Probabilistic Graphical Models

Belief Networks

Example: modeling dependent events

- Mr. Holmes leaves his house
- He observes that the lawn in front of his house is wet.
- This can have two reasons:
  - he left the sprinkler turned on,
    or
  - it rained during the night.
- Without any further information the probability of both events is increased.

Example: modeling dependent events

- Mr. Holmes leaves his house
- He observes that the lawn in front of his house is wet.
- This can have two reasons:
  - he left the sprinkler turned on,
    or
  - it rained during the night.
- Without any further information the probability of both events is increased.
- Now he also observes that his neighbor's lawn is also wet.
  - This raises the probability that is has rained
    and it lowers the probability that he left his sprinkler on.

Example: modeling dependent events

- Mr. Holmes leaves his house
- He observes that the lawn in front of his house is wet.
- This can have two reasons:
  - he left the sprinkler turned on,
    or
  - it rained during the night.
- Without any further information the probability of both events is increased.
- Now he also observes that his neighbor's lawn is also wet.
  - This raises the probability that is has rained
    and it lowers the probability that he left his sprinkler on.

Holmes knows that our knowledge about events influences our knowledge about other events.
How can we teach the computer to be as smart?

Example continued

- Let's formalize: there are four random variables
  - $R \in \{0, 1\}$, $R = 1$ means it has been **R**aining
  - $S \in \{0, 1\}$, $S = 1$ means the **S**prinkler was left on
  - $N \in \{0, 1\}$, $N = 1$ means **N**eighbours lawn is wet
  - $H \in \{0, 1\}$, $H = 1$ means **H**olmes lawn is wet

All of these carry information about each other $\rightarrow$ they are dependent

Example continued

- Let's formalize: there are four random variables
  - $R \in \{0,1\}$, $R = 1$ means it has been **R**aining
  - $S \in \{0,1\}$, $S = 1$ means the **S**prinkler was left on
  - $N \in \{0,1\}$, $N = 1$ means **N**eighbours lawn is wet
  - $H \in \{0,1\}$, $H = 1$ means **H**olmes lawn is wet

  All of these carry information about each other $\rightarrow$ they are dependent
- How many states to be specified for their joint distribution?

$$(R, S, N, H) \in \{0,1\} \times \{0,1\} \times \{0,1\} \times \{0,1\} \quad \text{has } 2^4 = 16 \text{ states}$$

$p(R, S, N, H)$ has 15 degrees of freedom (one less than states because of normalization)

Example continued

- Let's formalize: there are four random variables
  - $R \in \{0, 1\}$, $R = 1$ means it has been **R**aining
  - $S \in \{0, 1\}$, $S = 1$ means the **S**prinkler was left on
  - $N \in \{0, 1\}$, $N = 1$ means **N**eighbours lawn is wet
  - $H \in \{0, 1\}$, $H = 1$ means **H**olmes lawn is wet

  All of these carry information about each other $\rightarrow$ they are dependent
- How many states to be specified for their joint distribution?

$$(R, S, N, H) \in \{0, 1\} \times \{0, 1\} \times \{0, 1\} \times \{0, 1\} \quad \text{has } 2^4 = 16 \text{ states}$$

  $p(R, S, N, H)$ has 15 degrees of freedom (one less than states because of normalization)
- Maybe we can save something by a different parameterization?

$$p(R, S, N, H) = p(H \mid R, S, N)p(N \mid R, S)p(R \mid S)p(S)$$

Example continued

- Let's formalize: there are four random variables
  - $R \in \{0,1\}$, $R = 1$ means it has been **R**aining
  - $S \in \{0,1\}$, $S = 1$ means the **S**prinkler was left on
  - $N \in \{0,1\}$, $N = 1$ means **N**eighbours lawn is wet
  - $H \in \{0,1\}$, $H = 1$ means **H**olmes lawn is wet

  All of these carry information about each other $\to$ they are dependent
- How many states to be specified for their joint distribution?

$$(R, S, N, H) \in \{0,1\} \times \{0,1\} \times \{0,1\} \times \{0,1\} \quad \text{has } 2^4 = 16 \text{ states}$$

  $p(R, S, N, H)$ has 15 degrees of freedom (one less than states because of normalization)
- Maybe we can save something by a different parameterization?

$$p(R, S, N, H) = \underbrace{p(H \mid R, S, N)}_{2^3=8} \underbrace{p(N \mid R, S)}_{2^2=4} \underbrace{p(R \mid S)}_{2} \underbrace{p(S)}_{1}$$

  still $8 + 4 + 2 + 1 = 15$ values needed

Example – Conditional Independence

Holmes grass, Neighbours grass, Rain, Sprinkler

- As modeler of this problem we have prior knowledge: the dependencies / independencies between variables

Example – Conditional Independence

Holmes grass, Neighbours grass, Rain, Sprinkler

- As modeler of this problem we have prior knowledge: the dependencies / independencies between variables
- $p(R \mid S) = p(R)$

Example – Conditional Independence

Holmes grass, Neighbours grass, Rain, Sprinkler

- As modeler of this problem we have prior knowledge: the dependencies / independencies between variables
- $p(R \mid S) = p(R)$
- $p(N \mid R, S) = p(N \mid R)$

Example – Conditional Independence

Holmes grass, Neighbours grass, Rain, Sprinkler

- As modeler of this problem we have prior knowledge: the dependencies / independencies between variables
- $p(R \mid S) = p(R)$
- $p(N \mid R, S) = p(N \mid R)$
- $p(H \mid R, S, N) = p(H \mid R, S)$

Example – Conditional Independence

Holmes grass, Neighbours grass, Rain, Sprinkler

- As modeler of this problem we have prior knowledge: the dependencies / independencies between variables
- $p(R \mid S) = p(R)$
- $p(N \mid R, S) = p(N \mid R)$
- $p(H \mid R, S, N) = p(H \mid R, S)$
- In effect our model becomes

$$\begin{aligned} p(R, S, N, H) &= p(H \mid R, S, N) p(N \mid R, S) p(R \mid S) p(S) \\ &= p(H \mid R, S) p(N \mid R) p(R) p(S) \end{aligned}$$

Example – Conditional Independence

Holmes grass, Neighbours grass, Rain, Sprinkler

- As modeler of this problem we have prior knowledge: the dependencies / independencies between variables
- $p(R \mid S) = p(R)$
- $p(N \mid R, S) = p(N \mid R)$
- $p(H \mid R, S, N) = p(H \mid R, S)$
- In effect our model becomes

$$\begin{aligned} p(R, S, N, H) &= p(H \mid R, S, N)p(N \mid R, S)p(R \mid S)p(S) \\ &= p(H \mid R, S)p(N \mid R)p(R)p(S) \end{aligned}$$

- How many degrees of freedom?

Example – Conditional Independence

Holmes grass, Neighbours grass, Rain, Sprinkler

- As modeler of this problem we have prior knowledge: the dependencies / independencies between variables
- $p(R \mid S) = p(R)$
- $p(N \mid R, S) = p(N \mid R)$
- $p(H \mid R, S, N) = p(H \mid R, S)$
- In effect our model becomes

$$\begin{aligned} p(R, S, N, H) &= p(H \mid R, S, N) p(N \mid R, S) p(R \mid S) p(S) \\ &= \underbrace{p(H \mid R, S)}_{4} \underbrace{p(N \mid R)}_{2} \underbrace{p(R)}_{1} \underbrace{p(S)}_{1} \end{aligned}$$

- How many degrees of freedom? 8

Example – Conditional Independence

Holmes grass, Neighbours grass, Rain, Sprinkler

- As modeler of this problem we have prior knowledge: the dependencies / independencies between variables
- $p(R \mid S) = p(R)$
- $p(N \mid R, S) = p(N \mid R)$
- $p(H \mid R, S, N) = p(H \mid R, S)$
- In effect our model becomes

$$\begin{aligned} p(R, S, N, H) &= p(H \mid R, S, N)p(N \mid R, S)p(R \mid S)p(S) \\ &= \underbrace{p(H \mid R, S)}_{4}\underbrace{p(N \mid R)}_{2}\underbrace{p(R)}_{1}\underbrace{p(S)}_{1} \end{aligned}$$

- How many degrees of freedom? 8

Knowing (conditional) independencies can save us space/work!

Example – Inference

Holmes grass, Neighbours grass, Rain, Sprinkler

From the joint probabilities $p(R, S, N, H)$ we can answer all kind of questions.

Let's fix some values for the conditional probability table (CPT)

$$
\begin{aligned}
&p(R=1)=0.2, \qquad p(S=1)=0.1 \\
&p(N=1 \mid R=0)=0.2, \qquad p(N=1 \mid R=1)=1 \\
&p(H=1 \mid R=0, S=0)=0, \qquad p(H=1 \mid R=0, S=1)=0.9 \\
&p(H=1 \mid R=1, S=0)=1, \qquad p(H=1 \mid R=1, S=1)=1
\end{aligned}
$$

Example – Inference

Holmes grass, Neighbours grass, Rain, Sprinkler

Table of joint probabilities $p(R, S, N, H)$:

R	S	N	H	p(H,N,R,S)
0	0	0	0	0.5760
0	0	0	1	0.0000
0	0	1	0	0.1440
0	0	1	1	0.0000
0	1	0	0	0.0064
0	1	0	1	0.0576
0	1	1	0	0.0016
0	1	1	1	0.0144
1	0	0	0	0.0000
1	0	0	1	0.0000
1	0	1	0	0.0000
1	0	1	1	0.1800
1	1	0	0	0.0000
1	1	0	1	0.0000
1	1	1	0	0.0000
1	1	1	1	0.0200

Example – Inference

Holmes grass, Neighbours grass, Rain, Sprinkler

- What is the probability ... that Holmes' leaves his sprinkler on (in general)?

$$p(S=1) = \sum_{R\in\{0,1\},N\in\{0,1\},H\in\{0,1\}} p(R, S=1, N, H) = 0.1$$

Example – Inference

Holmes grass, Neighbours grass, Rain, Sprinkler

- What is the probability ...that Holmes' leaves his sprinkler on (in general)?

$$p(S=1) = \sum_{R\in\{0,1\},N\in\{0,1\},H\in\{0,1\}} p(R,S=1,N,H) = 0.1$$

- ...that Holmes' lawn is wet but his neighbor's is not?

$$p(N=0,H=1) = \sum_{R,S} p(R,S,N=0,H=1) = 0.0576$$

Example – Inference

Holmes grass, Neighbours grass, Rain, Sprinkler

- What is the probability . . . that Holmes' leaves his sprinkler on (in general)?

$$p(S=1) = \sum_{R\in\{0,1\},N\in\{0,1\},H\in\{0,1\}} p(R,S=1,N,H) = 0.1$$

- . . . that Holmes' lawn is wet but his neighbor's is not?

$$p(N=0,H=1) = \sum_{R,S} p(R,S,N=0,H=1) = 0.0576$$

- . . . that Holmes' sprinkler was on, given that his lawns is wet?

$$p(S=1|H=1) = \frac{p(S=1,H=1)}{p(H=1)} = \frac{\sum_{R,N} p(R,S=1,N,H=1)}{\sum_{R,S,N} p(R,S,N,H=1)} = \frac{0.092}{0.272} = 0.3382$$

Example – Inference

Holmes grass, Neighbours grass, Rain, Sprinkler

- What is the probability ... that Holmes' leaves his sprinkler on (in general)?

$$p(S=1) = \sum_{R\in\{0,1\},N\in\{0,1\},H\in\{0,1\}} p(R,S=1,N,H) = 0.1$$

- ... that Holmes' lawn is wet but his neighbor's is not?

$$p(N=0,H=1) = \sum_{R,S} p(R,S,N=0,H=1) = 0.0576$$

- ... that Holmes' sprinkler was on, given that his lawns is wet?

$$p(S=1|H=1) = \frac{p(S=1,H=1)}{p(H=1)} = \frac{\sum_{R,N} p(R,S=1,N,H=1)}{\sum_{R,S,N} p(R,S,N,H=1)} = \frac{0.092}{0.272} = 0.3382$$

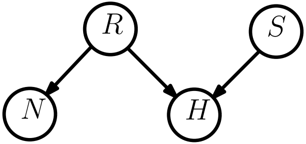
- ... that Holmes' sprinkler was on, given that both lawns are wet?

$$p(S=1|N=1,H=1) = \frac{p(S=1,H=1,N=1)}{p(H=1,N=1)} = \cdots = 0.1604$$

This example as a Belief Network

Holmes grass, Neighbours grass, Rain, Sprinkler

A directed graphical model or belief network (also: Bayesian network) is a way to graphically express how random variables interact with each other:

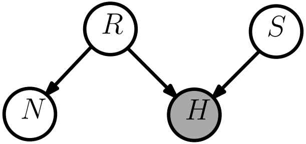


- random variables are circles
- observed random variables are shaded

This example as a Belief Network

Holmes grass, Neighbours grass, Rain, Sprinkler

A directed graphical model or belief network (also: Bayesian network) is a way to graphically express how random variables interact with each other:

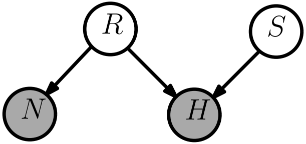


- random variables are circles
- observed random variables are shaded
  - observing Holmes' wet grass

This example as a Belief Network

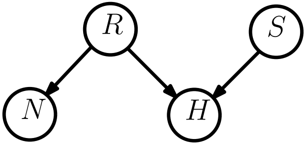
Holmes grass, Neighbours grass, Rain, Sprinkler

A directed graphical model or belief network (also: Bayesian network) is a way to graphically express how random variables interact with each other:



- random variables are circles
- observed random variables are shaded
  - observing Holmes' wet grass
  - also observing the neighbour's wet grass
- arrows encode a form of conditional dependence (later...)

Belief Networks



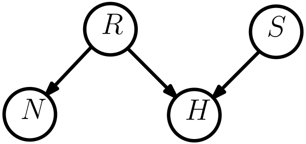
Belief network

A belief network specifies a distribution of the form

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)),$$

where $pa(x)$ denotes the parental variables of x

Belief Networks



Belief network

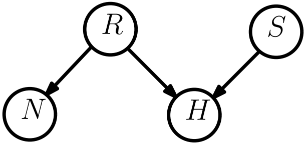
A belief network specifies a distribution of the form

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)),$$

where $pa(x)$ denotes the parental variables of x

- No cycles allowed! $\Rightarrow$ Directed uncyclic graph (DAG)

Belief Networks



Belief network

A belief network specifies a distribution of the form

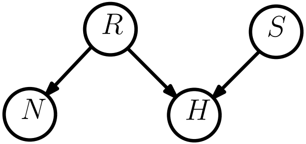
$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)),$$

where $pa(x)$ denotes the parental variables of x

- No cycles allowed! $\Rightarrow$ Directed uncyclic graph (DAG)

Quiz: What if the graph would have cycles?

Belief Networks



Belief network

A belief network specifies a distribution of the form

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)),$$

where $pa(x)$ denotes the parental variables of x

- No cycles allowed! $\Rightarrow$ Directed uncyclic graph (DAG)

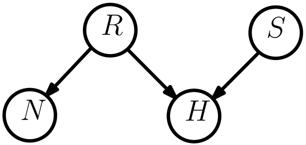
Quiz: What if the graph would have cycles? Product is not a valid probability distribution!

Sampling from a Bayesian network

Belief network

A belief network specifies a distribution of the form

$$p(x_1, \ldots, x_k) = \prod_{i=1}^{k} p(x_i \mid \mathrm{pa}(x_i))$$



For a distribution specified by a Bayesian network, it is easy to *generate samples*:

Sampling

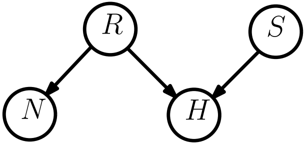
- bring random variables into an order, $i_1, \ldots, i_k$, such that every parent occurs before its children
- for $j = 1, \ldots, k$:
  - sample a value for x_{i_j} according to $p(x_{i_j} | \mathrm{pa}(x_{i_j}))$

Sampling from a Bayesian network

Belief network

A belief network specifies a distribution of the form

$$p(x_1, \ldots, x_k) = \prod_{i=1}^{k} p(x_i \mid \mathrm{pa}(x_i))$$



For a distribution specified by a Bayesian network, it is easy to *generate samples*:

Sampling

- bring random variables into an order, $i_1, \ldots, i_k$, such that every parent occurs before its children
- for $j = 1, \ldots, k$:
  - sample a value for x_{i_j} according to $p(x_{i_j} | \mathrm{pa}(x_{i_j}))$

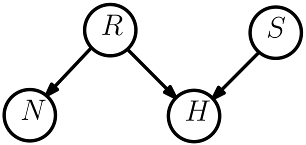
Quiz: What if the graph has cycles?

Sampling from a Bayesian network

Belief network

A belief network specifies a distribution of the form

$$p(x_1, \ldots, x_k) = \prod_{i=1}^{k} p(x_i \mid \mathrm{pa}(x_i))$$



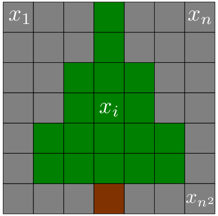
For a distribution specified by a Bayesian network, it is easy to *generate samples*:

Sampling

- bring random variables into an order, $i_1, \ldots, i_k$, such that every parent occurs before its children
- for $j = 1, \ldots, k$:
  - sample a value for x_{i_j} according to $p(x_{i_j} | \mathrm{pa}(x_{i_j}))$

Quiz: What if the graph has cycles? No such global order anymore!

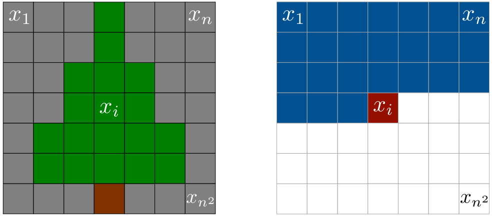
Example: Image generation with PixelCNNs [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]



- Let $p(x_1, \ldots, x_{n^2})$ be the the distribution of $n \times n$ (natural) images
  - very complex (high-dimensional, multi-modal, long-range dependencies between pixels, ...)
  - no good parametric models known

Images: [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

Example: Image generation with PixelCNNs [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

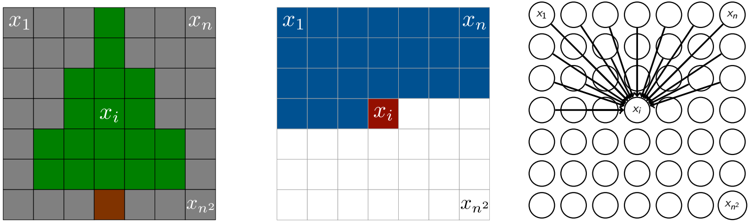


- Let $p(x_1, \ldots, x_{n^2})$ be the the distribution of $n \times n$ (natural) images
  - very complex (high-dimensional, multi-modal, long-range dependencies between pixels, ...)
  - no good parametric models known
- Factorize

$$p(x_1, \ldots, x_{n^2}) = \prod_{i=1}^{n^2} p(x_i | x_1, \ldots, x_{i-1})$$

Images: [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

Example: Image generation with PixelCNNs [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

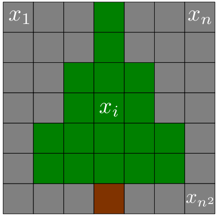


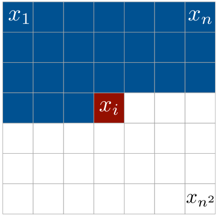
- Let $p(x_1, \ldots, x_{n^2})$ be the the distribution of $n \times n$ (natural) images
  - very complex (high-dimensional, multi-modal, long-range dependencies between pixels, ...)
  - no good parametric models known
- Factorize

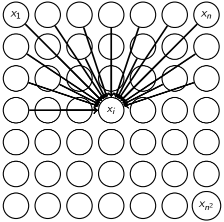
$$p(x_1, \ldots, x_{n^2}) = \prod_{i=1}^{n^2} p(x_i | x_1, \ldots, x_{i-1})$$

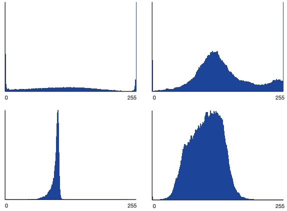
Images: [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

Example: Image generation with PixelCNNs [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]









- Let $p(x_1, \ldots, x_{n^2})$ be the the distribution of $n \times n$ (natural) images
  - very complex (high-dimensional, multi-modal, long-range dependencies between pixels, ...)
  - no good parametric models known
- Factorize

$$p(x_1, \ldots, x_{n^2}) = \prod_{i=1}^{n^2} p(x_i | x_1, \ldots, x_{i-1})$$

- For each factor in the product, learn an artificial neural network (later ...)

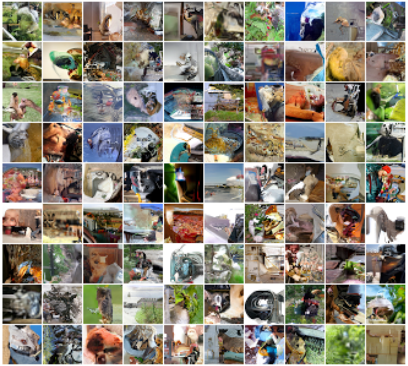
Images: [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

Example: Image generation with PixelCNNs [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

We can generate new images by sampling (pixel-by-pixel) from $p(x_1, \ldots, x_{n^2})$.

Images: [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

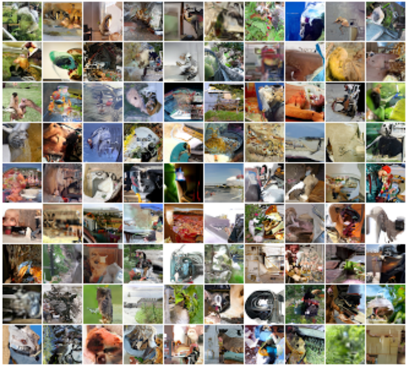
Example: Image generation with PixelCNNs [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

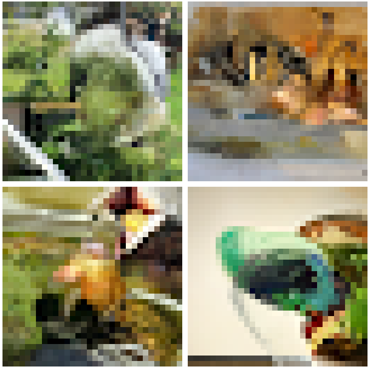


We can generate new images by sampling (pixel-by-pixel) from $p(x_1, \ldots, x_{n^2})$.

Images: [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

Example: Image generation with PixelCNNs [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]





We can generate new images by sampling (pixel-by-pixel) from $p(x_1, \ldots, x_{n^2})$.

Images: [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

Example: Image generation with PixelCNNs [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

We can generate new images by sampling (pixel-by-pixel) from $p(x_1, \ldots, x_{n^2})$.
We can also sample, conditioned on some of the pixels: $p(x_i, \ldots, x_{n^2} | x_1, \ldots, x_{i-1})$.

Images: [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

Example: Image generation with PixelCNNs [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

We can generate new images by sampling (pixel-by-pixel) from $p(x_1, ..., x_{n^2})$.
We can also sample, conditioned on some of the pixels: $p(x_i, ..., x_{n^2}|x_1, ..., x_{i-1})$.

Currently (*i.e.* as of December 2016), one of the state-of-the-art method for image generation.

Images: [Van den Oord et al, "Pixel Recurrent Neural Networks", ICML 2016]

Example: Time-Series

A time-series is an ordered sequence of (discrete or continuous) random variables

$$X_{a:b} = (X_a, X_{a+1}, \ldots, X_b) \qquad \text{for } a, b \in \mathbb{Z}$$

so that one can consider the ‘past’ and ‘future’ in the sequence.

Finance. Stock prices: identify anomalies, predict future behavior.

Climate research. Earth temperature, gas concentrations: analyze patterns, make forecasts.

Biology. DNA sequences: understand them, fill in gaps, cluster them, detect patterns.

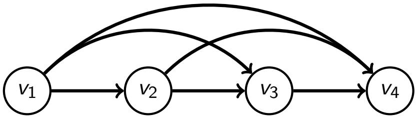
Surveillance. video stream: detect anomalies

Markov Models

For timeseries data $v_1, \dots, v_T$, we need a model $p(v_{1:T})$. For causal consistency, it is meaningful to consider the decomposition

$$p(v_{1:T}) = \prod_{t=1}^{T} p(v_t | v_{1:t-1})$$

with the convention $p(v_t | v_{1:t-1}) = p(v_1)$ for $t = 1$.



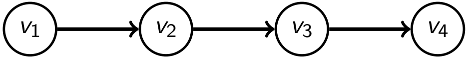
Independence assumptions. It is often natural to assume that the influence of the immediate past is more relevant than the remote past and in Markov models only a limited number of previous observations are required to predict the future.

Markov Chain

Only the recent past is relevant:

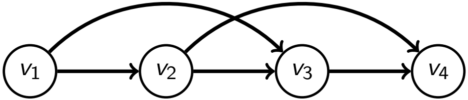
$$p(v_t | v_1, \ldots, v_{t-1}) = p(v_t | v_{t-L}, \ldots, v_{t-1})$$

where $L \geq 1$ is the order of the Markov chain.



$p(v_{1:T}) = p(v_1)p(v_2|v_1)p(v_3|v_2)\ldots p(v_T|v_{T-1})$

first order Markov chain ($L = 1$)



$p(v_{1:T}) = p(v_1)p(v_2|v_1)p(v_3|v_1, v_2)\ldots p(v_T|v_{T-2}, v_{T-1})$

second order Markov chain ($L = 2$)

We call a Markov chain stationary if the transitions $p(v_t = s | v_{(t-L):(t-1)} = S) = f(s, S)$ are time-independent ('homogeneous'). Otherwise it is called non-stationary ('inhomogeneous').

Examples

Examples of Markov chains

- backgammon: which positions can be reached next depends on the current position, not on earlier positions
- random walks:
  - a (very drunk) person walks around; each step is in a random direction
  - start with any graph; at each step, flip a random edge from present to absent or vice versa
- genetic drift: for clonal species, the DNA of the offspring depends only on the parent, not the grandparent
- trajectory of a constant speed moving object: position at previous time point is not enough, but the positions at two time points (as it can derive the speed from it)

Examples of Non-Markov chains

- German text: the probability of the next word can depend on arbitrarily long ago ones
- elephant behavior (because they have such good memories ;-)

Stationary Markov chains

A stationary Markov chains with finite state space, $\mathcal{X}_t = \{1, \ldots, K\}$, is described by

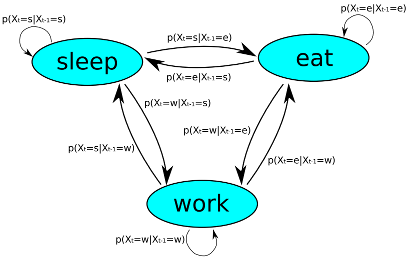
- initial distribution $a_i = p(x_1 = i)$,
- transition matrix: $A_{i',i} = p(x_{t+1} = i' | x_t = i) \in \mathbb{R}^{K \times K}$.

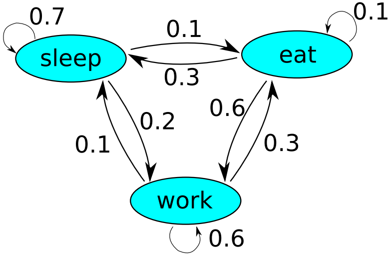
Stationary Markov chains

A stationary Markov chains with finite state space, $\mathcal{X}_t = \{1, \ldots, K\}$, is described by

- initial distribution $a_i = p(x_1 = i)$,
- transition matrix: $A_{i',i} = p(x_{t+1} = i' | x_t = i) \in \mathbb{R}^{K \times K}$.

We can visualize the transitions probabilities as a state diagram:



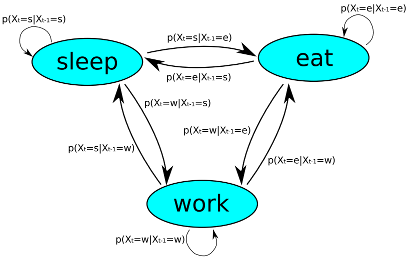


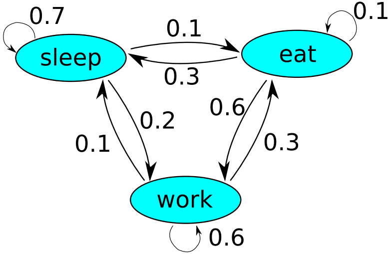
Stationary Markov chains

A stationary Markov chains with finite state space, $\mathcal{X}_t = \{1, \ldots, K\}$, is described by

- initial distribution $a_i = p(x_1 = i)$,
- transition matrix: $A_{i',i} = p(x_{t+1} = i' | x_t = i) \in \mathbb{R}^{K \times K}$.

We can visualize the transitions probabilities as a state diagram:



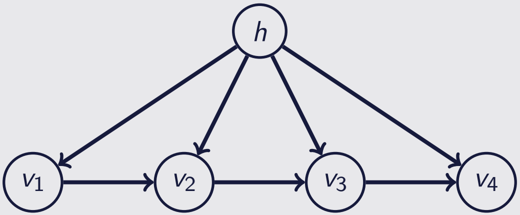


Beware: this is a common illustration, but not the graph of a Bayesian network.

Mixture of Markov models

The transitions of the Markov chain depends on a (discrete) variable $h \in \{1, \dots, K\}$.

Example: Mixture of first order Markov chains



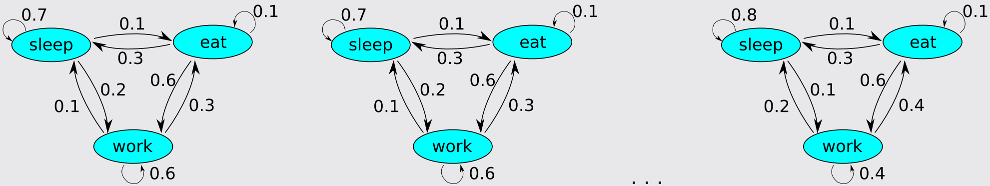
$$p(v_{1:T}, h) = p(v_1|h)p(v_2|v_1, h)p(v_3|v_2, h) \dots p(v_T|v_{T-1}, h)p(h)$$

- for any value of h, this is an ordinary Markov chain
- h is random $\to$ a set of samples will be a mixture of different Markov chains
- useful model, *e.g.*, for sequence clustering (h is the cluster identity)

Mixture of Markov models

Example: Mixture of first order Markov chains

Example: $h \in \{\texttt{Monday}, \texttt{Tueday}, \ldots, \texttt{Sunday}\}$

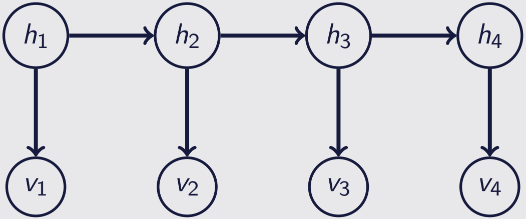


$h = \texttt{Monday}$ $\quad$ $h = \texttt{Tueday}$ $\quad \ldots \ldots \quad$ $h = \texttt{Sunday}$

Different transition probabilities on each day of the week.

Hidden Markov model (HMM)

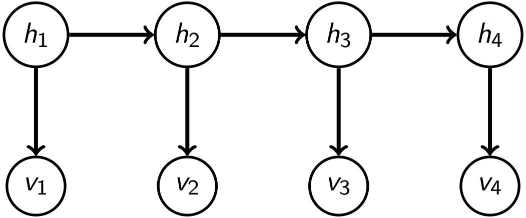
Example: Hidden Markov model



- joint distribution over 2T variables: $p(h_1, \ldots, h_T, x_1, \ldots, x_T)$
- h_t form a Markov chain, each x_t depends only on the corresponding h_t

- interpret: h_t is a state (of an object) at time $t = 1, \ldots, T$, *e.g.* a position
- interpret: x_t is an observations depending only the state, *e.g.* radar image

Hidden Markov model (HMM)



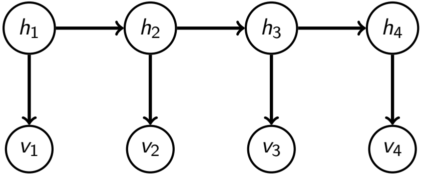
Example

- $h_t \in \{\text{sun, rain, snow}\}$: current weather
- $v_t \in \{\text{jogging, not jogging}\}$: my activity

Example

- $h_t \in \{\text{eat, sleep, work}\}$: my states
- $v_t \in \mathbb{R}$: my blood pressure

Hidden Markov model (HMM)



$$p(h_{1:T}, v_{1:T}) = p(v_1|h_1)p(h_1)\prod_{t=2}^{T} p(v_t|h_t)p(h_t|h_{t-1})$$

Most common: stationary HMM with discrete states $h_t \in \{1, \ldots, H\}$:

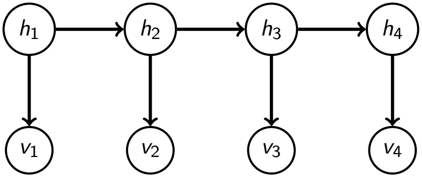
Transition Distribution. $p(h_t|h_{t-1})$ is defined by

- initial distribution $a_i = p(h_1 = i)$,
- transition matrix: $A_{i',i} = p(h_{t+1} = i'|h_t = i) \in \mathbb{R}^{H\times H}$.

Emission Distribution. $p(v_t|h_t)$

- for discrete states, $v_t \in \{1, \ldots, V\}$, matrix $B_{i,j} = p(v_t = i|h_t = j) \in \mathbb{R}^{V\times H}$
- for continuous states, h_t selects one of H possible output distributions $p(v_t|h_t)$.

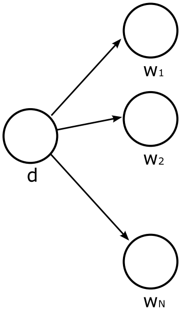
Very useful for reasoning with temporally changing data:



Model allows modeling dynamic processes and (efficiently) answering questions, such as

Filtering (Inferring the present)	$p(h_t \mid v_{1:t})$
Prediction (Inferring the future)	$p(h_t \mid v_{1:s})$ for $t > s$
Smoothing (Inferring the past)	$p(h_t \mid v_{1:u})$ for $t < u$
Predicting future observations	$p(v_t \mid v_{1:s})$ for $t > s$
Likelihood	$p(v_{1:T})$
Find most likely hidden path	$\operatorname{argmax}_{h_{1:T}} p(h_{1:T} \mid v_{1:T})$

A Generative Model of a Text Document: bag of words

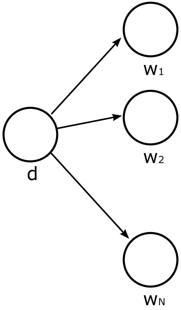


- text document consisting of N English words
- d: document id
- $w_1, \ldots, w_N \in \{\text{all English words}\}$: words

Model reflects how we imagine a corpus of documents could be generated:

- choose an document ID according to $p(d)$
- for $i = 1, \ldots, N$:
  - choose a word w_i according to $p(w|d)$
    (each document has its own preferred or non-preferred words)

A Generative Model of a Text Document: bag of words



- text document consisting of N English words
- d: document id
- $w_1, \ldots, w_N \in \{\text{all English words}\}$: words

Model reflects how we imagine a corpus of documents could be generated:

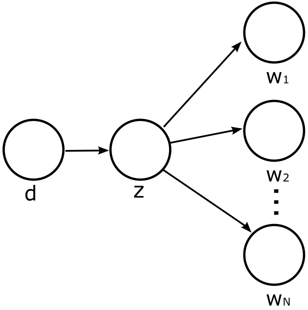
- choose an document ID according to $p(d)$
- for $i = 1, \ldots, N$:
  - choose a word w_i according to $p(w|d)$
    (each document has its own preferred or non-preferred words)

Knowing $p(d, w_1, \ldots, w_N)$ can we generate text documents by random sampling.

Not a particularly "realistic", though...

- *e.g.*, word order does not matter

A Generative Model of a Text Document: mixture of bag of words



- text document consisting of N English words
- d: document id
- $z \in \{1, \ldots, T\}$: topic id
- $w_1, \ldots, w_N \in$ {all English words}: words

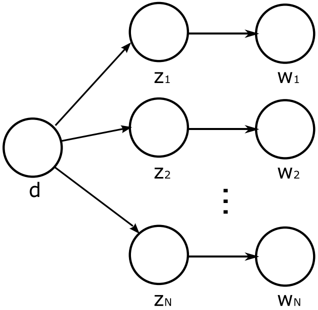
Generative model:

- choose an document ID according to $p(d)$
- pick a topic according to $p(z|d)$
- for $i = 1, \ldots, N$:
  - choose a word w_i according to $p(w|z)$
    (each topic has its own preferred or non-preferred words)

Can be used, *e.g.*, to *cluster* documents:

- estimate $p(z|d)$ and $p(w|z)$ from the data
- for each document, find the most likely topic: $z^* = \operatorname{argmax}_z p(z|d)$
- put documents into the same cluster if the they have the same topic

A Generative Model of a Text Document: probabilistic latent semantic analysis

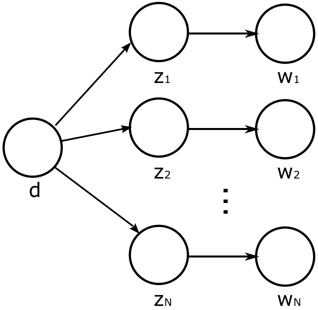


- text document consisting of N English words
- d: document id
- $w_1, \ldots, w_N \in \{\text{all English words}\}$: words
- $z_1, \ldots, z_N \in \{1, \ldots, K\}$: "topic" indicator. In which context/topic was this word used?

Generative model:

- choose an document ID according to $p(d)$
- for $i = 1, \ldots, N$:
  - choose a topic according to $p(z|d)$ (some documents prefer some topics z_i, other prefer others)
  - choose a word w_i according to $p(w|z_i)$ (each topics has its own preferred or non-preferred words)

A Generative Model of a Text Document: probabilistic latent semantic analysis



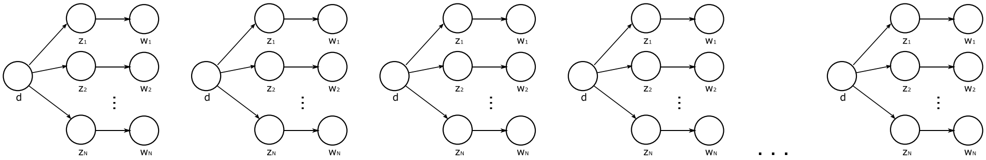
- text document consisting of N English words
- d: document id
- $w_1, \ldots, w_N \in \{\text{all English words}\}$: words
- $z_1, \ldots, z_N \in \{1, \ldots, K\}$: "topic" indicator. In which context/topic was this word used?

Generative model:

- choose an document ID according to $p(d)$
- for $i = 1, \ldots, N$:
  - choose a topic according to $p(z|d)$ (some documents prefer some topics z_i, other prefer others)
  - choose a word w_i according to $p(w|z_i)$ (each topics has its own preferred or non-preferred words)

Also a generative model, and a bit more interesting.

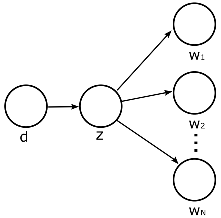
A Generative Model of A Text Corpus



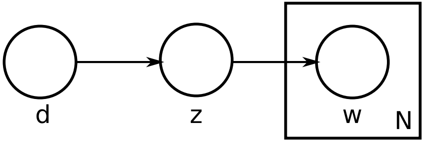
- text corpus: M documents

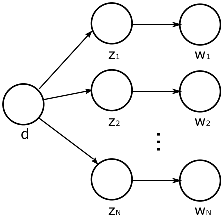
Plate Notation

For notational convenience, repeated elements are put into a box with a number in the corner indicating the number of repeats.



becomes





becomes

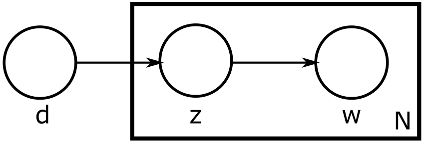
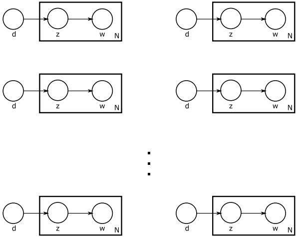
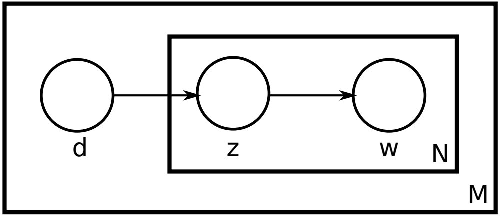


Plate Notation

For notational convenience, repeated elements are put into a box with a number in the corner indicating the number of repeats.



becomes



Probabilistic Latent Semantic Analysis (PLSA) [T. Hofmann, NIPS 2000]

Probabilistic

From $p(\text{documentID}, \text{topics}, \text{words})$ we can infer:

Most likely words per topic:

$$p(\text{words}|\text{topics}{=}1{,}2{,}3{,}4)$$

Topic 1	Topic 2	Topic 3	Topic 4
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Images: [Blei et al, "Latent Dirichlet Allocation", JMLR 2004]

Probabilistic

From $p(\text{documentID}, \text{topics}, \text{words})$ we can infer:

Most likely words per topic:

$$p(\text{words}|\text{topics}{=}1,2,3,4)$$

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Images: [Blei et al, "Latent Dirichlet Allocation", JMLR 2004]

Probabilistic

From $p(\text{documentID}, \text{topics}, \text{words})$ we can infer:

Most likely words per topic:

$$p(\text{words}|\text{topics}{=}1{,}2{,}3{,}4)$$

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Most likely topic per word:

$$p(\text{topics}|\text{word} = i, \text{documentID} = j)$$

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Images: [Blei et al, "Latent Dirichlet Allocation", JMLR 2004]

Latent Dirichlet Allocation (LDA)

- PLSA is a probabilistic model of exactly M text document
- LDA is a more flexible variant that allows generating new documents

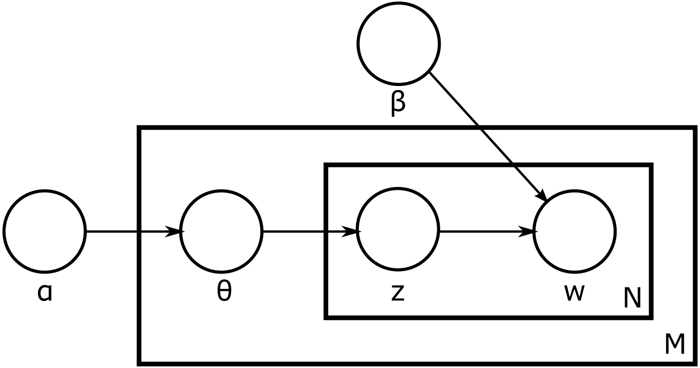
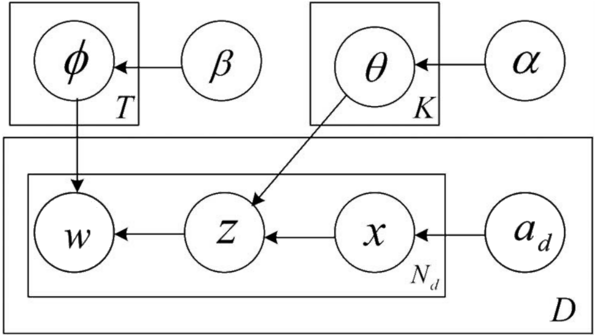


Image: By Bkkbrad - Own work, GFDL, https://commons.wikimedia.org/w/index.php?curid=3610403

Latent Dirichlet Allocation (LDA)

- LDA is a topic model: each word is generated according a word-topic distribution
- author-topic-model: allow for different authors, each has a word-topic distribution



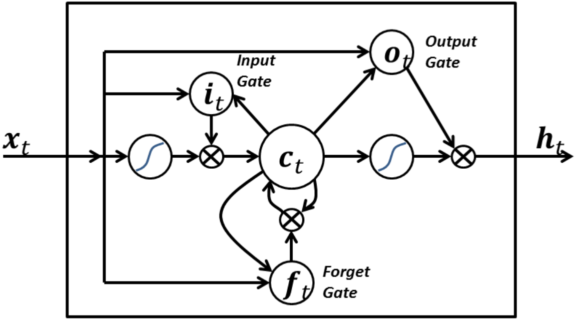
- allows questions such as *"Who wrote this paragraph?"* in an article

Image: [Luo, Li "Automatic Annotation of Multispectral Satellite Images Using Author-Topic Model" IEEE Geoscience, 2012]

Neural Networks for Text Generation

For generating text, Neural Networks can be used as well:

- Long short-term memory (LSTM) network [Hochreiter, Schmidhuber, Neural Computation 1997]
- can be seen as directed, non-Markov, Bayesian network that estimates
  - word sequences, $p(w_t | w_1, \ldots, w_{t-1})$
  - character sequences, $p(c_t | c_1, \ldots, c_{t-1})$



(neural network illustration, not a Bayesian network graph)

Neural Networks for Text Generation

Generating Shakespeare, character by character

KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

Neural Networks for Text Generation

Generating Shakespeare, character by character

KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

Generating Obama speeches

Good afternoon. God bless you.

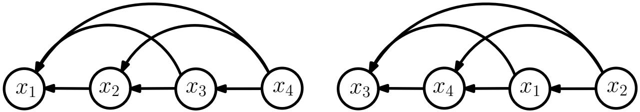
The United States will step up to the cost of a new challenges of the American people that will share the fact that we created the problem. They were attacked and so that they have to say that all the task of the final days of war that I will not be able to get this done.

Different factorizations

Which graph should we use for given random variables?

- graph specifies factorization: $p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i))$
- Any distribution can be written as such a product (in many ways):
- Two factorizations of four variables:

$$
\begin{aligned}
p(x_1, x_2, x_3, x_4) &= p(x_1 \mid x_2, x_3, x_4)p(x_2 \mid x_3, x_4)p(x_3 \mid x_4)p(x_4) \\
p(x_1, x_2, x_3, x_4) &= p(x_3 \mid x_1, x_2, x_4)p(x_4 \mid x_1, x_2)p(x_1 \mid x_2)p(x_2)
\end{aligned}
$$



- Which factorization we use matters if we know (conditional) independences

Belief Networks

- Structure of the DAG corresponds to a set of conditional independence assumptions
  - need to specify all $p(x \mid pa(x))$
  - which parents are sufficient to get the right joint distribution?

- Note: **it is not true that non-parental variables have no influence!**

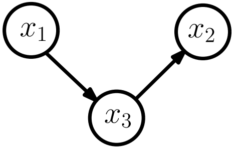
- Example: in distribution

$$p(x_1, x_2, x_3) = p(x_1)p(x_2 \mid x_3)p(x_3 \mid x_1)$$

we have

$$p(x_3 \mid x_1, x_2) \neq p(x_3 \mid x_1)$$

x_2 matters for x_3, even though they are not directly connected.



$x_3 \not\perp\!\!\!\perp x_2 \mid x_1$

Belief Networks

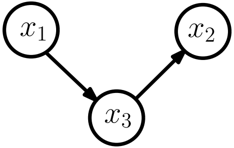
- Structure of the DAG corresponds to a set of conditional independence assumptions
  - need to specify all $p(x \mid pa(x))$
  - which parents are sufficient to get the right joint distribution?
- Note: it is **not** true that non-parental variables have no influence!
- Example: in distribution

$$p(x_1, x_2, x_3) = p(x_1)p(x_2 \mid x_3)p(x_3 \mid x_1)$$

we have

$$p(x_3 \mid x_1, x_2) \neq p(x_3 \mid x_1)$$

x_2 matters for x_3, even though they are not directly connected.

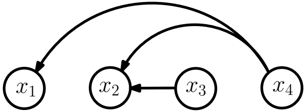


$x_3 \not\perp\!\!\!\perp x_2 \mid x_1$

Rule of thumb: if there is a connection (undirected path) there is some form of dependence.

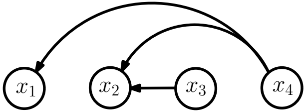
Conditional Independence

- Important task:
  - given graph, read off conditional independence statements



Conditional Independence

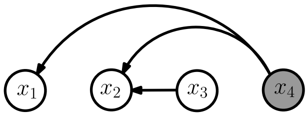
- Important task:
  - given graph, read off conditional independence statements



- Question:

Conditional Independence

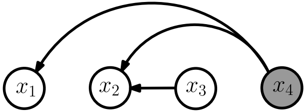
- Important task:
  - given graph, read off conditional independence statements



- Question:
  - are x_1 and x_2 conditionally independent given x_4?

Conditional Independence

- Important task:
  - given graph, read off conditional independence statements



- Question:
  - are x_1 and x_2 conditionally independent given x_4? Yes.

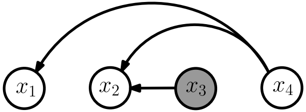
$$p(x_1, x_2, x_3, x_4) = p(x_1|x_4)p(x_2|x_3, x_4)p(x_3)p(x_4)$$

$$p(x_1, x_2|x_4) = \frac{p(x_1, x_2, x_4)}{p(x_4)} = \frac{\sum_{x_3} p(x_1, x_2, x_3, x_4)}{\sum_{x_1,x_2,x_3} p(x_1, x_2, x_3, x_4)} = \frac{\sum_{x_3} p(x_1|x_4)p(x_2|x_3, x_4)p(x_3)p(x_4)}{\sum_{x_1,x_2,x_3} p(x_1|x_4)p(x_2|x_3, x_4)p(x_3)p(x_4)}$$

$$= \frac{p(x_4)p(x_1|x_4)\sum_{x_3} p(x_2, x_3|x_4)}{p(x_4)\sum_{x_1} p(x_1|x_4)\sum_{x_2,x_3} p(x_2, x_3|x_4)} = \frac{p(x_4)p(x_1|x_4)p(x_2|x_4)}{p(x_4)} = p(x_1|x_4)p(x_2|x_4)$$

Conditional Independence

- Important task:
  - given graph, read off conditional independence statements



- Question:
  - are x_1 and x_2 conditionally independent given x_4? Yes.

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

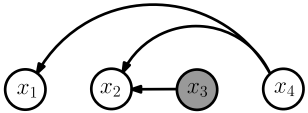
$$p(x_1, x_2|x_4) = \frac{p(x_1, x_2, x_4)}{p(x_4)} = \frac{\sum_{x_3} p(x_1, x_2, x_3, x_4)}{\sum_{x_1,x_2,x_3} p(x_1, x_2, x_3, x_4)} = \frac{\sum_{x_3} p(x_1|x_4)p(x_2|x_3, x_4)p(x_3)p(x_4)}{\sum_{x_1,x_2,x_3} p(x_1|x_4)p(x_2|x_3, x_4)p(x_3)p(x_4)}$$

$$= \frac{p(x_4)p(x_1|x_4)\sum_{x_3} p(x_2, x_3|x_4)}{p(x_4)\sum_{x_1} p(x_1|x_4)\sum_{x_2,x_3} p(x_2, x_3|x_4)} = \frac{p(x_4)p(x_1|x_4)p(x_2|x_4)}{p(x_4)} = p(x_1|x_4)p(x_2|x_4)$$

  - are x_1 and x_2 conditionally independent given x_3?

Conditional Independence

- Important task:
  - given graph, read off conditional independence statements



- Question:
  - are x_1 and x_2 conditionally independent given x_4? Yes.

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

$$p(x_1, x_2|x_4) = \frac{p(x_1, x_2, x_4)}{p(x_4)} = \frac{\sum_{x_3} p(x_1, x_2, x_3, x_4)}{\sum_{x_1,x_2,x_3} p(x_1, x_2, x_3, x_4)} = \frac{\sum_{x_3} p(x_1|x_4)p(x_2|x_3, x_4)p(x_3)p(x_4)}{\sum_{x_1,x_2,x_3} p(x_1|x_4)p(x_2|x_3, x_4)p(x_3)p(x_4)}$$

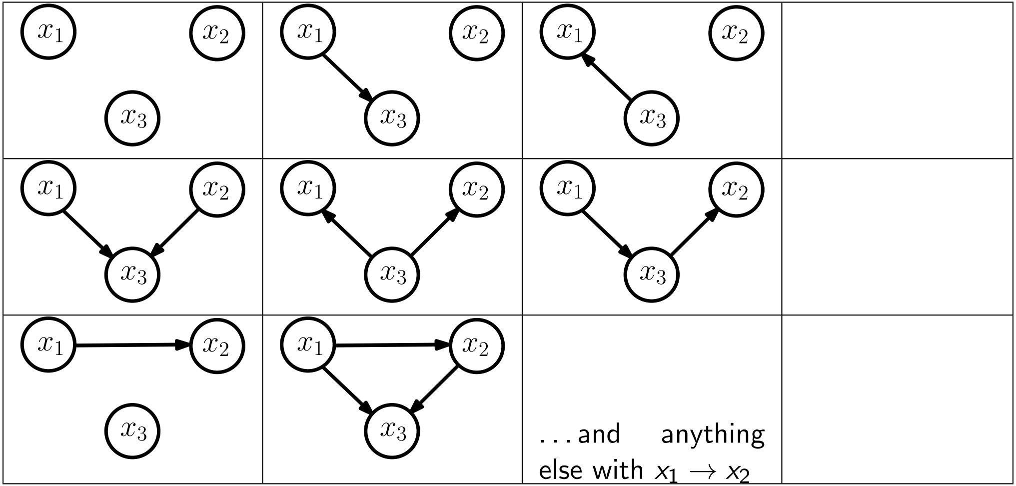
$$= \frac{p(x_4)p(x_1|x_4)\sum_{x_3} p(x_2, x_3|x_4)}{p(x_4)\sum_{x_1} p(x_1|x_4)\sum_{x_2,x_3} p(x_2, x_3|x_4)} = \frac{p(x_4)p(x_1|x_4)p(x_2|x_4)}{p(x_4)} = p(x_1|x_4)p(x_2|x_4)$$

  - are x_1 and x_2 conditionally independent given x_3? No.

Conditional Independences

Is there a way to check this just based on the graph?

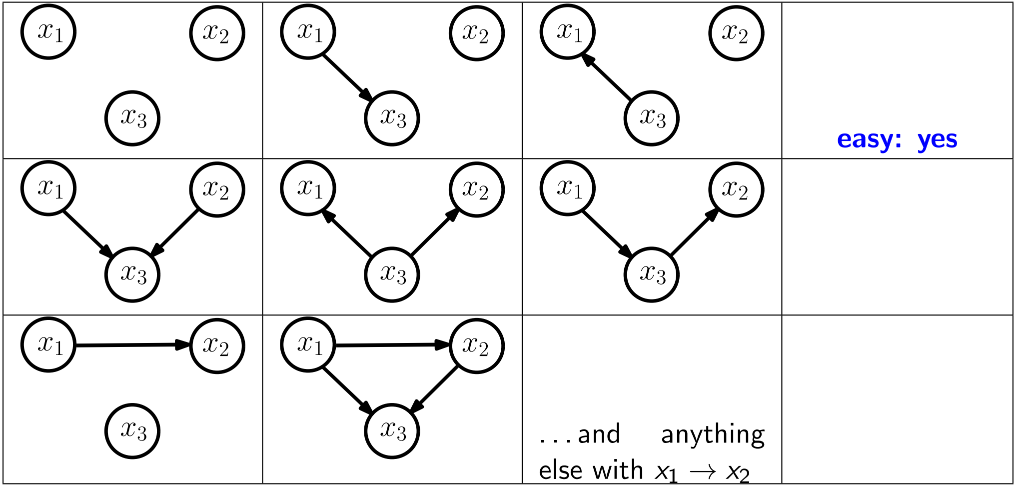
Simplest case: three variables. Are x_1 and x_2 conditionally independent given x_3?



Conditional Independences

Is there a way to check this just based on the graph?

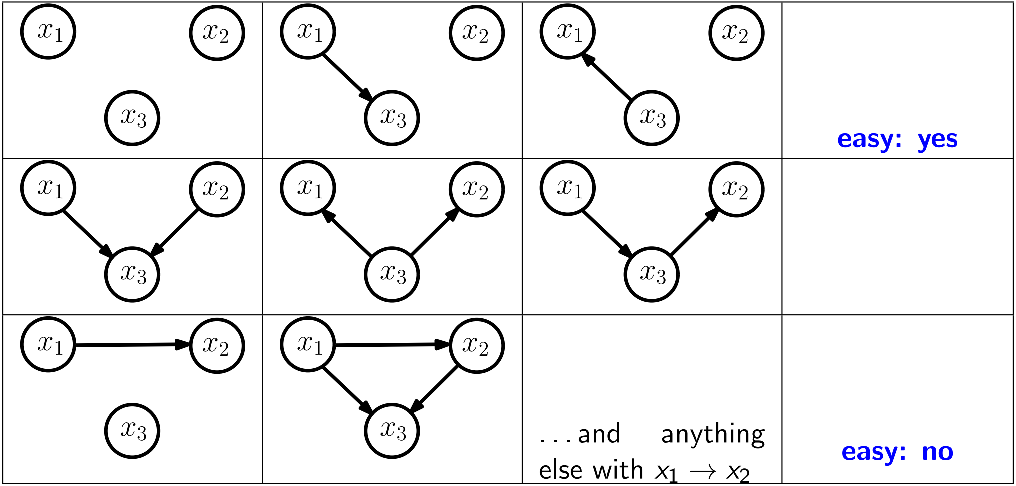
Simplest case: three variables. Are x_1 and x_2 conditionally independent given x_3?



Conditional Independences

Is there a way to check this just based on the graph?

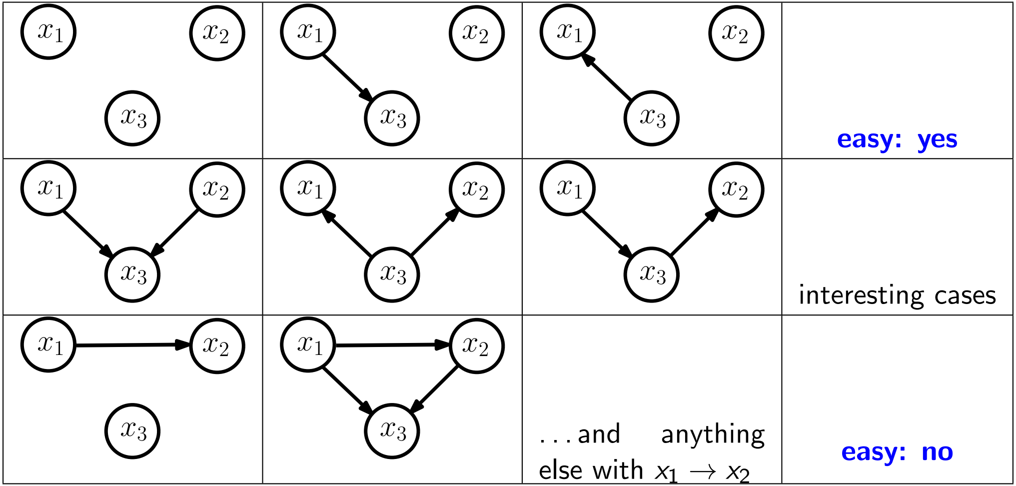
Simplest case: three variables. Are x_1 and x_2 conditionally independent given x_3?



Conditional Independences

Is there a way to check this just based on the graph?

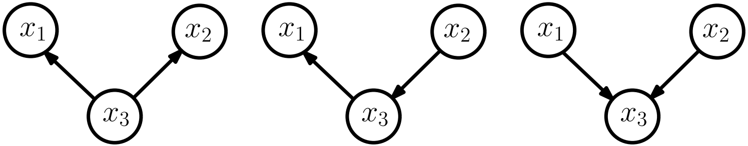
Simplest case: three variables. Are x_1 and x_2 conditionally independent given x_3?



Conditional Independences

Is there a way to check this just based on the graph?

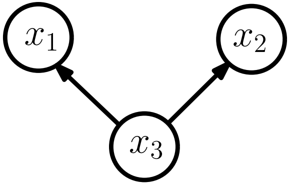
- Interesting cases: indirect connections



Definition: collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. ($a \rightarrow c \leftarrow b$)

Collider and conditional independence



Collision

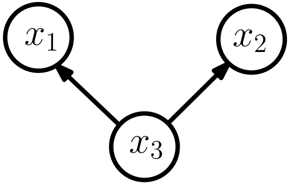
Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. ($a \rightarrow c \leftarrow b$)

- x_3 a collider ?
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$?

Collider and conditional independence

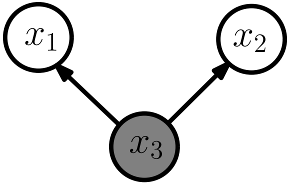
Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \to c \leftarrow b)$



- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$?

Collider and conditional independence



Collision

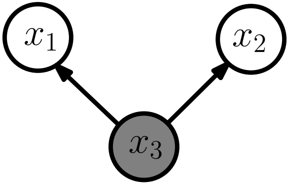
Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$

- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

Collider and conditional independence

Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$



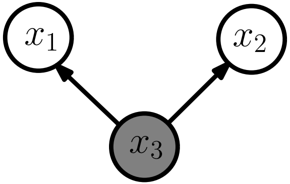
- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$
\begin{aligned}
p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\
&= \\
&=
\end{aligned}
$$

Collider and conditional independence

Collision

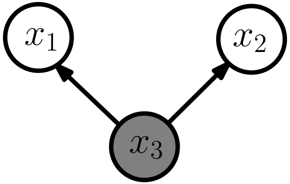
Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$



- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$
\begin{aligned}
p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\
&= p(x_1 \mid x_3)p(x_2 \mid x_3)p(x_3)/p(x_3) \\
&=
\end{aligned}
$$

Collider and conditional independence



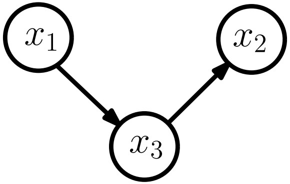
Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. ($a \to c \leftarrow b$)

- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$
\begin{aligned}
p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\
&= p(x_1 \mid x_3)p(x_2 \mid x_3)p(x_3)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_1 \mid x_3)
\end{aligned}
$$

Collider and conditional independence



Collision

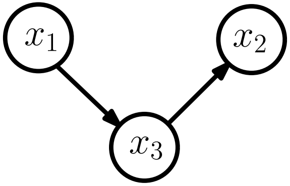
Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \to c \leftarrow b)$

- x_3 a collider ?
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$?

Collider and conditional independence

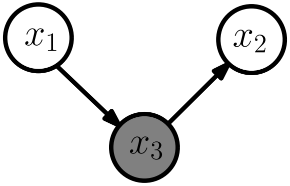
Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$



- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$?

Collider and conditional independence

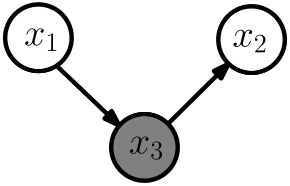


Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$

- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

Collider and conditional independence



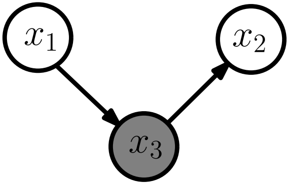
Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \to c \leftarrow b)$

- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$
\begin{aligned}
p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\
&= \\
&= \\
&=
\end{aligned}
$$

Collider and conditional independence



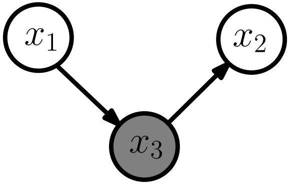
Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$

- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$\begin{aligned} p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\ &= p(x_2 \mid x_3)p(x_3 \mid x_1)p(x_1)/p(x_3) \\ &= \\ &= \end{aligned}$$

Collider and conditional independence



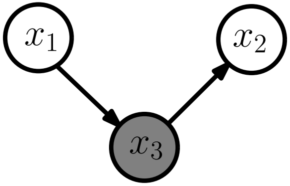
Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$

- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$
\begin{aligned}
p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_3 \mid x_1)p(x_1)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_1, x_3)/p(x_3) \\
&=
\end{aligned}
$$

Collider and conditional independence



Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$

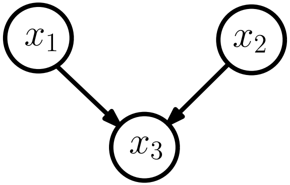
- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$
\begin{aligned}
p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_3 \mid x_1)p(x_1)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_1, x_3)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_1 \mid x_3)
\end{aligned}
$$

Collider and conditional independence

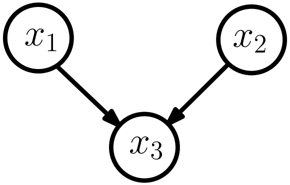
Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$



- x_3 a collider ?
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$?

Collider and conditional independence



Collision

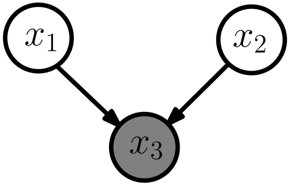
Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. ($a \rightarrow c \leftarrow b$)

- x_3 a collider ? yes
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$?

Collider and conditional independence

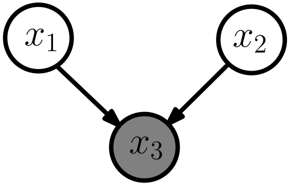
Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$



- x_3 a collider ? yes
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? no!

Collider and conditional independence



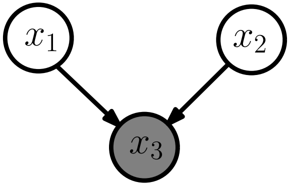
Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \rightarrow c \leftarrow b)$

- x_3 a collider ? yes
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? no!

$$\begin{aligned} p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\ &= p(x_1)p(x_2)\underbrace{p(x_3 \mid x_1, x_2)/p(x_3)}_{\neq 1 \text{ in general}} \end{aligned}$$

Collider and conditional independence



Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \to c \leftarrow b)$

- x_3 a collider ? yes
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? no!

$$\begin{aligned} p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\ &= p(x_1)p(x_2)\underbrace{p(x_3 \mid x_1, x_2)/p(x_3)}_{\neq 1 \text{ in general}} \end{aligned}$$

For three variables in which two are indirectly, but not directly connected: the two are conditionally independent conditioned on the third, if and only if the conditioned variable is not a collider.

Determining Conditional Independence

- Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be disjoint sets of random variables
- There is a general algorithm to check for conditional independence $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$ in any belief network, called "d-separation":

d-separation (the 'd' is for 'directional')

For every $x \in \mathcal{X}, y \in \mathcal{Y}$ check every undirected path U between x and y. A path is blocked if there is a node w on U such that either:

Theorem: If $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$, then $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$.

Determining Conditional Independence

- Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be disjoint sets of random variables
- There is a general algorithm to check for conditional independence $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$ in any belief network, called "d-separation":

d-separation (the 'd' is for 'directional')

For every $x \in \mathcal{X}, y \in \mathcal{Y}$ check every undirected path U between x and y. A path is blocked if there is a node w on U such that either:

1. w is a collider and neither w nor any of its descendant is in $\mathcal{Z}$

Theorem: If $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$, then $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$.

Determining Conditional Independence

- Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be disjoint sets of random variables
- There is a general algorithm to check for conditional independence $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$ in any belief network, called "d-separation":

d-separation (the 'd' is for 'directional')

For every $x \in \mathcal{X}, y \in \mathcal{Y}$ check every undirected path U between x and y. A path is blocked if there is a node w on U such that either:

1. w is a collider and neither w nor any of its descendant is in $\mathcal{Z}$
2. w is not a collider on U and w is in $\mathcal{Z}$

Theorem: If $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$, then $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$.

Determining Conditional Independence

- Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be disjoint sets of random variables
- There is a general algorithm to check for conditional independence $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$ in any belief network, called "d-separation":

d-separation (the 'd' is for 'directional')

For every $x \in \mathcal{X}, y \in \mathcal{Y}$ check every undirected path U between x and y. A path is blocked if there is a node w on U such that either:

1. w is a collider and neither w nor any of its descendant is in $\mathcal{Z}$
2. w is not a collider on U and w is in $\mathcal{Z}$

If all such paths are blocked then $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$

Theorem: If $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$, then $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$.

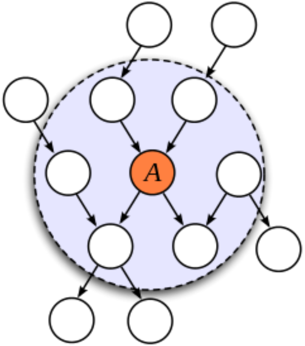
Determining Conditional Independence

Special case:

The distribution of A conditioned on all other variables depends only on the variables in the "Markov blanket".

The Markov blanket comprises:

- Parents
- Children
- Parents of children



Determining Conditional Independence

Other ways to check conditional independence exist, e.g. a detour via undirected graphs:

Given $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ how to determine whether $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$?

1. Let $\mathcal{D} = \{\mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}\}$
2. Build the Ancestral Graph
   - Remove all nodes that are $\notin \mathcal{D}$ and not an ancestor of a node in $\mathcal{D}$
   - Also remove all edges in or out of such nodes
3. Moralisation
   - Connect parents with common child
   - Remove directions
4. Separation
   - Remove links neighbouring $\mathcal{Z}$
   - If no path links a node in $\mathcal{X}$ to a node in $\mathcal{Y} \Rightarrow \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$

Definition: Markov equivalence (for directed and undirected graphs)

Markov equivalence

Two graphs are Markov equivalent if they represent the same set of conditional independence statements.

Definition: Markov equivalence (for directed and undirected graphs)

Markov equivalence

Two graphs are Markov equivalent if they represent the same set of conditional independence statements.

Skeleton

Graph resulting when removing all arrows of edges

Definition: Markov equivalence (for directed and undirected graphs)

Markov equivalence

Two graphs are Markov equivalent if they represent the same set of conditional independence statements.

Skeleton

Graph resulting when removing all arrows of edges

Immorality

Two or more parents of a child with no connection between them

Definition: Markov equivalence (for directed and undirected graphs)

Markov equivalence

Two graphs are Markov equivalent if they represent the same set of conditional independence statements.

Skeleton

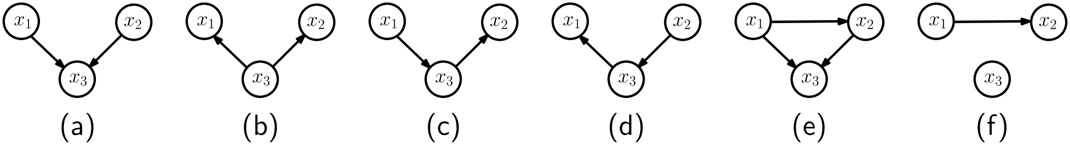
Graph resulting when removing all arrows of edges

Immorality

Two or more parents of a child with no connection between them

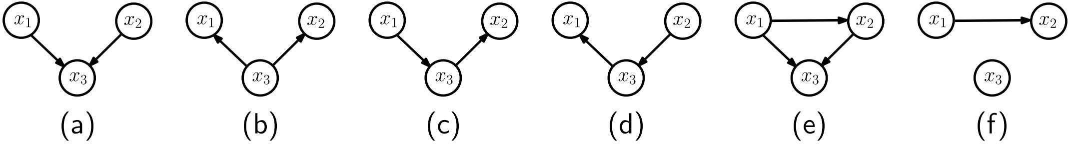
Theorem: Two graphs are Markov equivalent if and only if they have the same skeleton and same set of immoralities.

Three variable graphs revisited



- (a,b,c,d) have the same skeleton, (e) and (f) have different skeletons

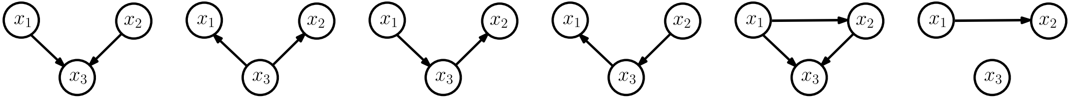
Three variable graphs revisited



- (a,b,c,d) have the same skeleton, (e) and (f) have different skeletons

  $\Rightarrow$ (e) and (f) are not equivalent to any of the others or each other

Three variable graphs revisited

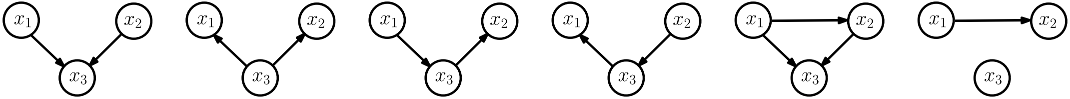


(a) (b) (c) (d) (e) (f)

- (a,b,c,d) have the same skeleton, (e) and (f) have different skeletons

  $\Rightarrow$ (e) and (f) are not equivalent to any of the others or each other
- (b,c,d) have no immoralities, (a) has immorality (x_1, x_2)

Three variable graphs revisited



(a) (b) (c) (d) (e) (f)

- (a,b,c,d) have the same skeleton, (e) and (f) have different skeletons

  $\Rightarrow$ (e) and (f) are not equivalent to any of the others or each other
- (b,c,d) have no immoralities, (a) has immorality (x_1, x_2)
  $\Rightarrow$ (b,c,d) are equivalent to each other, (a) is not equivalent to any of the others