

d-separation and Markov equivalence

Lecture 2 – Causality and Graphical Models

Christophe Ambroise

M2 Data Science – Université d'Évry

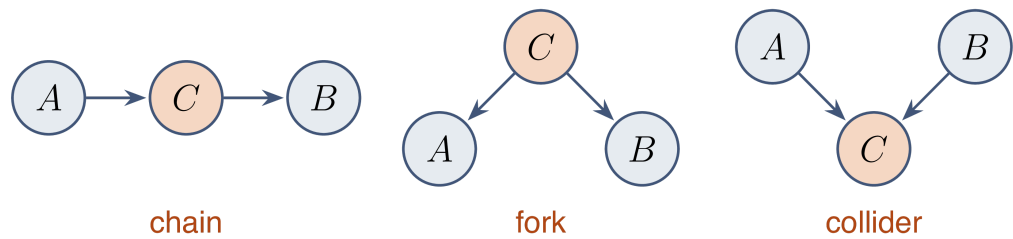
Lecture 1

- Ladder of causation: seeing $P(y \mid x)$, doing $P(y \mid do(x))$, imagining
- Simpson's paradox: confounder $\rightarrow$ reversal after aggregation
- Conditional independence: $p(x, y \mid z) = p(x \mid z) p(y \mid z)$
- DAG factorisation: $p(x) = \prod_t p(x_t \mid x_{pa(t)})$
- Gaussian Bayesian network: $x - \mu = W(x - \mu) + e$,
 $\Sigma = (I - W)^{-1} S^2 (I - W)^{-\top}$

Prior knowledge

Conditional independence in DAGs

Three elementary structures



Structure	Marginal	Given C
chain $A \rightarrow C \rightarrow B$	$A \not\perp B$	$A \perp B \mid C$
fork $A \leftarrow C \rightarrow B$	$A \not\perp B$	$A \perp B \mid C$
collider $A \rightarrow C \leftarrow B$	$A \perp B$	$A \not\perp B \mid C$

Collider: conditioning opens

Factorisation $p(a)p(b)p(c \mid a,b)$:

$$p(a,b) = \sum_c p(a)p(b)p(c \mid a,b) = p(a)p(b)$$

- $p(a,b \mid c) \propto p(a)p(b)p(c \mid a,b)$: no product form in general
- Same effect: conditioning on a **descendant** of C

Chain and fork: conditioning blocks

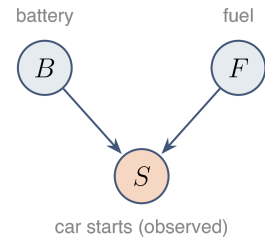
Chain, factorisation $p(a)p(c \mid a)p(b \mid c)$:

$$p(a,b \mid c) = \frac{p(a)p(c \mid a)}{p(c)}p(b \mid c) = p(a \mid c)p(b \mid c)$$

Fork, factorisation $p(c)p(a \mid c)p(b \mid c)$:

$$p(a,b \mid c) = p(a \mid c)p(b \mid c)$$

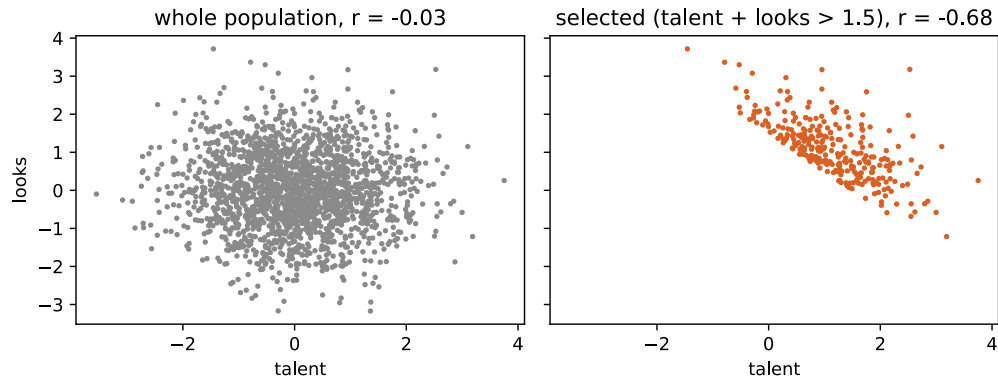
Explaining away



- Battery state $\perp$ fuel level
- Car not starting + battery fine $\rightarrow$ empty tank
- Two independent causes, dependent given the common effect

Pearl (1988), *Probabilistic Reasoning in Intelligent Systems*.

Selection on a collider: Berkson bias



Selection = conditioning on a collider $\rightarrow$ spurious negative correlation.

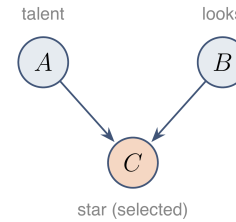
Blocked path

A path between two nodes is blocked by a set E when it contains at least one of the following nodes m .

chain $s \rightarrow m \rightarrow t$ or fork $s \leftarrow m \rightarrow t$, with $m \in E$
 collider $s \rightarrow m \leftarrow t$, with $(\{m\} \cup de(m)) \cap E = \emptyset$

Path not blocked = **active**: carries dependence.

Berkson bias as a collider



- A, B : independent causes of C
- Path blocked: $A \perp B$
- Selection on C : $A \not\perp B \mid C$
- Low-talent star $\rightarrow$ high looks: $r < 0$

Selecting stars conditions on the collider and couples the two causes.

$$p(a, b \mid C = 1) \propto p(a) p(b) \mathbb{1}\{a + b > 1.5\}$$

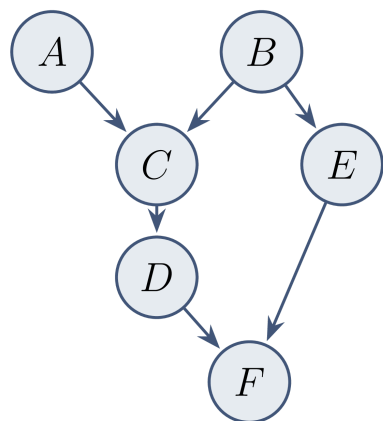
d-separation

Two sets A and B are d-separated by E when every path between a node of A and a node of B is blocked by E .

$$A \perp_G B \mid E \iff \forall \text{ path } a \rightsquigarrow b, a \in A, b \in B : \text{path blocked by } E$$

- Independence: **all** paths blocked
- Dependence: **one** active path suffices

d-separation: worked example



Query	Answer
$A \perp B$	yes
$A \perp B \mid F$	no: C opened via
$A \perp E \mid C$	no: $A \rightarrow C \leftarrow B \rightarrow$
$A \perp E \mid \{B, C\}$	yes
$D \perp E \mid \{B, C\}$	yes

Global Markov property

If p factorises according to G , every d-separation in G is a conditional independence of p .

$$p(x) = \prod_t p(x_t \mid x_{pa(t)}) \implies (A \perp_G B \mid E \Rightarrow X_A \perp X_B \mid X_E)$$

- Converse false in general: extra independences possible
- Graph = **safe proxy** for p

Local Markov property

Each node is independent of its non-descendants given its parents.

$$X_t \perp X_{nd(t) \setminus pa(t)} \mid X_{pa(t)}$$

Equivalent characterisations (DAG, density p)

factorisation $\iff$ local Markov $\iff$ global Markov

Lauritzen (1996), *Graphical Models*, th. 3.27.

Independence map

G is an I-map of p when every independence read on G holds in p .

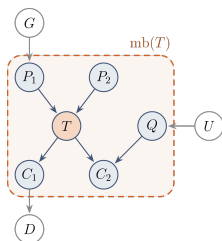
$$I(G) \subseteq I(p), \quad I(G) = \{(A, B, E) : A \perp_G B \mid E\}$$

- Complete DAG: I-map of every distribution, useless
- **Minimal I-map**: no edge removable without losing the I-map property

Markov blanket

In a DAG, the parents, children and co-parents of a node make it independent of all other nodes: this set is its Markov blanket.

$$mb(t) = \left(pa(t) \cup ch(t) \cup \bigcup_{c \in ch(t)} pa(c) \right) \setminus \{t\}, \quad X_t \perp X_{V \setminus (mb(t) \cup \{t\})} \mid X_m$$



17

Markov equivalence

Markov blanket: role of each part

- Parents: block paths towards non-descendants
- Children: block paths towards descendants
- Conditioning on children → colliders opened → **co-parents** needed
- Use: feature selection, Gibbs sampling, local structure learning

18

Faithfulness

A distribution is faithful to G when its conditional independences are exactly the d-separations of G .

$$X_A \perp X_B \mid X_E \iff A \perp_G B \mid E$$

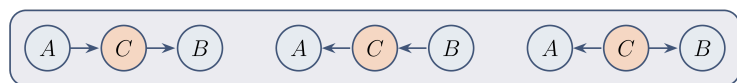
- Violation: exact cancellation of paths
- $X \rightarrow Y$ (weight a), $X \rightarrow Z \rightarrow Y$ (weights b, c), $a + bc = 0 \rightarrow X \perp Y$ despite the edges
- Cancellation: Lebesgue-measure zero set of parameters

19

20

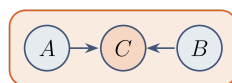
Three nodes, two equivalence classes

$$A \perp B \mid C$$



- Chain, reverse chain, fork: same independence $A \perp B \mid C$
- Collider: $A \perp B$, alone in its class

$$A \perp B$$



Markov equivalence

Two DAGs are Markov equivalent when they encode the same d-separations.

$$G_1 \sim G_2 \iff I(G_1) = I(G_2)$$

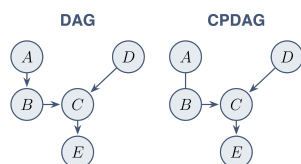
Verma–Pearl theorem (1990)

Two DAGs are Markov equivalent if and only if they have the same skeleton and the same v-structures.

v-structure: $a \rightarrow c \leftarrow b$ with a and b non-adjacent.

CPDAG: representation of a class

CPDAG (completed partially directed acyclic graph), or essential graph: a single graph standing for a whole Markov equivalence class, i.e. all DAGs encoding the same conditional independences.



- Directed edge: same orientation in every DAG of the class
- Undirected edge: orientation varies within the class
- $B \rightarrow C \leftarrow D$: v-structure; $C \rightarrow E$: forced, else new v-structure

Limit of observational data

- Observational data + faithfulness $\rightarrow$ CPDAG at best
- Gaussian linear models: no further information
- Orientation of undirected edges: extra ingredients

Beyond the equivalence class

Interventions, temporal order, or distributional assumptions (non-Gaussian noise, non-linear mechanisms).

Parameter learning, known graph

Discrete variables: conditional probability tables

Counts for node t , parent configuration c , state k :

$$N_{tck} = \sum_{i=1}^N \mathbb{1}(x_{it} = k, x_{i,pa(t)} = c), \quad \hat{\theta}_{tck} = \frac{N_{tck}}{\sum_{k'} N_{tck'}}$$

- Zero counts $\rightarrow$ zero probabilities
- Dirichlet prior α : $\hat{\theta}_{tck} = \frac{N_{tck} + \alpha}{\sum_{k'} N_{tck'} + K\alpha}$

Decomposition of the likelihood

With complete data, the likelihood is a product of one term per node and its parents.

$$\log p(\mathcal{D} \mid \theta) = \sum_{t \in V} \underbrace{\sum_{i=1}^N \log p(x_{it} \mid x_{i,pa(t)}, \theta_t)}_{\text{depends on } \theta_t \text{ only}}$$

$\rightarrow$ one independent estimation problem per node.

Gaussian variables: one regression per node

- x_t regressed on $x_{pa(t)}$: ordinary least squares
- $\hat{w}_t$: regression coefficients; $\hat{\sigma}_t^2$: residual variance
- Number of parameters: $\sum_t (|pa(t)| + 2)$