

Directed Graphical Models I

Definition and Basic Properties

Michael U. Gutmann

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

Spring Semester 2022

Recap

- ▶ We talked about reasonably weak assumption to facilitate the efficient representation of a probabilistic model
- ▶ Independence assumptions reduce the number of interacting variables, e.g.
 - ▶ $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x})p(\mathbf{y})p(\mathbf{z})$
 - ▶ $p(x_1, \dots, x_d) = p(x_d | x_{d-3}, x_{d-2}, x_{d-1})p(x_1, \dots, x_{d-1})$
- ▶ Parametric assumptions restrict the way the variables may interact.

Program

1. Visualising factorisations with directed acyclic graphs
2. Directed graphical models

Program

1. Visualising factorisations with directed acyclic graphs
 - Conditional independencies simplify factors in the chain rule
 - Visualisation as a directed acyclic graph
 - Graph concepts
2. 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.

Conditional independencies simplify the factors

- ▶ Given: a pdf/pmf that factorises as $p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \text{pre}_i)$ 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 .

- ▶ 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

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

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 $\pi_i \subseteq \text{pre}_i$.

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

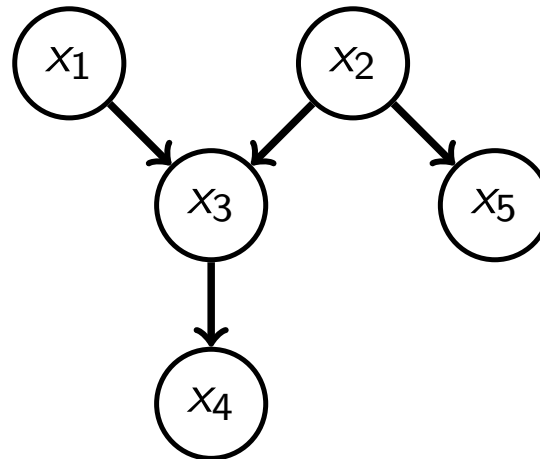
- ▶ What's the point?
 1. $p(x_i \mid \pi_i)$ involve fewer interacting variables than $p(x_i \mid \text{pre}_i)$.
 - ▶ Makes them easier to model.
 - ▶ If specified as a table, fewer numbers are needed for their representation (computational advantage).
 2. We can visualise the interactions between the variables with a graph.

Visualisation as a directed graph

Assume $p(\mathbf{x}) = \prod_{i=1}^d p(x_i | \pi_i)$ with $\pi_i \subseteq \text{pre}_i$. We 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:

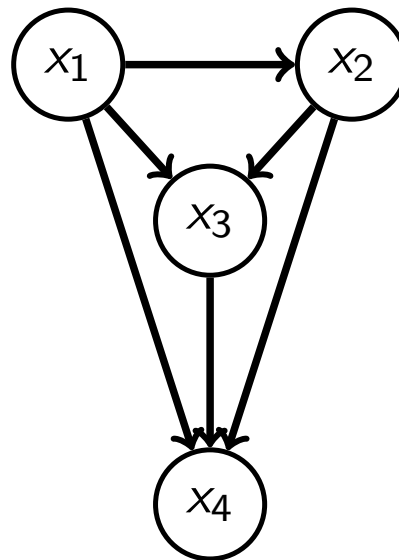
$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_2)$$



Visualisation as a directed graph

Example:

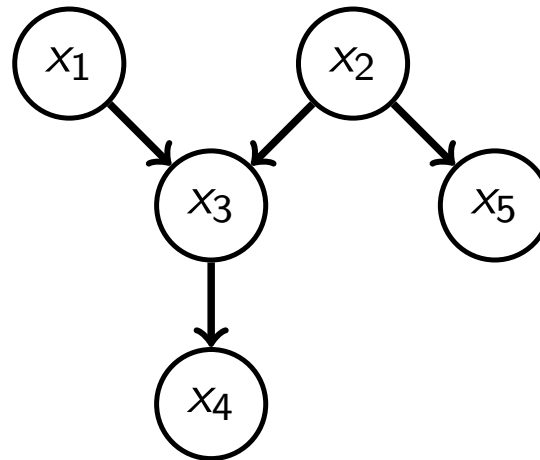
$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)$$



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

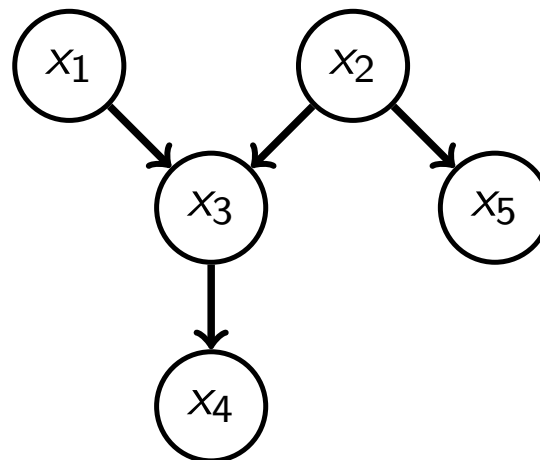
Graph concepts

- ▶ **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 $x_i \in \text{pa}(x_j)$, e.g. x_3 and x_5 are children of x_2 .



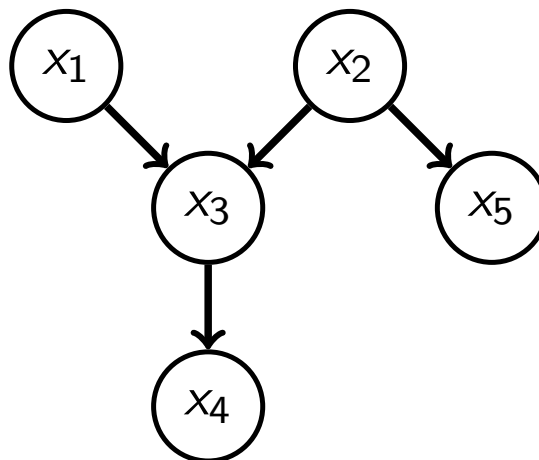
Graph concepts

- ▶ 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, $\text{nondesc}(x_3) = \{x_1, x_2, x_5\}$

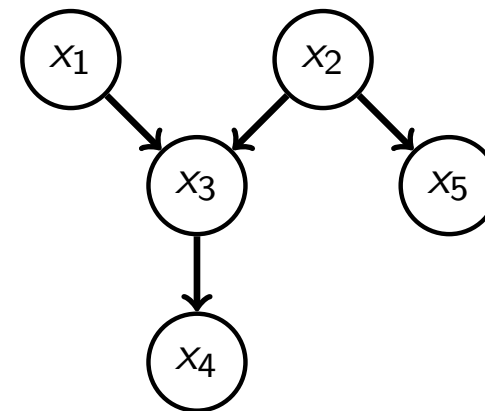


Graph concepts

- ▶ **Topological ordering:** an ordering (x_1, \dots, x_d) of some variables x_i is topological relative to a graph if parents come before their children in the ordering. (whenever there is a directed edge from x_i to x_j , x_i occurs prior to x_j in the ordering.)

- ▶ Examples for the graph on the right:

- ▶ x_1, x_2, x_3, x_4, x_5
- ▶ x_2, x_5, x_1, x_3, x_4
- ▶ x_2, x_1, x_3, x_5, x_4



- ▶ There is always at least one ordering that is topological to a DAG.
- ▶ The π_i in the factorisation are equal to the parents pa_i in the graph. We will call both sets the “parents” of x_i .

Program

1. Visualising factorisations with directed acyclic graphs
 - Conditional independencies simplify factors in the chain rule
 - Visualisation as a directed acyclic graph
 - Graph concepts
2. Directed graphical models

Program

1. Visualising factorisations with directed acyclic graphs

2. Directed graphical models

- Definition
- Conditionals, marginals, and ancestral sampling
- Examples

Directed graphical model (DGM)

- ▶ We started with a factorised pdf/pmf and associated a DAG with it.
- ▶ We can also go the other way around and start with a DAG.
- ▶ *Definition* A directed graphical model based on a DAG G with d nodes and associated random variables x_i is the set of pdfs/pmfs that factorise as

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

where the $k(x_i | \text{pa}_i)$ are conditional pdfs/pmfs. (they are sometimes called kernels or factors)

- ▶ Remark: a pdf/pmf $p(x_1, \dots, x_d)$ that can be written as above is said to “factorise over the graph G ”. We also say that it has property $F(G)$ (“F” for factorisation).

Why set of pdfs/pmfs?

- ▶ The directed graphical model corresponds to a **set of probability distributions** .
- ▶ This is because we did not specify any numerical values for the $k(x_i|pa_i)$. We only specified which variables the conditionals take as input (namely x_i and pa_i).
- ▶ The set includes all those distributions that you get by looping, for all variables x_i , over all possible $k(x_i|pa_i)$. (e.g. tables or parameter values in parametrised models)
- ▶ While a probability distribution corresponds to a probabilistic model, a set of probability distributions (probabilistic models) is often called a statistical model.
- ▶ Individual pdfs/pmf in the set are typically also called a directed graphical model.
- ▶ Other names for directed graphical models: belief network, Bayesian network, Bayes network.

The factors $k(x_i|\text{pa}_i)$ equal the conditionals $p(x_i|\text{pa}_i)$

- ▶ When we decomposed $p(\mathbf{x})$ with the chain rule and inserted conditional independencies, we obtained

$$p(\mathbf{x}) = \prod_i p(x_i|\pi_i)$$

where the $p(x_i|\pi_i)$ where the conditionals of x_i given π_i .

- ▶ We now show that the $k(x_i|\text{pa}_i)$ in the definition of the DGM are equal to the $p(x_i|\text{pa}_i)$.
- ▶ Assume $p(\mathbf{x})$ factorises over a DAG G and hence that $p(\mathbf{x}) = \prod_{i=1}^d k(x_i|\text{pa}_i)$. First step is to label the variables such that the ordering x_1, \dots, x_d is topologically to G .
- ▶ In a topological ordering, the parents come before the children. Hence $\text{pa}_i \subseteq \text{pre}_i = (x_1, \dots, x_{i-1})$

The factors $k(x_i|pa_i)$ equal the conditionals $p(x_i|pa_i)$

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

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

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

The factors $k(x_i|pa_i)$ equal the conditionals $p(x_i|pa_i)$

Hence:

$$p(x_d|x_1, \dots, x_{d-1}) = \frac{p(x_1, \dots, x_d)}{p(x_1, \dots, x_{d-1})} = \frac{\prod_{i=1}^d k(x_i|pa_i)}{\prod_{i=1}^{d-1} k(x_i|pa_i)} \quad (1)$$

$$= k(x_d|pa_d) \quad (2)$$

Split $(x_1, \dots, x_{d-1}) = \text{pre}_d$ into non-overlapping sets pa_d and $\tilde{\mathbf{x}} = \text{pre}_d \setminus pa_d$ so that $p(x_d|x_1, \dots, x_{d-1}) = p(x_d|\tilde{\mathbf{x}}, pa_d)$.

By the product rule, we have

$$p(x_d, \tilde{\mathbf{x}}|pa_d) = p(x_d|\tilde{\mathbf{x}}, pa_d)p(\tilde{\mathbf{x}}|pa_d) \quad (3)$$

$$= k(x_d|pa_d)p(\tilde{\mathbf{x}}|pa_d) \quad (4)$$

Next sum out $\tilde{\mathbf{x}}$ to obtain

$$p(x_d|pa_d) = \int p(x_d, \tilde{\mathbf{x}}|pa_d)d\tilde{\mathbf{x}} = k(x_d|pa_d) \int p(\tilde{\mathbf{x}}|pa_d)d\tilde{\mathbf{x}} \quad (5)$$

$$= k(x_d|pa_d) \quad (6)$$

where we have used that x_d and pa_d are not part of $\tilde{\mathbf{x}}$.

The factors $k(x_i|pa_i)$ equal the conditionals $p(x_i|pa_i)$

Hence:

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

Next, note that $p(x_1, \dots, x_{d-1})$ 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 for $p(\mathbf{x}) = \prod_{i=1}^d k(x_i|pa_i)$:

(1) $k(x_i|pa_i) = p(x_i|pa_i)$ for $i = 1, \dots, d$

(As desired!)

(2) $p(x_i|pre_i) = p(x_i|pa_i)$ for $i = 1, \dots, d$

(This means that the factorisation of the DGM implies independencies, see later)

(3) $p(x_1, \dots, x_k) = \prod_{i=1}^k k(x_i|pa_i)$ for $k = 1, \dots, d$

(The distr of the first k variables is given by the first k terms in the factorisation)

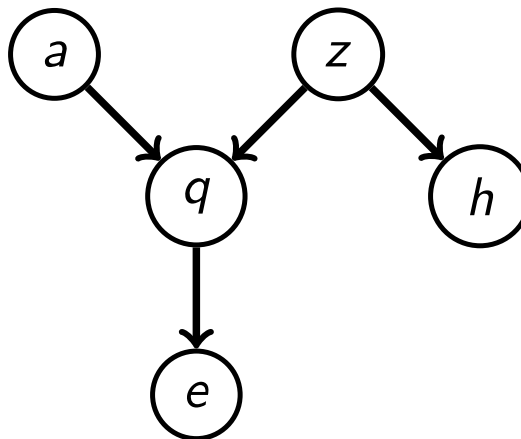
Note that (2) and (3) depend on the particular topological ordering chosen, e.g. it is “first k variables” in the chosen topological ordering.

Ancestral sampling

- ▶ This means that the DAG not only specifies the joint distribution $p(\mathbf{x}) = \prod_{i=1}^d k(x_i|\text{pa}_i)$ but also a sampling/data generating process.
- ▶ To generate data from $p(\mathbf{x})$:
 1. Pick an ordering x_1, \dots, x_d of the random variables that is topological to G .
 2. x_1 does not have any parents, i.e. $\text{pa}_1 = \emptyset$.
 3. Following the topological ordering, sample from $k(x_i|\text{pa}_i)$, $i = 1, \dots, d$.
- ▶ Moreover, from the results above:
 - ▶ $x_i|\text{pa}_i \sim p(x_i|\text{pa}_i)$
(The notation means that x_i follows or is sampled from $p(x_i|\text{pa}_i)$)
 - ▶ $(x_1, \dots, x_k) \sim p(x_1, \dots, x_k)$ for all k
(To e.g. sample from (x_1, x_2) , you can stop the sampling after $i = 2$.)
- ▶ It's called ancestral sampling because we sample the parents before the children, following the arrows in the DAG.

Example

DAG:



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

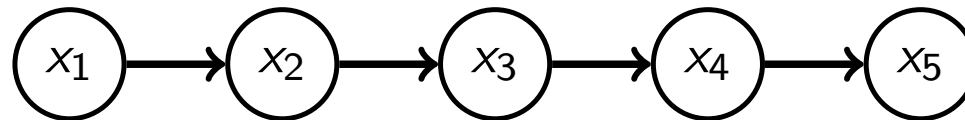
Parent sets: $pa_a = pa_z = \emptyset, pa_q = \{a, z\}, pa_e = \{q\}, pa_h = \{z\}$.

Directed graphical model: set of pdfs/pmfs $p(a, z, q, e, h)$ that factorise as:

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

Example: Markov chain

DAG:



Random variables: x_1, x_2, x_3, x_4, x_5

Parent sets:

$$pa_1 = \emptyset, pa_2 = \{x_1\}, pa_3 = \{x_2\}, pa_4 = \{x_3\}, pa_5 = \{x_4\}.$$

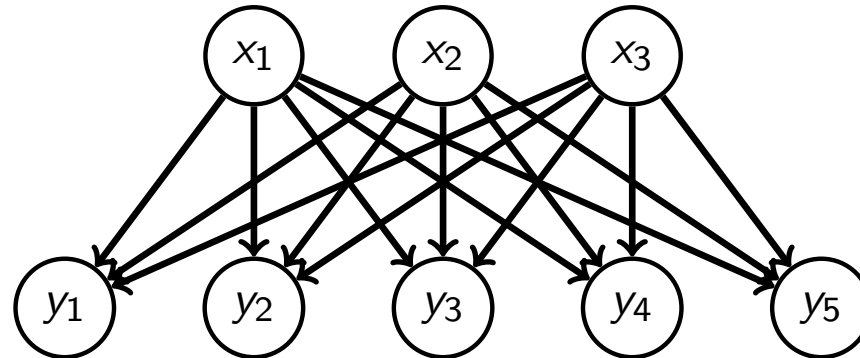
Directed graphical model: set of pdfs/pmfs $p(x_1, \dots, x_5)$ that factorise as:

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

Example: Probabilistic PCA, factor analysis, ICA

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

DAG:



Random variables: $x_1, x_2, x_3, y_1, \dots, y_5$

Parent sets: $\text{pa}(x_i) = \emptyset, \text{pa}(y_i) = \{x_1, x_2, x_3\}$ for all i .

Directed graphical model: set of pdfs/pmfs $p(x_1, x_2, x_3, y_1, \dots, y_5)$ that factorise as:

$$p(x_1, x_2, x_3, y_1, \dots, y_5) = p(x_1)p(x_2)p(x_3)p(y_1|x_1, x_2, x_3) \\ p(y_2|x_1, x_2, x_3) \dots p(y_5|x_1, x_2, x_3)$$

Program recap

1. Visualising factorisations with directed acyclic graphs
 - Conditional independencies simplify factors in the chain rule
 - Visualisation as a directed acyclic graph
 - Graph concepts

2. Directed graphical models
 - Definition
 - Conditionals, marginals, and ancestral sampling
 - Examples