

Directed Graphical Models

Michael Gutmann

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, University of Edinburgh

Spring semester 2018

Recap

- ▶ We talked about reasonably weak assumption to facilitate the efficient representation of a probabilistic model
- ▶ Independence assumptions reduce the number of interacting variables
- ▶ Parametric assumptions restrict the way the variables may interact.
- ▶ (Conditional) independence assumptions lead to a factorisation of the pdf/pmf, e.g.

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x})p(\mathbf{y})p(\mathbf{z})$$

$$p(x_1, \dots, x_d) = p(x_d | x_{d-3}, x_{d-2}, x_{d-1})p(x_1, \dots, x_{d-1})$$

Program

1. Equivalence of factorisation and ordered Markov property
2. Understanding models from their factorisation
3. Definition of directed graphical models
4. Independencies in directed graphical models

Program

1. Equivalence of factorisation and ordered Markov property
 - Chain rule
 - Ordered Markov property implies factorisation
 - Factorisation implies ordered Markov property
2. Understanding models from their factorisation
3. Definition of directed graphical models
4. Independencies in directed graphical models

Chain rule

Iteratively applying the product rule allows us to factorise any joint pdf (pmf) $p(\mathbf{x}) = p(x_1, x_2, \dots, x_d)$ into product of conditional pdfs.

$$\begin{aligned} p(\mathbf{x}) &= p(x_1)p(x_2, \dots, x_d|x_1) \\ &= p(x_1)p(x_2|x_1)p(x_3, \dots, x_d|x_1, x_2) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4, \dots, x_d|x_1, x_2, x_3) \\ &\vdots \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \dots p(x_d|x_1, \dots, x_{d-1}) \\ &= p(x_1) \prod_{i=2}^d p(x_i|x_1, \dots, x_{i-1}) \\ &= \prod_{i=1}^d p(x_i|\text{pre}_i) \end{aligned}$$

with $\text{pre}_i = \text{pre}(x_i) = \{x_1, \dots, x_{i-1}\}$, $\text{pre}_1 = \emptyset$ and $p(x_1|\emptyset) = p(x_1)$

The chain rule can be applied to any ordering x_{k_1}, \dots, x_{k_d} . Different orderings give different factorisations.

From (conditional) independence to factorisation

$p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \text{pre}_i)$ for the ordering x_1, \dots, x_d

- ▶ For each x_i , we condition on all previous variables in the ordering.
- ▶ Assume that, for each i , there is a minimal subset of variables $\pi_i \subseteq \text{pre}_i$ such that $p(\mathbf{x})$ satisfies

$$x_i \perp\!\!\!\perp (\text{pre}_i \setminus \pi_i) \mid \pi_i$$

for all i . The distribution is then said to satisfy the **ordered Markov property**.

- ▶ By definition of conditional independence:
 $p(x_i | x_1, \dots, x_{i-1}) = p(x_i | \text{pre}_i) = p(x_i | \pi_i)$
- ▶ With the convention $\pi_1 = \emptyset$, we obtain the factorisation

$$p(x_1, \dots, x_d) = \prod_{i=1}^d p(x_i | \pi_i)$$

- ▶ See later: the π_i correspond to the parents of x_i in graphs.

From (conditional) independence to factorisation

- ▶ Assume the variables are ordered as x_1, \dots, x_d , let $\text{pre}_i = \{x_1, \dots, x_{i-1}\}$ and $\pi_i \subseteq \text{pre}_i$.
- ▶ We have seen that

$$\begin{array}{l} \text{if} \quad x_i \perp\!\!\!\perp \text{pre}_i \setminus \pi_i \mid \pi_i \text{ for all } i \\ \text{then} \quad p(x_1, \dots, x_d) = \prod_{i=1}^d p(x_i \mid \pi_i) \end{array}$$

- ▶ The chain rule corresponds to the case where $\pi_i = \text{pre}_i$.
- ▶ Do we also have the reverse?

$$\begin{array}{l} \text{if} \quad p(x_1, \dots, x_d) = \prod_{i=1}^d p(x_i \mid \pi_i) \quad \text{with } \pi_i \subseteq \text{pre}_i \\ \text{then} \quad x_i \perp\!\!\!\perp \text{pre}_i \setminus \pi_i \mid \pi_i \text{ for all } i? \end{array}$$

From factorisation to (conditional) independence

- ▶ Let us first check whether $x_d \perp\!\!\!\perp \text{pre}_d \setminus \pi_d \mid \pi_d$ holds.
- ▶ We do that by checking whether

$$p(x_d \mid \overbrace{x_1, \dots, x_{d-1}}^{\text{pre}_d}) = p(x \mid \pi_d)$$

holds.

- ▶ Since

$$p(x_d \mid x_1, \dots, x_{d-1}) = \frac{p(x_1, \dots, x_d)}{p(x_1, \dots, x_{d-1})}$$

we start with computing $p(x_1, \dots, x_{d-1})$.

From factorisation to (conditional) independence

Assume that the x_i are ordered as x_1, \dots, x_d and that $p(x_1, \dots, x_d) = \prod_{i=1}^d p(x_i | \pi_i)$ with $\pi_i \subseteq \text{pre}_i$.

We compute $p(x_1, \dots, x_{d-1})$ using the sum rule:

$$\begin{aligned} p(x_1, \dots, x_{d-1}) &= \int p(x_1, \dots, x_d) dx_d \\ &= \int \prod_{i=1}^d p(x_i | \pi_i) dx_d \\ &= \int \prod_{i=1}^{d-1} p(x_i | \pi_i) p(x_d | \pi_d) dx_d \quad (x_d \notin \pi_i, i < d) \\ &= \prod_{i=1}^{d-1} p(x_i | \pi_i) \int p(x_d | \pi_d) dx_d \\ &= \prod_{i=1}^{d-1} p(x_i | \pi_i) \end{aligned}$$

From factorisation to (conditional) independence

Hence:

$$\begin{aligned} p(x_d | x_1, \dots, x_{d-1}) &= \frac{p(x_1, \dots, x_d)}{p(x_1, \dots, x_{d-1})} \\ &= \frac{\prod_{i=1}^d p(x_i | \pi_i)}{\prod_{i=1}^{d-1} p(x_i | \pi_i)} \\ &= p(x_d | \pi_d) \end{aligned}$$

And $p(x_d | x_1, \dots, x_{d-1}) = p(x_d | \pi_d)$ means that $x_d \perp\!\!\!\perp \text{pre}_d \setminus \pi_d \mid \pi_d$ as desired.

$p(x_1, \dots, x_{d-1})$ has the same form as $p(x_1, \dots, x_d)$: apply same procedure to all $p(x_1, \dots, x_k)$, for smaller and smaller $k \leq d - 1$

Proves that

- (1) $p(x_1, \dots, x_k) = \prod_{i=1}^k p(x_i | \pi_i)$ and that
- (2) factorisation implies $x_i \perp\!\!\!\perp \text{pre}_i \setminus \pi_i \mid \pi_i$ for all i

Brief summary

- ▶ Let $\mathbf{x} = (x_1, \dots, x_d)$ be a d -dimensional random vector with pdf/pmf $p(\mathbf{x})$.
- ▶ Denote the predecessors of x_i in the ordering by $\text{pre}(x_i) = \text{pre}_i = \{x_1, \dots, x_{i-1}\}$, and let $\pi_i \subseteq \text{pre}_i$.

$$p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \pi_i) \iff x_i \perp\!\!\!\perp \text{pre}_i \setminus \pi_i \mid \pi_i \text{ for all } i$$

- ▶ Equivalence of factorisation and ordered Markov property of the pdf/pmf

Why does it matter?

- ▶ Denote the predecessors of x_i in the ordering by $\text{pre}_i = \{x_1, \dots, x_{i-1}\}$, and let $\pi_i \subseteq \text{pre}_i$.

$$p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \pi_i) \iff x_i \perp\!\!\!\perp \text{pre}_i \setminus \pi_i \mid \pi_i \text{ for all } i$$

- ▶ Why does it matter?
 - ▶ Relatively strong result: It holds for sets of pdfs/pmfs and not only single instances
 - ▶ For all members of the set: Fewer numbers are needed for their representation
 - ▶ Given the independencies, we know what form $p(\mathbf{x})$ must have.
 - ▶ Increased understanding of the properties of the model (independencies and data generation mechanism)
 - ▶ Visualisation as a graph

Program

1. Equivalence of factorisation and ordered Markov property
 - Chain rule
 - Ordered Markov property implies factorisation
 - Factorisation implies ordered Markov property
2. Understanding models from their factorisation
3. Definition of directed graphical models
4. Independencies in directed graphical models

Program

1. Equivalence of factorisation and ordered Markov property
2. Understanding models from their factorisation
 - Ancestral sampling
 - Visualisation as a directed graph
 - Description of directed graphs and topological orderings
3. Definition of directed graphical models
4. Independencies in directed graphical models

Ancestral sampling

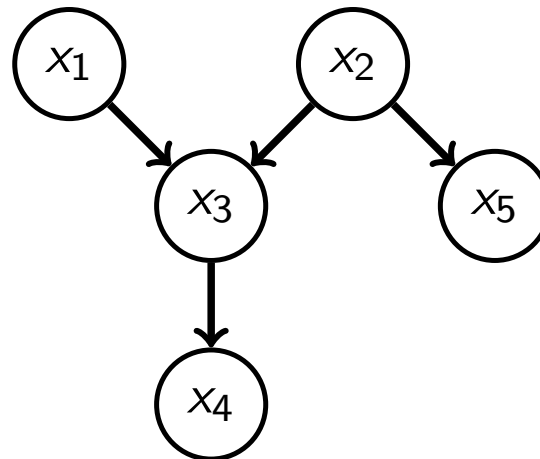
- ▶ Factorisation provides a recipe for data generation / sampling from $p(\mathbf{x})$
- ▶ Example:
$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_2)$$
- ▶ We can generate samples from the joint distribution $p(x_1, x_2, x_3, x_4, x_5)$ by sampling
 1. $x_1 \sim p(x_1)$
 2. $x_2 \sim p(x_2)$
 3. $x_3 \sim p(x_3|x_1, x_2)$
 4. $x_4 \sim p(x_4|x_3)$
 5. $x_5 \sim p(x_5|x_2)$
- ▶ Note: Helps in modelling and understanding of the properties of $p(\mathbf{x})$ but may **not** reflect causal relationships.

Visualisation as a directed graph

If $p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \pi_i)$ with $\pi_i \subseteq \text{pre}_i$ we can visualise the model as a graph with the random variables x_i as nodes, and directed edges that point from the $x_j \in \pi_i$ to the x_i . This results in a directed acyclic graph (DAG).

Example:

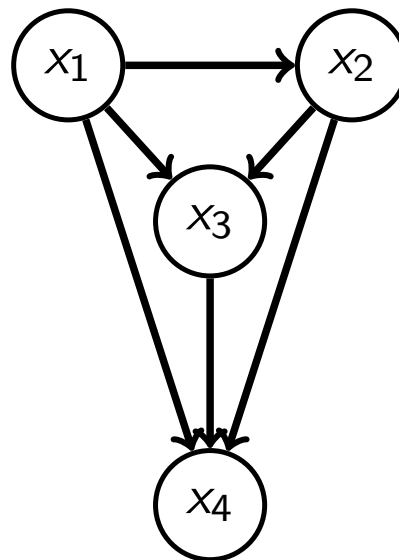
$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_2)$$



Visualisation as a directed graph

Example:

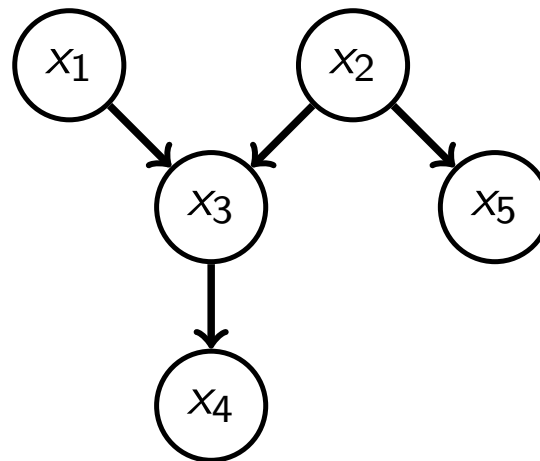
$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)$$



Factorisation obtained by chain rule \equiv fully connected directed acyclic graph.

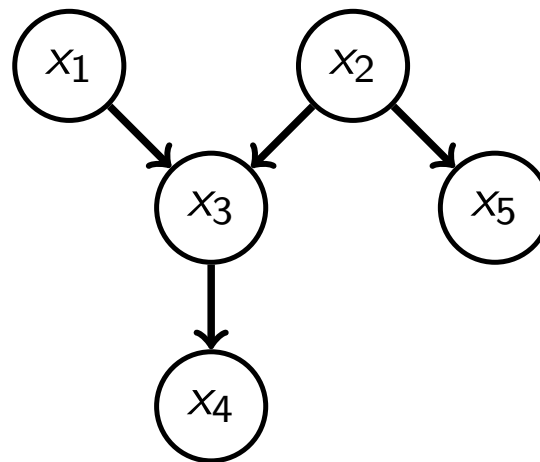
Graph concepts

- ▶ **Directed graph:** graph where all edges are directed
- ▶ **Directed acyclic graph (DAG):** by following the direction of the arrows you will never visit a node more than once
- ▶ x_i is a **parent** of x_j if there is a (directed) edge from x_i to x_j . The set of parents of x_i in the graph is denoted by $\text{pa}(x_i) = \text{pa}_i$, e.g. $\text{pa}(x_3) = \text{pa}_3 = \{x_1, x_2\}$.
- ▶ x_j is a **child** of x_i if $x_j \in \text{pa}(x_i)$, e.g. x_3 and x_5 are children of x_2 .



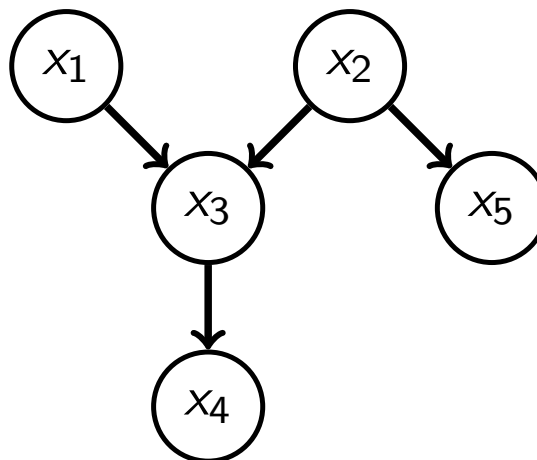
Graph concepts

- ▶ A **path** or **trail** from x_i to x_j is a sequence of distinct connected nodes starting at x_i and ending at x_j . The direction of the arrows does *not* matter. For example: x_5, x_2, x_3, x_1 is a trail.
- ▶ A **directed path** is a sequence of connected nodes where we follow the direction of the arrows. For example: x_1, x_3, x_4 is a directed path. But x_5, x_2, x_3, x_1 is not a directed path.



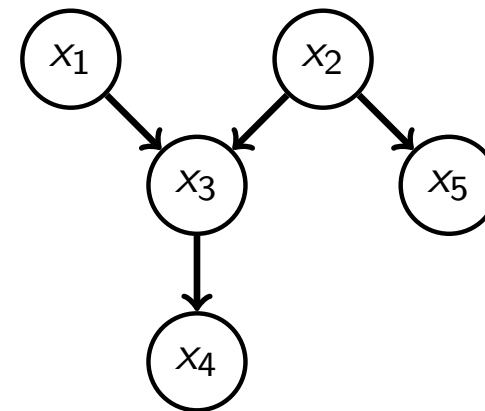
Graph concepts

- ▶ The **ancestors** $\text{anc}(x_i)$ of x_i are all the nodes where a directed path leads to x_i . For example, $\text{anc}(x_4) = \{x_1, x_3, x_2\}$.
- ▶ The **descendants** $\text{desc}(x_i)$ of x_i are all the nodes that can be reached on a directed path from x_i . For example, $\text{desc}(x_1) = \{x_3, x_4\}$.
(Note: sometimes, x_i is included in the set of ancestors and descendants)
- ▶ The **non-descendants** of x_i are all the nodes in a graph without x_i and without the descendants of x_i . For example, $\text{nondesc}(x_3) = \{x_1, x_2, x_5\}$



Graph concepts

- ▶ **Topological ordering:** an ordering (x_1, \dots, x_d) of some variables x_i is topological relative to a graph if, whenever there is a directed edge from x_i to x_j , x_i occurs prior to x_j in the ordering (“parents come before the children”). There is always at least one such ordering for DAGs.
- ▶ For a pdf $p(\mathbf{x})$, assume you order the random variables x_i in some manner and compute the corresponding factorisation, e.g. $p(\mathbf{x}) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_2)$
- ▶ When you visualise the factorised pdf as a graph, the graph is always such that the ordering used for the factorisation is topological to it.
- ▶ The π_i in the factorisation are equal to the parents pa_i in the graph. We may call both sets the “parents” of x_i .



Summary

1. Equivalence of factorisation and ordered Markov property
 - Chain rule
 - Ordered Markov property implies factorisation
 - Factorisation implies ordered Markov property
2. Understanding models from their factorisation
 - Ancestral sampling
 - Visualisation as a directed graph
 - Description of directed graphs and topological orderings

Program

1. Equivalence of factorisation and ordered Markov property
2. Understanding models from their factorisation
3. Definition of directed graphical models
 - Via factorisation according to the graph
 - Via ordered Markov property
 - Derive independencies from the ordered Markov property with different topological orderings
4. Independencies in directed graphical models

Directed graphical model

- ▶ We started with a pdf/pdf, wrote it in factorised form according to some ordering, and associated a DAG with it.
- ▶ We can also go the other way around and start with a DAG.
- ▶ *Definition (via factorisation property)* A directed graphical model based on a DAG with d nodes and associated random variables x_i is the set of pdfs/pmfs that factorise as

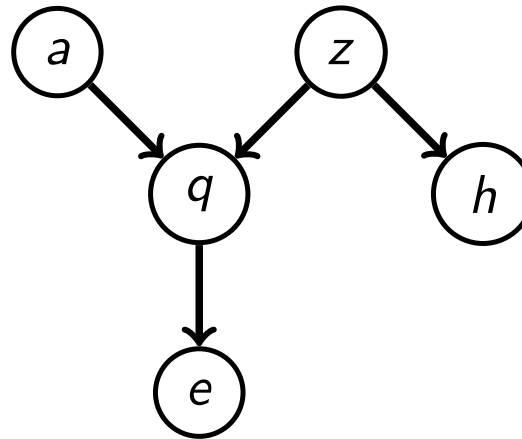
$$p(x_1, \dots, x_d) = \prod_{i=1}^d p(x_i | \text{pa}_i),$$

where pa_i denotes the parents of x_i in the graph.

- ▶ Other names for directed graphical models: belief network, Bayesian network, Bayes network.

Example

DAG:



Random variables: a, z, q, e, h

Parent sets: $pa_a = pa_z = \emptyset, pa_q = \{a, z\}, pa_e = \{q\}, pa_h = \{z\}$.

All models defined by the DAG factorise as:

$$p(a, z, q, e, h) = p(a)p(z)p(q|a, z)p(e|q)p(h|z)$$

Alternative definition of directed graphical models

- ▶ For any DAG with d nodes we can always find a topological ordering of the associated random variables. Re-label the nodes accordingly as x_1, \dots, x_d .
- ▶ In a topological ordering the parents come before the children.
- ▶ Hence: $\text{pa}_i \subseteq \text{pre}_i$ (recall: $\text{pre}_i = \{x_1, \dots, x_{i-1}\}$)
- ▶ Previous result on equivalence of factorisation and ordered Markov property gives

$$p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \text{pa}_i) \iff x_i \perp\!\!\!\perp \text{pre}_i \setminus \text{pa}_i \mid \text{pa}_i \text{ for all } i$$

- ▶ Provides an alternative definition of directed graphical models

Directed graphical model

- ▶ *Definition (via ordered Markov property)* A directed graphical model based on a DAG with d nodes and associated random variables x_i is the set of pdfs/pmfs that satisfy the ordered Markov property

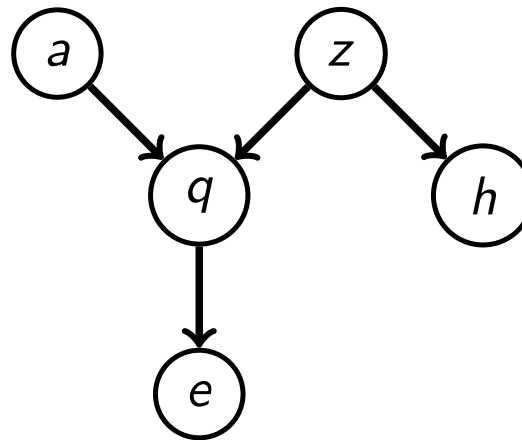
$$x_i \perp\!\!\!\perp \text{pre}_i \setminus \text{pa}_i \mid \text{pa}_i \text{ for all } i$$

for any topological ordering x_1, \dots, x_d of the x_i .

- ▶ Remark: the notation is as before:
 pre_i are the predecessors of x_i *in the topological ordering chosen*
 pa_i are the parents of x_i *in the graph*
- ▶ Remark: The missing edges in the graph cause the pa_i to be smaller than the pre_i , and thus lead to the independencies.

Example

DAG:



Random variables: a, z, q, e, h

Ordering: (a, z, q, e, h) (meaning: $x_1 = a, x_2 = z, x_3 = q, x_4 = e, x_5 = h$)

Predecessor sets for the ordering:

$\text{pre}_a = \emptyset, \text{pre}_z = \{a\}, \text{pre}_q = \{a, z\}, \text{pre}_e = \{a, z, q\}, \text{pre}_h = \{a, z, q, e\}$

Parent sets: as before

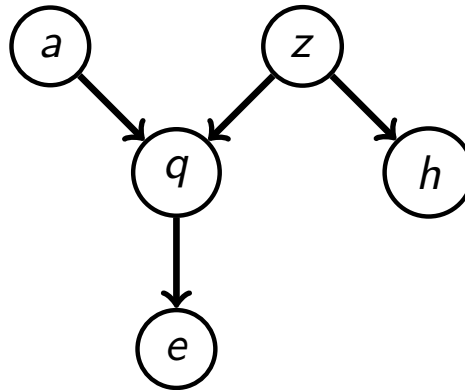
$\text{pa}_a = \text{pa}_z = \emptyset, \text{pa}_q = \{a, z\}, \text{pa}_e = \{q\}, \text{pa}_h = \{z\}$

All models in the set defined by the DAG satisfy $x_i \perp\!\!\!\perp \text{pre}_i \setminus \text{pa}_i \mid \text{pa}_i$:

$$z \perp\!\!\!\perp a \quad e \perp\!\!\!\perp \{a, z\} \mid q \quad h \perp\!\!\!\perp \{a, q, e\} \mid z$$

Example (different topological ordering)

DAG:



Ordering: (a, z, h, q, e)

Predecessor sets for the ordering:

$\text{pre}_a = \emptyset, \text{pre}_z = \{a\}, \text{pre}_h = \{a, z\}, \text{pre}_q = \{a, z, h\}, \text{pre}_e = \{a, z, h, q\}$

Parent sets: as before

$\text{pa}_a = \text{pa}_z = \emptyset, \text{pa}_h = \{z\}, \text{pa}_q = \{a, z\}, \text{pa}_e = \{q\}$

All models in the set defined by the DAG satisfy $x_i \perp\!\!\!\perp \text{pre}_i \setminus \text{pa}_i \mid \text{pa}_i$:

$$z \perp\!\!\!\perp a \quad h \perp\!\!\!\perp a \mid z \quad q \perp\!\!\!\perp h \mid a, z \quad e \perp\!\!\!\perp \{a, z, h\} \mid q$$

Note: the models also satisfy those obtained with the previous ordering:

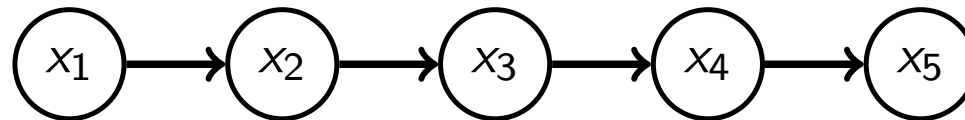
$$z \perp\!\!\!\perp a \quad e \perp\!\!\!\perp \{a, z\} \mid q \quad h \perp\!\!\!\perp \{a, q, e\} \mid z$$

Remarks

- ▶ The directed graphical model corresponds to a *set* of probability distributions. Two views according to the two definitions: The set includes all those distributions that you get
 - ▶ by looping over all possible conditionals $p(x_i|\text{pa}_i)$,
 - ▶ by retaining, from all possible joint distributions over the x_i , those that satisfy the ordered Markov property
- ▶ A directed graphical model with specified conditionals is typically also called a directed graphical model.
- ▶ By using different topological orderings you can generate possibly different independence relations satisfied by the model.
- ▶ We will see that the directed Markov properties obtained from one ordering induces all from the other orderings. This means that the directed graphical model can be specified via the directed Markov properties for one topological ordering only.

Example: Markov model

DAG:



All models in the set factorise as

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)p(x_5|x_4)$$

There is only one topological ordering: (x_1, x_2, \dots, x_5)

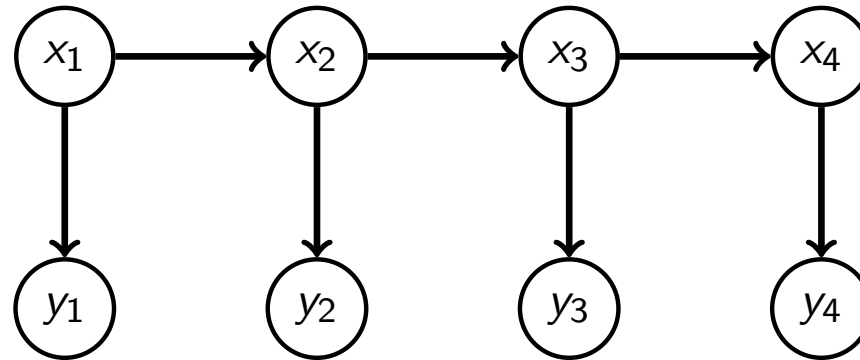
By ordered Markov property: all models in the set satisfy:

$$x_{i+1} \perp\!\!\!\perp x_1, \dots, x_{i-1} \mid x_i$$

(future independent of the past given the present)

Example: Hidden Markov model

DAG:



Called “hidden” Markov model because we typically assume to only observe the y_i and not the x_i that follow a Markov model.

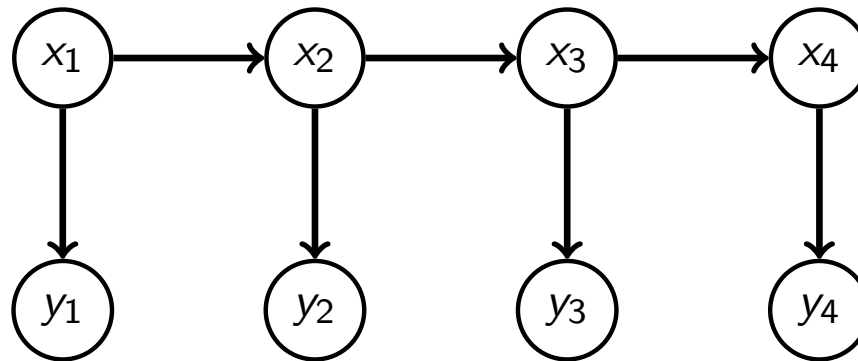
All models in the set factorise as $p(x_1, y_1, x_2, y_2, \dots, x_4, y_4) = p(x_1)p(y_1|x_1)p(x_2|x_1)p(y_2|x_2)p(x_3|x_2)p(y_3|x_3)p(x_4|x_3)p(y_4|x_4)$

With topological ordering $(x_1, y_1, x_2, y_2, \dots)$, the models in the set satisfy:

$$y_i \perp\!\!\!\perp x_1, y_1, \dots, x_{i-1}, y_{i-1} \mid x_i \quad x_i \perp\!\!\!\perp x_1, y_1, \dots, x_{i-2}, y_{i-2}, y_{i-1} \mid x_{i-1}$$

Example: Hidden Markov model

DAG:



With topological ordering $(x_1, x_2, \dots, x_4, y_1, \dots, y_4)$, the models in the set satisfy:

$$y_i \perp\!\!\!\perp x_1, \dots, x_{i-1}, y_1, \dots, y_{i-1} \mid x_i \quad x_i \perp\!\!\!\perp x_1, \dots, x_{i-2} \mid x_{i-1}$$

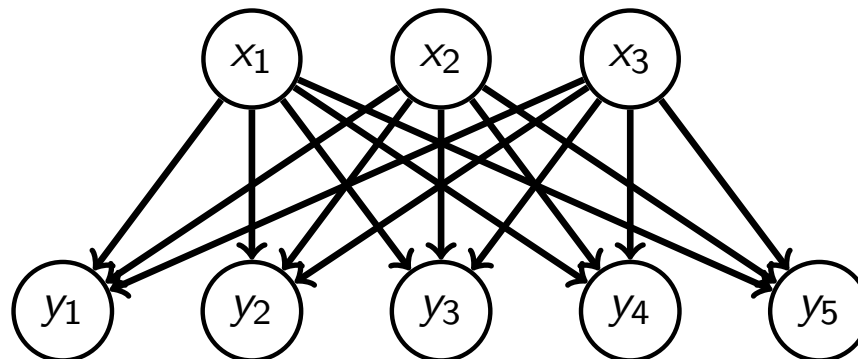
Independence relations obtained before:

$$y_i \perp\!\!\!\perp x_1, y_1, \dots, x_{i-1}, y_{i-1} \mid x_i \quad x_i \perp\!\!\!\perp x_1, y_1, \dots, x_{i-2}, y_{i-2}, y_{i-1} \mid x_{i-1}$$

Example: Probabilistic PCA, factor analysis, ICA

(PCA: principal component analysis; ICA: independent component analysis)

DAG:



Explains properties of (observed) y_i through fewer (unobserved) x_i .
Different further assumptions lead to different methods (more later).

All models in the set factorise as $p(x_1, x_2, x_3, y_1, \dots, y_5) = p(x_1)p(x_2)p(x_3)p(y_1|x_1, x_2, x_3)p(y_2|x_1, x_2, x_3) \dots p(y_5|x_1, x_2, x_3)$

With the ordering $(x_1, x_2, x_3, y_1, \dots, y_5)$: All satisfy:

$$\begin{aligned} x_i \perp\!\!\!\perp x_j & \quad y_2 \perp\!\!\!\perp y_1 \mid x_1, x_2, x_3 & \quad y_3 \perp\!\!\!\perp y_1, y_2 \mid x_1, x_2, x_3 \\ y_4 \perp\!\!\!\perp y_1, y_2, y_3 \mid x_1, x_2, x_3 & \quad y_5 \perp\!\!\!\perp y_1, y_2, y_3, y_4 \mid x_1, x_2, x_3 \end{aligned}$$

Program

1. Equivalence of factorisation and ordered Markov property
2. Understanding models from their factorisation
3. Definition of directed graphical models
 - Via factorisation according to the graph
 - Via ordered Markov property
 - Derive independencies from the ordered Markov property with different topological orderings
4. Independencies in directed graphical models

Program

1. Equivalence of factorisation and ordered Markov property
2. Understanding models from their factorisation
3. Definition of directed graphical models
4. **Independencies in directed graphical models**
 - Three canonical connections in a DAG and their properties
 - D-separation and I-map
 - Directed local Markov property
 - Equivalences of the different Markov properties and the factorisation
 - Markov blanket

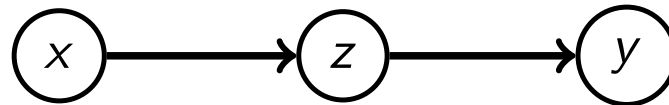
Further independence properties?

- ▶ Parent-child links in the graph encode (conditional) independence properties.
- ▶ Ordered Markov property yields sets of independence assertions.
- ▶ Questions:
 - ▶ Does the graph induce or impose additional independencies on any probability distribution that factorises over the graph?
 - ▶ For specific (x, y, z) , can we determine whether $x \perp\!\!\!\perp y|z$ holds?
- ▶ Important because
 - ▶ it yields increased understanding of the properties of the model
 - ▶ we can exploit the independencies e.g. for inference and learning
- ▶ Approach: Investigate how probabilistic evidence that becomes available at a node can “flow” through the DAG and influence our belief about another node (d-separation).

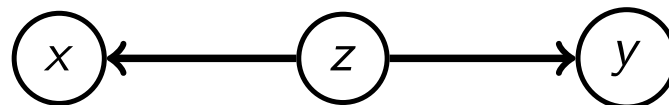
Three canonical connections in a DAG

In a DAG, two nodes x, y can be connected via a third node z in three ways:

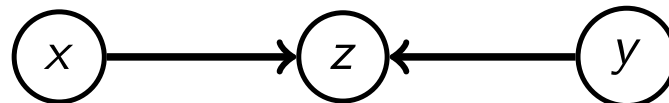
1. Serial connection (chain, head-tail or tail-head)



2. Diverging connection (fork, tail-tail)



3. Converging connection (collider, head-head, v-structure)



Note: in any case, the sequence x, z, y forms a trail

Serial connection

- ▶ Markov model is made up of serial connections
- ▶ Graph: x influences z , which in turn influences y but no direct influence from x to y .
- ▶ Factorisation: $p(x, z, y) = p(x)p(z|x)p(y|z)$
- ▶ Ordered Markov property: $y \perp\!\!\!\perp x \mid z$
If the state or value of z is known (i.e. if the random variable z is “instantiated”), evidence about x will not change our belief about y , and vice versa.

We say that the z node is “closed” and that the trail between x and y is “blocked” by the instantiated z . In other words, knowing the value of z blocks the flow of evidence *between* x and y .

Serial connection

- ▶ What can we say about the marginal distribution of (x, y) ?
- ▶ By sum rule, joint probability distribution of (x, y) is

$$\begin{aligned} p(x, y) &= \int p(x)p(z|x)p(y|z)dz \\ &= p(x) \int p(z|x)p(y|z)dz \\ &\neq p(x)p(y) \end{aligned}$$

- ▶ In a serial connection, if the state of z is unknown, then evidence or information about x will influence our belief about y , and the other way around. Evidence can flow through z between x and y .
- ▶ We say that the z node is “open” and the trail between x and y is “active”.

Diverging connection

- ▶ Graph for probabilistic PCA, factor analysis, ICA has such connections (z correspond to the latents, x and y to the observed)
- ▶ Graph: z influences both x and y . No directed connection between x and y .
- ▶ Factorisation: $p(x, y, z) = p(z)p(x|z)p(y|z)$
- ▶ Ordered Markov property (with ordering z, x, y): $y \perp\!\!\!\perp x \mid z$
If the state or value z is known, evidence about x will not change our belief about y , and vice versa.
- ▶ As in serial connection, knowing z closes the z node, which blocks the trail between x and y .

Diverging connection

- ▶ What can we say about the marginal distribution of (x, y) ?
- ▶ By sum rule, joint probability distribution of (x, y) is

$$p(x, y) = \int p(z)p(x|z)p(y|z)dz \\ \neq p(x)p(y)$$

- ▶ In a diverging connection, as in the serial connection, if the state of z is unknown, then evidence or information about x will influence our belief about y , and the other way around. Evidence can flow through z between x and y .
- ▶ The z node is open and the trail between x and z is active.

Converging connection

- ▶ Graph for probabilistic PCA, factor analysis, ICA has such connections (z corresponds to an observed, x and y to two latents)
- ▶ Graph: x and y influence z . No direction connection between x and y .
- ▶ Factorisation: $p(x, y, z) = p(x)p(y)p(z|x, y)$
- ▶ Ordered Markov property: $x \perp\!\!\!\perp y$
If nothing is known about z , except what might follow from knowledge of x and y , then evidence about x will not change our belief about y , and vice versa.

If no evidence about z is available, the z node is closed, which blocks the trail between x and y .

Converging connection

- ▶ This means that the marginal distribution of (x, y) factorises:
 $p(x, y) = p(x)p(y)$
- ▶ Conditional distribution of (x, y) given z ?

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x)p(y)p(z|x, y)}{\int p(x)p(y)p(z|x, y)dx dy}$$
$$\neq p(x|z)p(y|z)$$

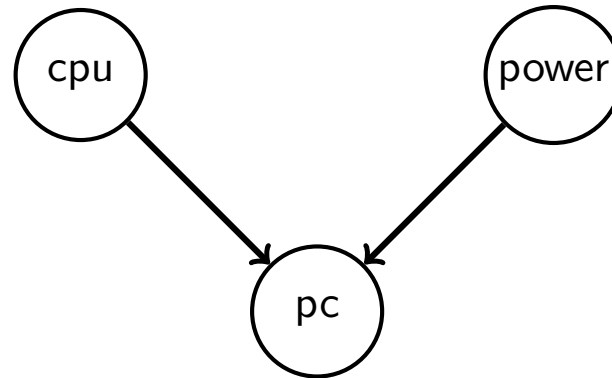
This means that $x \not\perp y | z$.

- ▶ If evidence or information about z is available, evidence about x will influence the belief about y , and vice versa.
- ▶ Information about z opens the z -node, and evidence can flow between x and y .
- ▶ Note: information about z means that **z or one of its descendants** is observed (see tutorials).

(A node w is a descendant of z if there is a directed path from z to w .)

Explaining away

Example:

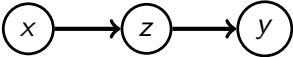
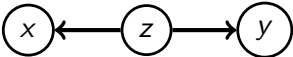



- ▶ One day your computer does not start and you bring it to a repair shop. You think the issue could be the power unit or the cpu.
- ▶ Investigating the power unit shows that it is damaged. Is the cpu fine?
- ▶ Without further information, finding out that the power unit is damaged typically reduces our belief that the cpu is damaged

$$\text{power} \not\perp \text{cpu} \mid \text{pc}$$

- ▶ Finding out about the damage to the power unit *explains away* the observed start-issues of the computer.

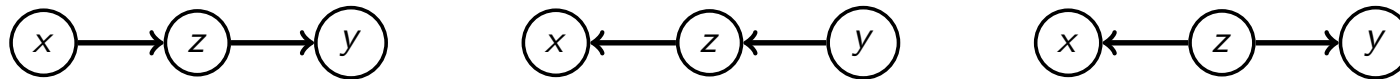
Summary

Connection	z node	$p(x, y)$	$p(x, y z)$
	default: open instantiated: closed	$x \not\perp y$	$x \perp y z$
	default: open instantiated: closed	$x \not\perp y$	$x \perp y z$
	default: closed with evidence: opens	$x \perp y$	$x \not\perp y z$

Think of the z node as a valve or gate through which evidence (probability mass) can flow. Depending on the type of the connection, it's default state is either open or closed. Instantiation/evidence acts as a switch on the valve.

I-equivalence

- ▶ Same independence assertions for



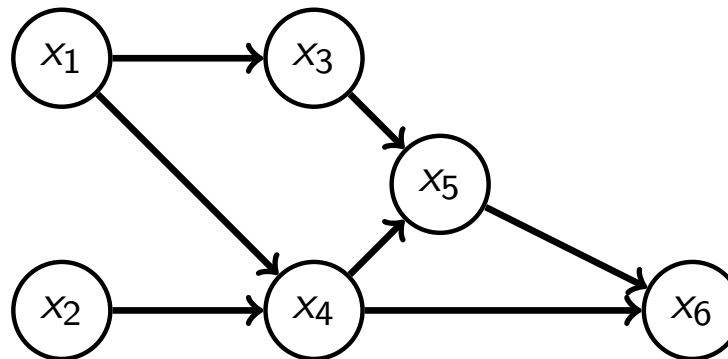
- ▶ The graphs have different causal interpretations
Consider e.g. $x \equiv$ rain; $z \equiv$ street wet; $y \equiv$ car accident
- ▶ This means that based on statistical dependencies (observational data) alone, we cannot select among the graphs and thus determine what causes what.
- ▶ The three directed graphs are said to be independence-equivalent (I-equivalent).

Program

1. Equivalence of factorisation and ordered Markov property
2. Understanding models from their factorisation
3. Definition of directed graphical models
4. **Independencies in directed graphical models**
 - Three canonical connections in a DAG and their properties
 - D-separation and I-map
 - Directed local Markov property
 - Equivalences of the different Markov properties and the factorisation
 - Markov blanket

Further independence relations

- ▶ Given the DAG below, what can we say about the independencies for the set of probability distributions that factorise over the graph?
- ▶ Is $x_1 \perp\!\!\!\perp x_2$? $x_1 \perp\!\!\!\perp x_2 \mid x_6$? $x_2 \perp\!\!\!\perp x_3 \mid \{x_1, x_4\}$?
- ▶ Ordered Markov properties give some independencies.
- ▶ Limitation: only allows us to condition on parent sets.
- ▶ Directed separation (d-separation) gives further independencies.



D-separation

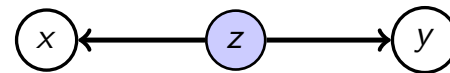
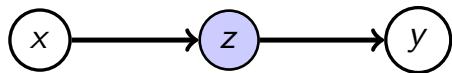
Let $X = \{x_1, \dots, x_n\}$, $Y = \{y_1, \dots, y_m\}$, and $Z = \{z_1, \dots, z_r\}$ be three disjoint sets of nodes in the graph. Assume all z_i are observed (instantiated).

- ▶ Two nodes x_i and y_j are said to be d-separated by Z if all trails between them are blocked by Z .
- ▶ The sets X and Y are said to be d-separated by Z if every trail from any variable in X to any variable in Y is blocked by Z .

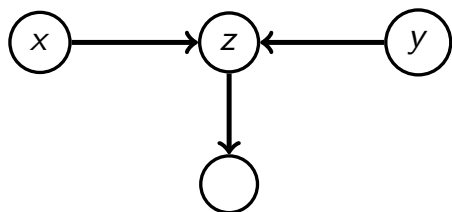
D-separation

A trail is blocked by Z if there is a node on it such that

1. either the node is part of a tail-tail or head-tail connection along the trail and the node is in Z ,



2. or the node is part of a head-head (collider) connection along the trail and neither the node itself nor any of its descendants are in Z .



D-separation and conditional independence

Theorem: If X and Y are d-separated by Z then $X \perp\!\!\!\perp Y \mid Z$ for all probability distributions that factorise over the DAG.

For those interested: A proof can be found in Section 2.8 of *Bayesian Networks – An Introduction* by Koski and Noble (not examinable)

Important because:

1. the theorem allows us to read out (conditional) independencies from the graph
2. no restriction on the sets X, Y, Z
3. the theorem shows that d-separation does not indicate false independence relations. It's independence assertions are sound (“soundness of d-separation”).

D-separation and conditional independence

Theorem: If X and Y are not d-separated by Z then $X \not\perp\!\!\!\perp Y \mid Z$ in **some** probability distributions that factorise over the DAG.

For those interested: A proof sketch can be found in Section 3.3.1 of *Probabilistic Graphical Models* by Koller and Friedman (not examinable).

“not d-separated” is also called “d-connected”

$\not\perp\!\!\!\perp$ means statistically dependent

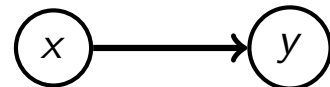
D-separation and conditional independence

- ▶ It can also be that d-connected variables are independent for some distributions.
- ▶ Example (Koller, Example 3.3): $p(x, y)$ with $x, y \in \{0, 1\}$ and

$$p(y = 0|x = 0) = a \quad p(y = 0|x = 1) = a$$

for $a > 0$ and some non-zero $p(x = 0)$.

- ▶ Graph has arrow from x to y . Variables are not d-separated.



- ▶ $p(y = 0) = ap(x = 0) + ap(x = 1) = a$,
which is $p(y = 0|x)$ for all x .
- ▶ $p(y = 1) = (1 - a)p(x = 0) + (1 - a)p(x = 1) = 1 - a$,
which is $p(y = 1|x)$ for all x .
- ▶ Hence: $p(y|x) = p(y)$ so that $x \perp\!\!\!\perp y$.

D-separation and conditional independence

- ▶ This means that d-separation does generally not reveal all independencies in all probability distributions that factor over the graph.
- ▶ In other words, individual probability distributions that factor over the graph may have further independencies not included in the set obtained by d-separation.
- ▶ We say that d-separation is not “complete”.

I-map

- ▶ A graph is said to be an independency map (I-map) for a set of independencies \mathcal{I} if the independencies asserted by the graph are part of \mathcal{I} .
- ▶ For a directed graph G , let $\mathcal{I}(G)$ be all the independencies that we can derive via d-separation.
- ▶ Denote the independencies that a distribution p satisfies by $\mathcal{I}(p)$.
- ▶ The previous results on d-separation can thus be written as

$$\mathcal{I}(G) \subseteq \mathcal{I}(p) \quad \text{for all } p \text{ that factorise over } G$$

- ▶ As we have seen, we generally do not have $\mathcal{I}(G) = \mathcal{I}(p)$. If we have equality, the graph is said to be a perfect map (P-map) for $\mathcal{I}(p)$.

Recipe to determine whether two nodes are d-separated

1. Determine all trails between x and y (note: direction of the arrows does here not matter).
2. For each trail:
 - i Determine the default state of all nodes on the trail.
 - ▶ open if part of a tail-head or a tail-tail connection
 - ▶ closed if part of a head-head connection
 - ii Check whether the set of observed nodes Z switches the state of the nodes on the trail.
 - iii The trail is blocked if it contains a closed node.
3. The nodes x and y are d-separated if all trails between them are closed.

Example: Are x_1 and x_2 d-separated?

Follows from ordered Markov property, but let us answer it with d-separation.

1. Determine all trails between x_1 and x_2

2. For trail x_1, x_4, x_2

i default state

ii conditioning set is empty

iii \Rightarrow Trail is blocked

For trail x_1, x_3, x_5, x_4, x_2

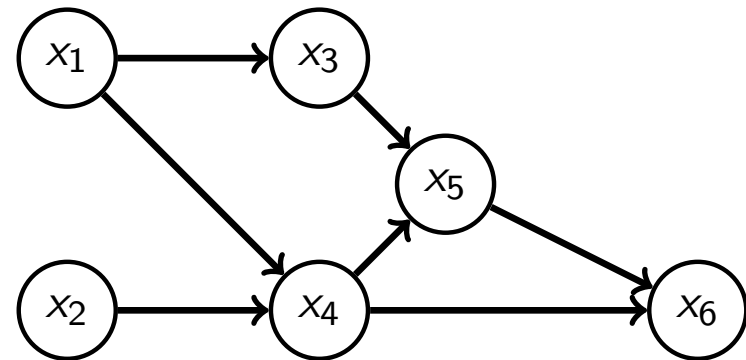
i default state

ii conditioning set is empty

iii \Rightarrow Trail is blocked

Trail $x_1, x_3, x_5, x_6, x_4, x_2$ is blocked too (same arguments).

3. $\Rightarrow x_1$ and x_2 are d-separated.

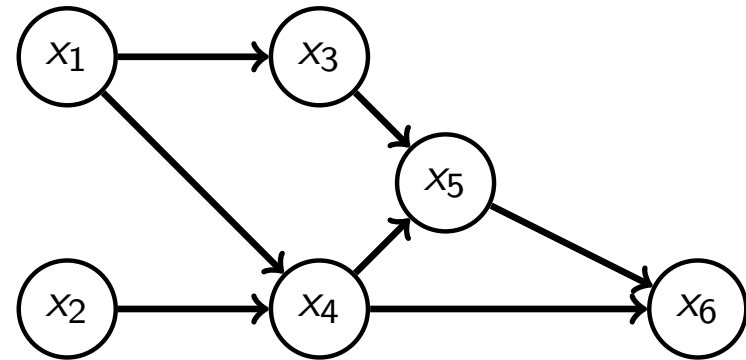


$x_1 \perp\!\!\!\perp x_2$ for all probability distributions that factorise over the graph.

Example: Are x_1 and x_2 d-separated by x_6 ?

1. Determine all trails between x_1 and x_2
2. For trail x_1, x_4, x_2
 - i default state
 - ii influence of x_6
 - iii \Rightarrow Trail not blocked

No need to check the other trails: x_1 and x_2 are not d-separated by x_6



$x_1 \perp\!\!\!\perp x_2 \mid x_6$ does generally not hold for probability distributions that factorise over the graph.

Example: Are x_2 and x_3 d-separated by x_1 and x_4 ?

1. Determine all trails between x_2 and x_3

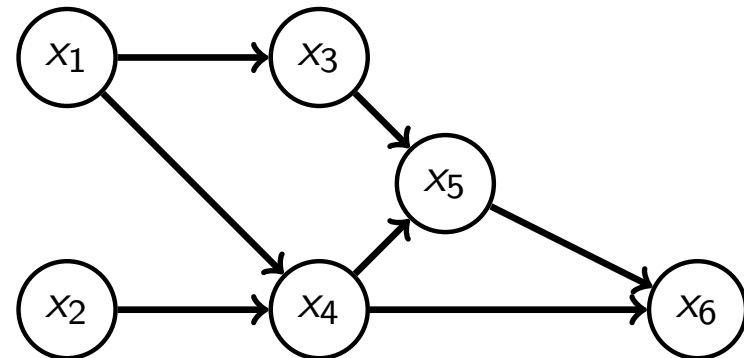
2. For trail x_3, x_1, x_4, x_2
- i default state
 - ii influence of $\{x_1, x_4\}$
 - iii \Rightarrow Trail blocked

For trail x_3, x_5, x_4, x_2

- i default state
- ii influence of $\{x_1, x_4\}$
- iii \Rightarrow Trail blocked

Trail x_3, x_5, x_6, x_4, x_2 is blocked too (same arguments).

3. $\Rightarrow x_2$ and x_3 are d-separated by x_1 and x_4 .



$x_2 \perp\!\!\!\perp x_3 \mid \{x_1, x_4\}$ for all probability distributions that factorise over the graph.

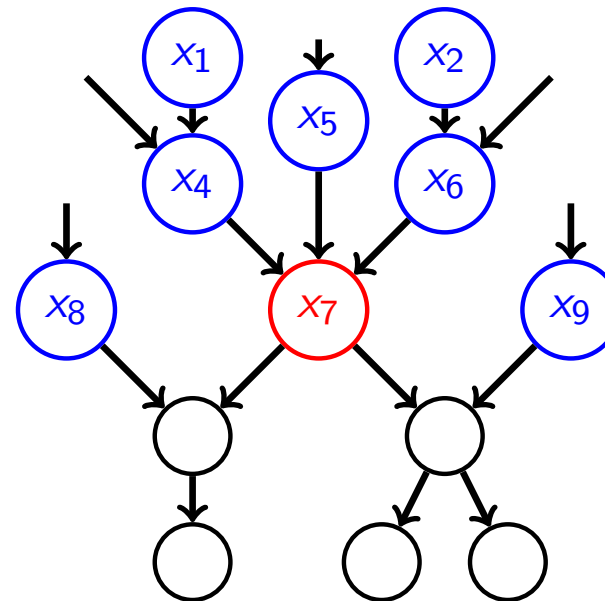
Directed local Markov property

- ▶ The independencies from the ordered Markov property depend on the topological ordering chosen.
- ▶ We now use d-separation to derive a similarly local Markov property that does not depend on the ordering, and show the equivalence for any topological ordering:

$$x_i \perp\!\!\!\perp \text{pre}_i \setminus \text{pa}_i \mid \text{pa}_i \iff x_i \perp\!\!\!\perp \text{nondesc}(x_i) \setminus \text{pa}_i \mid \text{pa}_i$$

where $\text{nondesc}(x_i)$ denotes the non-descendants of x_i .

$x_i \equiv x_7$
 $\text{pa}_7 = \{x_4, x_5, x_6\}$
 $\text{pre}_7 = \{x_1, x_2, \dots, x_6\}$
 $\text{nondesc}(x_7)$ in blue



Directed local Markov property

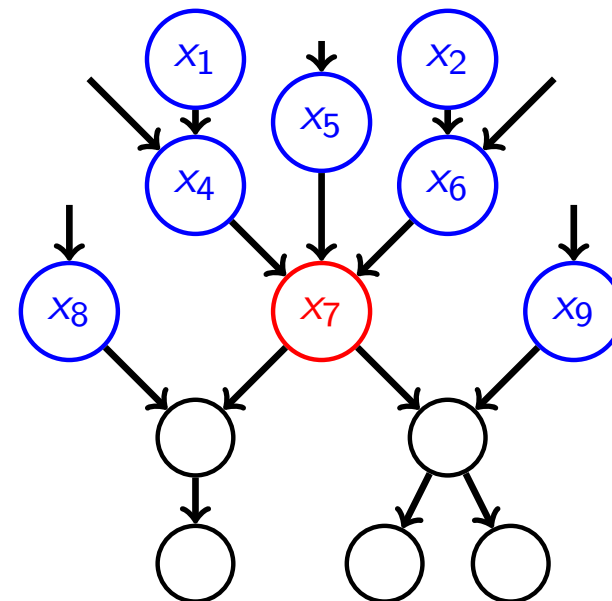
$x_i \perp\!\!\!\perp \text{pre}_i \setminus \text{pa}_i \mid \text{pa}_i \iff x_i \perp\!\!\!\perp \text{nondesc}(x_i) \setminus \text{pa}_i \mid \text{pa}_i$ follows because $\{x_1, \dots, x_{i-1}\} \subseteq \text{nondesc}(x_i)$ for all topological orderings

For \Rightarrow consider all trails from x_i to $\{\text{nondesc}(x_i) \setminus \text{pa}_i\}$.

Two cases: move against or with the arrows:

- (1) upward trails are blocked by the parents
- (2) downward trails must contain a head-head (collider) connection because x_i is a non-descendant. These paths are blocked because the collider node or its descendants are never part of pa_i .

The result now follows because all paths from x_i to all elements in $\{\text{nondesc}(x_i) \setminus \text{pa}_i\}$ are blocked.



Remarks

- ▶ The local Markov independencies do not depend on a topological ordering. They can be directly read from the graph.
- ▶ The direction “local Markov property \Rightarrow ordered Markov property” explains why models that satisfy one ordered Markov property also have to satisfy all other ordered Markov properties obtained with different topological orderings.
- ▶ The union of all ordered Markov independencies is generally not equal to the set of directed Markov independencies.

Summary of the equivalences

Factorisation		$p(\mathbf{x}) = \prod_{i=1}^d p(x_i \text{pa}_i)$
	\Leftrightarrow	
ordered Markov property		$x_i \perp\!\!\!\perp \text{pre}_i \setminus \text{pa}_i \mid \text{pa}_i$
	\Leftrightarrow	
local directed Markov property		$x_i \perp\!\!\!\perp \text{nondesc}(x_i) \setminus \text{pa}_i \mid \text{pa}_i$
	\Leftrightarrow	
global directed Markov property		all independencies by d-separation

Broadly speaking, the graph serves two related purposes:

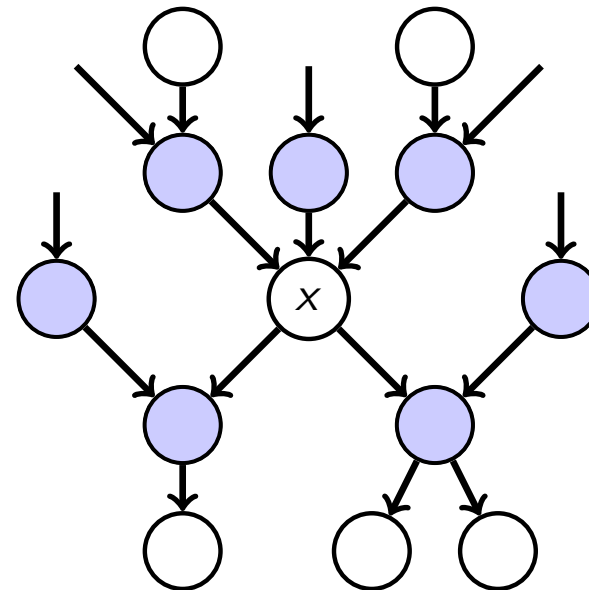
1. it tells us how distributions factorise
2. it represents the independence assumptions made

Markov blanket

What is the minimal set of variables such that knowing their values makes x independent from the rest?

From d-separation:

- ▶ Isolate x from its ancestors
⇒ condition on parents
- ▶ Isolate x from its descendants
⇒ condition on children
- ▶ Deal with collider connection
⇒ condition on co-parents
(other parents of the children of x)



In a directed graphical model, the parents, children, and co-parents of x are called its Markov blanket, denoted by $MB(x)$. We have

$$x \perp\!\!\!\perp \{ \text{all variables} \setminus x \setminus MB(x) \} \mid MB(x).$$

Program recap

1. Equivalence of factorisation and ordered Markov property
 - Chain rule
 - Ordered Markov property implies factorisation
 - Factorisation implies ordered Markov property
2. Understanding models from their factorisation
 - Ancestral sampling
 - Visualisation as a directed graph
 - Description of directed graphs and topological orderings
3. Definition of directed graphical models
 - Via factorisation according to the graph
 - Via ordered Markov property
 - Derive independencies from the ordered Markov property with different topological orderings
4. Independencies in directed graphical models
 - Three canonical connections in a DAG and their properties
 - D-separation and I-map
 - Directed local Markov property
 - Equivalences of the different Markov properties and the factorisation
 - Markov blanket