Directed Graphical Models I

Definition and Basic Properties

Chris Williams
(based on slides by Michael U. Gutmann)

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

Spring Semester 2023

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.
  - $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x})p(\mathbf{y})p(\mathbf{z})$
  - $p(x_1, \ldots, x_d) = p(x_d | x_{d-3}, x_{d-2}, x_{d-1})p(x_1, \ldots, x_{d-1})$
- 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) $p(\mathbf{x}) = p(x_1, x_2, \ldots, x_d)$ into product of conditional pdfs.

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

with $\mathrm{pre}_i = \mathrm{pre}(x_i) = \{x_1, \ldots, x_{i-1}\}$, $\mathrm{pre}_1 = \varnothing$ and $p(x_1|\varnothing) = p(x_1)$

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

Conditional independencies simplify the factors

- Given: a pdf/pmf that factorises as $p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i|\mathrm{pre}_i)$ for the ordering $x_1, \ldots, x_d$.
- 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 $\pi_i \subseteq \mathrm{pre}_i$ such that $p(\mathbf{x})$ satisfies

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

  for all i.
- By definition of conditional independence: $p(x_i|x_1, \ldots, x_{i-1}) = p(x_i|\mathrm{pre}_i) = p(x_i|\pi_i)$
- With the convention $\pi_1 = \varnothing$, we obtain the factorisation

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

Why does it matter?

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

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

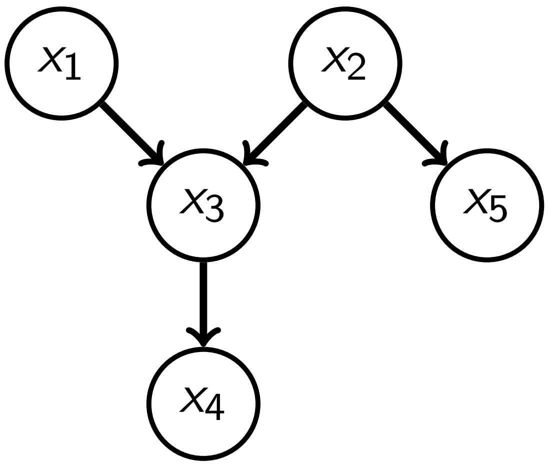
- What's the point?
  1. $p(x_i|\pi_i)$ involve fewer interacting variables than $p(x_i|\mathrm{pre}_i)$.
     - Makes them easier to model.
     - If specified as a table, fewer numbers are needed for their representation (computational advantage).
  2. We can visualise the interactions between the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

Visualisation as a directed graph

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

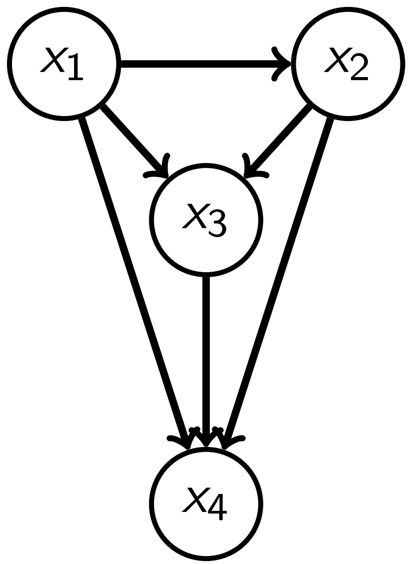
$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_2)$$



Visualisation as a directed graph

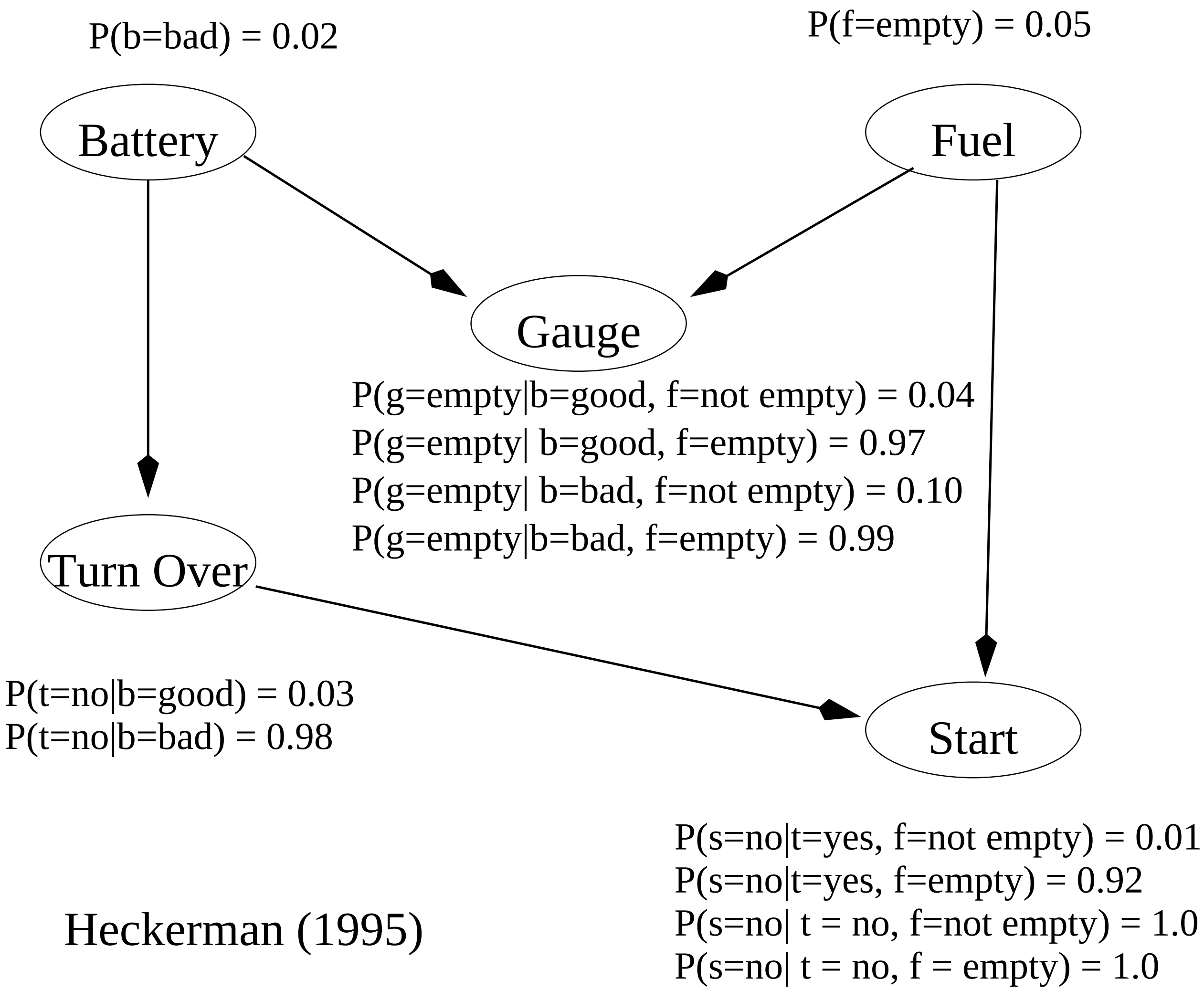
Example:

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



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

Example: Car start belief network



Heckerman (1995)

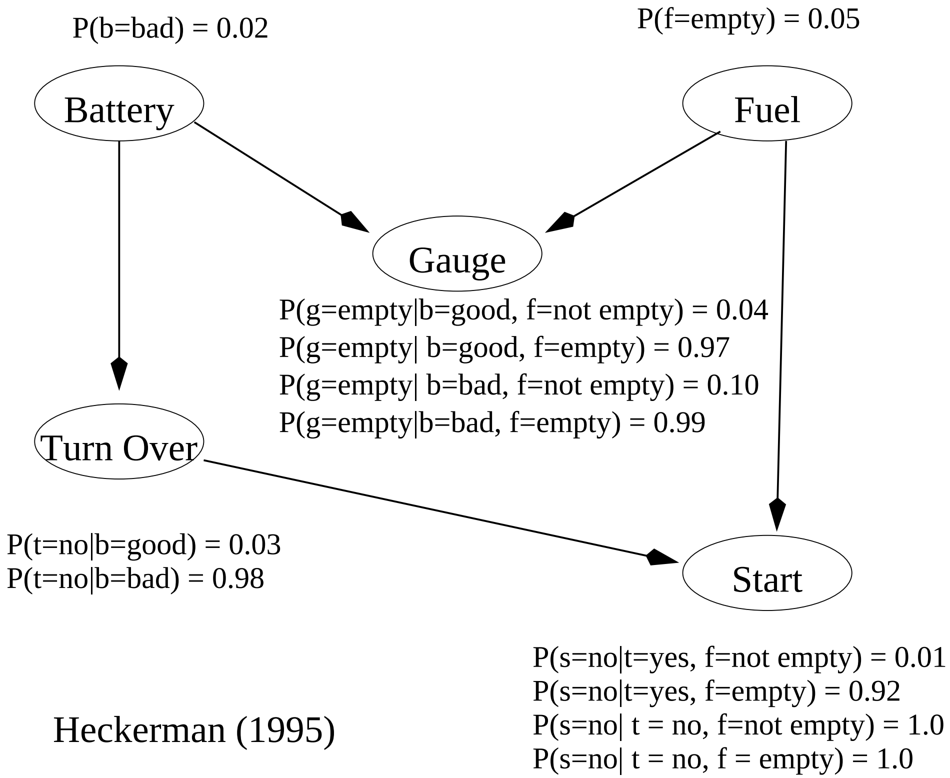
- Unstructured joint distribution requires $2^5 - 1 = 31$ numbers to specify it. Here can use 12 numbers
- Take the ordering b, f, g, t, s. Joint can be expressed as

$$p(b, f, g, t, s) = p(b)p(f|b)p(g|b, f)p(t|b, f, g)p(s|b, f, g, t)$$

- Conditional independences (missing links) give

$$p(b, f, g, t, s) = p(b)p(f)p(g|b, f)p(t|b)p(s|t, f)$$

Example: Car start belief network



What is probability of

$p(b = good, t = no, g = empty, f = not\ empty, s = no)$?

Example: Linear-Gaussian networks

- Let the x's be real-valued

$$p(x_i|\pi_i) = N(x_i|\mathbf{w}_i^T \mathbf{x}_{\pi_i} + b_i, \sigma_i^2)$$

- $p(\mathbf{x})$ is jointly Gaussian
- Exact inference can be carried out
  - (i) by first constructing the joint and conditioning, or
  - (ii) by exploiting the graphical structure
- Example: factor analysis (see later)

Constructing belief networks

1. Choose a relevant set of variables $\{x_i\}$ that describe the domain
2. Choose an ordering for the variables
3. While there are variables left
    (a) Pick a variable x_i and add it to the network
    (b) Set $\mathit{Parents}(x_i)$ to some minimal set of nodes already in the net
    (c) Define the conditional probability table for x_i

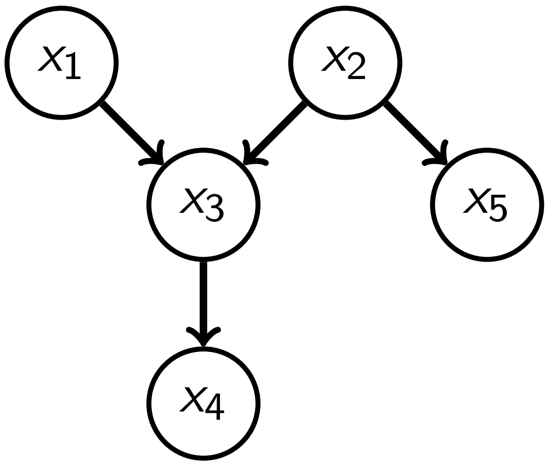
- This procedure is guaranteed to produce a DAG
- To ensure maximum sparsity, add “root causes” first, then the variables they influence and so on, until leaves are reached. Leaves have no direct causal influence over other variables
- **Example**: Construct DAG for the car example using the ordering s, t, g, f, b
- “Wrong” ordering will give same joint distribution, but will require the specification of more numbers than otherwise necessary

Specifying conditional probability distributions

- CPDs: conditional probability distributions
- CPTs: conditional probability tables for discrete variables
- Where do the numbers come from? Can be elicited from experts, or learned (see later)
- CPTs can still be very large (and difficult to specify) if there are many parents for a node. Can use combination rules such as the logistic regression form

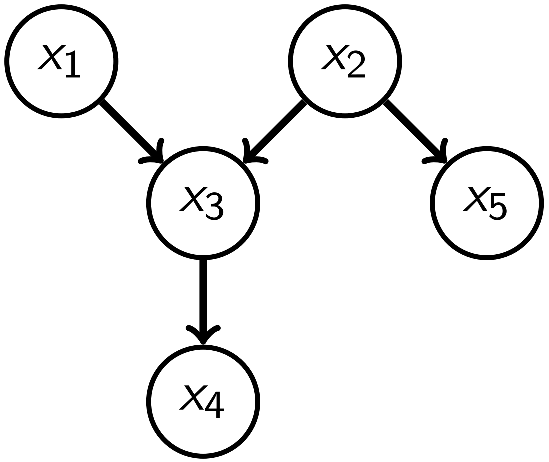
Graph concepts

- 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 $\mathrm{pa}(x_i) = \mathrm{pa}_i$, e.g. $\mathrm{pa}(x_3) = \mathrm{pa}_3 = \{x_1, x_2\}$.
- x_j is a child of x_i if $x_i \in \mathrm{pa}(x_j)$, e.g. x_3 and x_5 are children of x_2.



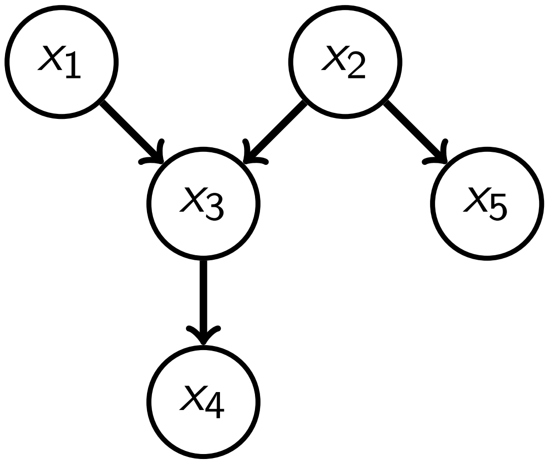
Graph concepts

- 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

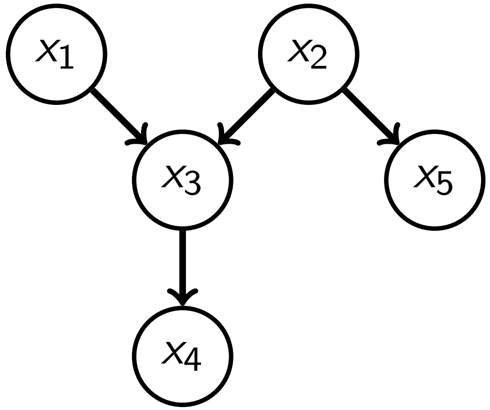
- The ancestors $\mathrm{anc}(x_i)$ of x_i are all the nodes where a directed path leads to x_i. For example, $\mathrm{anc}(x_4) = \{x_1, x_3, x_2\}$.
- The descendants $\mathrm{desc}(x_i)$ of x_i are all the nodes that can be reached on a directed path from x_i. For example, $\mathrm{desc}(x_1) = \{x_3, x_4\}$.
  (Note: sometimes, x_i is included in the set of ancestors and descendants)
- The non-descendents of x_i are all the nodes in a graph except x_i and the descendants of x_i. For example, $\mathrm{nondesc}(x_3) = \{x_1, x_2, x_5\}$



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 relative 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

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

  where the $k(x_i|\mathrm{pa}_i)$ are some conditional pdfs/pmfs. (they are sometimes called kernels or factors)
- Remark: a pdf/pmf $p(x_1, \ldots, x_d)$ 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 $k(x_i|\mathrm{pa}_i)$. 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 $k(x_i|\mathrm{pa}_i)$.
  (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 $k(x_i|\mathrm{pa}_i)$ equal the conditionals $p(x_i|\mathrm{pa}_i)$

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

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

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

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

$$\begin{aligned}
p(x_1, \ldots, x_{d-1}) &= \int p(x_1, \ldots, x_d)\mathrm{d}x_d \\
&= \int \prod_{i=1}^{d} k(x_i|\mathrm{pa}_i)\mathrm{d}x_d \\
&= \int \prod_{i=1}^{d-1} k(x_i|\mathrm{pa}_i)k(x_d|\mathrm{pa}_d)\mathrm{d}x_d \quad (x_d \notin \mathrm{pa}_i, i < d) \\
&= \prod_{i=1}^{d-1} k(x_i|\mathrm{pa}_i) \int k(x_d|\mathrm{pa}_d)\mathrm{d}x_d \\
&= \prod_{i=1}^{d-1} k(x_i|\mathrm{pa}_i)
\end{aligned}$$

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

Hence:

$$\begin{aligned} p(x_d|x_1,\ldots,x_{d-1}) &= \frac{p(x_1,\ldots,x_d)}{p(x_1,\ldots,x_{d-1})} = \frac{\prod_{i=1}^{d} k(x_i|\mathrm{pa}_i)}{\prod_{i=1}^{d-1} k(x_i|\mathrm{pa}_i)} \\ &= k(x_d|\mathrm{pa}_d) \end{aligned}$$

Split $(x_1,\ldots,x_{d-1}) = \mathrm{pre}_d$ into non-overlapping sets pa_d and $\tilde{\mathbf{x}}_d = \mathrm{pre}_d \setminus \mathrm{pa}_d$ so that $p(x_d|x_1,\ldots,x_{d-1}) = p(x_d|\tilde{\mathbf{x}}_d, \mathrm{pa}_d)$.
By the product rule, we have

$$\begin{aligned} p(x_d, \tilde{\mathbf{x}}_d|\mathrm{pa}_d) &= p(x_d|\tilde{\mathbf{x}}_d, \mathrm{pa}_d)p(\tilde{\mathbf{x}}_d|\mathrm{pa}_d) \\ &= k(x_d|\mathrm{pa}_d)p(\tilde{\mathbf{x}}_d|\mathrm{pa}_d) \end{aligned}$$

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

$$\begin{aligned} p(x_d|\mathrm{pa}_d) &= \int p(x_d, \tilde{\mathbf{x}}_d|\mathrm{pa}_d)\mathrm{d}\tilde{\mathbf{x}}_d = k(x_d|\mathrm{pa}_d)\int p(\tilde{\mathbf{x}}_d|\mathrm{pa}_d)\mathrm{d}\tilde{\mathbf{x}}_d \\ &= k(x_d|\mathrm{pa}_d) \end{aligned}$$

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

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

Hence:

$$p(x_d|x_1, \ldots, x_{d-1}) = p(x_d|\mathrm{pa}_d) = k(x_d|\mathrm{pa}_d)$$

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

Proves that for $p(\mathbf{x}) = \prod_{i=1}^{d} k(x_i|\mathrm{pa}_i)$:

(1) $k(x_i|\mathrm{pa}_i) = p(x_i|\mathrm{pa}_i)$ for $i = 1, \ldots, d$
(As desired!)

(2) $p(x_i|\mathrm{pre}_i) = p(x_i|\mathrm{pa}_i)$ for $i = 1, \ldots, d$
(This means that the factorisation of the DGM implies independencies, see later)

(3) $p(x_1, \ldots, x_k) = \prod_{i=1}^{k} k(x_i|\mathrm{pa}_i)$ fo $k = 1, \ldots, d$
(The distr of the first k variables is given by the first k terms in the factorisation)

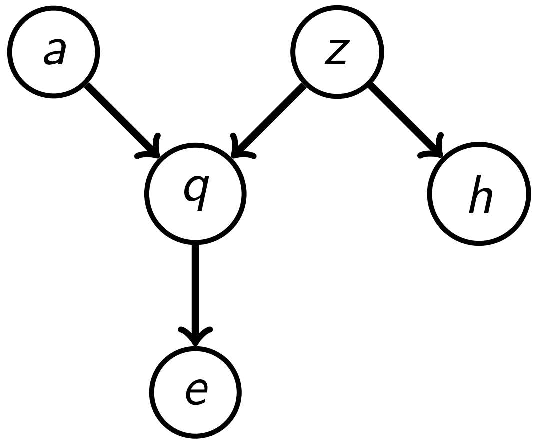
Note that (2) and (3) depend on the particular topological ordering chosen, e.g. it is "first k variables" in the chosen topological ordering.

Ancestral sampling

- This means that the DAG not only specifies the joint distribution $p(\mathbf{x}) = \prod_{i=1}^{d} k(x_i|\mathrm{pa}_i)$ but also a sampling/data generating process.
- To generate data from $p(\mathbf{x})$:
  1. Pick an ordering $x_1, \ldots, x_d$ of the random variables that is topological to G.
  2. x_1 does not have any parents, i.e. $\mathrm{pa}_1 = \varnothing$.
  3. Following the topological ordering, sample from $k(x_i|\mathrm{pa}_i)$, $i = 1, \ldots, d$.
- Moreover, from the results above:
  - $x_i|\mathrm{pa}_i \sim p(x_i|\mathrm{pa}_i)$
    (The notation means that x_i follows or is sampled from $p(x_i|\mathrm{pa}_i)$)
  - $(x_1, \ldots, x_k) \sim p(x_1, \ldots, x_k)$ 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: $\mathrm{pa}_a = \mathrm{pa}_z = \varnothing, \mathrm{pa}_q = \{a, z\}, \mathrm{pa}_e = \{q\}, \mathrm{pa}_h = \{z\}$.

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:

$$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5$$

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

Parent sets:
$\mathrm{pa}_1 = \varnothing, \mathrm{pa}_2 = \{x_1\}, \mathrm{pa}_3 = \{x_2\}, \mathrm{pa}_4 = \{x_3\}, \mathrm{pa}_5 = \{x_4\}$.

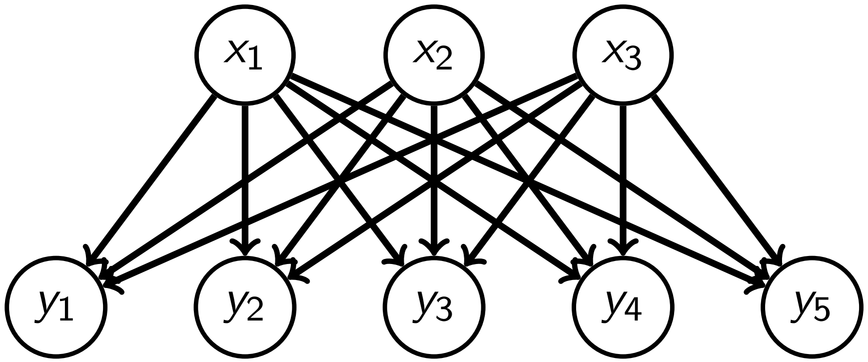
Directed graphical model: set of pdfs/pmfs $p(x_1, \ldots, x_5)$ that factorise as:

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

Example: Probabilistic PCA, factor analysis, ICA

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

DAG:



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

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

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

$$
\begin{aligned}
p(x_1, x_2, x_3, y_1, \ldots, y_5) = & p(x_1)p(x_2)p(x_3)p(y_1|x_1, x_2, x_3) \\
& p(y_2|x_1, x_2, x_3) \ldots p(y_5|x_1, x_2, x_3)
\end{aligned}
$$

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

Credits

These slides are modified from ones produced by Michael Gutmann, made available under Creative Commons licence CC BY 4.0.