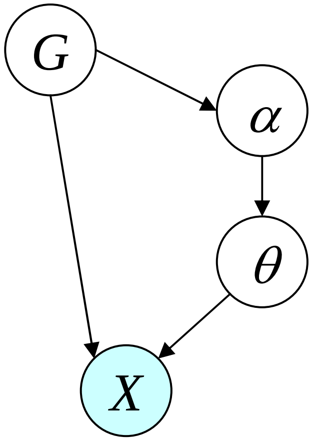
Score-based Structure Learning

Discrete variables: $X_1, \ldots, X_m$

Data: $D = \{(X_1^{(1)}, \ldots, X_m^{(1)}), \ldots, (X_1^{(N)}, \ldots, X_m^{(N)})\}$

- Model:

  When a directed graph G is specified, multinomial distribution is assumed with Dirichlet prior.



$$p(X \mid \theta) = \prod_{b=1}^{m} p(X_b \mid X_{pa(b)}, \theta_b)$$

$\theta_b = (\theta_{b,i}^j)$ $\quad$ i : multi-index for $pa(b)$

$\theta_{b,i}^j = P(X_b = j \mid X_{pa(b)} = i) \quad \theta_{b,i}^j \geq 0, \sum_{j=1}^{K_b} \theta_{b,i}^j = 1.$

$$p(D \mid \theta) = \prod_{n=1}^{N} \prod_{b=1}^{m} p(X_b^{(n)} \mid X_{pa(b)}^{(n)}, \theta_b)$$

Dirichlet prior:

$$\theta_{b,i} = (\theta_{b,i}^1, \ldots, \theta_{b,i}^{K_b}) \sim \mathrm{Dir}(\theta_{b,i} \mid \alpha_{b,i}^1, \ldots, \alpha_{b,i}^{K_b}) = \frac{\Gamma(\sum_j \alpha_{b,i}^j)}{\prod_j \Gamma(\alpha_{b,i}^j)} \prod_{j=1}^{K_b} (\theta_{b,i}^j)^{\alpha_{b,i}^j - 1}$$

Score-based Structure Learning

- Marginal likelihood:

$$\text{Score}(G) \equiv \text{Marginal log likelihood of } G$$

$$= \log \int P(D \mid \theta, G)\, p(\theta \mid G, \alpha)\, d\theta \qquad \alpha = (\alpha_{b,i}^{j})$$

$$= \log \int \prod_{b=1}^{m} \prod_{i=1}^{\#pa(b)} \prod_{j=1}^{K_b} (\theta_{b,i}^{j})^{N_{b,i}^{j}} \frac{\Gamma(\sum_j \alpha_{b,i}^{j})}{\prod_j \Gamma(\alpha_{b,i}^{j})} \prod_{j=1}^{K_b} (\theta_{b,i}^{j})^{\alpha_{b,i}^{j}-1} d\theta_{b,i}$$

$$= \sum_{b=1}^{m} \sum_{i=1}^{\#pa(b)} \left[\log \Gamma(\sum_j \alpha_{b,i}^{j}) - \sum_{j=1}^{K_b} \Gamma(\alpha_{b,i}^{j}) - \log \Gamma(\sum_j \tilde{\alpha}_{b,i}^{j}) + \sum_{j=1}^{K_b} \Gamma(\tilde{\alpha}_{b,i}^{j}) \right]$$

where $\quad \tilde{\alpha}_{b,i}^{j} = N_{b,i}^{j} + \alpha_{b,i}^{j}$

Score-based Structure Learning

- Prior to the models

  We can use a prior distribution $P(G)$ on the graphs.

$$\text{Score}(G) = \log P(D \mid G) + \log P(G)$$

- Optimization over the graphs

  The space is very huge → greedy search.

  Start from a graph G

  Repeat the following process:

  Update the graph by deleting, inserting, or reversing an edge.

  Accept the new graph G' if Score(G') > Score(G).

- Many others
  - MDL / BIC, MCMC, etc.

  See D. Heckerman "A tutorial on learning with Bayesian networks" in Learning in Graphical Models (M. Jordan ed.) 1998.

Marginal Log Likelihood / ABIC

- Bayesian method for model selection

Maximum a posteriori model given data

$$\hat{G} = \arg\max P(G \mid D)$$

Note:

$$P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)} \propto P(D \mid G)P(G)$$ as a function of model

$\Rightarrow$

$$\hat{G} = \arg\max\left[\log P(D \mid G) + \log P(G)\right]$$

If $P(G)$ is uniform over the models,

$$\hat{G} = \arg\max \log P(D \mid G)$$

$$= \arg\max \log \int P(D \mid \theta, G)P(\theta \mid G)d\theta$$

Marginal log likelihood

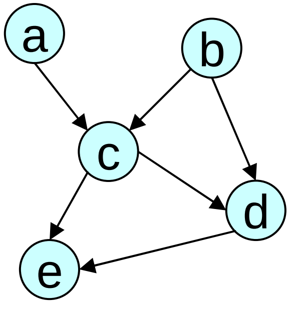
(ABIC: Akaike's Bayesian information criterion)

Mini-Summary on score-based method

- Use a global score to match a graph and data.
  - Marginal log likelihood (ABIC), MDL, etc.
- Optimization in huge search space.
  - Some techniques are needed. e.g. greedy search.
- Able to use informative prior on graphs.
- Usually, discrete or Gaussian variables are assumed.
  - For non-Gaussian continuous variables, we need some techniques such as discretization.
- Also known as Bayesian structure learning

Causal Learning

- Directed graph as causal graph
  - A directed graph can be regarded as the expression of causal relationships among variables.



Causal direction = Edge-direction

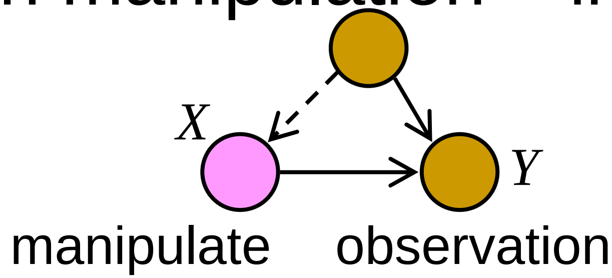
$$p(X) = p(X_a)p(X_b)p(X_c \mid X_a, X_b) \\ \times p(X_d \mid X_b, X_c)p(X_e \mid X_c, X_d)$$

  - Causal learning: learning of the directed graph from data.

Assumption: the data is given by the probability factorizing w.r.t. the directed graph.

Causal Leaning from Data

- With manipulation – intervention



X is a cause of Y?

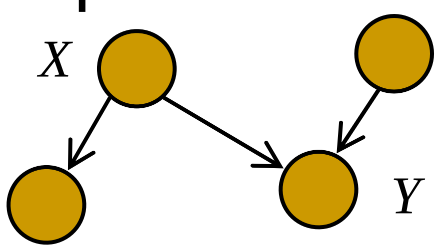
Easier. (*do*-calculus, Pearl 1995)

- No manipulation / with temporal information

$X(t)$ $Y(t)$: observed time series

$X(1), \ldots, X(t)$ are a cause of $Y(t+1)$?

- No manipulation / no temporal information



Causal inference is harder.

Causal Learning without Manipulation

- Difficulty of causal inference from non-experimental data
  - Widely accepted view till 80's

    Causal inference is impossible without manipulating some variables.

    e.g.) *"No causation without manipulation"* (Holland 1986, JASA)

  - Temporal information is very helpful, but not decisive.

    e.g.) The barometer falls before it rains, but it does not cause the rain.

  - Many philosophical discussions, but not discussed here.

    See Pearl (2000) and the references therein.

Addendum: Causality and Correlation

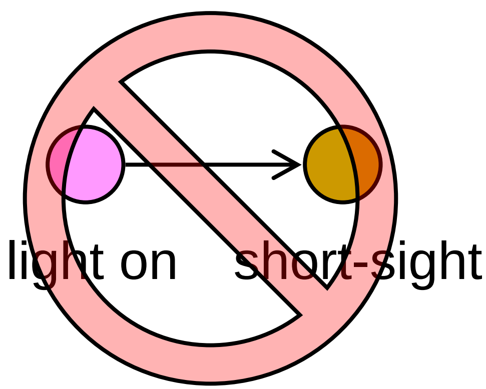
- Correlation (dependence) and causality

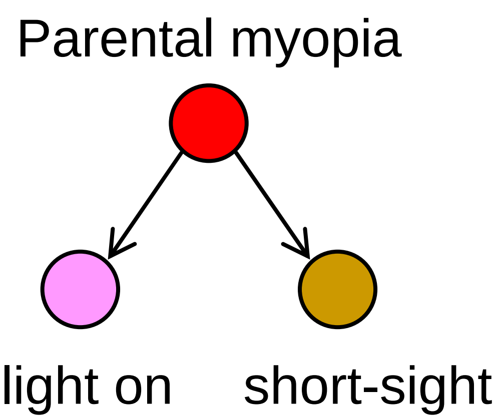
Do not confuse causality with dependence (or correlation)!

Example)

A study shows:

Young children who sleep with the light on are much more likely to develop myopia in later life. (*Nature* 1999)





(*Nature* 2000)

Hidden common cause

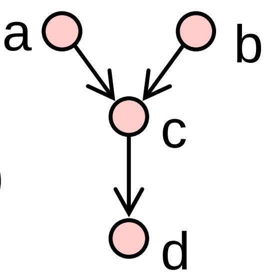
Causal Learning without Manipulation

- Fundamental assumptions
  - Causal Markov condition

    The probability generating data is associated with a DAG.

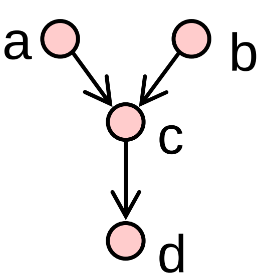
$$p(X) = \prod_{i=1}^{n} p(X_i \mid \mathrm{pa}(i))$$

$$p(X) = p(X_a)\,p(X_b)\,p(X_c \mid X_a, X_b)\,p(X_d \mid X_c)$$

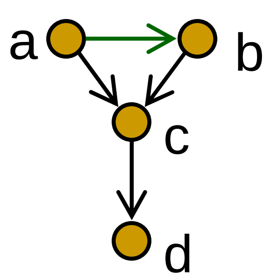


  - Causal Faithfulness Condition

    The inferred DAG (causal structure) must express all the independence relations.



true



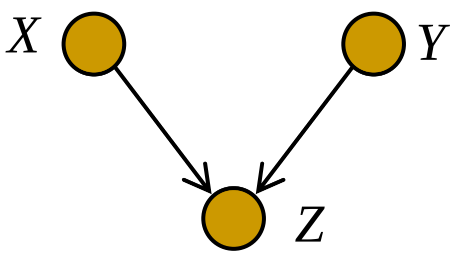
unfaithful

This includes the true probability as a special case, but the structure does not express $a \perp\!\!\!\perp b$

Causal Learning without Manipulation

- Why is it possible?
  - DAG of chain X – Z – Y

V-structure



$$\Longleftrightarrow$$

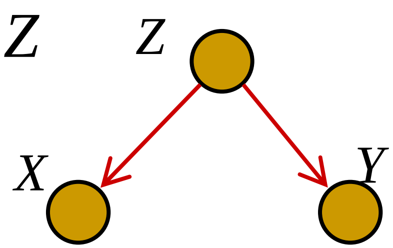
$$X \perp\!\!\!\perp Y$$

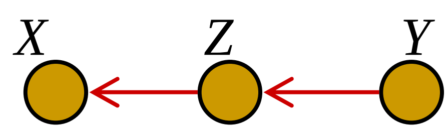
and

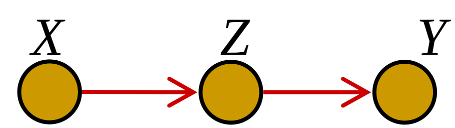
$$X \not\!\perp\!\!\!\perp Y \mid Z$$

  - This is the only detectable directed graph of three variables.
  - The following structures cannot be distinguished from the probability.

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







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

Constraint-based Causal Learning

- IC algorithm (Verma&Pearl 90)

  Input – V: set of variables, D: dataset of the variables.

  Output – Partial DAG (specifies an equivalence class, directed partially)

1. For each $(a,b) \in V \times V \ \ (a \neq b)$, search for $S_{ab} \subset V \setminus \{a,b\}$ such that

   $$X_a \perp\!\!\!\perp X_b \mid S_{ab}$$

   Construct an undirected graph (skeleton) by making an edge between a and b if and only if no set S_{ab} can be found.

2. For each nonadjacent pair (a,b) with $a - c - b$, direct the edges by $a \rightarrow c \leftarrow b$ if $c \notin S_{ab}$

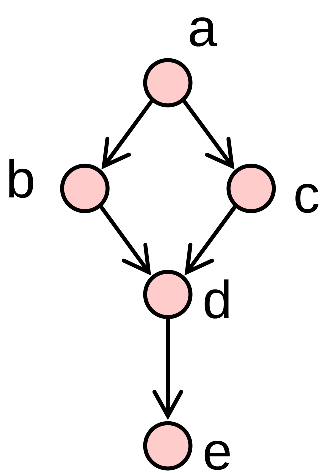
3. Orient as many of undirected edges as possible on condition that neither new v-structures nor directed cycles are created.

– Implemented in PC algorithm (Spirtes & Glymour) efficiently.

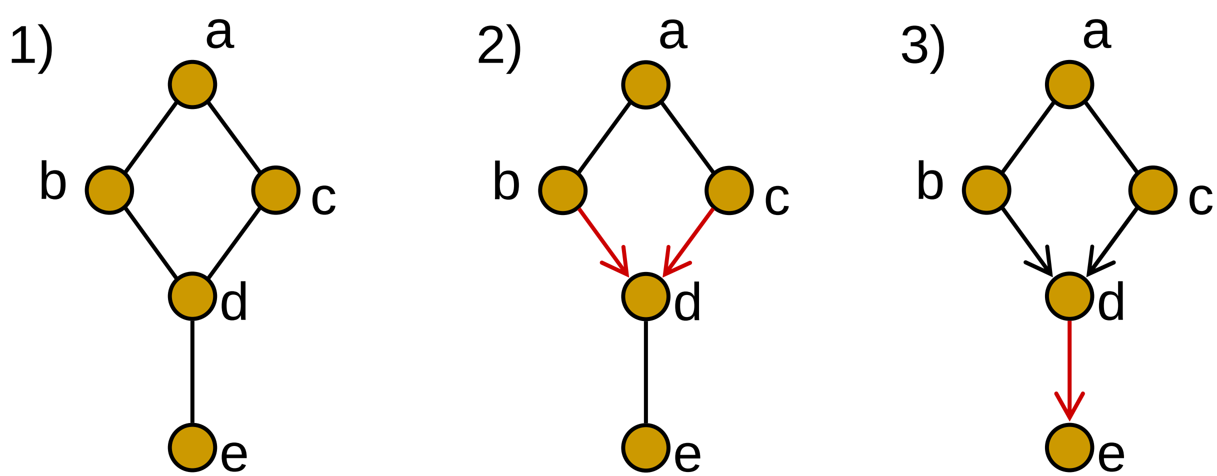
Constraint-based Causal Learning

- Example

True structure



The output from each step of IC algorithm



$$S_{ad} = \{b, c\}$$
$$S_{ae} = \{d\}$$
$$S_{bc} = \{a\}$$
$$S_{be} = S_{ce} = \{d\}$$

For other pairs, S does not exist.

For (b,c), $d \notin S_{bc}$

Direction of some edges may be left undetermined.

Mini-summary on constraint-based method

- Determine the conditional independence of the underlying probability by statistical tests.
- Many statistical tests are required.
  - Problems:
    - Errors in statistical tests.
    - Computational costs.
    - Multiple comparison – difficult to set critical regions
- Effects of hidden variables are important to consider (not discussed here).
- Often discussed in the context of causal learning.

Summary: Structure learning

- Two major approaches
  - Score-based Bayesian structure learning
    - There are many methods how to define score function.
    - Marginal likelihood, MDL, etc.
  - Constraint-based causal learning
    - Testing conditional independence.

- More recent approach
  - Sparse network by Lasso
    - Meinshausen and Buhlmann [*Ann. Statist.* **34** (2006) 1436–1462]

- Further readings
  - D. Heckerman. A tutorial on learning with Bayesian networks. in *Learning in Graphical Models*. (ed. M.Jordan) pp.301-354. MIT Press (1999)
    - This book contains various advanced topics.
  - J. Pearl. *Causality*. Cambridge University Press (2000)
  - 宮川雅巳 「統計的因果推論」朝倉書店(2004)
  - 宮川雅巳 「グラフィカルモデリング」朝倉書店(1997)