Bayesian network	Learning
• directed acyclic graph $(V, \mathcal{E})$ - **structure**	learning structure
• to each node a variable X_v and a conditional probability distribution $P(X_v \mid X_{pa(v)})$ - **parameters**	learning parameters

Standard statistical modeling approach:

- $P(X_V) \in \mathcal{P}$, $\mathcal{P}$ is a family of probability distributions
- each $P(X_V)$ is characterized by the value of a parameter θ.

Two methods:

- Maximum likelihood estimation (based on frequencies of x_V in data).
- Bayesian inference (quantifies uncertainty about θ in probabilistic terms, incorporates additional information).

Parametrizing a Bayesian network:

$$\theta_j^{v,\rho} \stackrel{def}{=} P(X_v = j | X_{pa(v)} = \rho), \quad v \in V,\ \rho \in \mathcal{X}_{pa(v)},\ j \in \mathcal{X}_v$$

$x \in \mathcal{X}$ a typical configuration of $X = (X_1, \ldots, X_{|V|})$

n number of samples that were independently sampled from identical probability distribution $P(X)$.

Probability of a sample x for given parameter θ

$$P(x|\theta) = \prod_{v \in V} P(X_v = j | X_{pa(v)} = \rho)$$

Let $n(x)$ be multiplicity of x. Probability of $n(x)$ samples of x is

$$P(x|\theta)^{n(x)} = \prod_{v \in V} P(X_v = j | X_{pa(v)} = \rho)^{n(x)}$$

$n(x_A)$ number of samples observed to have $X_A = x_A, A \subseteq V$ (marginal counts)

$fa(v)$ family of node v $fa(v) = v \cup pa(v)$.

Probability of a database D of samples is

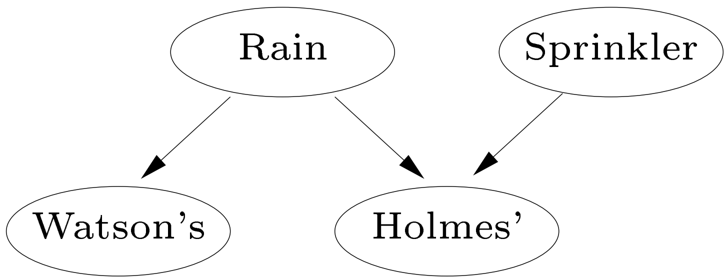
$$P(D|\theta) = \prod_{v \in V} \prod_{x_{fa(v)}} P(X_v = x_v | X_{pa(v)} = x_{pa(v)})^{n(fa(v))} .$$

Probability of θ given an observed data

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

Maximum likelihood parameter $\hat{\theta} = \arg\max_\theta P(D|\theta)$. It is known that

$$\hat{\theta}_{x_v}^{v, x_{pa(v)}} \stackrel{def}{=} P(X_v = x_v | X_{pa(v)} = x_{pa(v)}) = \frac{n(x_{fa(v)})}{n(x_{pa(v)})}$$



$$\begin{aligned} fa(W) &= \{W, R\} \\ fa(R) &= \{R\} \\ fa(H) &= \{H, R, S\} \\ fa(S) &= \{S\} \end{aligned}$$

W	R	H	S
yes	yes	yes	no
no	no	yes	yes
no	no	no	no
yes	yes	yes	yes

$$\begin{aligned} P(W = w \mid R = r) &= \frac{n(w, r)}{n(r)} \\ P(R = r) &= \frac{n(r)}{n} \\ P(H = h \mid R = r, S = s) &= \frac{n(h, r, s)}{n(r, s)} \\ P(S = s) &= \frac{n(s)}{n} \end{aligned}$$

W	R	H	S
yes	yes	yes	no
no	no	yes	yes
no	no	no	no
yes	yes	yes	yes

$n(w, r) =$

	W=y	W=n
R=y	2	0
R=n	0	2

$n(r) =$

R=y	R=n
2	2

$n(h, r, s) =$

	R=y		R=n	
	S=y	S=n	S=y	S=n
H=y	1	1	1	0
H=n	0	0	0	1

$n(r, s) =$

	S=y	S=n
R=y	1	1
R=n	1	1

$$P(W = w|R = r) = \frac{n(w, r)}{n(r)} =$$

	W=y	W=n
R=y	1	0
R=n	0	1

$$P(R = r) = \frac{n(r)}{n} =$$

R=y	R=n
1/2	1/2

$$P(H = h|R = r, S = s) = \frac{n(h, r, s)}{n(r, s)} =$$

	R=y		R=n	
	S=y	S=n	S=y	S=n
H=y	1	1	1	0
H=n	0	0	0	1

$$P(S = s) = \frac{n(s)}{n} =$$

S=y	S=n
1/2	1/2

An incomplete case = one or more variables does not have a value associated with it. Data are missing or unobserved or censored.

W	R	H	S
no	no	?	?

There may be a structure in the manner in which data are missing. If probability of observing the particular set of variables

- does not depend upon the values of those variables $\longrightarrow$ Missing completely at random (MCAR)
- depends upon the values of those variables $\longrightarrow$ Missing at random (MAR)

Assume *age* and *income* of respondents are recorded.

- If the probability that *income* is recorded is the same for all individuals, regardless their *age* or *income* $\longrightarrow$ Missing completely at random (MCAR)
- If the probability that *income* is recorded varies according to the *age* of the respondent but does not depend on values of *income* $\longrightarrow$ Missing at random (MAR)
- If the probability that *income* is recorded varies according to the values of *income* $\longrightarrow$ neither MCAR nor MAR. In this case the censoring mechanism is *non-ignorable* or *informative*.

The EM-algorithm computes *maximum likelihood* estimates involving *missing data* or *incomplete information.* Assume

$f(x\|\theta)$	the underlying PDF for complete data
y	observed values of a complete data record $x \in \mathcal{X}$
$\mathcal{X}(y)$	set of all $x \in \mathcal{X}$ that can generate y

Given an initial estimate of θ the EM-algorithm repeats:

- **E-step**: compute current expected value of log-likelihood

$$Q(\theta'|\theta) \quad = \quad E_{\theta}\{\log f(x|\theta') \mid \text{observed data}\}$$

- **M-step**: set θ'' to θ' that maximize $Q(\theta'|\theta)$.
- If the difference of θ and θ'' is bigger than a threshold then set $\theta = \theta''$ and go to the E-step again.

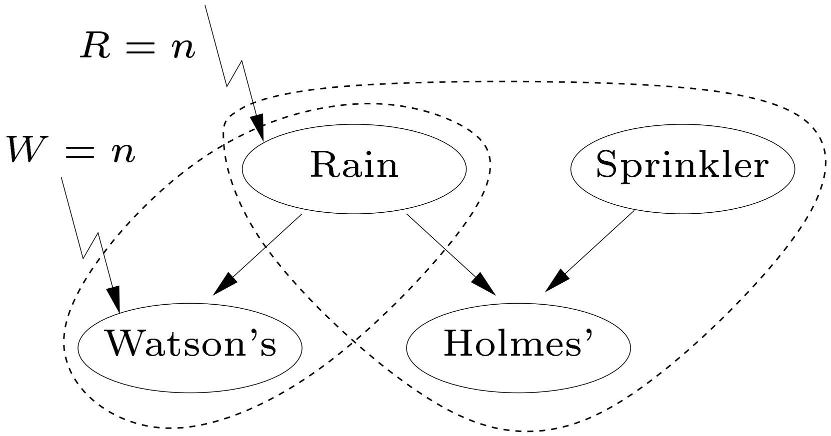
Let P be an initial estimate of the probability distribution of a Bayesian network. Let $a \in \mathcal{A}, \mathcal{A} = \{fa(v) : v \in V\}$. Then the E-step is computing of the expected marginal count of x_a given observed incomplete data.

$$n^\star(x_a) \quad = \quad E_P\{n(x_a) \mid \text{observed data}\}$$

Assume we have observed incomplete observations $y^1, \ldots, y^N$.

Define $\chi^\nu(x_a) = \begin{cases} 1 & \text{if } \exists x' : x' \in \mathcal{X}(y^\nu) \text{ and } x'_a = x_a \quad \text{(a)} \\ 0 & \text{otherwise} \quad \text{(b)} \end{cases}$

$$\begin{aligned} n^\star(x_a) &= E_P\{\sum_{\nu=1}^N \chi^\nu(x_a) \mid y^1, \ldots, y^N\} = \sum_{\nu=1}^N E_P\{\chi^\nu(x_a) \mid y^\nu\} \\ &= \sum_{\nu=1}^N \begin{cases} P(x_a \mid y^\nu) \cdot 1 & \text{if (a)} \\ P(x_a \mid y^\nu) \cdot 0 & \text{if (b)} \end{cases} = \sum_{\nu=1}^N \begin{cases} P(x_a \mid y^\nu) & \text{if (a)} \\ 0 \cdot 0 & \text{if (b)} \end{cases} \\ &= \sum_{\nu=1}^N P(x_a \mid y^\nu) \end{aligned}$$



	W	R	H	S
$y^1 =$	(yes	yes	yes	no)
$y^2 =$	(no	no	yes	yes)
$y^3 =$	(no	no	no	no)
$y^4 =$	(yes	yes	yes	yes)
$\rightarrow$ $y^5 =$	(**no**	**no**	**?**	**?**)

$w, r \in \{yes, no\}$:

$$\begin{aligned} n^\star(w,r) &= n(w,r) + P(W = w, R = r \mid W = no, R = no) \\ n^\star(r,s,h) &= n(r,s,h) + P(R = r, S = s, H = h \mid W = no, R = no) \end{aligned}$$

Recall, that P is an initial distribution.

Let the initial distribution P be a uniform distribution then

$n^\star(w, r) =$

	W=y	W=n
R=y	2+0	0+0
R=n	0+0	2+1

$n^\star(h, r, s) =$

	R=y		R=n	
	S=y	S=n	S=y	S=n
H=y	1+0	1+0	1+1/4	0+1/4
H=n	0	0	0+1/4	1+1/4

For each case $y^\nu, \nu = 1, \ldots, N$ the corresponding evidence is entered into the Bayesian network encoding P and propagated.
In each clique the updated probability tables are accumulated so that we get $n^\star(x_a), a \in \mathcal{A}$.

M-step is equivalent to the case of complete data, i.e. new conditional probability distributions are computed as

$$P'(x_v|x_{pa(v)}) \quad = \quad \frac{n^\star(x_{fa(v)})}{n^\star(x_{pa(v)})}$$

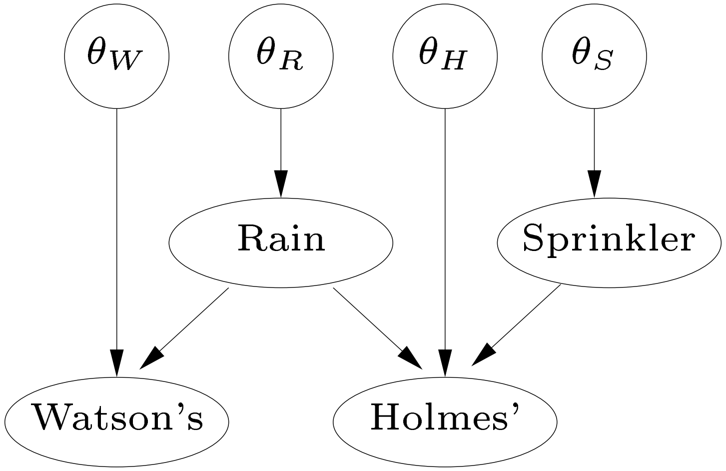
$P'(w\|r)$	W=y	W=n
R=y	$\frac{2}{2}$	0
R=n	0	$\frac{3}{3}$

$P'(r)$	R=y	R=n
	2/5	3/5

$P'(s)$	S=y	S=n
	$\frac{2+1/2}{5}$	$\frac{2+1/2}{5}$

$P'(h\|r,s)$	R=y		R=n	
	S=y	S=n	S=y	S=n
H=y	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{1+1/4}{1+1/2}$	$\frac{0+1/4}{1+1/2}$
H=n	0	0	$\frac{0+1/4}{1+1/2}$	$\frac{1+1/4}{1+1/2}$

- :-) EM-algorithm converges to the probability distribution that is the maximum likelihood estimate given the missing data.
- :-) In both steps the complexity is proportional to the size of the largest clique, S. Lauritzen (1995).
- :-(Often, EM-algorithm converges quite slowly.
- :-) It can be speeded up by the gradient descent search near the maximum, B. Thieson (1997).



- It allows second order uncertainty - uncertainty about conditional probabilities.
- With complete data both batch and sequential updating are straightforward. In case of a discrete model $\longrightarrow$ updating Dirichlet distributions.
- With incomplete data an exact Bayesian inference would require consideration of all possible completions of the incomplete cases. Even with initial distribution being Dirichlet the updating results in a mixture of Dirichlets with enormous number of terms.

We will use the Bayesian inference (updating) for *adaptation of parameters of Bayesian networks.* Assumptions:

- *global independence*: $\theta_X \perp\!\!\!\perp \theta_Y$ for all variables $X \neq Y$
- *local independence*: e.g.
  the uncertainty θ_H on $P(H|S = yes, R = yes)$ is independent of the uncertainty θ_H on $P(H|S = no, R = no)$.

We can express our certainty in a distribution by an fictitious sample size s

$$P(H|R = y, S = y) = \left(\frac{n(H = y)}{s_H}, \frac{n(H = n)}{s_H} \right) ,$$

such that $s_H = n(H = y) + n(H = n)$.

- A new case $e = (H = y, R = y, S = y)$:

$$P(H|R = y, S = y) \quad := \quad \left(\frac{n(H = y) + 1}{s_H + 1}, \quad \frac{n(H = n)}{s_H + 1}\right)$$

- A new case $e = (R = y, S = y)$ with $P(H|R = y, S = y) = (y(H = y), y(H = n))$:

$$P(H|R = y, S = y) \quad := \quad \left(\frac{n(H = y) + y(H = y)}{s_H + 1}, \quad \frac{n(H = n) + y(H = n)}{s_H + 1}\right)$$

- A new case e with $P(R = y, S = y|e) = z(R = y, S = y)$:

$$P(H|R = y, S = y) \quad := \quad \begin{pmatrix} \dfrac{n(H = y) + z(R = y, S = y) \cdot y(H = y)}{s_H + z(R = y, S = y)}, \\ \\ \dfrac{n(H = n) + z(R = y, S = y) \cdot y(H = n)}{s_H + z(R = y, S = y)} \end{pmatrix}$$

Reasons for *fading*:

- adaptation to a changing environment
- preventing an unbounded growth of the sample size

Assume a new case $e = (H = y, R = y, S = y)$. Let $q \in (0, 1)$ be fading factor. Then instead of

$$P(H|R = y, S = y) \quad := \quad \left(\frac{n(H = y) + 1}{s_H + 1}, \quad \frac{n(H = n)}{s_H + 1} \right) \quad \text{we perform}$$

$$P(H|R = y, S = y) \quad := \quad \left(\frac{q \cdot n(H = y) + 1}{q \cdot s_H + 1}, \quad \frac{q \cdot n(H = n)}{q \cdot s_H + 1} \right)$$

If all counts are 1 then the past fade exponentially.

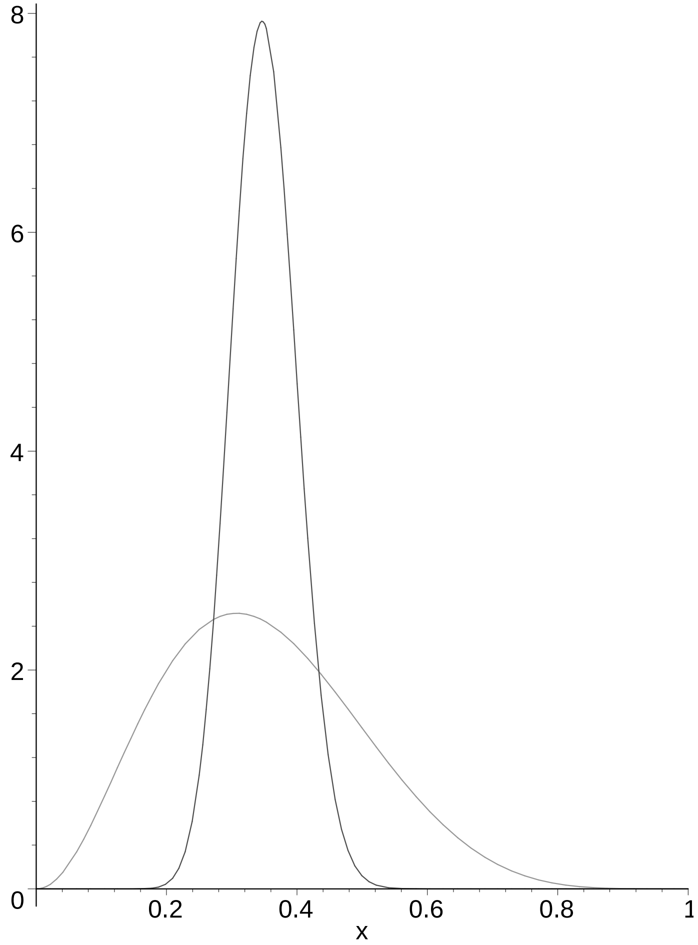
$s^\star = \frac{1}{1-q}$ is effective sample size.

Observe that $s := q \cdot s^\star + 1 = \frac{q}{1-q} + 1 = \frac{1}{1-q} = s^\star$.

Transformation of uncertainty of a parameter from an interval $[x, y]$ to counts n and initial sample size s:

- assume: searching a beta-distribution with mean at $1/2(x + y)$ and with 90% of density mass inside $[x, y]$
- mean $\mu = n/s$ and variance $\sigma^2 = \dfrac{\mu(1-\mu)}{s+1}$ of a beta-distribution.
- 90% of density mass is inside $[\mu - \sigma, \mu - \sigma]$.
- Therefore $n/s = \mu = 1/2(x + y)$ and $\sigma = 1/2(x - y)$, which implies that

$$s = \frac{-2 \cdot (-x + x^2 - y + y^2)}{(x-y)^2} \text{ and } n = \frac{-(x+y) \cdot (-x + x^2 - y + y^2)}{(x-y)^2}.$$



$$\begin{aligned} [x,y] &= [0.3, 0.4] \\ n &= 31.5 \\ s &= 90 \end{aligned} \qquad \begin{aligned} [x,y] &= [0.2, 0.5] \\ n &= 3.2 \\ s &= 9.1 \end{aligned}$$