Causal Inference from Graphical Models — II

Steffen Lauritzen, University of Oxford

Graduate Lectures

Oxford, October 2013
An probability distribution $P$ of $X_v, v \in V$ satisfies the local Markov property w.r.t. a directed acyclic graph $\mathcal{D}$ if

$$(L) : \quad \forall \alpha \in V : \alpha \perp \{\text{nd}(\alpha) \setminus \text{pa}(\alpha)\} | \text{pa}(\alpha).$$

It factorizes over $\mathcal{D}$ if its density or probability mass function $f$ has the form

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

It satisfies the global Markov property w.r.t. $\mathcal{D}$ if

$$(G) : \quad A \perp_d B | S \Rightarrow A \perp B | S.$$

These directed Markov properties are equivalent:

$$(G) \iff (L) \iff (F).$$
A node $\gamma$ in a trail $\tau$ is a *collider* if edges meet head-to-head at $\gamma$:

A trail $\tau$ from $\alpha$ to $\beta$ in $\mathcal{D}$ is *active relative to $S$* if both conditions below are satisfied:

- all its colliders are in $S \cup \text{an}(S)$
- 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 $A \perp_d B \mid S$. 

Steffen Lauritzen, University of Oxford

Causal Inference from Graphical Models — II
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.
The *moral graph* $D^m$ of a DAG $D$ is obtained by adding undirected edges between unmarried parents and subsequently dropping directions, as in the example below:
To resolve query involving three sets $A$, $B$, $S$:

1. Reduce to subgraph induced by ancestral set $\mathcal{D}_{\text{An}}(A \cup B \cup S)$ of $A \cup B \cup S$;
2. Moralize to form $(\mathcal{D}_{\text{An}}(A \cup B \cup S))^m$;
3. Say that $S$ $m$-separates $A$ from $B$ and write $A \perp_m B \mid S$ if and only if $S$ separates $A$ from $B$ in this undirected graph.

It then holds that $A \perp_m B \mid S$ if and only if $A \perp_d B \mid S$.

Proof in Lauritzen (1996) needs to allow self-intersecting paths to be correct.
Forming ancestral set

The subgraph induced by all ancestors of nodes involved in the query $4 \perp_m 6 \mid 3, 5$?
Adding links between unmarried parents

Adding an undirected edge between 2 and 3 with common child 5 in the subgraph induced by all ancestors of nodes involved in the query $4 \perp_m 6 \mid 3, 5$?
Since \( \{3, 5\} \) separates 4 from 6 in this graph, we can conclude that 
\( 4 \perp_m 6 \mid 3, 5 \)
Standard causal interpretation of any probabilistic model (Spirtes et al., 1993; Pearl, 2000) emphasizes distinction between conditioning by observation and conditioning by intervention. We use special notations for this

\[ P(X = x \mid Y \leftarrow y) = P\{X = x \mid \text{do}(Y = y)\} = p(x \mid y), \]  

whereas

\[ p(y \mid x) = p(Y = y \mid X = x) = P\{Y = y \mid \text{is}(X = x)\}. \]

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*

\[
p(x \mid \mathbf{x}_A^*) = \prod_{v \in V \setminus A} p(x_v \mid x_{pa(v)}) \bigg|_{x_A=x_A^*}
\]

*For* \( A = \emptyset \) *we obtain standard factorisation.*

Note that *conditional distributions* \( p(x_v \mid x_{pa(v)}) \) *are stable under interventions* which do not involve \( x_v \). *Such assumption must be justified in any given context.*
A linear structural equation system for this network is

\[
\begin{align*}
X_1 & \leftarrow \alpha_1 + U_1 \\
X_2 & \leftarrow \alpha_2 + \beta_{21}x_1 + U_2 \\
X_3 & \leftarrow \alpha_3 + \beta_{31}x_1 + U_3 \\
X_4 & \leftarrow \alpha_4 + \beta_{42}x_2 + U_4 \\
X_5 & \leftarrow \alpha_5 + \beta_{52}x_2 + \beta_{53}x_3 + U_5 \\
X_6 & \leftarrow \alpha_6 + \beta_{63}x_3 + \beta_{65}x_5 + U_6 \\
X_7 & \leftarrow \alpha_7 + \beta_{74}x_4 + \beta_{75}x_5 + \beta_{76}x_6 + U_7.
\end{align*}
\]
After *intervention by replacement*, the system changes to

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

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

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

where $\delta_{xy}$ 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 $F_v \neq \phi$. 

Steffen Lauritzen, University of Oxford
It now holds in the *extended* DAG, i.e. the intervention diagram that

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

but also

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

In particular it holds that *if* \( \text{pa}(v) = \emptyset \), *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 $\Delta$ to indicate that intervention at $\delta$ may depend on states of $\text{pa}(\delta)$. A strategy $\sigma$ yields a conditional distribution of decisions, given their parents to yield

$$f(x \mid \sigma) = \prod_{v \in V} f(x_v \mid x_{\text{pa}(v)}) \prod_{\delta \in \Delta} \sigma(x_\delta \mid x_{\text{pa}(\delta)})$$

where now $\text{pa}(v)$ 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).
Treatment variable $t$, response $r$, set of observed covariates $C$, unobserved variables $U$.

*When and how can $p(X_r \mid \mid x_t)$ 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.
\(D'\) denotes \(D\) augmented with \(F_t\).

Assume \(C \supseteq C_0\), where \(C_0\) satisfies

\[
(BD1) \quad \text{Covariates in } C_0 \text{ are unaffected by an intervention: } C_0 \perp_{D'} F_t;
\]

\[
(BD2) \quad \text{Intervention only affects response through chosen treatment: } R \perp_{D'} F_t \mid C_0 \cup \{t\}.
\]

Then \(C\) identifies the effect of the treatment \(t\) on \(R\) as

\[
p(x_r \mid x_t^*) = \sum_{x_{C_0}} p(x_r \mid x_{C_0}, x_t^*)p(x_{C_0}).
\]
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.
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.
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.
In this case $c$ is the \textit{agent} through which the treatment effects the response. Then one can show

$$p(x_r \mid x^*_t) = \sum_{x_c} p(x_c \mid x^*_t) \sum_{x_t} p(x_r \mid x_c, x_t)p(x_t).$$
I is an \textit{instrument} (Durbin, 1954; Bowden and Turkington, 1984; Angrist et al., 1996) if

\begin{align*}
F_t & \quad u \\
\text{Back-door criterion and formula} & \\
\text{Classic cases} & \\
\text{Front-door formula} & \\
\text{Instrumental variable} & \\
\end{align*}

The graph to the right reveals that $r \perp_{\mathcal{D}} F_i \mid \{i\}$ 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.
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.

But linearity and additivity of errors are very strong assumptions. Bounds are available in the general case using linear programming methods (Balke and Pearl, 1997; Dawid, 2003).
Mendelian randomization

Same as instrumental variable

$g$ is gene assigned, $x$ could be exposure or expression.

Bounds for exposure effects are available.
It holds
\[
\max_{x_t} \sum_{x_r} \max_{x_i} p(x_r, x_t \mid x_i) p(x_r) \leq 1, \tag{2}
\]

This *instrumental inequality* was first derived by Pearl (1995). Can be used to falsify that something is an instrument (Ramsahai and Lauritzen, 2011).
A *standard chain graph* is a mixed graph with no multiple edges, no bi-directed edges, and *no directed or semi-directed cycles* i.e. no cycles with all arrows on the cycle pointing in the same direction.

![Chain Graph Example](attachment:chain_graph_example.png)

The graph to the left is a chain graph, with *chain components* (connected components after removing arrows) \{A\}, \{B\}, \{C, D\}, \{E\}. The graph to the right is *not* a chain graph, due to the semi-directed cycle \(\langle A \rightarrow C \leftarrow D \rightarrow B \rightarrow A \rangle\).
A chain graph with no undirected edges is a *directed acyclic graph* or *DAG*.

A chain graph with no directed edges is an *undirected graph* or *UG*.

The *chain components* $\mathcal{T}$ of a chain graph are connected components of subgraph induced by undirected edges.

In a DAG, all chain components are singletons and in an undirected graph, the chain components are the connected components.
The chain graph Markov has an \textit{outer factorization}

\[ f(x) = \prod_{\tau \in \mathcal{T}} f(x_\tau | x_{pa(\tau)}) , \]  

(3)

where each factor further factorizes w.r.t. the graph $\mathcal{G}^*(\tau)$ as

\[ f(x_\tau | x_{pa(\tau)}) = Z^{-1}(x_{pa(\tau)}) \prod_{A \in A(\tau)} \phi_A(x_A), \]  

(4)

where $A(\tau)$ are the complete sets in $\mathcal{G}^*(\tau)$ and

\[ Z(x_{pa(\tau)}) = \sum x_\tau \prod_{A \in A(\tau)} \phi_A(x_A). \]

The graph $\mathcal{G}^*(\tau)$ is obtained from $\mathcal{G}_{\tau \cup pa(\tau)}$ by dropping directions on edges and adding edges between any pair of members of $pa(\tau)$. Matched by a \textit{global Markov property} as for DAGs and UGs.
Chain components \{A\}, \{B\}, \{C, D\}, \{E\}.

Outer factorization:

\[
f(x) = f(x_A)f(x_B)f(x_{CD} | x_{AB})f(x_E | x_{CD})
\]

Inner factorization:

\[
f(x_{CD} | x_{AB}) = Z^{-1}(x_{AB})\phi(x_{AC})\phi(x_{BD})\phi(x_{CD}).
\]
Recapitulating
Bayesian networks
Causal Bayesian networks
Structural equation systems
Computation of effects
Identifiability of causal effects
Chain graph models
References

Definition
Factorization and Markov property
Causal interpretation in undirected graphs
Causal chain graphs

A
\rightarrow
B
\rightarrow
D
\rightarrow
E
\rightarrow
C

\mathcal{G}_{AB}^m
\mathcal{G}_{ABCD}^m
\mathcal{G}_m

Chain components \{A\}, \{B\}, \{C, D\}, \{E\}.
Conditional independence read from sequence of moral graphs

A \perp \perp B, \quad C \perp \perp B \mid \{A, D\}, \quad D \perp \perp A \mid \{B, C\}, \quad E \perp \perp \{A, B\} \mid \{C, D\}
Intervention conditioning in an undirected graph, corresponding to ferromagnetism, is made by

\[
f(x_{V\setminus B} \mid x_B^*) = (Z^*)^{-1} \prod_{A \in A} \phi_A(x_A) \bigg|_{x_B = x_B^*} = f(x_{V\setminus B} \mid x_B^*).\]

Hence this corresponds to standard conditioning. More generally, the system can be affected by new potentials

\[
f(x_V \mid \sigma) = (Z^*)^{-1} \prod_{a \in A} \phi_A(x_A) \prod_{B \in B} \sigma_B(x_B)
\]

where the atomic interventions above correspond to some of the new potentials being Dirac delta functions, known as \textit{quenching} in Physics.
There is a *similar intervention calculus for chain graphs*

\[ f(x) = \prod_{\tau \in \mathcal{T}} f(x_\tau \mid x_{\text{pa}(\tau)}) \prod_{\delta \in \Delta} \sigma(x_\delta \mid x_{\text{pa}(\delta)}) \]

where each factor in the left product further factorizes according to the graph \( \mathcal{G}^*(\tau) \) as before. Also pa refer to parents in the extended graph, hence may include intervention nodes.

To make sense, the extended diagram must be a chain graph.

This form of LIMIDs was discussed in Cowell et al. (1999).
LIMID for a chain graph

The exact same interpretation can be given to a chain graph.
Atomic intervention conditioning in a chain graph now leads to

\[
f(x_{V\setminus A} \mid x_{A}^*) = \frac{f(x)}{\prod_{\tau \in T} f(x_{\tau \cap A} \mid x_{\text{pa}(\tau)})} \bigg|_{x_A = x_A^*}.
\]

This specializes to standard conditioning in undirected graphs and intervention conditioning in DAGs (Lauritzen and Richardson, 2002).


Recapitulating Bayesian networks
Causal Bayesian networks
Structural equation systems
Computation of effects
Identifiability of causal effects
Chain graph models


