Causal Inference from Graphical Models - I

Steffen Lauritzen, University of Oxford

Graduate Lectures

Oxford, October 2013

A *graphical model* is a set of distributions, satisfying a set of conditional independence relations encoded by a graph. This encoding is known as a *Markov property.*

In many graphical models, the Markov property is matched by a corresponding *factorization property* of the associated densities or probability mass functions.

This lecture is mostly concerned with graphical models based on *directed acyclic graphs* as these allow particularly simple causal interpretations.

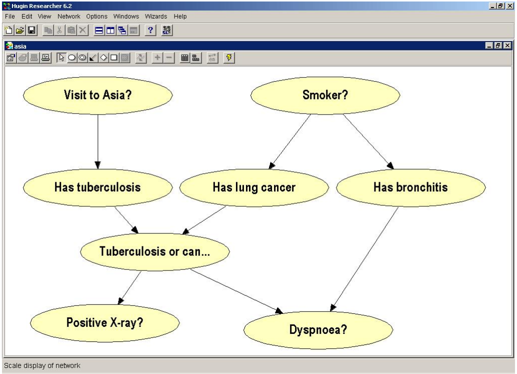
Such models are also known as *Bayesian networks*, a term coined by Pearl (1986). There is nothing Bayesian about them.

A *directed acyclic graph* $\mathcal{D}$ over a finite set V is a simple graph with all edges directed and *no directed cycles.* We use DAG for brevity.

Absence of directed cycles means that, *following arrows in the graph, it is impossible to return to any point.*

Bayesian networks have proved fundamental and useful in a wealth of interesting applications, including expert systems, genetics, complex biomedical statistics, causal analysis, and machine learning.

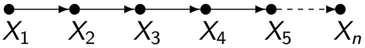
Example of a directed graphical model



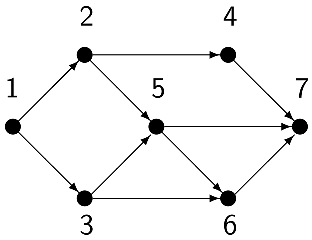
An probability distribution P of random variables $X_v, v \in V$ satisfies *the local Markov property* (L) w.r.t. a directed acyclic graph $\mathcal{D}$ if

$$\forall \alpha \in V : \alpha \perp\!\!\!\perp \{\mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha)\} \mid \mathrm{pa}(\alpha).$$

Here $\mathrm{nd}(\alpha)$ are the *non-descendants* of α.
A well-known example is a Markov chain:



with $X_{i+1} \perp\!\!\!\perp (X_1, \ldots, X_{i-1}) \mid X_i$ for $i = 3, \ldots, n$.



For example, the local Markov property says

$4 \perp\!\!\!\perp \{1, 3, 5, 6\} \mid 2,$

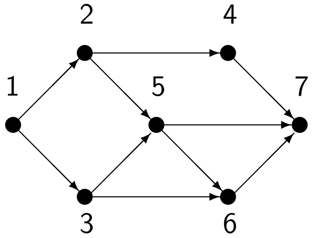
$5 \perp\!\!\!\perp \{1, 4\} \mid \{2, 3\}$

$3 \perp\!\!\!\perp \{2, 4\} \mid 1.$

A probability distribution P over $\mathcal{X} = \mathcal{X}_V$ *factorizes* over a DAG $\mathcal{D}$ if its density or probability mass function f has the form

$$(\mathrm{F}): \quad f(x) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v)}).$$

Example of DAG factorization



The above graph corresponds to the factorization

$$\begin{aligned} f(x) &= f(x_1)f(x_2 \mid x_1)f(x_3 \mid x_1)f(x_4 \mid x_2) \\ &\times f(x_5 \mid x_2, x_3)f(x_6 \mid x_3, x_5)f(x_7 \mid x_4, x_5, x_6). \end{aligned}$$

Separation in DAGs

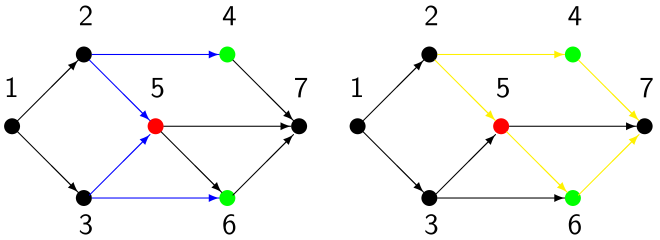
A node γ in a *trail* τ is a *collider* if edges meet head-to head at γ:

A trail τ from α to β in $\mathcal{D}$ is *active relative to S* if both conditions below are satisfied:

- all its colliders are in $S \cup \mathrm{an}(S)$
- all its non-colliders are outside S

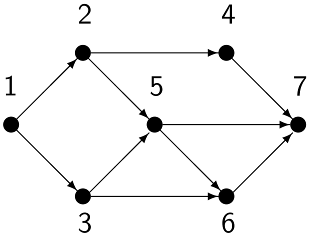
A trail that is not active is *blocked.* Two subsets A and B of vertices are *d-separated by S* if all trails from A to B are blocked by S. We write $A \perp_d B \mid S$.

Separation by example



For $S = \{5\}$, the trail $(4, 2, 5, 3, 6)$ is *active*, whereas the trails $(4, 2, 5, 6)$ and $(4, 7, 6)$ are *blocked.*
For $S = \{3, 5\}$, they are all blocked.

Returning to example



Hence $4 \perp_d 6 \mid 3, 5$, but it is *not* true that $4 \perp_d 6 \mid 5$ nor that $4 \perp_d 6$.

Equivalence of Markov properties

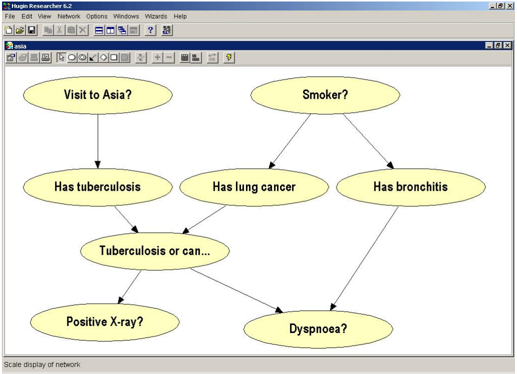
A probability distribution P satisfies the *global Markov property* (G) w.r.t. $\mathcal{D}$ if

$$A \perp_d B \mid S \Rightarrow A \perp\!\!\!\perp B \mid S.$$

It holds for any DAG $\mathcal{D}$ and any distribution P that these three directed Markov properties are equivalent:

$$(\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{F}).$$

Example is compelling for causal reasons



The now rather standard causal interpretation of a DAG (Spirtes et al., 1993; Pearl, 2000) emphasizes the distinction between *conditioning by observation* and *conditioning by intervention*.

We use special notations for this

$$P(X = x \mid Y \leftarrow y) = P\{X = x \mid \mathrm{do}(Y = y)\} = p(x \,||\, y), \qquad (1)$$

whereas

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

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

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

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.

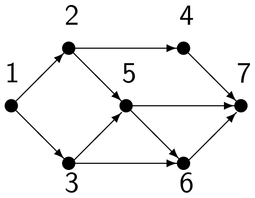
Contrast the formula for intervention conditioning with that for observation conditioning:

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

whereas

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

An example



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

whereas

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

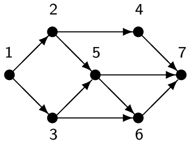
DAG $\mathcal{D}$ can also represent structural equation system:

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

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* if intervention is not made on node v. Hence the equation is *structural*.



A linear structural equation system for this network is

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

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

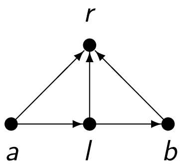
Intervention by replacement in structural equation system implies $\mathcal{D}$ causal for distribution of $X_v, v \in V$.

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.

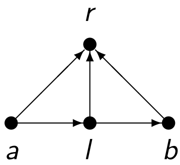
Structural equations more expressive when choice of g_v and U_v can be externally justified.



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



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

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 $F_v \neq \phi$.

It now holds in the extended intervention diagram that

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

but also

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

In particular it holds that *if* $\mathrm{pa}(v) = \emptyset$, *then* $p(x \mid x^*_v) = p(x_v \,\|\, x^*_v)$.

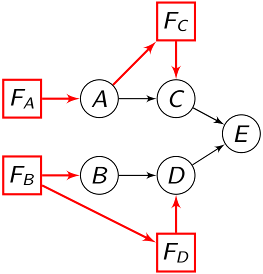
More generally we can explicitly join decision nodes $\delta \in \Delta$ to the DAG as parents of nodes which they affect.

Further, each of these can have parents in $\mathcal{D}$ or in Δ to indicate that intervention at δ may depend on states of $\mathrm{pa}(\delta)$. A *strategy* σ yields a conditional distribution of decisions, given their parents to yield

$$f(x \,||\, \sigma) = \prod_{v \in V} f(x_v \,|\, x_{\mathrm{pa}(v)}) \prod_{\delta \in \Delta} \sigma(x_\delta \,|\, x_{\mathrm{pa}(\delta)})$$

where now $\mathrm{pa}(v)$ refer to parents in the *extended diagram,* which must be a DAG to make sense.

This formally corresponds to the notion of LIMIDs (Lauritzen and Nilsson, 2001).



LIMID for a causal interpretation of a DAG. Red nodes represent (external) forces or interventions that affect the conditional distributions of their children. Note that interventions can be allowed to depend on other variables (treatment strategies).

Lauritzen, S. L. (2001). Causal inference from graphical models. In Barndorff-Nielsen, O. E., Cox, D. R., and Klüppelberg, C., editors, *Complex Stochastic Systems*, pages 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*, 47:1238–1251.

Pearl, J. (1986). Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, 29:241–288.

Pearl, J. (2000). *Causality*. Cambridge University Press, Cambridge.

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