COMP9418: Advanced Topics in Statistical Machine Learning

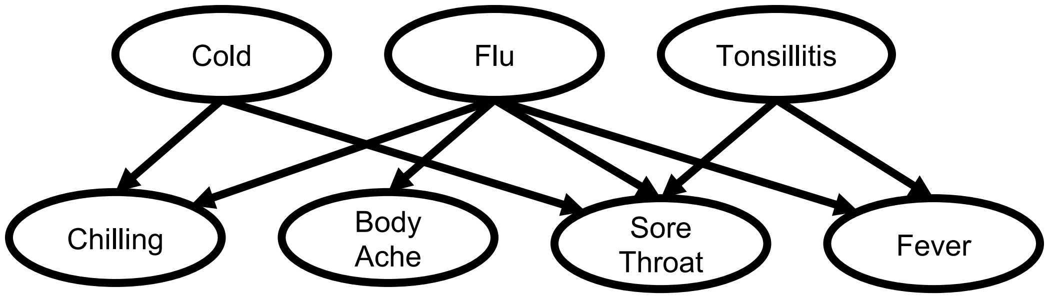
Learning Bayesian Network Parameters with Maximum Likelihood

Instructor: Gustavo Batista

University of New South Wales

Introduction

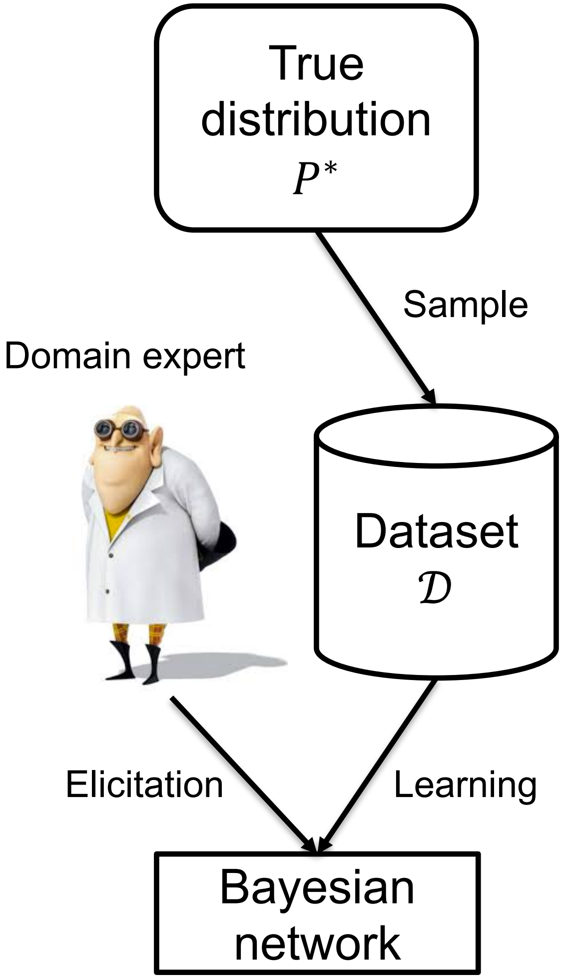
- Consider this Bayesian network structure and dataset
  - Each row in the dataset is called a case and represent a medical record for a patient
  - Some cases are incomplete, where "?" indicates unavailability
- Therefore, the dataset is said to be *incomplete* due to these missing values
  - Otherwise it is called *complete*
- The objective of this lecture is to provide techniques for estimating parameters of a network structure from data
  - Given both complete and incomplete datasets



Case	Cold	Flu	Tonsillitis	Chilling	Body ache	Sore throat	Fever
1	T	?	T	T	F	F	F
2	F	F	T	T	T	F	T
3	?	T	F	F	?	T	F
…	…	…	…	…	…	…	…

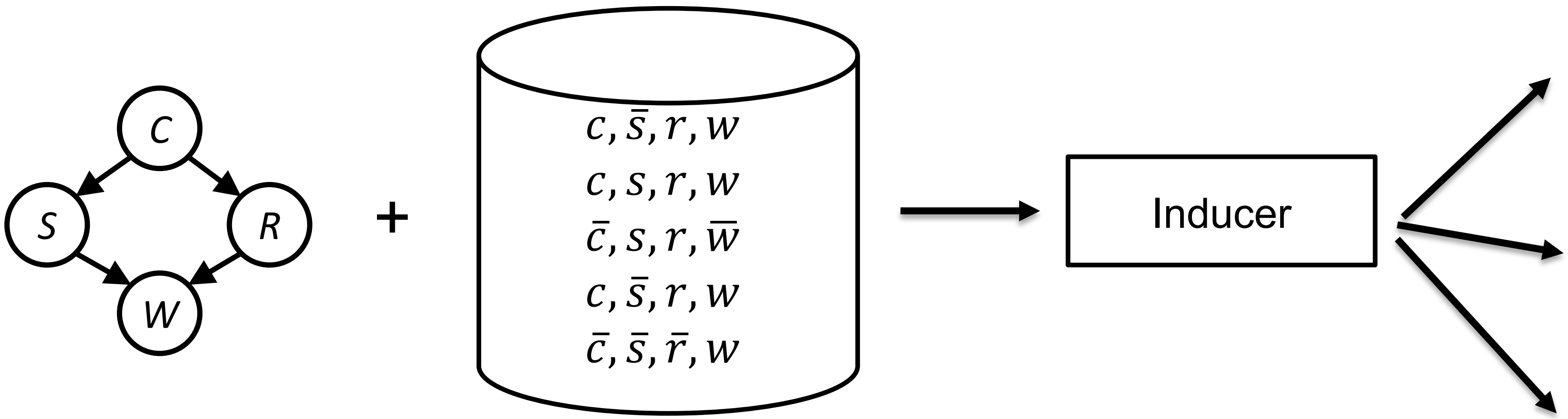
Introduction

- We can construct a network structure by
  - Design information
  - Working with domain experts
- In this lecture, we discuss techniques to estimate the CPTs from data
- Also, we will discuss techniques for learning the network structure itself
  - Although we focus on the complete datasets for this subtask
- The next slides list some possible learning tasks



Known Structure, Complete Data

- This is the simplest setting
  - Given a network that factorizes P^*
  - Dataset with IID samples from P^*
  - We need to output the CPTs



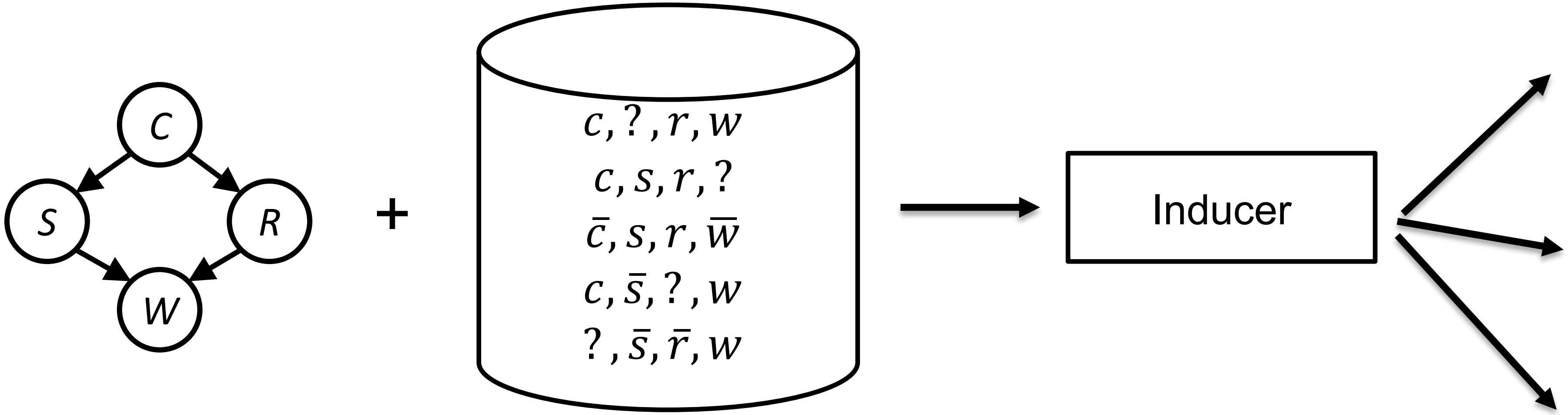
C	$P(C)$
c	0.5
$\bar{c}$	0.5

C	R	$P(R\|C)$
c	r	0.8
c	$\bar{r}$	0.2
$\bar{c}$	r	0.2
$\bar{c}$	$\bar{r}$	0.8

…

Known Structure, Incomplete Data

- Incomplete data complicates the problem considerably
  - Given a network that factorizes P^*
  - Dataset with IID samples from P^* with unknown values
  - We need to output the CPTs



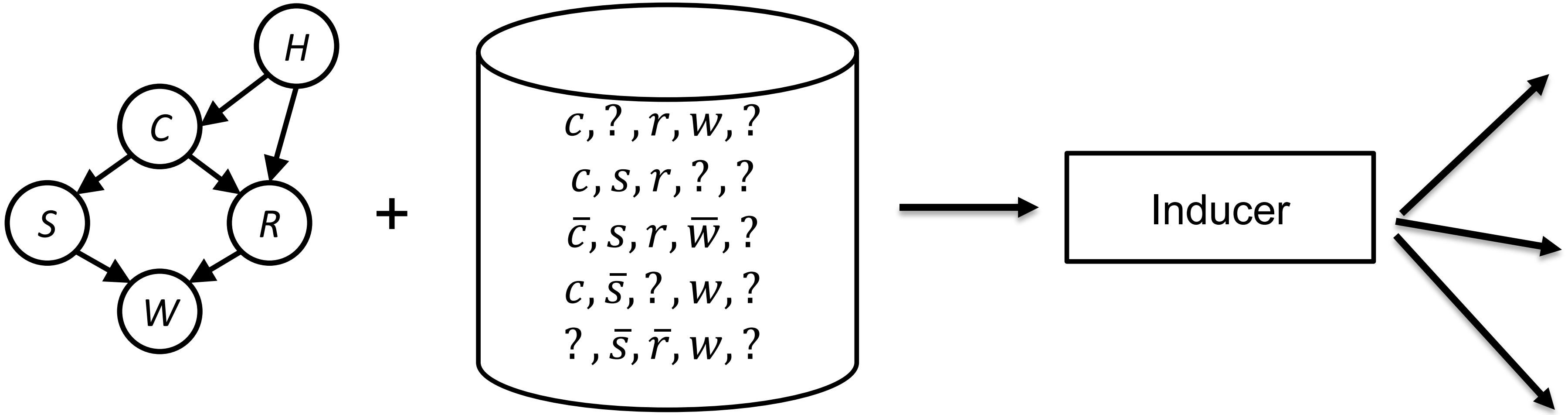
C	$P(C)$
c	0.5
$\bar{c}$	0.5

C	R	$P(R \mid C)$
c	r	0.8
c	$\bar{r}$	0.2
$\bar{c}$	r	0.2
$\bar{c}$	$\bar{r}$	0.8

…

Known Structure, Latent Variables

- Latent variables are not recorded in data
  - Given a network that factorizes P^*
  - Dataset with IID samples from P^* with unknown values and latent variables
  - We need to output the CPTs



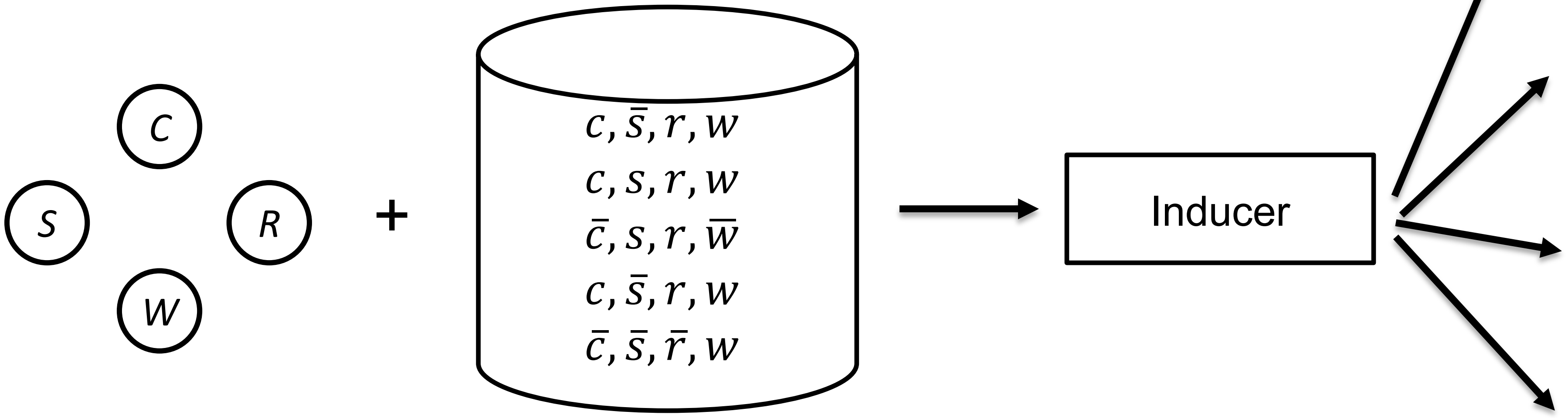
C	$P(C)$
c	0.5
$\bar{c}$	0.5

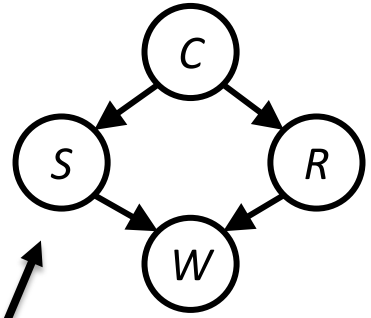
C	R	$P(R\|C)$
c	r	0.8
c	$\bar{r}$	0.2
$\bar{c}$	r	0.2
$\bar{c}$	$\bar{r}$	0.8

…

Unknown Structure, Complete Data

- We may also want to learn the network structure
  - Given a set of random variables
  - Dataset with IID samples from P^*
  - We need to output the edges connectivity and CPTs





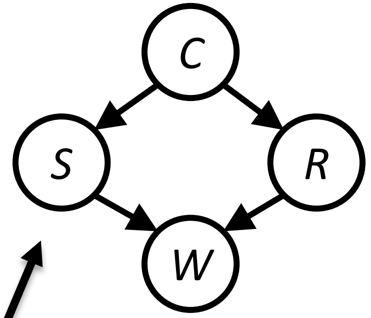
C	$P(C)$
c	0.5
$\bar{c}$	0.5

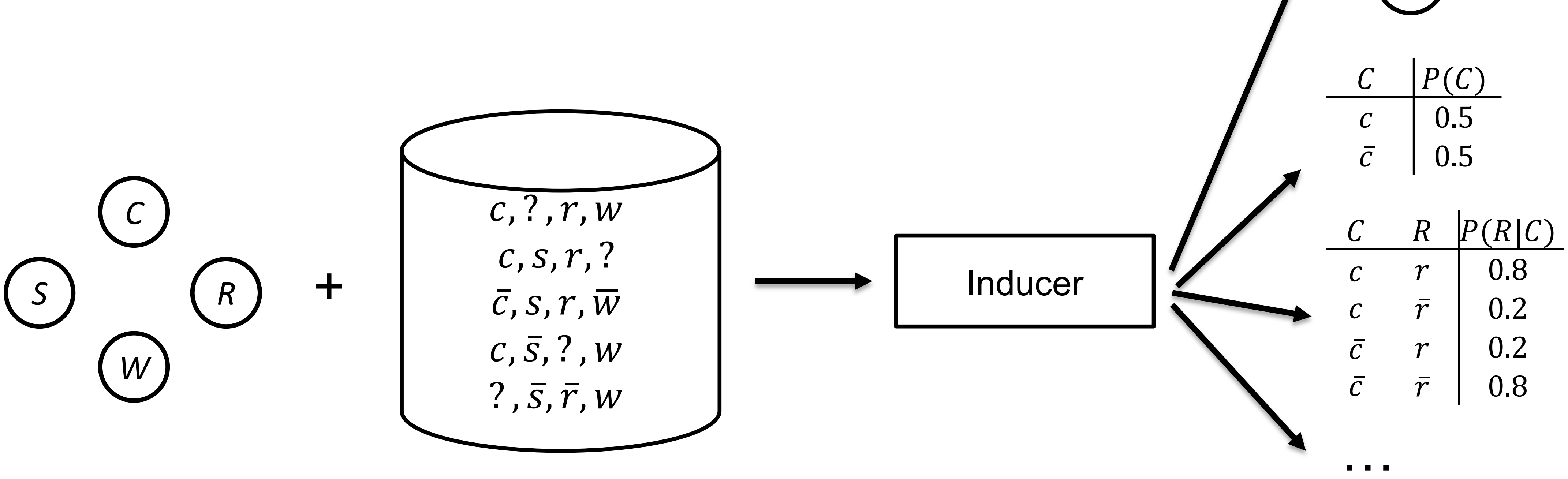
C	R	$P(R\|C)$
c	r	0.8
c	$\bar{r}$	0.2
$\bar{c}$	r	0.2
$\bar{c}$	$\bar{r}$	0.8

…

Unknown Structure, Incomplete Data

- A challenging scenario
  - Given a set of random variables
  - Dataset with IID samples from P^* with unknown values
  - We need to output the edges connectivity and CPTs





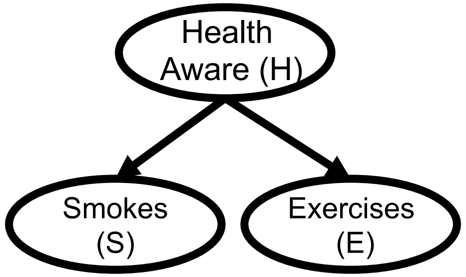
C	$P(C)$
c	0.5
$\bar{c}$	0.5

C	R	$P(R\|C)$
c	r	0.8
c	$\bar{r}$	0.2
$\bar{c}$	r	0.2
$\bar{c}$	$\bar{r}$	0.8

Estimating Parameter from Complete Data

- Consider this simple network
  - Our goal is to estimate its parameters from the data
- Our assumption are
  - These cases are generated independently
  - According to their true probabilities
- Under these assumptions
  - We can define an empirical distribution $P_{\mathcal{D}}$
  - According to this distribution, the empirical probability of an instantiation is simply its frequency of occurrence

Case	H	S	E
1	h	$\bar{s}$	e
2	h	$\bar{s}$	e
3	$\bar{h}$	s	$\bar{e}$
4	$\bar{h}$	$\bar{s}$	e
5	h	$\bar{s}$	$\bar{e}$
6	h	$\bar{s}$	e
7	$\bar{h}$	$\bar{s}$	$\bar{e}$
8	h	$\bar{s}$	e
9	h	$\bar{s}$	e
10	$\bar{h}$	$\bar{s}$	e
11	h	$\bar{s}$	e
12	h	s	e
13	h	$\bar{s}$	e
14	h	s	e
15	h	$\bar{s}$	e
16	h	$\bar{s}$	e



H	S	E	$P_{\mathcal{D}}(.)$
h	s	e	$2/16$
h	s	$\bar{e}$	$0/16$
h	$\bar{s}$	e	$9/16$
h	$\bar{s}$	$\bar{e}$	$1/16$
$\bar{h}$	s	e	$0/16$
$\bar{h}$	s	$\bar{e}$	$1/16$
$\bar{h}$	$\bar{s}$	e	$2/16$
$\bar{h}$	$\bar{s}$	$\bar{e}$	$1/16$

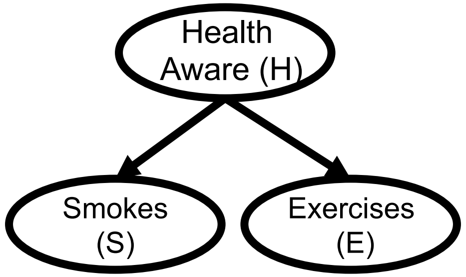
Estimating Parameter from Complete Data

- Empirical distribution $P_{\mathcal{D}}$

$$P_{\mathcal{D}}(h,s,e) = \frac{\mathcal{D}\#(h,s,e)}{N}$$

- where
  - $\mathcal{D}\#(h,s,e)$ is the number of cases in dataset $\mathcal{D}$ that satisfies instantiation h,s,e
  - N is the dataset size

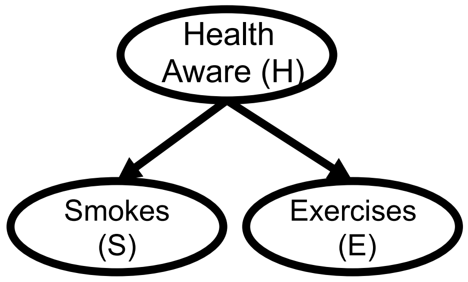
Case	H	S	E
1	h	$\bar{s}$	e
2	h	$\bar{s}$	e
3	$\bar{h}$	s	$\bar{e}$
4	$\bar{h}$	$\bar{s}$	e
5	h	$\bar{s}$	$\bar{e}$
6	h	$\bar{s}$	e
7	$\bar{h}$	$\bar{s}$	$\bar{e}$
8	h	$\bar{s}$	e
9	h	$\bar{s}$	e
10	$\bar{h}$	$\bar{s}$	e
11	h	$\bar{s}$	e
12	h	s	e
13	h	$\bar{s}$	e
14	h	s	e
15	h	$\bar{s}$	e
16	h	$\bar{s}$	e



H	S	E	$P_{\mathcal{D}}(.)$
h	s	e	2/16
h	s	$\bar{e}$	0/16
h	$\bar{s}$	e	9/16
h	$\bar{s}$	$\bar{e}$	1/16
$\bar{h}$	s	e	0/16
$\bar{h}$	s	$\bar{e}$	1/16
$\bar{h}$	$\bar{s}$	e	2/16
$\bar{h}$	$\bar{s}$	$\bar{e}$	1/16

Estimating Parameter from Complete Data

- We can now estimate parameters based on the empirical distribution
- For example, the parameter $\theta_{s|h}$
  - Corresponds to $P_{\mathcal{D}}(s|h)$
  - Probability a person will smoke given they are health-aware



H	S	E	$P_{\mathcal{D}}(.)$
h	s	e	$2/16$
h	s	$\bar{e}$	$0/16$
h	$\bar{s}$	e	$9/16$
h	$\bar{s}$	$\bar{e}$	$1/16$
$\bar{h}$	s	e	$0/16$
$\bar{h}$	s	$\bar{e}$	$1/16$
$\bar{h}$	$\bar{s}$	e	$2/16$
$\bar{h}$	$\bar{s}$	$\bar{e}$	$1/16$

$$P_{\mathcal{D}}(s|h) = \frac{P_{\mathcal{D}}(s,h)}{P_{\mathcal{D}}(h)} = \frac{2/16}{12/16} = \frac{1}{6}$$

Empirical Distribution: Definition

- A dataset $\mathcal{D}$ for variables $\boldsymbol{X}$ is a vector $\boldsymbol{d}_1, \dots, \boldsymbol{d}_N$ where each $\boldsymbol{d}_i$ is called a case and represents a partial instantiation of variables $\boldsymbol{X}$
  - The dataset is *complete* if each case is a complete instantiation of variables $\boldsymbol{X}$
  - Otherwise, the dataset is incomplete

- The empirical distribution for a complete dataset D is defined as

$$P_{\mathcal{D}}(\alpha) \stackrel{\text{def}}{=} \frac{\mathcal{D}\#(\alpha)}{N}$$

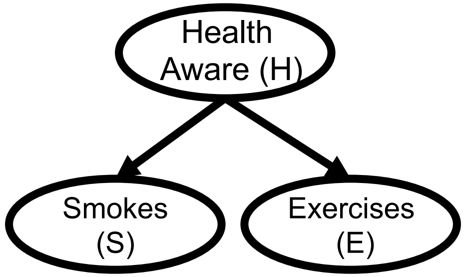
- where
  - $\mathcal{D}\#(\alpha)$ is the number of cases $\boldsymbol{d}_i$ in the dataset $\mathcal{D}$ that satisfy the event α

Complete Data Parameter Estimation: Definition

- We can estimate the parameter $\theta_{x|\boldsymbol{u}}$ by the empirical probability

$$\theta^{ml}_{x|\boldsymbol{u}} \stackrel{\text{def}}{=} P_{\mathcal{D}}(x|\boldsymbol{u}) = \frac{\mathcal{D}\#(x,\boldsymbol{u})}{\mathcal{D}\#(\boldsymbol{u})}$$

- The count $\mathcal{D}\#(x,\boldsymbol{u})$ is called a *sufficient statistic*
  - More generally, any function of the data is called a *statistic*
  - A sufficient statistic is a statistic that contains all the information in the data needed for a particular estimation task
- Considering the network structure and corresponding dataset
  - We have the following parameter estimates



H	S	E	$P_{\mathcal{D}}(.)$
h	s	e	$2/16$
h	s	$\bar{e}$	$0/16$
h	$\bar{s}$	e	$9/16$
h	$\bar{s}$	$\bar{e}$	$1/16$
$\bar{h}$	s	e	$0/16$
$\bar{h}$	s	$\bar{e}$	$1/16$
$\bar{h}$	$\bar{s}$	e	$2/16$
$\bar{h}$	$\bar{s}$	$\bar{e}$	$1/16$

H	θ^{ml}_{H}
h	$3/4$
$\bar{h}$	$1/4$

H	S	$\theta^{ml}_{S\|H}$
h	s	$1/6$
h	$\bar{s}$	$5/6$
$\bar{h}$	s	$1/2$
$\bar{h}$	$\bar{s}$	$1/2$

H	E	$\theta^{ml}_{E\|H}$
h	e	$11/12$
h	$\bar{e}$	$1/12$
$\bar{h}$	e	$1/2$
$\bar{h}$	$\bar{e}$	$1/2$

Complete Data Parameter Estimation: Definition

- We expect the variance of $\theta^{ml}_{x|\boldsymbol{u}}$ will decrease as the dataset increases in size
  - If the dataset is an IID sample of a distribution P
  - The Central Limit Theorem tells us $\theta^{ml}_{x|\boldsymbol{u}}$ is asymptotically Normal
  - It can be approximated by a Normal distribution with
    - Mean
    - Variance
- The variance depends on N, $P(\boldsymbol{u})$ and $P(x|\boldsymbol{u})$
  - It vey sensitive to $P(\boldsymbol{u})$, and it is difficult to estimate this parameter when this probability is small
  - Small $P(\boldsymbol{u})$ and not large enough N leads to the problem of *zero counts*
  - We have seen this problem before in the Naïve Bayes lecture and will return to it when we discuss Bayesian learning

$$P(X|\boldsymbol{u})$$

$$\frac{P(x|\boldsymbol{u})\big(1 - P(x|\boldsymbol{u})\big)}{NP(\boldsymbol{u})}$$

Maximum Likelihood (ML) Estimates

- Let θ be the set of all parameter estimates for a network G
  - P_θ be the probability distribution induced by G and θ
- We define the likelihood of these estimates as
  - That is, the likelihood of estimates θ is the probability of observing the dataset D under these estimates

$$L(\theta;\mathcal{D}) \stackrel{\text{def}}{=} \prod_{i=1}^{N} P_\theta(\boldsymbol{d}_i)$$

- We can show that given a complete dataset $\mathcal{D}$, the parameters $\theta_{x|\boldsymbol{u}}^{ml}$ are the only estimates that maximize the likelihood function
  - For this reason, these estimates are called *maximum likelihood (ML) estimates*
  - They are denoted by θ^{ml}

$$\theta^* = argmax_\theta\, L(\theta;\mathcal{D})$$

iff

$$\theta_{x|\boldsymbol{u}}^* = P_{\mathcal{D}}(x|\boldsymbol{u})$$

$$\theta^{ml} = argmax_\theta\, L(\theta;\mathcal{D})$$

ML Estimates and KL Divergence

- ML estimates also minimize the KL divergence between the learned Bayesian network and the empirical distribution
  - For a complete dataset $\mathcal{D}$ and variables $\boldsymbol{X}$

$$\operatorname{argmax}_\theta L(\theta; \mathcal{D}) = \operatorname{argmin}_\theta \mathrm{KL}(P_{\mathcal{D}}(\boldsymbol{X}), P_\theta(\boldsymbol{X}))$$

- ML estimates are unique for a given structure G and complete dataset $\mathcal{D}$
  - Therefore, the likelihood of these parameters is a function of G and $\mathcal{D}$
  - We define the likelihood of structure G given $\mathcal{D}$ as

$$L(G; \mathcal{D}) \stackrel{\text{def}}{=} L(\theta^{ml}; \mathcal{D})$$

  - Where θ^{ml} are the ML estimates for structure G and dataset $\mathcal{D}$

Log-Likelihood

- Often, it is more convenient to work with the logarithm of the likelihood function

$$LL(\theta;\mathcal{D}) \stackrel{\text{def}}{=} \log L(\theta;\mathcal{D}) = \sum_{i=1}^{N} \log P_\theta(\boldsymbol{d}_i)$$

- The log-likelihood of structure G is defined similarly

$$LL(G;\mathcal{D}) \stackrel{\text{def}}{=} \log L(G;\mathcal{D})$$

- Maximizing the log-likelihood is equivalent to maximizing the likelihood function
  - Although likelihood is ≥ 0 and log-likelihood is ≤ 0
  - We use $\log_2$ for the log-likelihood but suppress the base 2

Log-Likelihood

- A key property of log-likelihood function is that it decomposes into several components
  - One for each family in the Bayesian network structure

- Let G be a structure and $\mathcal{D}$ a complete dataset of size N. If $X\boldsymbol{U}$ ranges over the families of structure G, then

$$LL(G;\mathcal{D}) = -N\sum_{X\boldsymbol{U}} H_{\mathcal{D}}(X|\boldsymbol{U})$$

  - Where $H_{\mathcal{D}}(X|\boldsymbol{U})$ is the conditional entropy, defined as

$$H_{\mathcal{D}}(X|\boldsymbol{U}) = -\sum_{x\boldsymbol{u}} P_{\mathcal{D}}(x\boldsymbol{u}) \log_2 P_{\mathcal{D}}(x|\boldsymbol{u})$$

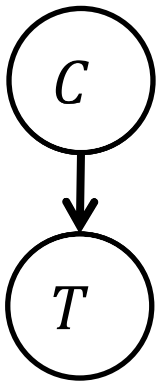
Estimating Parameters from Incomplete Data

- The parameter estimates considered so far have a number of interesting properties
  - They are unique, asymptotically Normal, and maximize the probability of data
  - They are easy to compute with a single pass on the dataset

- Given these properties, we could seek for maximum likelihood estimates for incomplete data as well
  - However, the properties of these estimates will depend on the nature of incompleteness

Incomplete Data: Example

- For example, consider the network structure on the right
  - C is a medical condition and T a test for detecting this condition
  - Let's also suppose the true parameters are given by the tables
  - Hence, we have $P(ve) = P(\overline{ve}) = .5$



C	θ_c
yes	.25
no	.75

C	T	$\theta_{t\|c}$
yes	ve	.80
yes	$\overline{ve}$	.20
no	ve	.40
no	$\overline{ve}$	.60

- Consider now the following incomplete datasets

$\mathcal{D}^1$	C	T
1	?	ve
2	?	ve
3	?	$\overline{ve}$
4	?	$\overline{ve}$
5	?	$\overline{ve}$
6	?	ve
7	?	ve
8	?	$\overline{ve}$

$\mathcal{D}^2$	C	T
1	yes	ve
2	yes	ve
3	yes	$\overline{ve}$
4	no	?
5	yes	$\overline{ve}$
6	yes	ve
7	no	?
8	no	$\overline{ve}$

$\mathcal{D}^3$	C	T
1	yes	ve
2	yes	ve
3	?	$\overline{ve}$
4	no	?
5	yes	$\overline{ve}$
6	?	ve
7	no	?
8	no	$\overline{ve}$

Incomplete Data: Example

- Let us consider the first dataset $\mathcal{D}^1$
  - The cases split equally between ve and $\overline{ve}$ values of T
  - We expect this to be true in the limit given the distribution of this data

- We can show the ML estimates are not unique for this dataset
  - The ML estimates for $\mathcal{D}^1$ are characterized by the following equation

$$\theta_{T=ve|C=yes}\,\theta_{C=yes} + \theta_{T=ve|C=no}\theta_{C=no} = \frac{1}{2}$$

- The true parameters satisfy this equation
  - But the following estimates do as well

$$\theta_{C=yes} = 1, \qquad \theta_{T=ve|C=yes} = \frac{1}{2}$$

  - With $\theta_{T=ve|C=no}$ taking any value

C	θ_c
yes	$.25$
no	$.75$

C	T	$\theta_{t\|c}$
yes	ve	$.80$
yes	$\overline{ve}$	$.20$
no	ve	$.40$
no	$\overline{ve}$	$.60$

$\mathcal{D}^1$	C	T
1	?	ve
2	?	ve
3	?	$\overline{ve}$
4	?	$\overline{ve}$
5	?	$\overline{ve}$
6	?	ve
7	?	ve
8	?	$\overline{ve}$

Incomplete Data: Example

- Therefore, ML estimates are not unique for this dataset
  - This is not surprising since incomplete datasets may not contain enough information to pin down the true parameters
  - The nonuniqueness of ML estimates is a desirable property

C	θ_c
yes	$.25$
no	$.75$

C	T	$\theta_{t\|c}$
yes	ve	$.80$
yes	$\overline{ve}$	$.20$
no	ve	$.40$
no	$\overline{ve}$	$.60$

$\mathcal{D}^1$	C	T
1	?	ve
2	?	ve
3	?	$\overline{ve}$
4	?	$\overline{ve}$
5	?	$\overline{ve}$
6	?	ve
7	?	ve
8	?	$\overline{ve}$

Incomplete Data: Example

- Consider now dataset $\mathcal{D}^2$ to illustrate why data may be missing:
  - People who do not suffer from the condition tend to not take the test. That is, the data is missing because the test is not performed
  - People who test negative tend not to report the result. That is, the test is performed but its value is not recorded

- These two scenarios are different in a fundamental way
  - In the second scenario, the missing value provides some evidence its true value must be negative
  - ML estimates give the intended results for the first scenario but not for the second one as it does not integrate all the information about the second scenario
  - However, we return to this topic later to show that ML can still be applied under the second scenario but requires some explication of the missing data mechanism

$\mathcal{D}^2$	C	T
1	yes	ve
2	yes	ve
3	yes	$\overline{ve}$
4	no	$?$
5	yes	$\overline{ve}$
6	yes	ve
7	no	$?$
8	no	$\overline{ve}$

Expectation Maximization (EM)

- Consider the Bayesian network on the right
  - Suppose our goal is to find ML estimates for the dataset $\mathcal{D}$
  - We start with initial estimates θ^0 with the following likelihood

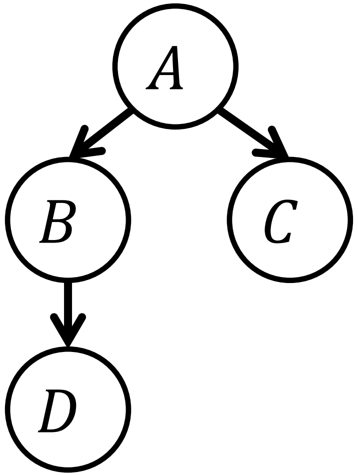
$$L(\theta;\mathcal{D}) = \prod_{i=1}^{5} P_{\theta^0}(\boldsymbol{d}_i)$$

$$= P_{\theta^0}(b,\bar{c}) P_{\theta^0}(b,\bar{d})\ P_{\theta^0}(\bar{b},c,d)\ P_{\theta^0}(\bar{b},c,d)\ P_{\theta^0}(b,\bar{d})$$

$$= (.135)(.184)(.144)(.144)(.184) = 9.5\times10^{-5}$$

- Evaluating the terms in this product generally requires inference on the Bayesian network
  - Contrary, the complete data case each term can be evaluated using the chain rule for the Bayesian network

A	B	$\theta^0_{b\|a}$
a	b	.75
a	$\bar{b}$	.25
$\bar{a}$	b	.10
$\bar{a}$	$\bar{b}$	.90



A	C	$\theta^0_{c\|a}$
a	c	.50
a	$\bar{c}$	.50
$\bar{a}$	c	.25
$\bar{a}$	$\bar{c}$	.75

A	θ^0_a
a	.20
$\bar{a}$	.80

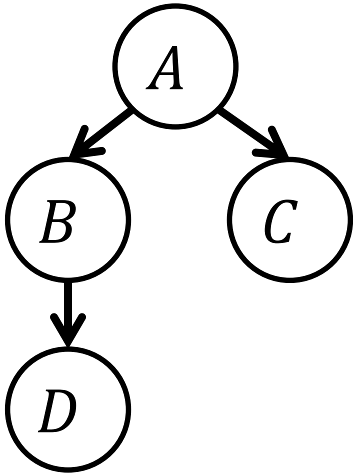
$\mathcal{D}$	A	B	C	D
1	?	b	$\bar{c}$	?
2	?	b	?	$\bar{d}$
3	?	$\bar{b}$	c	d
4	?	$\bar{b}$	c	d
5	?	b	?	$\bar{d}$

B	D	$\theta^0_{b\|d}$
b	d	.20
b	$\bar{d}$	.80
$\bar{b}$	d	.70
$\bar{b}$	$\bar{d}$	.30

Expectation Maximization (EM)

- The *expectation maximization* (EM) algorithm is based on the complete data method
  - EM first completes the dataset, inducing an empirical distribution
  - Then it estimates parameters using ML
  - The new set of parameters are guaranteed to have no less likelihood than the initial parameters
  - This process is repeated until some convergence condition is met
- For instance, the first case of dataset $\mathcal{D}$ has variables A and D with missing values
  - There are four possible completions for this case
  - Although we do not know which one is correct, we can compute the probability of each completion based on the initial set of parameters

A	B	$\theta^0_{b\|a}$
a	b	$.75$
a	$\bar{b}$	$.25$
$\bar{a}$	b	$.10$
$\bar{a}$	$\bar{b}$	$.90$



A	C	$\theta^0_{c\|a}$
a	c	$.50$
a	$\bar{c}$	$.50$
$\bar{a}$	c	$.25$
$\bar{a}$	$\bar{c}$	$.75$

A	θ^0_a
a	$.20$
$\bar{a}$	$.80$

$\mathcal{D}$	A	B	C	D
1	?	b	$\bar{c}$	?
2	?	b	?	$\bar{d}$
3	?	$\bar{b}$	c	d
4	?	$\bar{b}$	c	d
5	?	b	?	$\bar{d}$

B	D	$\theta^0_{b\|d}$
b	d	$.20$
b	$\bar{d}$	$.80$
$\bar{b}$	d	$.70$
$\bar{b}$	$\bar{d}$	$.30$

Expected Empirical Dist

- This tables lists for each case $\boldsymbol{d}_i$
  - The probability of each completion, $P_{\theta^0}(\boldsymbol{c}_i|\boldsymbol{d}_i)$
  - Where, $\boldsymbol{C}_i$ are the variables with missing values in $\boldsymbol{d}_i$
- The completed dataset defines an (expected) empirical distribution
  - The probability of an instantiation is computed considering all its occurrences in the completed dataset
  - However, instead of counting the number of occurrences, we add up the probabilities
- For instance, there are 3 occurrences of instantiation $a, b, \bar{c}, \bar{d}$ in cases $\boldsymbol{d}_1$, $\boldsymbol{d}_2$ and $\boldsymbol{d}_5$

$\mathcal{D}$	A	B	C	D	$P_{\theta^0}(\boldsymbol{C}_i\|\boldsymbol{d}_i)$
$\boldsymbol{d}_1$	?	b	$\bar{c}$	?	
	a	b	$\bar{c}$	d	$.111 = P_{\theta^0}(a, d\|b, \bar{c})$
	a	b	$\bar{c}$	$\bar{d}$	.444
	$\bar{a}$	b	$\bar{c}$	d	.089
	$\bar{a}$	b	$\bar{c}$	$\bar{d}$	.356
$\boldsymbol{d}_2$	?	b	?	$\bar{d}$	
	a	b	c	$\bar{d}$	$.326 = P_{\theta^0}(a, c\|b, \bar{d})$
	a	b	$\bar{c}$	$\bar{d}$	.326
	$\bar{a}$	b	c	$\bar{d}$	.087
	$\bar{a}$	b	$\bar{c}$	$\bar{d}$	.261
$\boldsymbol{d}_3$	?	$\bar{b}$	c	d	
	a	$\bar{b}$	c	d	$.122 = P_{\theta^0}(a\|\bar{b}, c, d)$
	$\bar{a}$	$\bar{b}$	c	d	.878
$\boldsymbol{d}_4$	?	$\bar{b}$	c	d	
	a	$\bar{b}$	c	d	$.122 = P_{\theta^0}(a\|\bar{b}, c, d)$
	$\bar{a}$	$\bar{b}$	c	d	.878
$\boldsymbol{d}_5$	?	b	?	$\bar{d}$	
	a	b	c	$\bar{d}$	$.326 = P_{\theta^0}(a, c\|b, \bar{d})$
	a	b	$\bar{c}$	$\bar{d}$	.326
	$\bar{a}$	b	c	$\bar{d}$	.087
	$\bar{a}$	b	$\bar{c}$	$\bar{d}$	.261

Expected Empirical Dist

- The probability, $P(a, b, \bar{c}, \bar{d})$, of seeing these completions is

$$\frac{P_{\theta^0}(a, \bar{d}|b, \bar{c}) + P_{\theta^0}(a, \bar{c}|b, \bar{d}) + P_{\theta^0}(a, \bar{c}|b, \bar{d})}{N}$$

$$= \frac{.444 + .326 + .326}{5} = .219$$

- We can define the *expected empirical distribution* of dataset $\mathcal{D}$ under parameters θ^k as

$$P_{\mathcal{D},\theta^k}(\alpha) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{\boldsymbol{d}_i, \boldsymbol{c}_i \vDash \alpha} P_{\theta^k}(\boldsymbol{c}_i|\boldsymbol{d}_i)$$

  - Where α is an event and $\boldsymbol{C}_i$ are the variables with missing values in case $\boldsymbol{d}_i$
  - $\boldsymbol{d}_i, \boldsymbol{c}_i \vDash \alpha$ means that event α is satisfied by complete case $\boldsymbol{d}_i, \boldsymbol{c}_i$

$\mathcal{D}$	A	B	C	D	$P_{\theta^0}(\boldsymbol{C}_i\|\boldsymbol{d}_i)$
$\boldsymbol{d}_1$	?	b	$\bar{c}$	?	
	a	b	$\bar{c}$	d	$.111 = P_{\theta^0}(a, d\|b, \bar{c})$
	a	b	$\bar{c}$	$\bar{d}$	.444
	$\bar{a}$	b	$\bar{c}$	d	.089
	$\bar{a}$	b	$\bar{c}$	$\bar{d}$	.356
$\boldsymbol{d}_2$	?	b	?	$\bar{d}$	
	a	b	c	$\bar{d}$	$.326 = P_{\theta^0}(a, c\|b, \bar{d})$
	a	b	$\bar{c}$	$\bar{d}$	.326
	$\bar{a}$	b	c	$\bar{d}$	.087
	$\bar{a}$	b	$\bar{c}$	$\bar{d}$	.261
$\boldsymbol{d}_3$	?	$\bar{b}$	c	d	
	a	$\bar{b}$	c	d	$.122 = P_{\theta^0}(a\|\bar{b}, c, d)$
	$\bar{a}$	$\bar{b}$	c	d	.878
$\boldsymbol{d}_4$	?	$\bar{b}$	c	d	
	a	$\bar{b}$	c	d	$.122 = P_{\theta^0}(a\|\bar{b}, c, d)$
	$\bar{a}$	$\bar{b}$	c	d	.878
$\boldsymbol{d}_5$	?	b	?	$\bar{d}$	
	a	b	c	$\bar{d}$	$.326 = P_{\theta^0}(a, c\|b, \bar{d})$
	a	b	$\bar{c}$	$\bar{d}$	.326
	$\bar{a}$	b	c	$\bar{d}$	.087
	$\bar{a}$	b	$\bar{c}$	$\bar{d}$	.261

Expected Empirical Distribution

- Given the definition of expected empirical distribution we can compute $P_{\mathcal{D},\theta^0}$ for all instantiations of variables A, B, C and D
- When the dataset is complete
  - $P_{\mathcal{D},\theta^k}(.)$ reduces to the empirical probability $P_{\mathcal{D}}(.)$, which is independent of parameter θ^k
  - Moreover, $NP_{\mathcal{D},\theta^k}(\boldsymbol{x})$ is called *expected count* of instantiation $\boldsymbol{x}$
- We can use the expected empirical distribution to estimate parameters
  - Similarly we did for the complete data
  - For instance, for the parameter $\theta_{c|\bar{a}}$

$$\theta^1_{c|\bar{a}} = P_{\mathcal{D},\theta^0}(c|\bar{a}) = \frac{P_{\mathcal{D},\theta^0}(c,\bar{a})}{P_{\mathcal{D},\theta^0}(\bar{a})} \approx .666$$

A	B	C	D	$P_{\mathcal{D},\theta^0}(.)$
a	b	c	d	0
a	b	c	$\bar{d}$	.130
a	b	$\bar{c}$	d	.022
a	b	$\bar{c}$	$\bar{d}$	.219
a	$\bar{b}$	c	d	.049
a	$\bar{b}$	c	$\bar{d}$	0
a	$\bar{b}$	$\bar{c}$	d	0
a	$\bar{b}$	$\bar{c}$	$\bar{d}$	0
$\bar{a}$	b	c	d	0
$\bar{a}$	b	c	$\bar{d}$	.035
$\bar{a}$	b	$\bar{c}$	d	.018
$\bar{a}$	b	$\bar{c}$	$\bar{d}$	.176
$\bar{a}$	$\bar{b}$	c	d	.351
$\bar{a}$	$\bar{b}$	c	$\bar{d}$	0
$\bar{a}$	$\bar{b}$	$\bar{c}$	d	0
$\bar{a}$	$\bar{b}$	$\bar{c}$	$\bar{d}$	0

Expectation Maximization (EM)

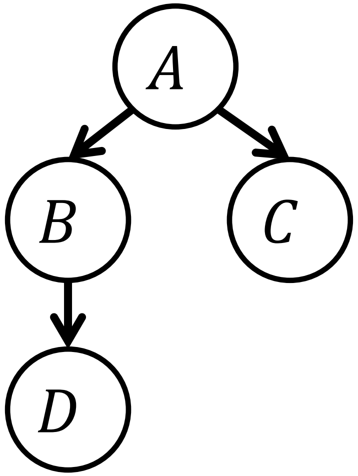
- The figure on the right shows all parameter estimates based on $P_{\mathcal{D},\theta^0}$ leading to new estimates θ^1
- The new estimates θ^1 have the following likelihood for dataset $\mathcal{D}$

$$L(\theta^1;\mathcal{D}) = \prod_{i=1}^{5} P_{\theta^1}(\boldsymbol{d}_i)$$
$$= (.290)(.560)(.255)(.255)(.560)$$
$$= 5.9\times 10^{-3} > L(\theta^0|\mathcal{D})$$

- Therefore, we can define the EM estimates for a dataset $\mathcal{D}$ and parameters θ^k as

$$\theta^{k+1}_{x|\boldsymbol{u}} \stackrel{\text{def}}{=} \mathrm{P}_{\mathcal{D},\theta^k}(x|\boldsymbol{u})$$

A	B	$\theta^1_{b\|a}$
a	b	$.883$
a	$\bar{b}$	$.117$
$\bar{a}$	b	$.395$
$\bar{a}$	$\bar{b}$	$.605$



A	C	$\theta^1_{c\|a}$
a	c	$.426$
a	$\bar{c}$	$.574$
$\bar{a}$	c	$.666$
$\bar{a}$	$\bar{c}$	$.334$

A	θ^1_{a}
a	$.420$
$\bar{a}$	$.580$

$\mathcal{D}$	A	B	C	D
1	?	b	$\bar{c}$	?
2	?	b	?	$\bar{d}$
3	?	$\bar{b}$	c	d
4	?	$\bar{b}$	c	d
5	?	b	?	$\bar{d}$

B	D	$\theta^1_{b\|d}$
b	d	$.067$
b	$\bar{d}$	$.933$
$\bar{b}$	d	1
$\bar{b}$	$\bar{d}$	0

Expectation Maximization (EM)

- EM estimates can be computed without constructing the expected empirical distribution
  - The expected empirical distribution of dataset $\mathcal{D}$ given parameters θ^k can be computed as
  - That is, we simply iterate over the dataset cases computing the probability of α for each case
  - The EM estimates can now be computes as

$$P_{\mathcal{D},\theta^k}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} P_{\theta^k}(\alpha|\boldsymbol{d}_i)$$

$$\theta_{x|\boldsymbol{u}}^{k+1} = \frac{\sum_{i=1}^{N} P_{\theta^k}(x\boldsymbol{u}|\boldsymbol{d}_i)}{\sum_{i=1}^{N} P_{\theta^k}(\boldsymbol{u}|\boldsymbol{d}_i)}$$

- This equation computes EM estimates performing inference in a Bayesian network parametrizes by θ^k. For example

$$\theta_{c|\bar{a}}^{1} = \frac{\sum_{i=1}^{5} P_{\theta^0}(c,\bar{a}|\boldsymbol{d}_i)}{\sum_{i=1}^{N} P_{\theta^0}(\bar{a}|\boldsymbol{d}_i)} = \frac{0+.087+.878+.878+.087}{.444+.348+.878+.878+.348} = .666$$

$\mathcal{D}$	A	B	C	D
1	?	b	$\bar{c}$	?
2	?	b	?	$\bar{d}$
3	?	$\bar{b}$	c	d
4	?	$\bar{b}$	c	d
5	?	b	?	$\bar{d}$

EM: Algorithm

$k \leftarrow 0$

$\theta^k \leftarrow$ initial parameter values

while convergence criterion is not met **do**

  $c_{x\boldsymbol{u}} \leftarrow 0$ for each family instantiation $x\boldsymbol{u}$

  for $i \leftarrow 1$ to N **do**

    for each family instantiation $x\boldsymbol{u}$ **do**

      $c_{x\boldsymbol{u}} \leftarrow c_{x\boldsymbol{u}} + P_{\theta^k}(x\boldsymbol{u}|\boldsymbol{d}_i)$ # requires inference on network (G, θ^k)

  $\theta^{k+1}_{x|\boldsymbol{u}} \leftarrow c_{x\boldsymbol{u}} / \sum_{x^*} c_{x^*\boldsymbol{u}}$

  $k \leftarrow k + 1$

return θ^k

Note:

- The stop criterion usually employed is a small difference between θ^k and θ^{k+1} or a small change in log-likelihood

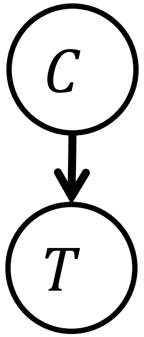
EM Algorithm: Observations

- There are a few observations about the behaviour of the EM algorithm
  - The algorithm may converge to different parameters depending on the initial estimate θ^0
  - It is common to run the algorithm multiple times, starting with different estimates in each iteration
  - In this case, we return the best estimates across all iterations

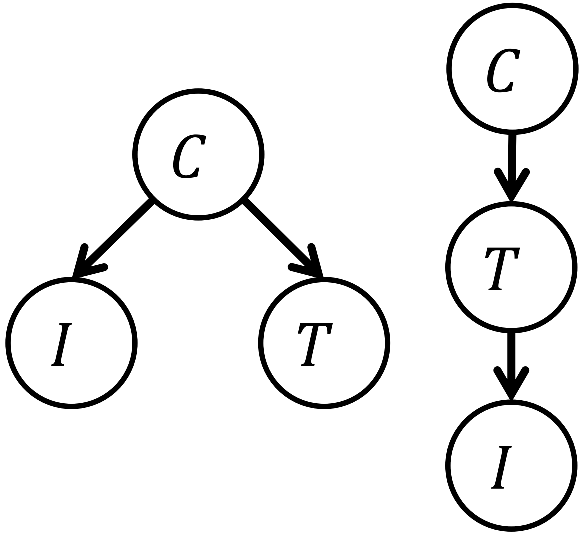
- Each iteration of the EM algorithm will have to perform inference on a Bayesian network
  - In each iteration, the algorithm computes the probability of each instantiation $x\boldsymbol{u}$ given each case $\boldsymbol{d}_i$ as evidence
  - These computations correspond to posterior marginals over network families
  - Therefore, we can use an algorithm such as the jointree that efficiently computes family marginals

Missing Data Mechanism

- Let us consider again the network where C represents a medical condition and T a test for detecting this condition
  - We depict two extended network structures for this problem
  - Each includes an additional variable I that indicates whether the test result is missing in the dataset

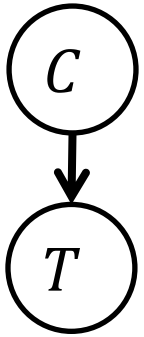


- In the left network, the missing data depends on the condition
  - E.g., people who do not suffer from the condition tend not to take the test
- In the right network, the missing data depends on the test result
  - E.g., individuals who test negative tend not to report the result
- Hence, these networks structures explicate different dependencies between missing data missingness
  - We say the structures explicate different *missing data mechanisms*

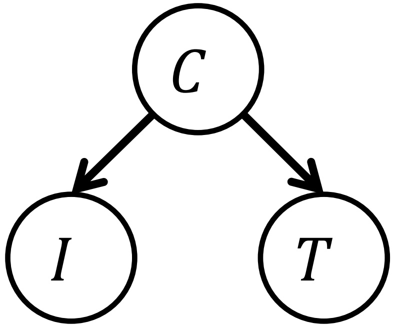


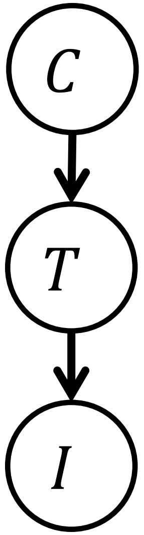
Missing Data Indicator

- Our goal is to discuss ML estimates that we would obtain with respect to structures that explicate missing data mechanisms
  - And compare these estimates with those obtained when ignoring such mechanisms
  - E.g., when we use the simpler structure on the top



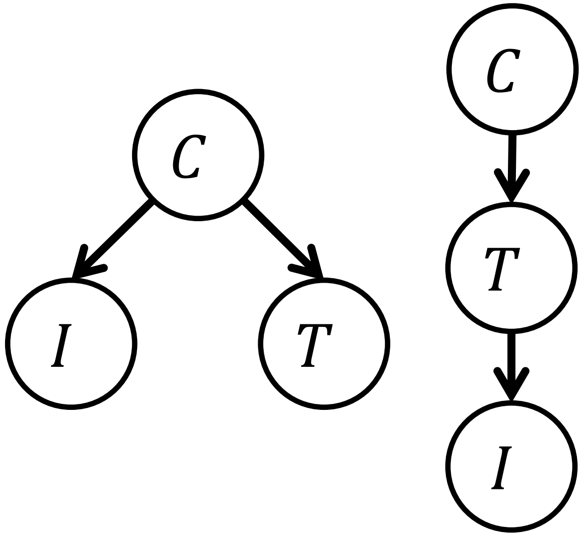
- Let $\boldsymbol{M}$ be the variables of a network G that have missing values in the data set
  - We define $\boldsymbol{I}$ as a set of variables called *missing data indicators* that are in one-to-one correspondence with variables $\boldsymbol{M}$
  - A network structure that results from adding variables $\boldsymbol{I}$ as leaf nodes to G is said to explicate the *missing data mechanism* and is denoted by G_I





Missing Data Indicator

- In these figures, variable I is the *missing data indicator*
  - It corresponds to variable T
  - I is always observed, as its value is determined by whether the value of T is missing
  - We use D_I to denote an extension of the dataset D that includes missing data indicators

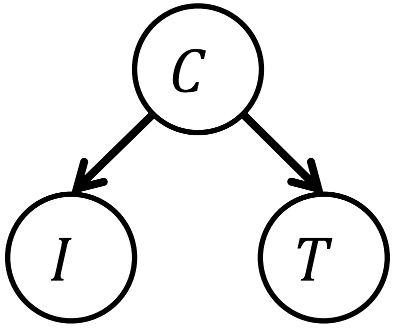


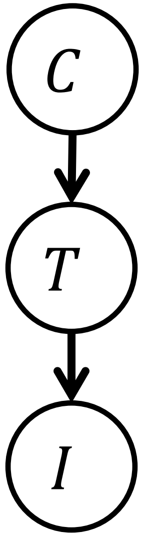
- We can apply the ML approach in three different ways
  - To the original structure $C \rightarrow T$ and the original dataset $\mathcal{D}$
  - To the extended structure on the left and dataset $\mathcal{D}_I$
  - To the extended structure on the right and dataset $\mathcal{D}_I$

$\mathcal{D}_I$	C	T	I
1	yes	ve	no
2	yes	ve	no
3	yes	$\overline{ve}$	no
4	no	?	yes
5	yes	$\overline{ve}$	no
6	yes	ve	no
7	no	?	yes
8	no	$\overline{ve}$	no

Missing Data Indicator

- We are ignoring the missing data mechanism in the first case and accounting for it in the remaining ones
  - All three approaches yield estimates for C and T
  - The question is whether ignoring the missing data mechanism will change the ML estimates
- It turns out the first and second approaches yield identical estimates
  - These estimates are different from the second approach
  - This suggests that missing data mechanism can be ignored in the second case but not in the third one

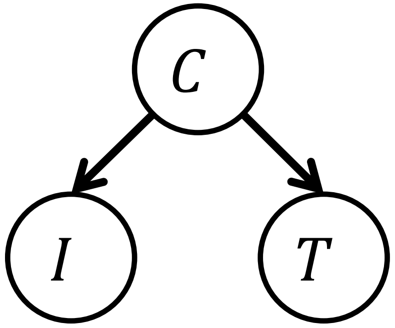


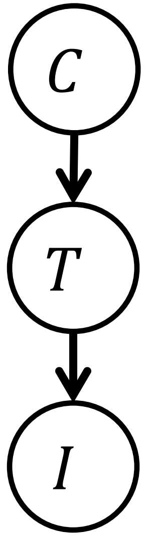


$\mathcal{D}_I$	C	T	I
1	yes	ve	no
2	yes	ve	no
3	yes	$\overline{ve}$	no
4	no	$?$	yes
5	yes	$\overline{ve}$	no
6	yes	ve	no
7	no	$?$	yes
8	no	$\overline{ve}$	no

Missing at Random (MAR)

- Let G_I be a network structure that explicates the missing data mechanism of structure G and data set $\mathcal{D}$
  - Let $\boldsymbol{O}$ be variables that are always observed in data set D
  - Let $\boldsymbol{M}$ be the variables that have missing values in the data set
  - We say that G_I satisfies the *missing at random (MAR)* assumption if $\boldsymbol{I}$ and $\boldsymbol{M}$ are d-separated by $\boldsymbol{O}$ in structure G_I





- Intuitively, G_I satisfies MAR assumption if once we know the values of variables $\boldsymbol{O}$, the specific values of $\boldsymbol{M}$ become irrelevant to whether these values are missing in the dataset
  - For the left network, once we know the condition, the test value becomes irrelevant to whether the test is missing
  - For the right network, even if we know the condition, the test result may still be relevant to whether it will be missing

$\mathcal{D}_I$	C	T	I
1	yes	ve	no
2	yes	ve	no
3	yes	$\overline{ve}$	no
4	no	$?$	yes
5	yes	$\overline{ve}$	no
6	yes	ve	no
7	no	$?$	yes
8	no	$\overline{ve}$	no

Missing at Random (MAR)

- If the MAR assumption holds, the missing data mechanism can be ignored
  - Under MAR assumption we obtain the same ML estimates θ if we include or ignore the missing data mechanism

$$argmax_{\theta} LL(\theta; \mathcal{D}) = argmax_{\theta} \max_{\theta_{\boldsymbol{I}}} LL(\theta, \theta_{\boldsymbol{I}}; \mathcal{D}_{\boldsymbol{I}})$$

Conclusion

- In this lecture, we discussed approaches based on Maximum Likelihood for parameter estimation
  - When the dataset is complete, the problem is easy
    - We can estimate the parameters using the empirical distribution
    - The algorithm is simple and efficient. We can compute all parameters with a single pass over the data
  - When the dataset is incomplete, the problem involves inference in the Bayesian network
    - A common approach is to use Expectation Maximization
    - This approach estimates the parameter using an expected empirical distribution
    - The algorithm is more intricate. It requires inference over the Bayesian network since we need to compute condition probabilities $P(\boldsymbol{C}_i|\boldsymbol{d}_i)$ for missing variables