Graphical models for causal inference

Steffen Lauritzen, University of Oxford

Graphical Models and Inference, Lecture 14, Michaelmas Term 2009

November 26, 2009

Consider a directed acyclic graph $\mathcal{D}$ and associate for every vertex a random variable X_v. Consider now the equation system

$$X_v \leftarrow \alpha_v^\top X_{\mathrm{pa}(v)} + \beta_v + U_v, v \in V \tag{1}$$

where $U_v, v \in V$ are independent random disturbances with $U_v \sim \mathcal{N}(0, \sigma_v^2)$.

Such an equation system is known as a *recursive structural equation system*.

Structural equation systems are used heavily in social sciences and in economics. The term *structural* refers to the fact that the equations are assumed to be *stable under intervention* so that fixing a value of x_v^* would change the system only by removing the line in the equation system (1) defining x_v^*.

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

$$P(X = x \mid Y \leftarrow y) = P\{X = x \mid \mathrm{do}(Y = y)\} = p(x \,||\, y), \qquad (2)$$

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

$$P(X = x \mid Y = y) = P\{X = x \mid \mathrm{is}(Y = y)\} = p(x \mid y) = p(x, y)/p(y).$$

A causal interpretation of a Bayesian network involves giving (2) a simple form.

[Also distinguish $p(x \mid y)$ from $P\{X = x \mid \mathrm{see}(Y = y)\}$. Observation/sampling bias.]

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.

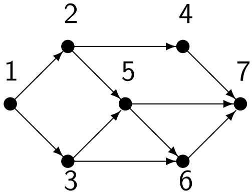
Contrast the formula for intervention conditioning with that for observation conditioning:

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

whereas

$$p(x \,|\, x_A^*) = \frac{\prod_{v \in V} p(x_v \,|\, x_{\mathrm{pa}(v)})}{p(x_A)} \Bigg|_{x_A = x_A^*}.$$

An example



$$p(x \,||\, x_5^*) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2) \times p(x_6 \mid x_3, x_5^*)p(x_7 \mid x_4, x_5^*, x_6)$$

whereas

$$p(x \mid x_5^*) \propto p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2) \times p(x_5^* \mid x_2, x_3)p(x_6 \mid x_3, x_5^*)p(x_7 \mid x_4, x_5^*, x_6)$$

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

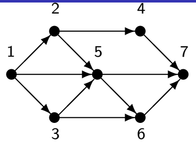
$$X_v \leftarrow g_v(x_{\mathrm{pa}(v)}, U_v), v \in V, \tag{3}$$

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' (3).

Corresponds to *g_v and U_v being unaffected by the intervention* if intervention is not made on node v. Hence the equation is *structural*.

Example revisited



For the network shown, we get

$$
\begin{aligned}
X_1 &\leftarrow \alpha_1 + U_1\\
X_2 &\leftarrow \alpha_2 + \beta_{21}x_1 + U_2\\
X_3 &\leftarrow \alpha_3 + \beta_{31}x_1 + U_3\\
X_4 &\leftarrow \alpha_4 + \beta_{42}x_2 + U_4\\
X_5 &\leftarrow \alpha_5 + \beta_{52}x_2 + \beta_{53}x_3 + U_5\\
X_6 &\leftarrow \alpha_6 + \beta_{63}x_3 + \beta_{65}x_5 + U_6\\
X_7 &\leftarrow \alpha_7 + \beta_{74}x_4 + \beta_{75}x_5 + \beta_{76}x_6 + U_7.
\end{aligned}
$$

After *intervention by replacement,* the system changes to

$$
\begin{aligned}
X_1 &\leftarrow \alpha_1 + U_1\\
X_2 &\leftarrow \alpha_2 + \beta_{21}x_1 + U_2\\
X_3 &\leftarrow \alpha_3 + \beta_{31}x_1 + U_3\\
X_4 &\leftarrow x_4\\
X_5 &\leftarrow \alpha_5 + \beta_{52}x_2 + \beta_{53}x_3 + U_5\\
X_6 &\leftarrow \alpha_6 + \beta_{63}x_3 + \beta_{65}x_5 + U_6\\
X_7 &\leftarrow \alpha_7 + \beta_{74}x_4^* + \beta_{75}x_5 + \beta_{76}x_6 + U_7.
\end{aligned}
$$

Justification of causal models by structural equations

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

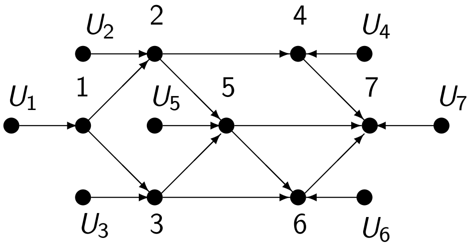
Occasionally used for *justification* of CBN.

Ambiguity in choice of g_v and U_v makes this problematic.

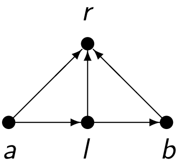
May take *stability of conditional distributions* as a primitive rather than structural equations.

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:



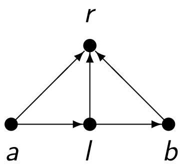
Links then represent *deterministic* relationships and all randomness is in error terms.



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



$$p(1_r \,||\, 1_a, 1_b) = \sum_{x_l} p(1_r, x_l \,||\, 1_a, 1_b)$$

$$= \sum_{x_l} p(1_r \mid x_l, 1_a, 1_b) p(x_l \mid 1_a).$$

More complex interventions

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

$$p(x \,||\, \sigma) = \prod_{v \in A} \pi_v(x_v \,|\, x_{\mathrm{pa}^*(v)}) \prod_{v \in V \setminus A} p(x_v \,|\, x_{\mathrm{pa}(v)}). \qquad (4)$$

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.

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 $F_v \neq \phi$.

In more general setup, F_v can have parents and decision policies π can be specified.
It now holds in the extended intervention diagram that

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

but also

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

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?

In this case we could say that C is a *identifier* for assessing the effect of T on R.

Answer can be found by analysing intervention diagram.

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

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

Assume $C \supseteq C_0$, where C_0 satisfies

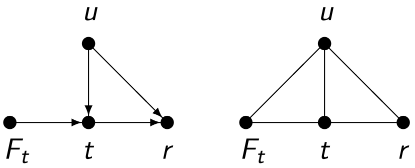
(BD1) Covariates in C_0 are unaffected by an intervention: $C_0 \perp_{\mathcal{D}'} F_t$;

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

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

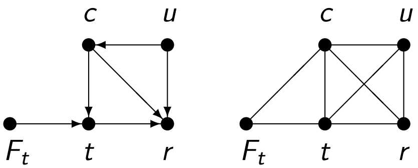
$$p(x_r \,||\, x_t^*) = \sum_{x_{C_0}} p(x_r \mid x_{C_0}, x_t^*) p(x_{C_0}).$$

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, so treatment effect is not identifiable.

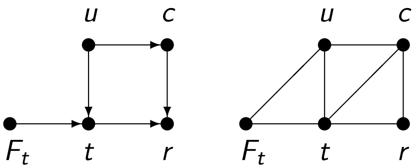
Randomisation



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

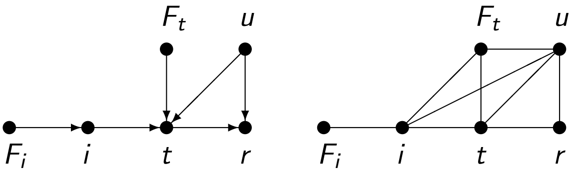
Now $F_t \perp_{\mathcal{D}'} r \mid \{c, t\}$ and *c is an identifier.*

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 *c is an identifier.*

Instrumental variable



i is an instrumental variable as it affects t and it is uncorrelated with the confounders.

Graph to the right shows $r \perp_{\mathcal{D}'} F_i \,|\, \{i, t\}$ so *the effect of the instrument can be identified.*

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

Note that *in the linear case, the effect of t on r can be found* as the ratio of effects of i on r and the effect of i on t, both of which are identified.

In the linear case, many more effects can be identified. But linearity and additivity of errors are very strong assumptions.

Bounds are available in the general case