Bayesian networks with likelihood evidence in R

Nordstat 2018, Tartu, Estonia

June, 2016

Søren Højsgaard

Department of Mathematical Sciences, Aalborg University, Denmark

Printed: June 28, 2018 File: bayesnet-slides.tex

Contents

Outline

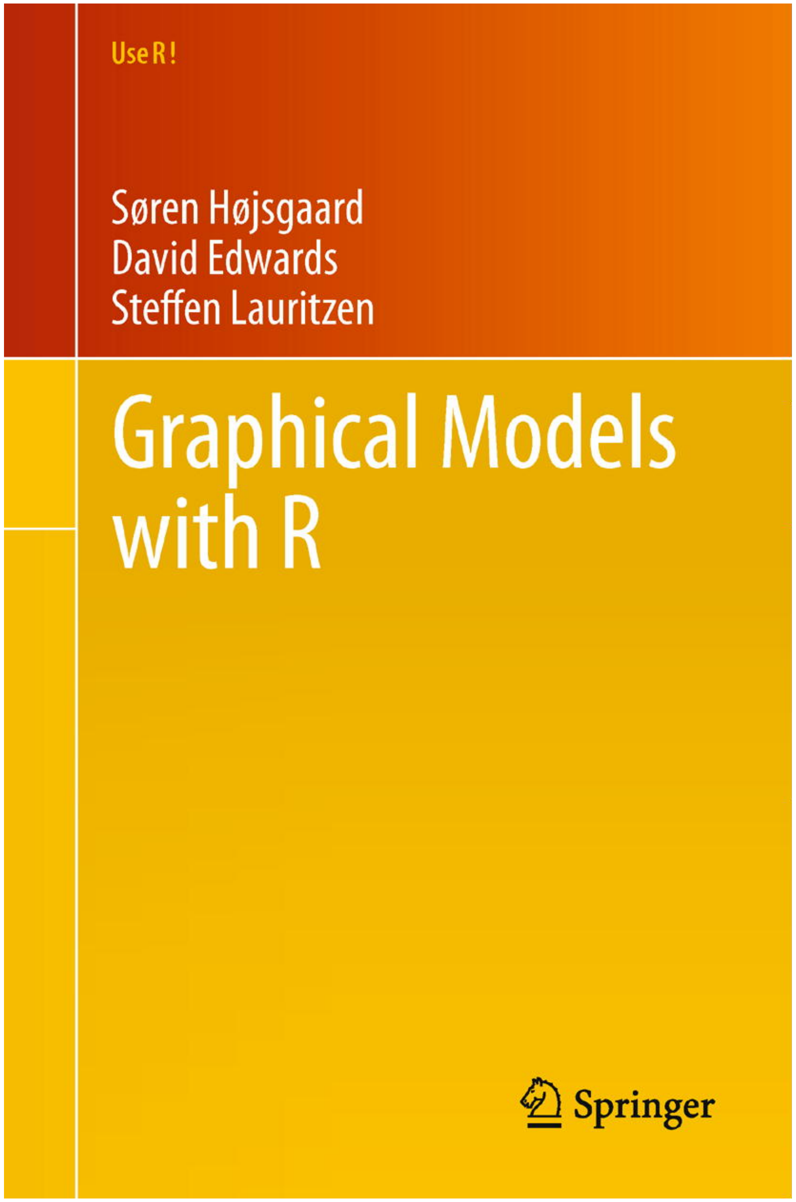
Since presentation is in a “miscellaneous topics” session, the plan is

- Gentle introduction to Bayesian network.
- Probability propagation; conditional independence restrictions and dependency graphs
- Different types of evidence.

The real agenda:

- The **gRain** package handles Bayesian network with discrete variables only.
- A FAQ: Does **gRain** handle other type of variables?
- Short answer: No
- Slightly longer answer: Yes, in some cases by using a little trick.

Book: Graphical Models with R



1 Bayesian networks (BN) basics

- What is a BN? There is no canoical definition - so here is one:
- A probabilistic model / a density $p_X(x)$ for a d dimensional random vector $X = (X_1, \dots, X_d)$.
- Often - but not always - $p_X(\dots)$ is specified by help of a directed acyclic graph (*DAG*).
- Often - but not always - $p_X(\dots)$ has a simplifying structure that allows for simplifying computations (*conditional independence restrictions*).

- Split X in subvectors $X = (X_U, X_V, X_W)$. Often - but not always - interest is in computing marginal / conditional distributions in an efficient way; e.g.

$$P_U(x_U); p_{U|V}(x_U | x_V = x_V^*)$$

Call $x_V = x_V^*$ for *hard evidence*

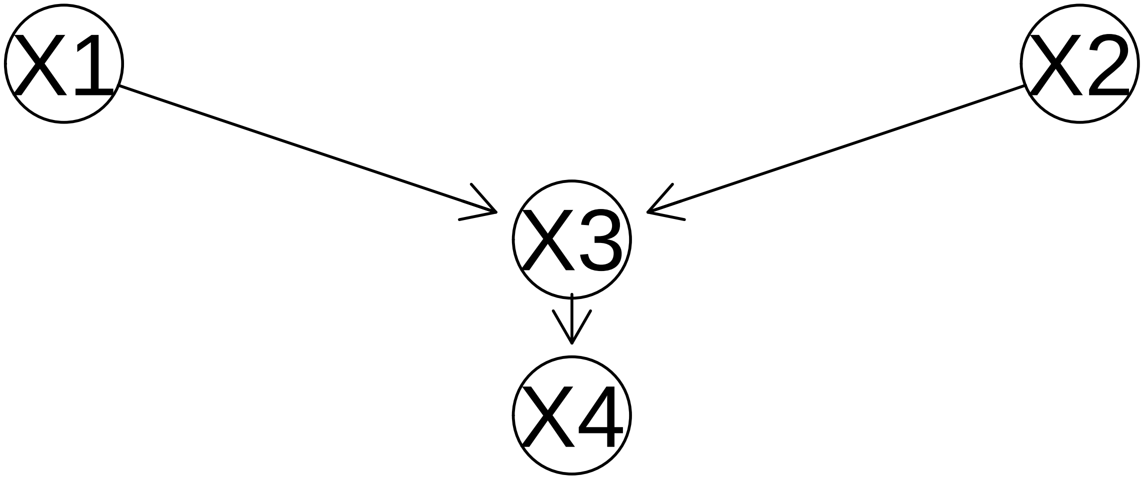
- Sometimes interest is in

$$p_{U|V}(x_U | x_V \approx x_V^*)$$

Call $x_V \approx x_V^*$ for *likelihood evidence* or *soft evidence*.

- Likelihood evidence is topic of talk; using this we can (sometimes) handle other type of variables.

1.1 A small example



- $X_1 \sim bern(.3)$;
- $X_2 \sim poi(5)$;
- $X_3|X_1 = x_1, X_2 = x_2 \sim N(x_1 + x_2, 1)$;
- $X_4|X_3 = x_3 \sim poi(\exp(x_3))$.
- $p_X(x_1, x_2, x_3, x_4) = q_1(x_1)q_2(x_2)q_3(x_3|x_1, x_2)q_4(x_4|x_3)$

- Structure $p_X(x_1, x_2, x_3, x_4) = q_1(x_1)q_2(x_2)q_3(x_3|x_1, x_2)q_4(x_4|x_3)$ implies various things:

- A *conditional independence*: $X_4 \perp\!\!\!\perp X_1, X_2 | X_3$.

$$p_{4|321}(x_4|x_3, x_2, x_1) = q_4(x_4|x_3) \text{ independently of } x_2, x_1$$

- A *marginal independence*: $X_1 \perp\!\!\!\perp X_2$

$$p_{21}(x_2, x_1) = q_1(x_1)q_2(x_2)$$

- Structure $p_X(x_1,x_2,x_3,x_4) = q_1(x_1)q_2(x_2)q_3(x_3|x_1,x_2)q_4(x_4|x_3)$ implies various things:

- Computation of e.g. $p_{12|4}(x_1,x_2|x_4^*)$ can be made *locally* and WITHOUT ever forming the joint density $p_X(x_1,x_2,x_3,x_4)$.

  1. Set $u_4(x_3) = q_4(x_4^*|x_3)$
  2. Set $u_3(x_1,x_2,x_3) = q_3(x_3|x_1,x_2)u_4(x_3)$
  3. Set $u_2(x_1,x_2) = \int u_3(x_1,x_2,x_3)dx_3$
  4. Set $c = \int q_1(x_1)q_2(x_2)u_2(x_1,x_2)dx_1dx_2$ and we have
  5. $p_{12|4}(x_1,x_2|x_4^*) = q_1(x_1)q_2(x_2)u_2(x_1,x_2)/c$.

- Often computations above can not be made analytically and we resort to simulations (BUGS, JAGS, STAN, ...)

- But in important special cases, closed form expressions can be obtained.

- One such case is when all variables are discrete with a finite state space.

1.2 The **gRain** package

When all variables are discrete with a finite state space,

- the **gRain** package will do all computations efficiently.
- all conditional densities are represented by conditional probability tables (CPTs).
- From the perspective of this talk, **gRain** is a calculator. For details on computations, see references.

FAQ:

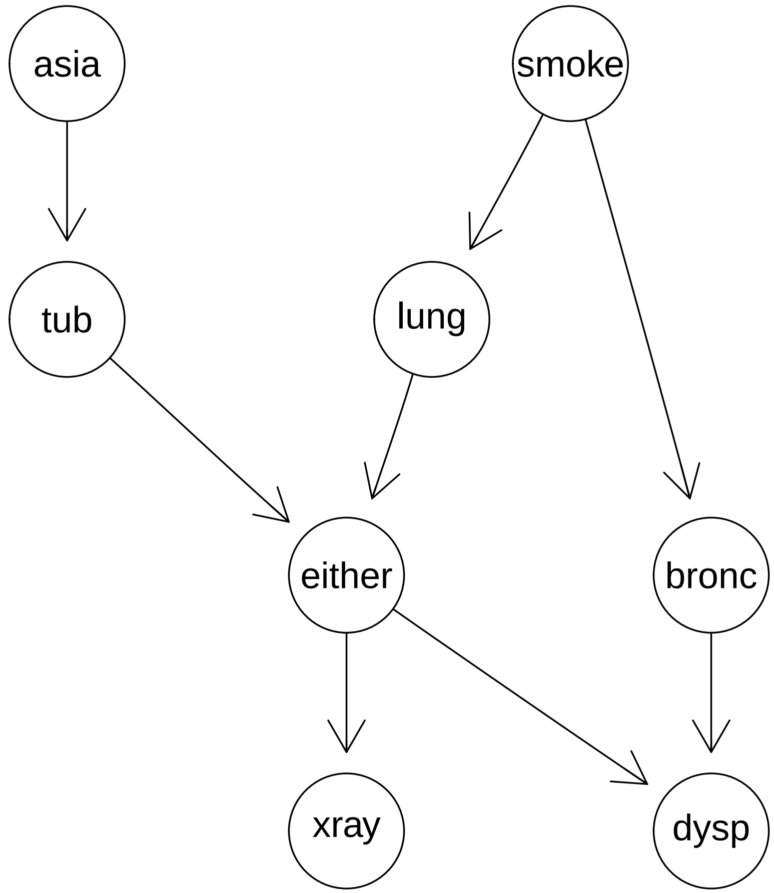
Q: Will **gRain** handle variables that are not discrete?

A: No, not directly, but there is a small trick that allows for non-discrete variables in certain cases.

1.3 Example: The chest clinic narrative

Lauritzen and Spiegehalter (1988) present the following narrative:

- "Shortness–of–breath (*dyspnoea*) may be due to *tuberculosis*, *lung cancer* or *bronchitis*, or none of them, or more than one of them.
- A recent visit to *Asia* increases the chances of tuberculosis, while *smoking* is known to be a risk factor for both lung cancer and bronchitis.
- The results of a single chest *X–ray* do not discriminate between lung cancer and tuberculosis, as n*either* does the presence or absence of *dyspnoea*."


asia
smoke
tub
lung
either
bronc
xray
dysp

```r
yn <- c("yes","no")
a    <- cptable(~asia,          values=c(1,99), levels=yn)
t.a  <- cptable(~tub | asia,    values=c(5,95, 1,99), levels=yn)
s    <- cptable(~smoke,         values=c(5,5), levels=yn)
l.s  <- cptable(~lung | smoke,  values=c(1,9, 1,99), levels=yn)
b.s  <- cptable(~bronc | smoke, values=c(6,4, 3,7), levels=yn)
e.lt <- cptable(~either | lung:tub,
                values=c(1,0, 1,0, 1,0, 0,1), levels=yn)
x.e  <- cptable(~xray | either,
                values=c(98,2, 5,95), levels=yn)
d.be <- cptable(~dysp | bronc:either,
                values=c(9,1, 7,3, 8,2, 1,9), levels=yn)
```

```r
cpt.list <- compileCPT(list(a, t.a, s, l.s, b.s, e.lt, x.e, d.be))
cpt.list$tub

##        asia
## tub    yes no
##   yes    5  1
##   no    95 99

bn <- grain(cpt.list)
bn

## Independence network: Compiled: FALSE Propagated: FALSE
##   Nodes: chr [1:8] "asia" "tub" "smoke" "lung" "bronc" "either" "xray" ...
```

Marginal distributions:

```
qgrain(bn, nodes=c("lung", "tub"))

## $tub
## tub
##    yes     no
## 0.0104 0.9896
##
## $lung
## lung
##    yes     no
## 0.0182 0.9818
```

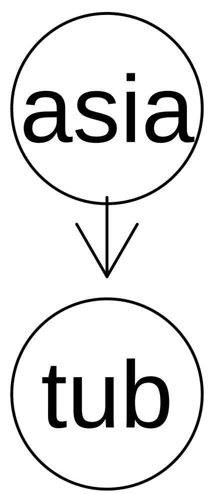
Conditional distributions given hard evidence:

```
qgrain(bn, nodes=c("lung", "tub"), evidence=list(asia="yes", smoke="no", dysp="yes"))

## $tub
## tub
##   yes    no
## 0.113 0.887
##
## $lung
## lung
##    yes     no
## 0.0226 0.9774
```

2 Hard and soft/likelihood/virtual evidence

Consider the following excerpt of the chest clinic network:

```
yn <- c("yes","no")
a    <- cptable(~asia, values=c(1,99),levels=yn)
t.a  <- cptable(~tub|asia, values=c(5,95, 1,99),levels=yn)

plist1 <- compileCPT(list(a, t.a))
chest1 <- grain(plist1)
plot(chest1)
```



2.1 Hard evidence

A person has recently been to Asia so `asia="yes"`. We compute $p(tub)$ and $p(tub|asia = yes)$.

```
qgrain(chest1, nodes="tub")

## $tub
## tub
##    yes     no
## 0.0104 0.9896

qgrain(chest1, nodes="tub", evidence=list(asia="yes"))

## $tub
## tub
##  yes   no
## 0.05 0.95
```

2.2 Likelihood/virtual/soft evidence

Suppose we do not know with certainty whether a patient has recently been to Asia or not

- Perhaps the patient is too ill to tell
- However the patient (a Caucasian Dane) may be unusually tanned. This lends support to the hypothesis of a recent visit to Asia.

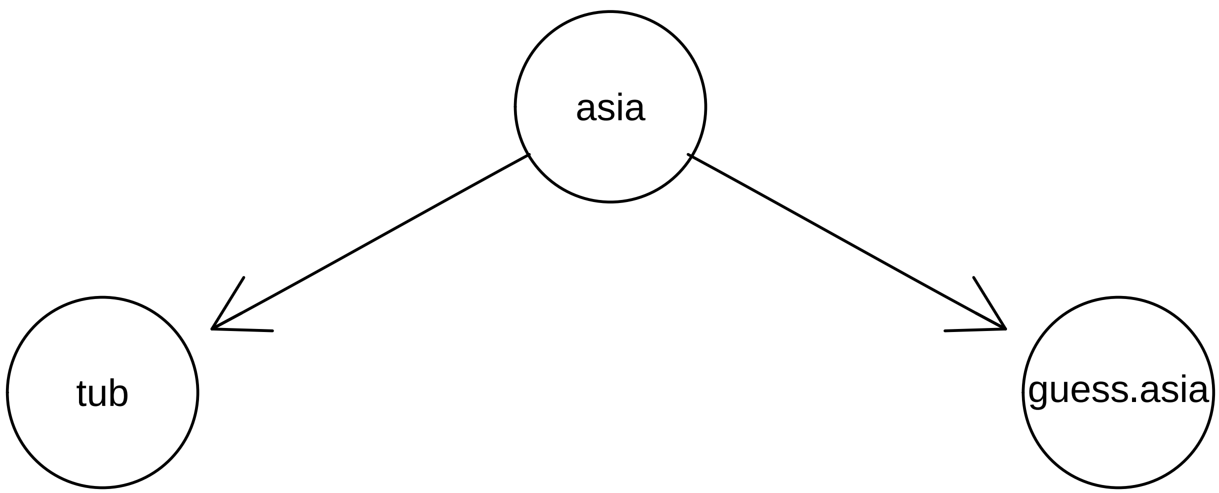
To accommodate we can create an extended network with an extra node for which we enter evidence.

We can then introduce a new variable `guess.asia` with `asia` as its only parent.

- If recently in Asia we would guess so in 80% of the times
- If not recently in Asia we would guess so in 90% of the times

```
g.a <- cptable(~ guess.asia|asia, levels=yn,
               values=c(.8,.2, .1,.9))
```

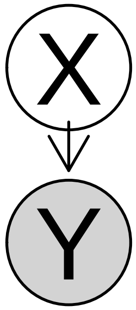
```
plist2 <- compileCPT(list(a, t.a, g.a))
plist2$guess.asia

##             asia
## guess.asia yes  no
##        yes 0.8 0.1
##        no  0.2 0.9

chest2 <- grain(plist2)
plot(chest2)
```



Now specify different type of information on visit to Asia:

```
qgrain(chest2, nodes="tub")

## $tub
## tub
##    yes     no
## 0.0104 0.9896

qgrain(chest2, nodes="tub", evidence=list(guess.asia="yes"))

## $tub
## tub
##   yes    no
## 0.013 0.987

qgrain(chest2, nodes="tub", evidence=list(asia="yes"))

## $tub
## tub
##  yes   no
## 0.05 0.95
```

2.3 Likelihood evidence



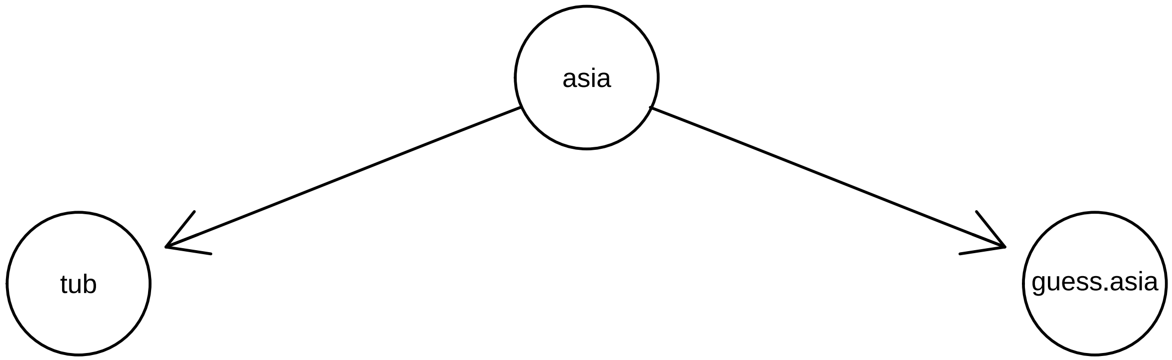
Very simple network

- "Prior": X: binary; levels="yes"/"no"
- "Likelihood": $Y|X=x$: $N(\mu_x, 1)$
- Joint: $p(y, x) = q_1(x)q_2(y|x)$

The effect of observing $y = y^*$ is to modify prior by contribution from likelihood:

- Set $q_1^*(x) \leftarrow q_1(x)q_2(y^*|x)$
- Normalize $p(x|y = y^*) = q_1^*(x)/\sum_{x'=yes,no} q_1^*(x')$

Same argument applies to small chest clinic network:



$$p(asia, tub, guess.asia) = q_1(asia)q_2(tub|asia)q_3(guess.asia|tub)$$

Same with the effect of `guess.asia="yes"`: Absorb likelihood $q_3(guess.asia = "yes"|asia)$ information into $q_1(asia)$

- Set $q_1^*(asia) \leftarrow q_1(asia)$
- Then $p(asia, tub|guess.asia = "yes") \propto q_1^*(asia)q_2(tub|asia)$
- Normalize and we are done

2.4 Specifying virtual evidence

Hence we can absorb likelihood inforation directly into existing network (without expanding with extra nodes):

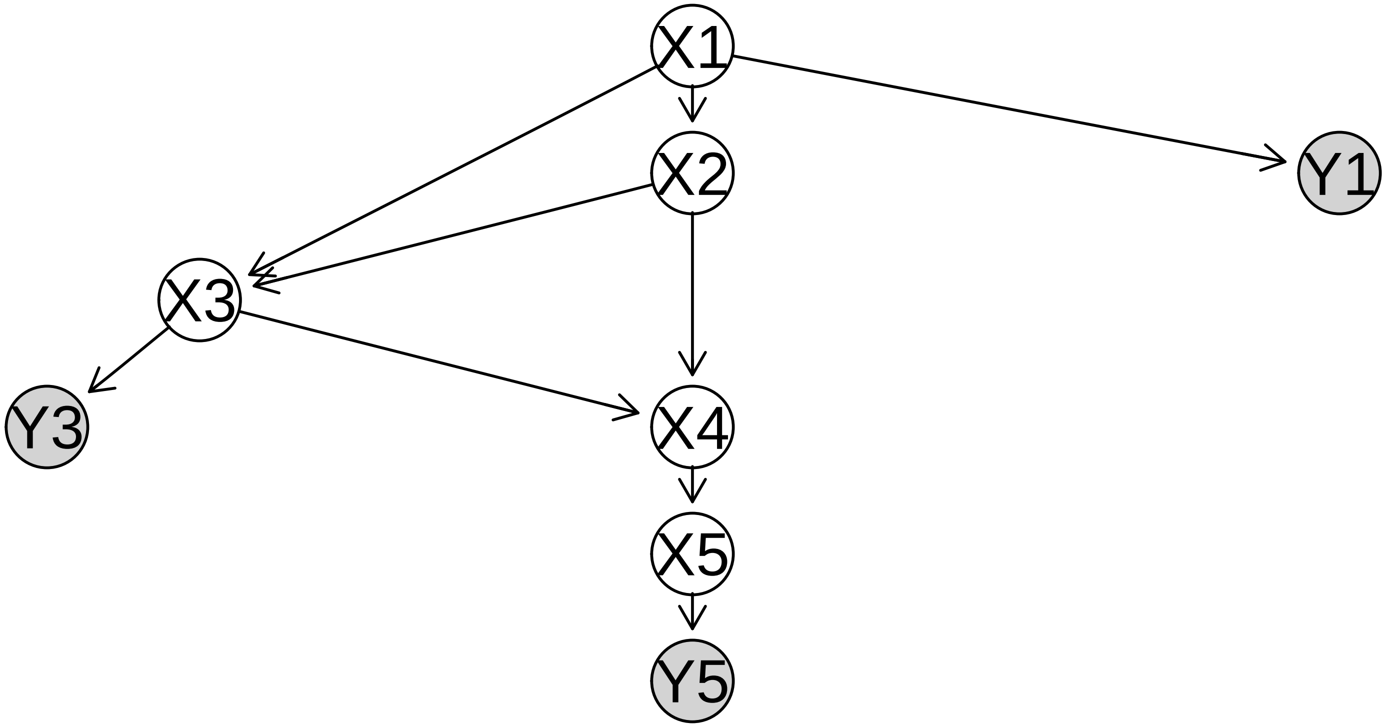
```
qgrain(chest1, nodes="tub", evidence=list(asia=c(.8, .1)))

## $tub
## tub
##   yes    no
## 0.013 0.987
```

This also means that hard evidence e.g. **asia='yes'** can be entered as

```
qgrain(chest1, nodes="tub", evidence=list(asia=c(1, 0)))

## $tub
## tub
##  yes   no
## 0.05 0.95
```

3 Winding up

The likelihood evidence trick will handle situations like



Thank you for your attention!

Package versions

For installation information, please go to:

http://people.math.aau.dk/~sorenh/software/gR

```
packageVersion("gRain")

## [1] '1.3.0.1'
```