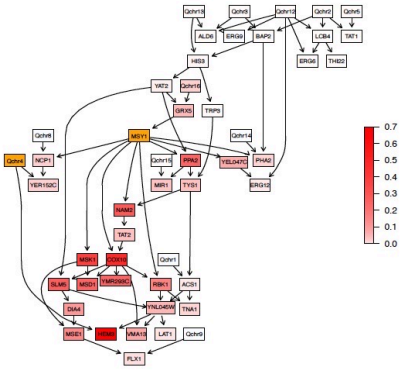
Introduction to Bayesian Networks in R!



Rachael Hageman Blair
March 28, 2023

1

Workshop Objectives

- Motivation for Bayesian Networks (BNs)
- Graphs and basic properties
- Exposure to three tasks for working with BNs:
  1. Parameter Learning
  2. Structural Learning
  3. Inference and Reasoning

Three sections – broken up with computation!

2

Related Summer Offerings

CDA 500: Bayesian Networks in R (Dr. Rachael Hageman Blair
This course gives an overview of Bayesian Networks with application in R. The focus will be Bayesian network modeling, from structural learning to parameter learning and inference. Classic discrete, Gaussian, and conditional Gaussian networks will be described. Applications will showcase the wealth of R packages dedicated to learning and inference.
(Monday, Wednesday, and Friday 1:00pm to 3:30pm • 5/30/2023 TO 6/09/2023)

CDA 500: Introduction to Network Analysis (Dr. Sarah Muldoon)
Networks are everywhere. From the internet to our social interactions and even in our brain, it has become clear that we are surrounded by complex systems of interconnected elements. In this course, we will explore how one can better understand the world around us by thinking about systems as networks and examining the specific structure of how network elements are connected. What do social networks and brain networks have in common? Why is that important? How would we design a power grid so that it's less likely to fail? We will initially focus on learning and applying network statistics that can measure and quantify static network structure and then introduce temporal and multilayer frameworks that can capture richer features of systems. An undergraduate level understanding of linear algebra will be expected.
(tba)

https://www.buffalo.edu/ai-data-science/academics/grad/summer-courses.html#title_0_copy

3

Getting Started

Install Jupyter Notebook with R:

https://jupyter.org/install
https://towardsdatascience.com/how-to-run-r-scripts-in-jupyter-15527148d2a

Alternatively, install R:

https://www.r-project.org/

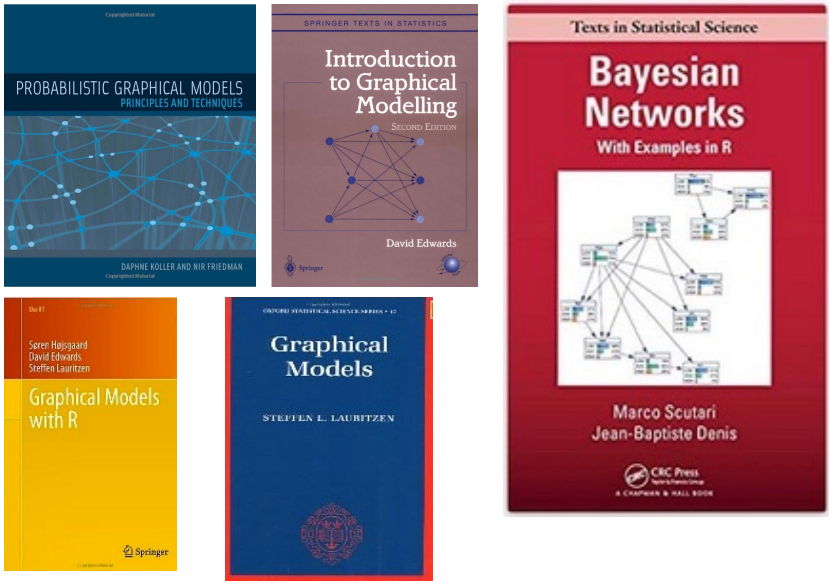
Or RStudio:

https://posit.co/download/rstudio-desktop/

Packages we will be using (>install.packages("bnlearn")):
bnlearn,

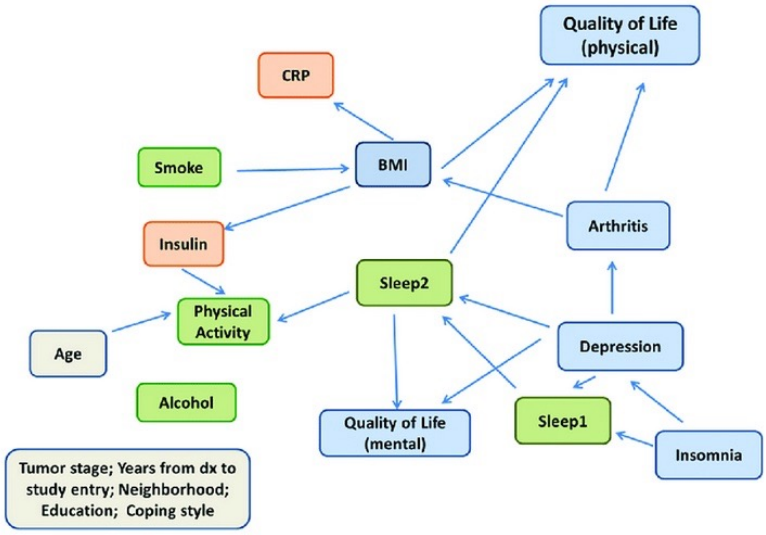
4

Motivation



5

Motivation



Bayesian network of total physical activity, other lifestyle factors and health in the Reach for Health cohort. Lifestyle factors (green), biomarkers (orange), and physical and mental health outcomes (blue). https://doi.org /10.1371/journal.pone.0202923.g001

https://doi.org/10.1371/journal.pone.0202923.g001

6

Motivation

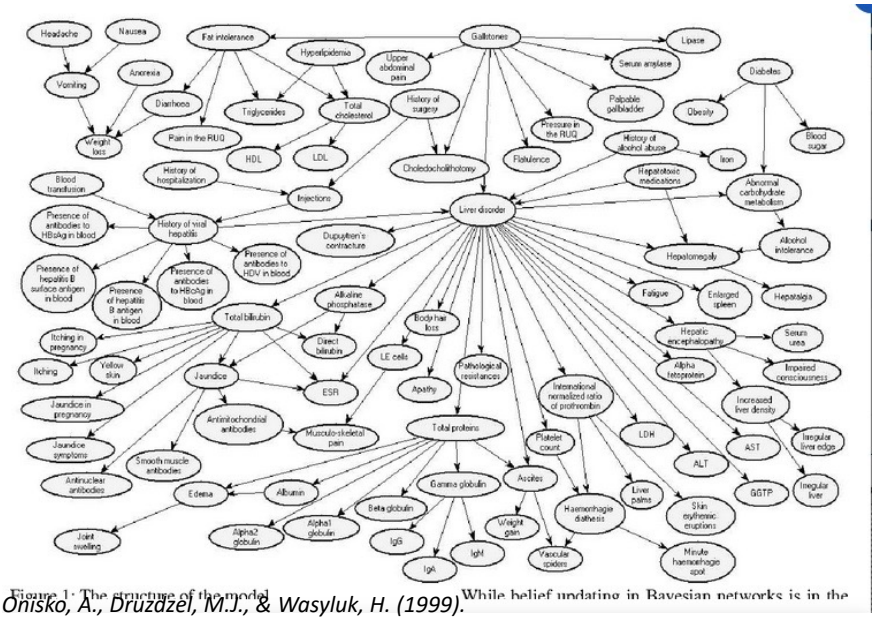
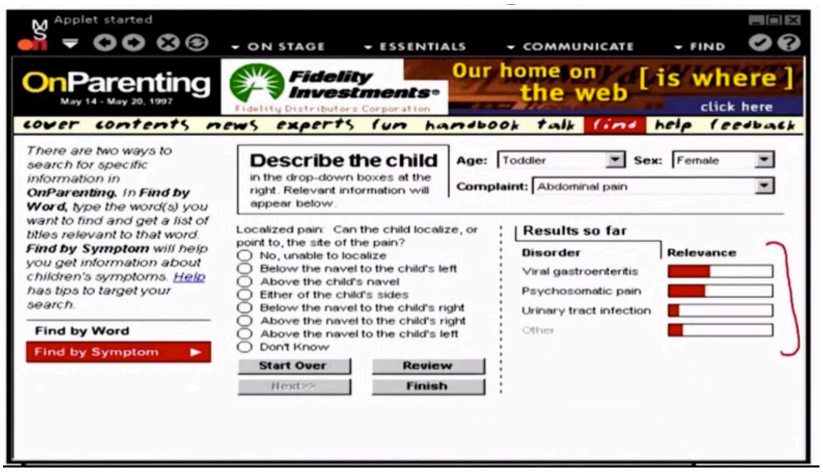


Figure 1: The structure of the model. While belief updating in Bayesian networks is in the

Oniśko, A., Druzdzel, M.J., & Wasyluk, H. (1999). A Bayesian Network Model for Diagnosis of Liver Disorders.

7

Motivation



8

Motivation

Uncertainty:

- Partial knowledge of the state of the world.
- Partial data available.
- Phenomena not covered by the model.
- Noisy observations.
- Modeling limitations of complicated systems.

Probabilistic Graphical Models:

- Representation in terms of random variables and distributions.
- Powerful reasoning patterns.

9

Section 1: Graphs and Basic Properties

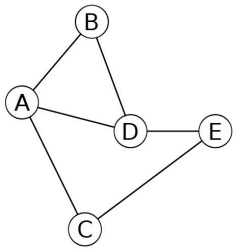
The first component of a BN is a graph. A graph $\mathcal{G}$ is a mathematical object with:

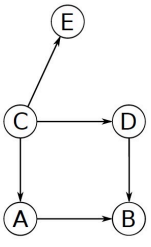
- a set of nodes $\mathbf{V} = \{v_1, \ldots, v_N\}$;
- a set of arcs A which are identified by pairs for nodes in $\mathbf{V}$, e.g. $a_{ij} = (v_i, v_j)$.

Given $\mathbf{V}$, a graph is uniquely identified by A. The arcs in A can be:

- undirected if (v_i, v_j) is an unordered pair and the arc $v_i - v_j$ has no direction;
- directed if $(v_i, v_j) \neq (v_j, v_i)$ is an ordered pair and the arc has a specific direction $v_i \rightarrow v_j$.

The assumption is that there is at most one arc between a pair of nodes.



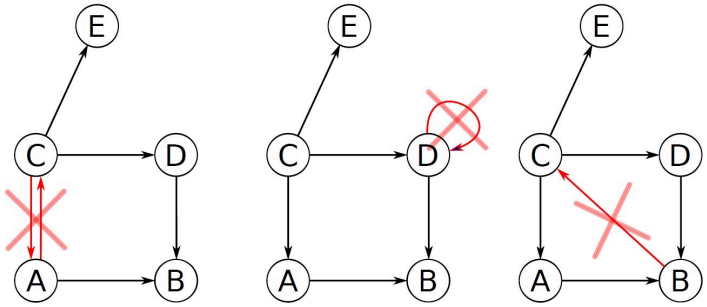


10

Section 1: Graphs and Basic Properties

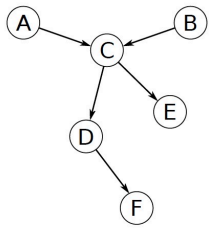
BNs use a specific kind of graph called a directed acyclic graph, that:

- contains only directed arcs;
- does not contain any loop (e.g. an arc $v_i \to v_i$ from a node to itself);
- does not contain any cycle (e.g. a sequence of arcs $v_i \to v_j \to \ldots \to v_k \to v_i$ that starts and ends in the same node).



11

Section 1: Graphs and Basic Properties



DAG

Graphical separation

$$
\begin{aligned}
&A \perp\!\!\!\perp_G B \\
&A \perp\!\!\!\perp_G D \mid C \\
&B \perp\!\!\!\perp_G D \mid C \\
&A \perp\!\!\!\perp_G E \mid C \\
&B \perp\!\!\!\perp_G E \mid C \\
&D \perp\!\!\!\perp_G E \mid C \\
&C \perp\!\!\!\perp_G F \mid D \\
&\ldots
\end{aligned}
$$

Probabilistic independence

$$
\begin{aligned}
&A \perp\!\!\!\perp_P B \\
&A \perp\!\!\!\perp_P D \mid C \\
&B \perp\!\!\!\perp_P D \mid C \\
&A \perp\!\!\!\perp_P E \mid C \\
&B \perp\!\!\!\perp_P E \mid C \\
&D \perp\!\!\!\perp_P E \mid C \\
&C \perp\!\!\!\perp_P F \mid D \\
&\ldots
\end{aligned}
$$

Formally, the DAG is an independence map of the probability distribution of $\mathbf{X}$, with graphical separation ($\perp\!\!\!\perp_G$) implying probabilistic independence ($\perp\!\!\!\perp_P$).

12

Section 1: Graphs and Basic Properties

A trail $X_1 - X_2 - \ldots - X_k$ is **active** if it has no v-structures:

$$X_{i-1} \rightarrow X_i \leftarrow X_{i+1}.$$

General edge type up/down

A block in the trail

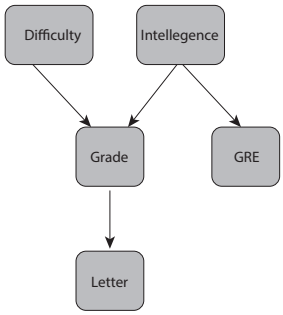
Therefore, information can flow freely though the network, in the "active" sense, unless a v-structure arises.

13

Section 1: Graphs and Basic Properties

Three types of reasoning:

1. Causal
2. Evidential
3. Inter-causal



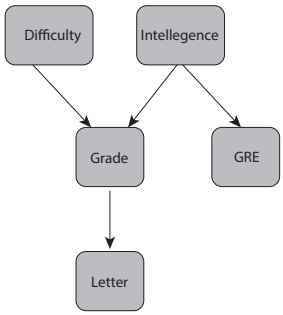
What happens when we 'absorb' new evidence

14

Section 1: Graphs and Basic Properties

Three types of reasoning:

1. Causal
2. Evidential
3. Inter-causal



***can information flow between X and Y, if W is given.*

$X \rightarrow Y$

$X \leftarrow Y$

$X \rightarrow W \rightarrow Y$

$X \leftarrow W \leftarrow Y$

$X \leftarrow W \rightarrow Y$

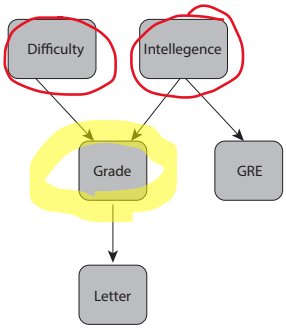
$X \rightarrow W \leftarrow Y$

15

Section 1: Graphs and Basic Properties

Three types of reasoning:

1. Causal
2. Evidential
3. Inter-causal



***can information flow between X and Y, if W is given.*

$X \rightarrow Y$

$X \leftarrow Y$

$X \rightarrow W \rightarrow Y$

$X \leftarrow W \leftarrow Y$

$X \leftarrow W \rightarrow Y$

$X \rightarrow W \leftarrow Y$

16

Section 1: Graphs and Basic Properties

If we use d-separation as our definition of graphical separation, assuming that the DAG is an I-map leads to the general formulation of the decomposition of the global distribution $\mathrm{P}(\mathbf{X})$:

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid \Pi_{X_i}) \tag{1}$$

into the local distributions for the X_i given their parents Π_{X_i}. If X_i has two or more parents it depends on their joint distribution, because each pair of parents forms a convergent connection centred on X_i and we cannot establish their independence. This decomposition is preferable to that obtained from the chain rule,

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid X_{i+1}, \ldots, X_N) \tag{2}$$

because the conditioning sets are typically smaller.

17

Computing Part 1

Computer Time!

Expert defined Bayesian Network:
Structure input
Graph and properties output

Packages: bnlearn, Rgrahviz

18

Computing Part 1

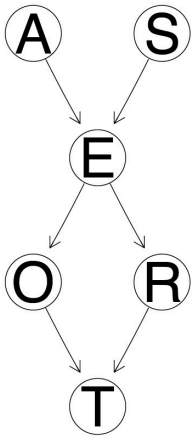
Consider a simple, hypothetical survey whose aim is to investigate the usage patterns of different means of transport, with a focus on cars and trains.

- Age (A): *young* for individuals below 30 years old, *adult* for individuals between 30 and 60 years old, and *old* for people older than 60.
- Sex (S): *male* or *female*.
- Education (E): *up to high school* or *university degree*.
- Occupation (O): *employee* or *self-employed*.
- Residence (R): the size of the city the individual lives in, recorded as either *small* or *big*.
- Travel (T): the means of transport favoured by the individual, recorded either as *car*, *train* or *other*.

The nature of the variables recorded in the survey suggests how they may be related with each other.

19

Computing Part 1



That is a prognostic view of the survey as a BN:

1. the blocks in the experimental design on top (e.g. stuff from the registry office);
2. the variables of interest in the middle (e.g. socio-economic indicators);
3. the object of the survey at the bottom (e.g. means of transport).

Variables that can be thought as "causes" are on above variables that can be considered their "effect", and confounders are on above everything else.

20

Computing Part 1

21

Section 2: Local Distributions

If we use d-separation as our definition of graphical separation, assuming that the DAG is an I-map leads to the general formulation of the decomposition of the global distribution $\mathrm{P}(\mathbf{X})$:

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid \Pi_{X_i}) \tag{1}$$

into the local distributions for the X_i given their parents Π_{X_i}. If X_i has two or more parents it depends on their joint distribution, because each pair of parents forms a convergent connection centred on X_i and we cannot establish their independence. This decomposition is preferable to that obtained from the chain rule,

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid X_{i+1}, \ldots, X_N) \tag{2}$$

because the conditioning sets are typically smaller.

22

Section 2: Local Distributions

Another result along the same lines is called the local Markov property, which can be combined with the chain rule above to get the decomposition into local distributions.

> *Each node X_i is conditionally independent of its non-descendants (e.g., nodes X_j for which there is no path from X_i to X_j) given its parents.*

Compared to the previous decomposition, it highlights the fact that parents are not completely independent from their children in the BN; a trivial application of Bayes' theorem to invert the direction of the conditioning shows how information on a child can change the distribution of the parent.

23

Section 2: Local Distributions

If we use d-separation as our definition of graphical separation, assuming that the DAG is an I-map leads to the general formulation of the decomposition of the global distribution $\mathrm{P}(\mathbf{X})$:

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid \Pi_{X_i}) \tag{1}$$

into the local distributions for the X_i given their parents Π_{X_i}. If X_i has two or more parents it depends on their joint distribution, because each pair of parents forms a convergent connection centred on X_i and we cannot establish their independence. This decomposition is preferable to that obtained from the chain rule,

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid X_{i+1}, \ldots, X_N) \tag{2}$$

because the conditioning sets are typically smaller.

24

Section 2: Local Distributions

The second component of a BN is the probability distribution $\mathrm{P}(\mathbf{X})$. The choice should such that the BN:

- can be learned efficiently from data;
- is flexible (distributional assumptions should not be too strict);
- is easy to query to perform inference.

The three most common choices in the literature (by far), are:

- discrete BNs (DBNs), in which $\mathbf{X}$ and the $X_i \mid \Pi_{X_i}$ are multinomial;
- Gaussian BNs (GBNs), in which $\mathbf{X}$ is multivariate normal and the $X_i \mid \Pi_{X_i}$ are univariate normal;
- conditional linear Gaussian BNs (CLGBNs), in which $\mathbf{X}$ is a mixture of multivariate normals and the $X_i \mid \Pi_{X_i}$ are either multinomial, univariate normal or mixtures of normals.

It has been proved in the literature that exact inference is possible in these three cases, hence their popularity.

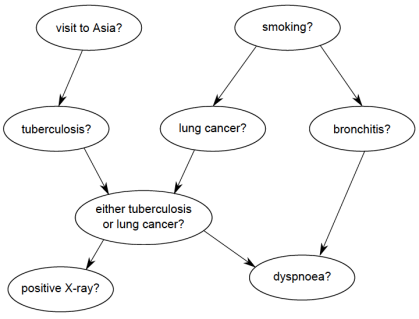
25

Section 2: Local Distributions

Definitions

UNIVERSITY OF OXFORD

Discrete BNs



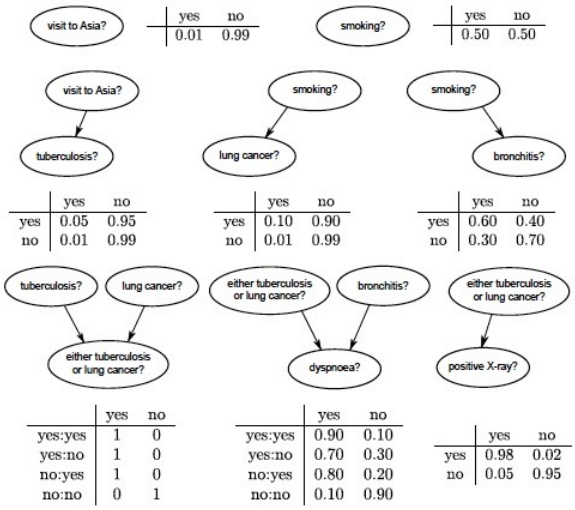
A classic example of DBN is the ASIA network from Lauritzen & Spiegelhalter (1988), which includes a collection of binary variables. It describes a simple diagnostic problem for tuberculosis and lung cancer.

Total parameters of $\mathbf{X}$: $2^8 - 1 = 255$

Marco Scutari Probabilistic Graphical Models IV University of Oxford

26

Section 2: Local Distributions



The local distributions $X_i \mid \Pi_{X_i}$ take the form of conditional probability tables for each node given all the configurations of the values of its parents.

Overall parameters of the $X_i \mid \Pi_{X_i}$: 18

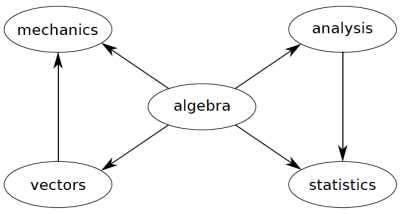
27

Section 2: Local Distributions

Definitions

UNIVERSITY OF OXFORD

Gaussian BNs



A classic example of GBN is the MARKS networks from Mardia, Kent & Bibby (1979), which describes the relationships between the marks on 5 math-related topics.

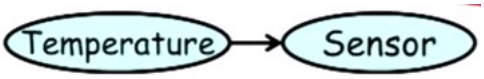
Assuming $X \sim N(\boldsymbol{\mu}, \Sigma)$, we can compute $\Omega = \Sigma^{-1}$. Then $\Omega_{ij} = 0$ implies $X_i \perp\!\!\!\perp_P X_j \mid \mathbf{X} \setminus \{X_i, X_j\}$. The absence of an arc $X_i \to X_j$ in the DAG implies $X_i \perp\!\!\!\perp_G X_j \mid \mathbf{X} \setminus \{X_i, X_j\}$, which in turn implies $X_i \perp\!\!\!\perp_P X_j \mid \mathbf{X} \setminus \{X_i, X_j\}$.

Total parameters of $\mathbf{X}$: $5 + 15 = 20$

Marco Scutari Probabilistic Graphical Models IV University of Oxford

28

Section 2: Local Distributions



$$S \sim N\left(T, \sigma_S^2\right)$$

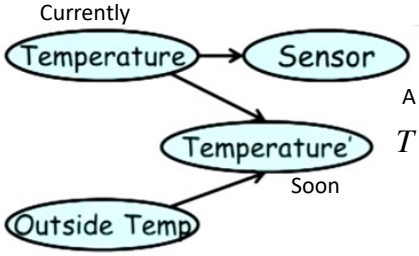
We have defined a probability around every value of T, which depends very compactly on parameters of the normal distribution.

Linear Gaussian!

29

Section 2: Local Distributions

Currently

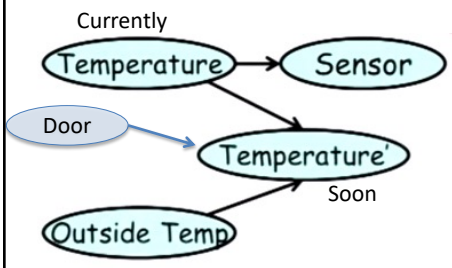


A possible Model:

$$T \sim N\left(\alpha T + (1-\alpha)O;\ \sigma_T^2\right)$$

30

Section 2: Local Distributions



Two possible Models:

$$T' \sim N\left(\alpha_0 T + (1-\alpha_0)O;\ \sigma^2_{0T}\right),\ \text{if}\ D=0$$

$$T' \sim N\left(\alpha_1 T + (1-\alpha_1)O;\ \sigma^2_{1T}\right),\ \text{if}\ D=1$$

Conditional Gaussian!

31

Computing Part 2

32

Section 3: Probabilistic Queries

Probabilistic reasoning on BNs works in the framework of Bayesian statistics and focuses on the computation of posterior probabilities or densities.
For example, suppose we have learned a BN $\mathcal{B}$ with DAG G and parameters Θ. We want to use $\mathcal{B}$ to investigate the effects of a new piece of evidence $\mathbf{E}$ using the knowledge encoded in $\mathcal{B}$, that is, to investigate the posterior distribution

$$\mathrm{P}(\mathbf{X} \mid \mathbf{E}, \mathcal{B}) = \mathrm{P}(\mathbf{X} \mid \mathbf{E}, G, \Theta).$$

Questions that can be asked are called queries and are typically an event of interest. The two most common queries are conditional probability (CPQ) and maximum a posteriori (MAP) queries, also known as most probable explanation (MPE) queries.

33

Section 3: Probabilistic Queries

- Hard evidence: an instantiation of one or more variables in the BN. In other words,

$$\mathbf{E} = \{X_{i_1} = e_1, X_{i_2} = e_2, \ldots, X_{i_k} = e_k\},$$

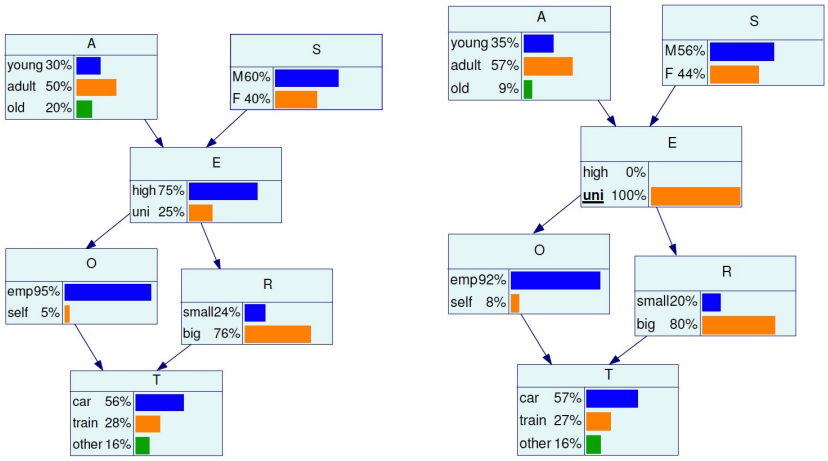
  which ranges from the value of a single variable X_i to a complete specification for $\mathbf{X}$ (such a new partial or complete observation).
- Soft evidence: a new distribution for one or more variables in the network. Since both the network structure and the distributional assumptions are treated as fixed, soft evidence is usually specified as a new set of parameters,

$$\mathbf{E} = \left\{X_{i_1} \sim (\Theta_{X_{i_1}}), X_{i_2} \sim (\Theta_{X_{i_2}}), \ldots, X_{i_k} \sim (\Theta_{X_{i_k}})\right\}.$$

  This new distribution may be, for instance, the null distribution in a hypothesis testing problem.

34

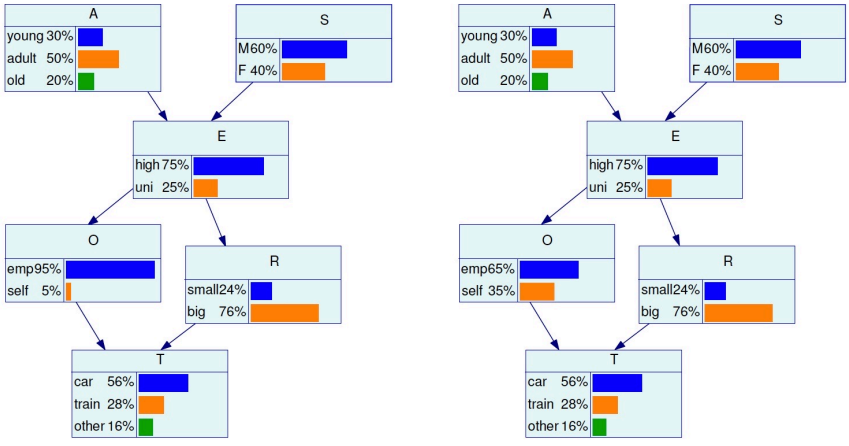
Section 3: Probabilistic Queries



The original survey BN (left), and the posterior BN with hard evidence on Education (right).

35

Section 3: Probabilistic Queries



The original survey BN (left), and the posterior BN with soft evidence on Employment (right).

36

Section 3: Probabilistic Queries

Conditional probability queries are concerned with

$$\mathrm{CPQ}(\mathbf{Q} \mid \mathbf{E}, \mathcal{B}) = \mathrm{P}(\mathbf{Q} \mid \mathbf{E}, G, \Theta) = \mathrm{P}(X_{j_1}, \ldots, X_{j_l} \mid \mathbf{E}, G, \Theta),$$

for some query variables $\mathbf{Q}$ given some hard evidence $\mathbf{E}$ on other variables, that is, the marginal posterior probability distribution of $\mathbf{Q}$,

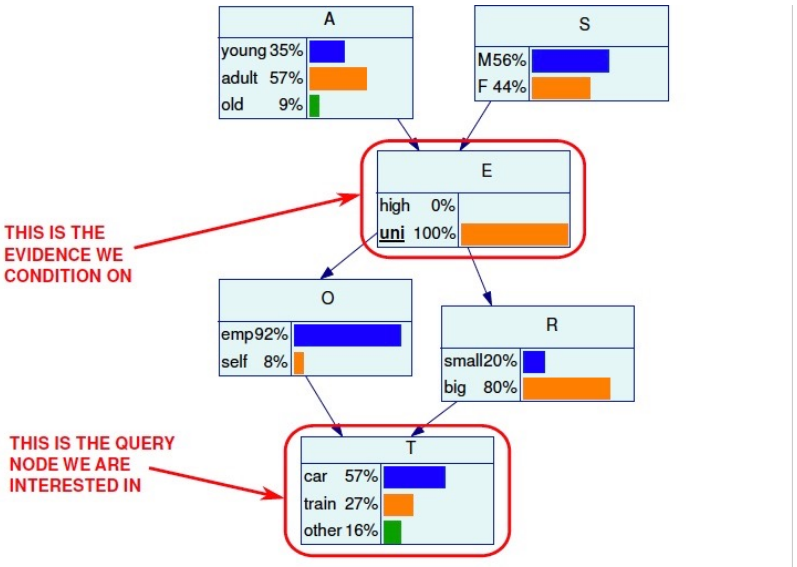
$$\mathrm{P}(\mathbf{Q} \mid \mathbf{E}, G, \Theta) = \int \mathrm{P}(\mathbf{X} \mid \mathbf{E}, G, \Theta)\, d(\mathbf{X} \setminus \mathbf{Q}).$$

This class of queries has many useful applications due to their versatility.

For instance, we can assess the odds of an unfavourable outcome $\mathbf{Q}$ can for different sets of hard evidence $\mathbf{E}_1$, $\mathbf{E}_2$, ..., $\mathbf{E}_m$ of one or more related variables.

37

Section 3: Probabilistic Queries



38

Section 3: Probabilistic Queries

Maximum a posteriori queries are concerned with finding the configuration $\mathbf{q}^*$ of $\mathbf{Q}$ that has the highest posterior probability (for discrete BNs) or the maximum posterior density (for GBNs and CLGBNs),

$$\text{MAP}(\mathbf{Q} \mid \mathbf{E}, \mathcal{B}) = \mathbf{q}^* = \underset{\mathbf{q}}{\text{argmax}}\, \text{P}(\mathbf{Q} = \mathbf{q} \mid \mathbf{E}, G, \Theta). \quad (4)$$

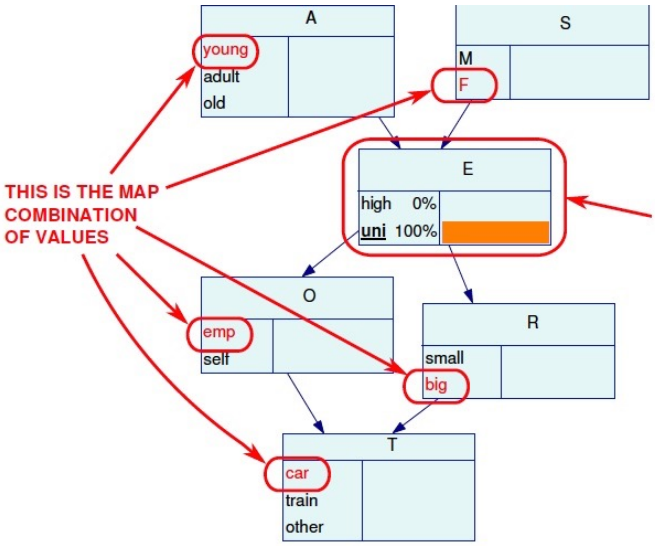
Two main applications:

- imputing missing data, where the variables in $\mathbf{Q}$ are not observed and are imputed from those in $\mathbf{E}$;
- comparing $\mathbf{q}^*$ with the observed values for the variables in $\mathbf{Q}$.

NOTE: $\mathbf{q}^*$ is not the collection of the values with the highest posterior in each posterior marginal distribution, those distributions are not independent!

39

Section 3: Probabilistic Queries



40

Section 3: Probabilistic Queries

The act of propagating the effects of evidence is called belief updating or belief propagation: our belief on $\mathbf{X}$ as encoded by the BN $\mathcal{B}$ is updated in the face of new evidence $\mathbf{E}$. This task is computationally feasible because we rely on local computations (only using local distributions):

$$\begin{aligned} \mathrm{P}(\mathbf{Q} \mid \mathbf{E}, G, \Theta) &= \int \mathrm{P}(\mathbf{X} \mid \mathbf{E}, G, \Theta)\, d(\mathbf{X} \setminus \mathbf{Q}) \\ &= \int \left[\prod_{i=1}^{p} \mathrm{P}(X_i \mid \mathbf{E}, \Pi_{X_i}, \Theta_{X_i}) \right] d(\mathbf{X} \setminus \mathbf{Q}) \\ &= \prod_{i:X_i \in \mathbf{Q}} \int \mathrm{P}(X_i \mid \mathbf{E}, \Pi_{X_i}, \Theta_{X_i})\, dX_i. \end{aligned}$$

The correspondence between d-separation and conditional independence can also be used to further reduce the dimension of the problem (e.g. to the Markov blanket).

41

Section 3: Probabilistic Queries

Algorithms for belief updating can be classified either as

- Exact: algorithms that combine repeated applications of Bayes theorem with local computations to obtain the exact value of $\mathrm{P}(\mathbf{Q} \mid \mathbf{E}, G, \Theta)$. The two best known are
  - variable elimination; and
  - belief updates based on junction trees.
- Approximate: algorithms that use Monte Carlo simulations to sample from the global distribution and thus estimate $\mathrm{P}(\mathbf{Q} \mid \mathbf{E}, G, \Theta)$. In computer science, these random samples are often called particles, and the algorithms that make use of them are known as particle filters. The two best known are
  - logic sampling; and
  - likelihood weighting.

Approximate algorithms tend to scale better to larger number of variables since they are usually embarrassingly parallel; exact algorithms tend to be more sequential and iterative in nature.

42

Computing Part 3

43

References

Nagarajan, Radhakrishnan, Marco Scutari, and Sophie Lèbre. "Bayesian networks in R." *Springer* 122 (2013): 125-127.

Slides adopted from:
(https://dipartimenti.unicatt.it/scienze-statistiche-23-25-1-17ScutariSlides.pdf)

44