Graphical Models for Causal Inference

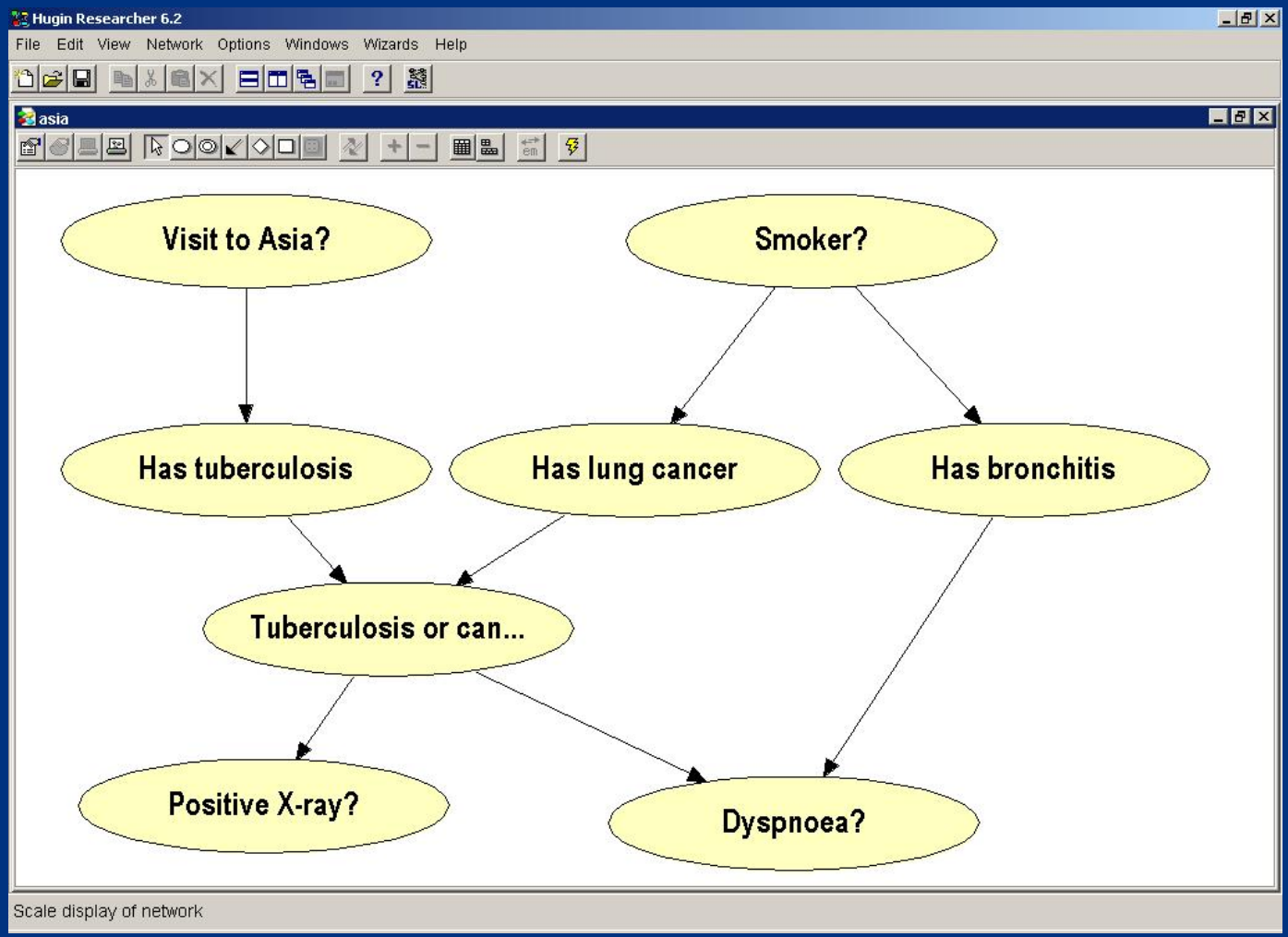
Aarhus University, Fall 2003
Lectures 11 and 12

Steffen L. Lauritzen, Aalborg University

Overview

- Causal interpretation of Bayesian networks
- Structural equation systems
- Assessment of treatment effects
- Intervention diagrams and LIMIDS
- Identifiability of causal effects
- Potential responses and mapping variables
- Discovery of (causal) structure

Why are Bayesian networks sensible?



Causal interpretation!

Intervention vs. observation

Causal interpretations are tied to the notion of *conditioning by intervention*

$$P(X = x \,|\, Y \leftarrow y) = p(x \,||\, y), \tag{1}$$

which in general is quite different from conventional conditioning or *conditioning by observation* which is

$$P(X = x \,|\, Y = y) = p(x \,|\, y) = p(x, y)/p(y).$$

A causal interpretation of a Bayesian network involves giving (1) a simple form.

Causal Bayesian network

We say that a BN is *causal w.r.t. atomic interventions at* $B \subseteq V$ if it holds for any $A \subseteq B$ that

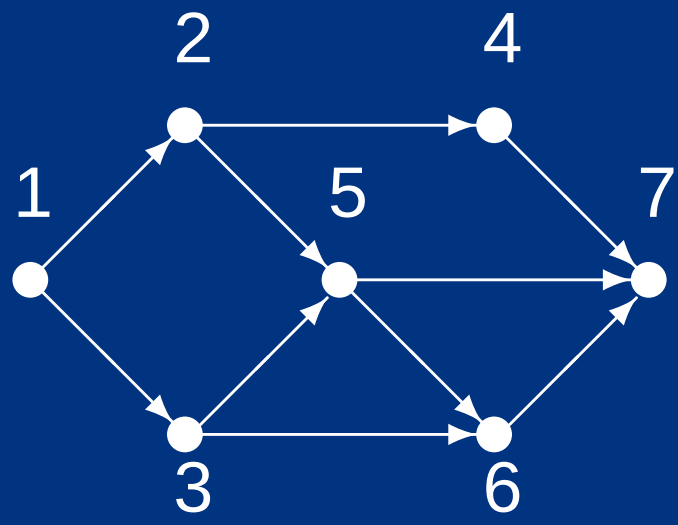
$$p(x \,||\, x^*_A) = \prod_{v \in V \setminus A} p(x_v \,|\, x_{\mathrm{pa}(v)}) \Big|_{x_A = x^*_A}$$

For $A = \emptyset$ we obtain standard factorisation.

Note that *conditional distributions* $p(x_v \,|\, x_{\mathrm{pa}(v)})$ are *stable under interventions* which do not involve x_v.

Such assumption must be justified in any given context.

Intervention vs. observation in example



$$
\begin{aligned}
p(x \,||\, x_5^*) &= p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_1)p(x_4 \,|\, x_2) \\
&\times \; p(x_6 \,|\, x_3, x_5^*)p(x_7 \,|\, x_4, x_5^*, x_6)
\end{aligned}
$$

whereas

$$
\begin{aligned}
p(x \,|\, x_5^*) &\propto p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_1)p(x_4 \,|\, x_2) \\
&\times \; p(x_5^* \,|\, x_2, x_3)p(x_6 \,|\, x_3, x_5^*)p(x_7 \,|\, x_4, x_5^*, x_6)
\end{aligned}
$$

Structural equation systems

DAG $\mathcal{D}$ can also represent structural equation system:

$$X_v \leftarrow g_v(x_{\mathrm{pa}(v)}, U_v), v \in V, \tag{2}$$

where g_v are fixed functions and U_v are independent random disturbances.

Intervention in structural equation system can be made by *replacement*, i.e. so that $X_v \leftarrow x_v^*$ is replacing the corresponding line in 'program' (2).

Corresponds to g_v and U_v being unaffected by the intervention.

Justification by structural equations

Intervention by replacement in structural equation system implies $\mathcal{D}$ causal for distribution of $X_v, v \in V$.

Occasionally used for *justification* of CBN.

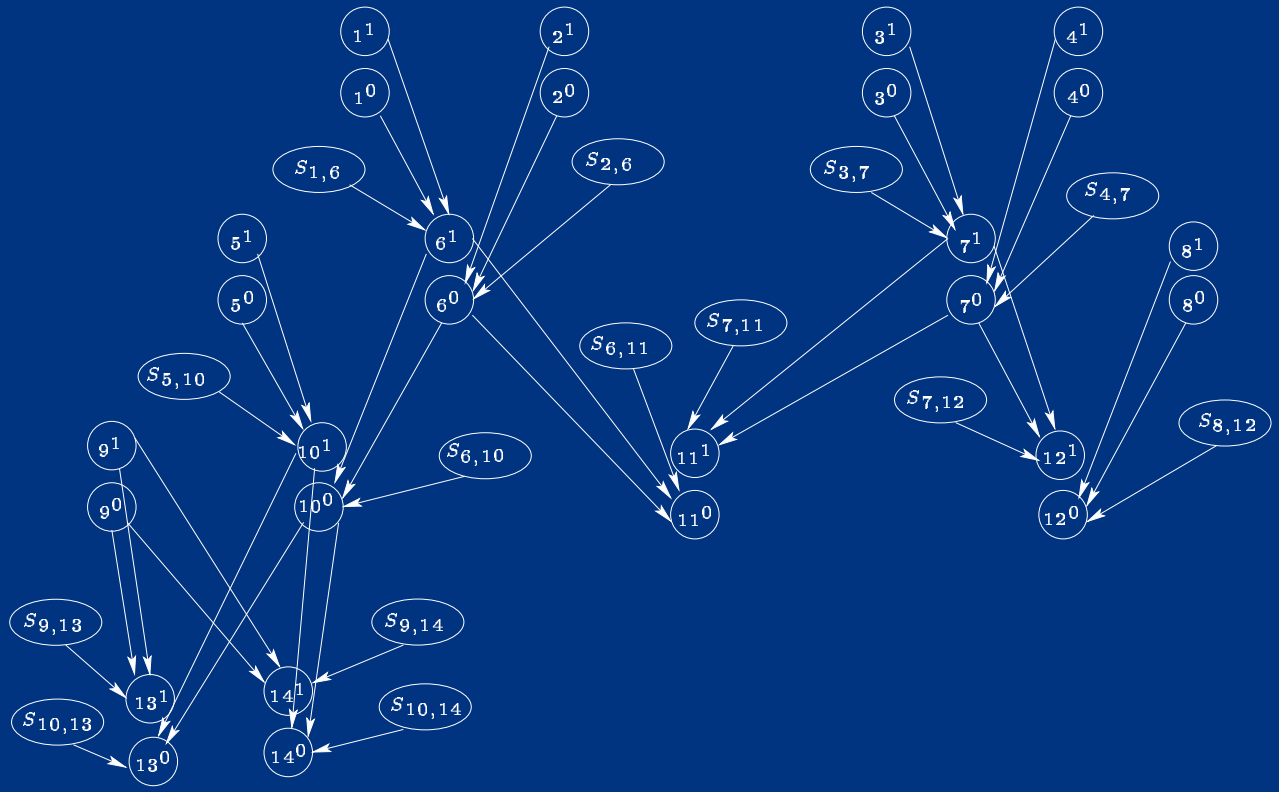
Ambiguity in choice of g_v and U_v makes this problematic.

May take *stability of conditional distributions* as a primitive rather than structural equations.

Structural equations more expressive when choice of g_v and U_v can be externally justified.

Nodes $U_v, v \in A$ can be adjoined to the network as additional parents of X_v.

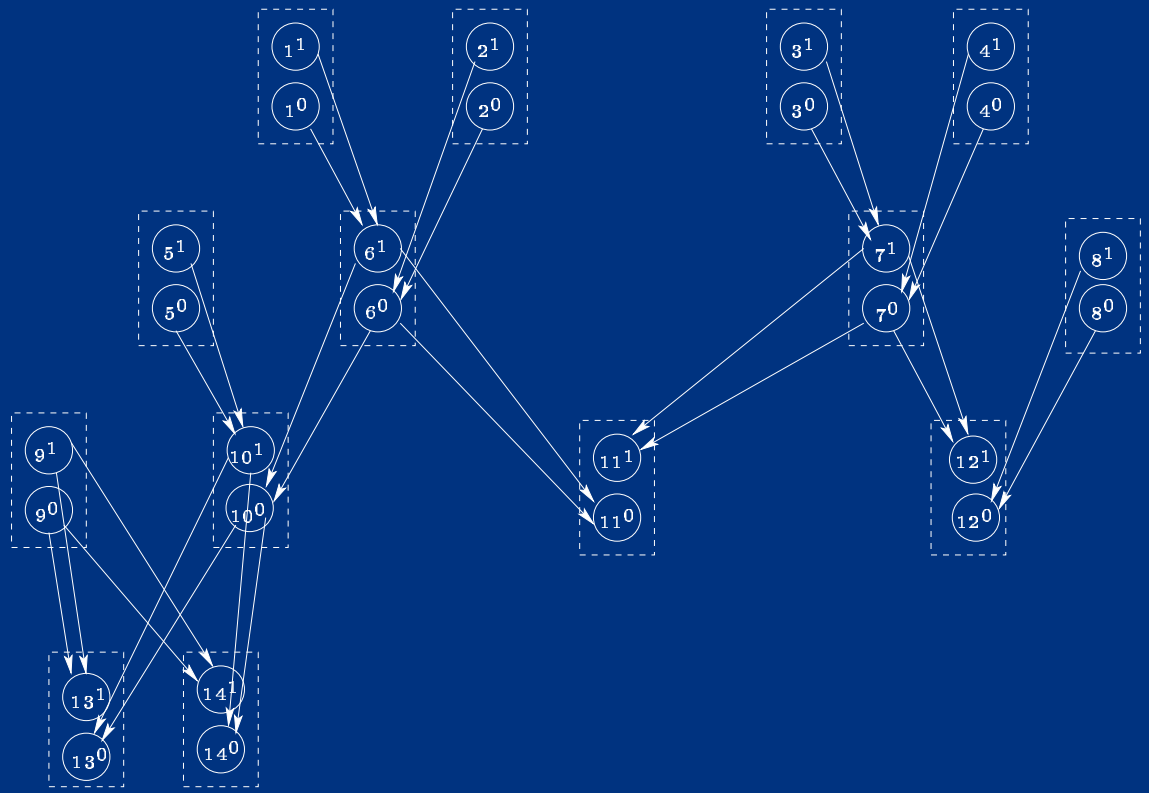
Genetic segregation network



Circles represent *alleles*. Ovals represent *segregation indicators:* 1 for paternal transmission, 0 for maternal.

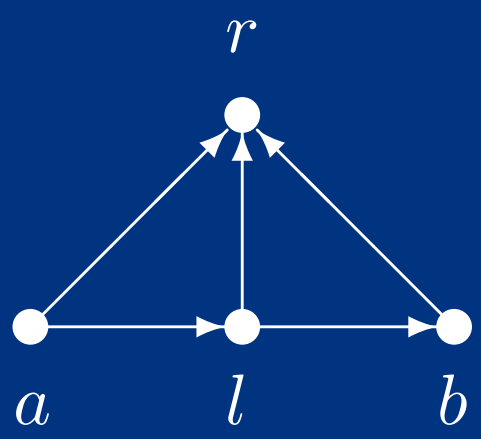
Relationships deterministic!

Allele network



Causal Markov property follows from deterministic representation as segregation network, equivalent to structural equation model.

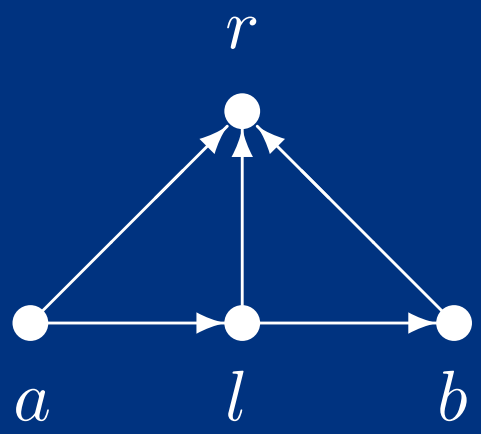
Assessment of treatment effect



a - treatment with AZT; l - intermediate response (possible lung disease); b - treatment with antibiotics; r - survival after a fixed period.

Predict survival if $X_a \leftarrow 1$ and $X_b \leftarrow 1$, assuming stable conditional distributions.

G-computation



$$
\begin{aligned}
p(1_r \,||\, 1_a, 1_b) &= \sum_{x_l} p(1_r, x_l \,||\, 1_a, 1_b) \\
&= \sum_{x_l} p(1_r \,|\, x_l, 1_a, 1_b) p(x_l \,|\, 1_a).
\end{aligned}
$$

More complex interventions

Intervene with *strategy* $\sigma_A = \{\pi_v, v \in A\}$ for choosing the 'treatments' $x_v, v \in A$ depending on the outcome of other variables in $\mathrm{pa}^*(v)$.

Stability of conditional distributions gives

$$p(x \,||\, \sigma) = \prod_{v \in A} \pi_v(x_v \,|\, x_{\mathrm{pa}^*(v)}) \prod_{v \in V \setminus A} p(x_v \,|\, x_{\mathrm{pa}(v)}).$$

Typically, $\mathrm{pa}^*(v) \neq \mathrm{pa}(v)$. Graph $\mathcal{D}^* = (V, E^*)$ must be DAG for intervention to make sense.

Variables in $\mathrm{pa}^*(v)$ must be observed before intervention on X_v is implemented.

Limited Memory Influence Diagrams

A *Limited Memory Influence Diagram* (LIMID) is a BN of *chance nodes*, *decision nodes* and *utility nodes*.

- Chance nodes Γ represented with circles
- Decision nodes Δ represented with squares.
- Utility nodes Υ represented with diamonds.
- Parents of decision nodes are observed before decision taken.

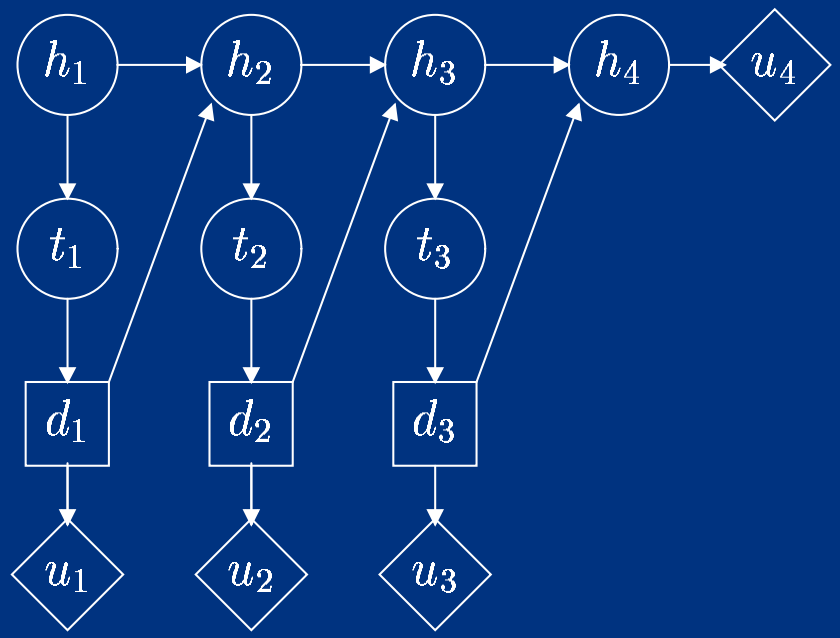
Relaxes traditional assumptions of influence diagrams, where decisions are taken in specified order and previous decisions and observations remembered.

Limited Memory Influence Diagram

h is health

t is test

d is treat or not



t_1 observed when d_1 is taken. Then t_2 is observed and d_2 is taken, etc.

Intervention diagram

Augment each node $v \in A$ where intervention is contemplated with additional parent variable F_v.

F_v has state space $\mathcal{X}_v \cup \{\phi\}$ and conditional distributions in the intervention diagram are

$$p'(x_v \mid x_{\mathrm{pa}(v)}, f_v) = \begin{cases} p(x_v \mid x_{\mathrm{pa}(v)}) & \text{if } f_v = \phi \\ \delta_{x_v, x_v^*} & \text{if } f_v = x_v^*, \end{cases}$$

where δ_{xy} is Kronecker's symbol

$$\delta_{xy} = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise.} \end{cases}$$

F_v is *forcing* the value of X_v when $F_v \neq \phi$.

Intervention diagrams

In more general setup, F_v can have parents and decision policies π can be specified.

Intervention diagrams similar to LIMIDS, but without utility nodes.

F_v correspond to *decision nodes* in LIMIDS, only with special relation to its child v.

When F_v has no parents it holds that

$$p(x) = p'(x \,|\, F_v = \phi, v \in A),$$

but also

$$\begin{aligned} p(x \,||\, x^*_B) &= P(X = x \,|\, X_B \leftarrow x^*_B) \\ &= P'(x \,|\, F_v = x^*_v, v \in B, F_v = \phi, v \in B \setminus A), \end{aligned}$$

Identifiability of causal effects

Treatment variable t, response r, set of observed covariates C, unobserved variables U.

When and how can $p(X_r \,||\, x_t)$ *be calculated from* $p(x_t, x_r, x_C)$, the latter in principle being observable from data?

Answer can be found by analysing intervention diagram.

Simplest cases known as *back-door* and *front-door* criteria and formulae.

Back-door criterion and formula

$\mathcal{D}'$ denotes $\mathcal{D}$ augmented with F_t.

Assume $C \supseteq C_0$, where C_0 satisfies

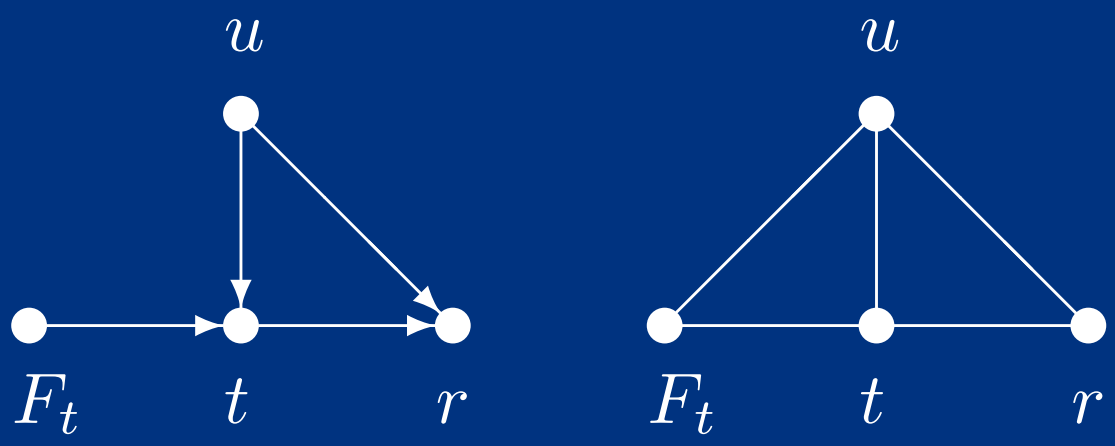
(BD1) Covariates in C_0 are unaffected by an intervention: $C_0 \perp_{\mathcal{D}'} F_t$;

(BD2) Intervention only affects response through the treatment it chooses: $R \perp_{\mathcal{D}'} F_t \,|\, C_0 \cup \{t\}$.

Then C identifies the effect of the treatment t on R as

$$p(x_r \,||\, x_t^*) = \sum_{x_{C_0}} p(x_r \,|\, x_{C_0}, x_t^*) p(x_{C_0}).$$

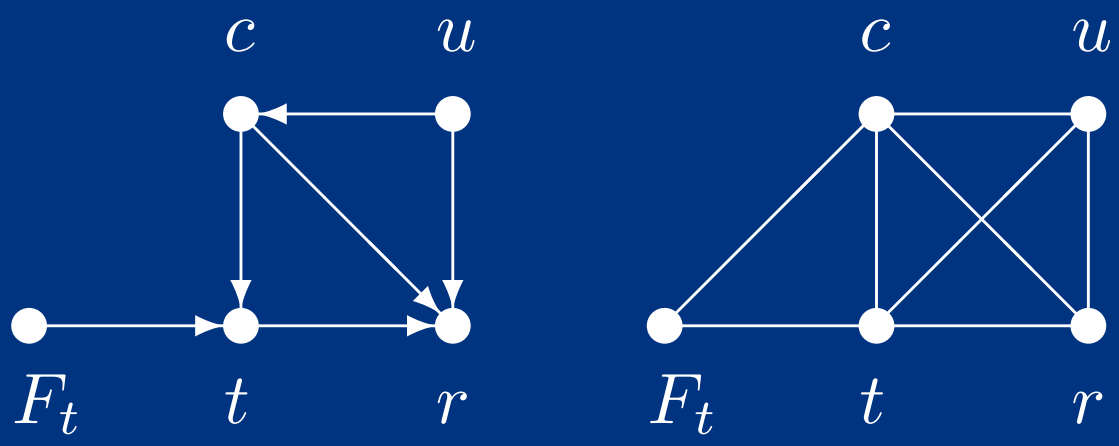
Confounding



The unobserved *confounder* X_u is affecting both treatment and response.

BD2 is violated; graph to the right reveals that F_t is *not* d-separated from r by t.

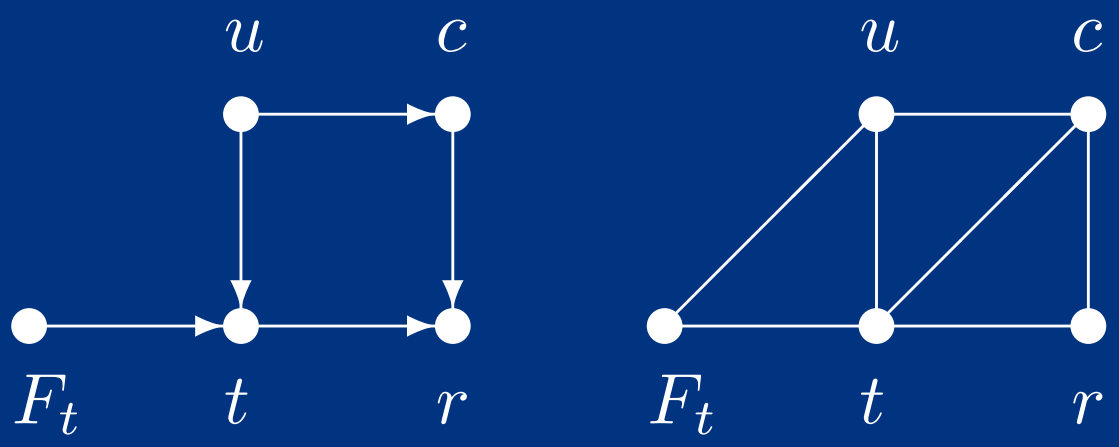
Randomisation



When X_t is randomised, possibly depending on observed covariate c, confounding is resolved.

Now $F_t \perp_{\mathcal{D}'} r \mid \{c, t\}$ and the treatment effect is identifiable.

Sufficient covariate

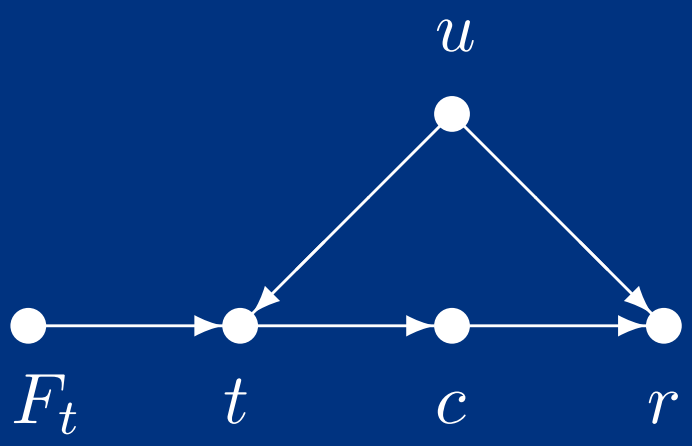


Alternatively, an observed covariate c can ‘screen away’ the confounding effect on the treatment.

Also here, $F_t \perp_{\mathcal{D}'} r \mid \{c, t\}$ and the treatment effect is identifiable.

Assumption slightly more dubious.

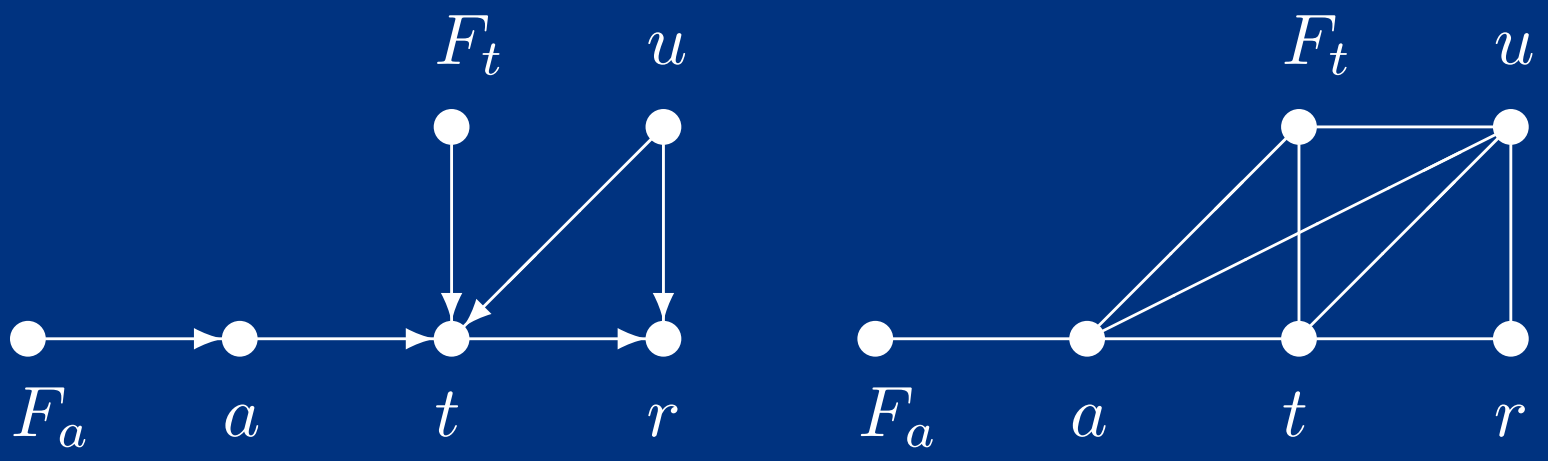
Front-door formula



In this case c is the *agent* through which the treatment effects the response. Then one can show

$$p(x_r \,||\, x_t^*) = \sum_{x_c} p(x_c \,|\, x_t^*) \sum_{x_t} p(x_r \,|\, x_c, x_t) p(x_t).$$

Partial compliance



a is treatment assigned, t is treatment taken.

The graph to the right reveals that $r \perp_{\mathcal{D}'} F_a \mid \{a, t\}$ so the effect of the treatment assignment is identified.

However, r is not d-separated from F_t by t so the effect of the treatment itself cannot be identified.

Mapping variables

In a structural equation system

$$X_v \leftarrow g_v(x_{\mathrm{pa}(v)}, U_v),$$

each (g_v, u_v) defines a map $\omega_v : \mathcal{X}_{\mathrm{pa}(v)} \to \mathcal{X}_v$ as

$$\omega_v(x_{\mathrm{pa}(v)}) = g_v(x_{\mathrm{pa}(v)}, u_v)$$

Different u_v may lead to same map.

If some of $\mathrm{pa}(v)$ are unobserved, we may consider them as part of U_v, just losing the independence among U_v.

Conversely, from *mapping variables* ω_v, we can define g_v^*

$$g_v^*(x_{\mathrm{pa}(v)}, \omega_v) = \omega_v(x_{\mathrm{pa}(v)}).$$

Potential responses

Since now

$$g_v(x_{\mathrm{pa}(v)}, u_v) = g_v^*(x_{\mathrm{pa}(v)}, \omega_v) = \omega_v(x_{\mathrm{pa}(v)}),$$

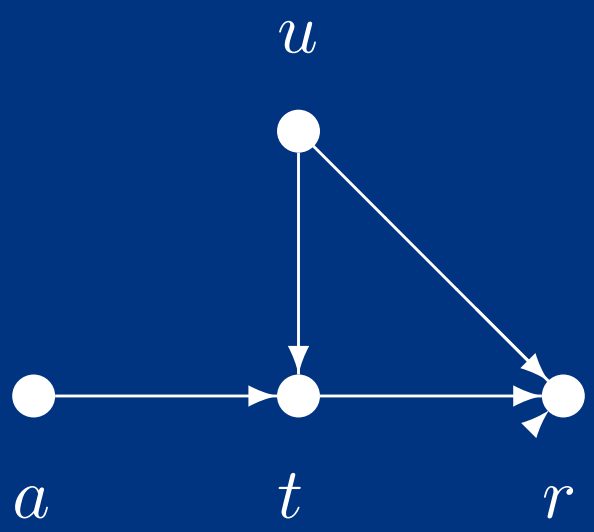
we obtain an observationally equivalent structural equation system

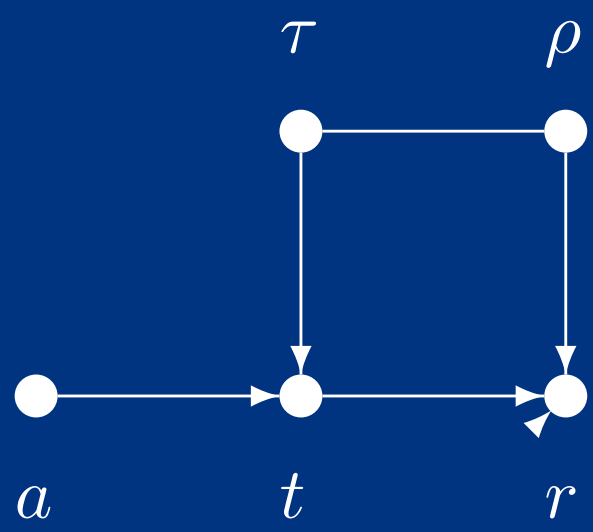
$$X_v \leftarrow g_v^*(X_{\mathrm{pa}(v)}, \Omega_v), v \in V,$$

for random maps Ω_v, a system of *canonical functional form.*

Mapping variables $\omega_v(x_{\mathrm{pa}(v)})$ describe the *potential responses*, i.e. the values of X_v that would have been observed, had the parent configuration been $x_{\mathrm{pa}(v)}$.

Partial compliance and mapping variables





$$\omega_\tau : \mathcal{X}_a \to \mathcal{X}_t, \quad X_t(x_a, \omega_\tau) \leftarrow \omega_\tau(x_a) = g_t(x_a, x_u, U_t)$$

$$\omega_\rho : \mathcal{X}_t \to \mathcal{X}_r, \quad X_r(x_t, \omega_\rho) \leftarrow \omega_\rho(x_t) = g_r(x_t, x_u, U_r).$$

Undirected link between τ and ρ indicates possible dependence.

Possible maps

Four possible maps of each if all observed variables are binary:

The maps ω_τ may well be called

$$\{\textit{always taker}, \textit{never taker}, \textit{complier}, \textit{defier}\},$$

so that

$$\textit{always taker}\,(x_a) = 1, \quad \textit{complier}\,(x_a) = x_a, \text{ etc.}$$

Similarly the four values of ω_ρ may be called

$$\{\textit{always cured}, \textit{never cured}, \textit{beneficial}, \textit{damaging}\}.$$

Causal discovery and structural learning

V variables. DAG $\mathcal{D}$ unknown and P given.

Assume P *faithful* to $\mathcal{D}$:

$$X_A \perp\!\!\!\perp X_B \mid X_S \iff A \perp_{\mathcal{D}} B \mid S$$

Most distributions are faithful

Find $\mathcal{D}$ matching conditional independences of P.

$\mathcal{D}$ and $\mathcal{D}'$ are *Markov equivalent* if the separation relations $\perp_{\mathcal{D}}$ and $\perp_{\mathcal{D}'}$ are identical.

$\mathcal{D}$ can only be determined up to Markov equivalence.
Only “causal” aspect is causal motivation for looking for DAGs.

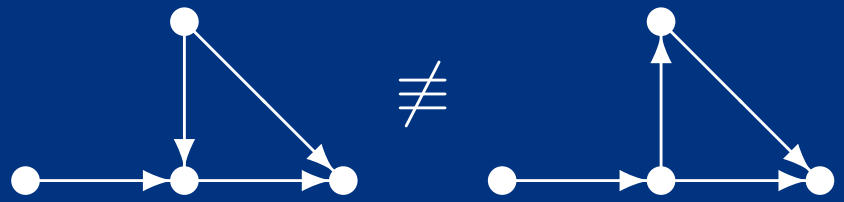
Markov equivalence

$\mathcal{D}$ and $\mathcal{D}'$ are equivalent if and only if:

1. $\mathcal{D}$ and $\mathcal{D}'$ have same *skeleton* (ignoring directions)
2. $\mathcal{D}$ and $\mathcal{D}'$ have same unmarried parents

so

but



Constraint-based search

Step 1: Identify skeleton, using that, for a faithful distribution

$$u \not\sim v \iff \exists S \subseteq V \setminus \{u, v\} : X_u \perp\!\!\!\perp X_v \mid X_S.$$

Begin with complete graph and check first for $S = \emptyset$ and remove edges when independence holds. Then continue for increasing cardinality of S.

PC-algorithm exploits that only S with $S \subseteq \mathrm{ne}(u)$ or $S \subseteq \mathrm{ne}(v)$ needs checking, where ne refers to current skeleton graph.

Step 2: Identify directions to be consistent with independence relations found in Step 1.

Exact properties of PC-algorithm

If P is faithful to DAG $\mathcal{D}$, PC-algorithm finds $\mathcal{D}'$ equivalent to $\mathcal{D}$.

It uses N independence checks where N is at most

$$N \leq 2\binom{|V|}{2}\sum_{i=0}^{d}\binom{|V|-1}{i} \leq \frac{|V|^{d+1}}{(d-1)!},$$

where d is the maximal degree of any vertex in $\mathcal{D}$.

So worst case complexity is exponential, but algorithm fast for sparse graphs.

Equivalence class searches

Searches directly in equivalence classes of DAGS.

Define *score function* $\sigma(P, \mathcal{D})$, measuring the adequacy of $\mathcal{D}$ for P with the property that

$$\mathcal{D} \equiv \mathcal{D}' \implies \sigma(P, \mathcal{D}) = \sigma(P, \mathcal{D}').$$

Typically the score function will penalise $\mathcal{D}$ with unnecessary many links.

Equivalence class with maximal score is sought.

Greedy equivalence class search

1. Initialize with empty DAG
2. Repeatedly search among equivalence classes with a single additional edge and go to class with highest score - until no improvement.
3. Repeatedly search among equivalence classes with a single edge less and move to one with highest score - until no improvement.

For suitable score functions, this algorithm identifies correct equivalence class for P. (Chickering 2002)

Data uncertainty and structural learning

Situation less clear if P is not known, but estimated:

Constraint-based: Independence checks may randomly give errors.

Algorithms more robust than PC exist.

Most checks are made with separation set S small, so 'power' high.

Asymptotically correct if e.g. marginal BIC used in checks.

Greedy equivalence search: Asymptotically correct if using BIC or fully Bayesian approach.

Latent variables and selection

More serious that *one would rarely expect all causally relevant variables to be measured.* Selection effects are also an issue.

More relevant to assume data obtained from P by *marginalisation* to subset V and *conditioning* with subset C so $W = V \cup U \cup C$, data represents P_V^C, where P is faithful to some DAG $\mathcal{D}$.

Graphs that describe independence relations in such cases are *Maximal Ancestral Graphs* (Richardson and Spirtes 2002)

Constraint-based methods for identifying MAGs exist.

Bayesian approach seems out of hand.

References

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, **3**, 507–54.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**, 161–89.

Lauritzen, S. L. (2001). Causal inference from graphical models. In *Complex stochastic systems*, (ed. O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg), pp. 63–107. Chapman and Hall/CRC Press, London/Boca Raton.

Lauritzen, S. L. and Nilsson, D. (2001). Representing and solving decision problems with limited information. *Management Science*, **47**, 1238–51.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press, Cambridge, UK.

Richardson, T. S. and Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, **30**, 963–1030.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, prediction and search*. Springer-Verlag, New York. Reprinted by MIT Press.