Topics in Statistical Sciences 1 – Exam exercise 3

Søren Højsgaard and Torben Tvedebrink

6th November 2016

This exercise is about the Bayesian networks as as discussed in lectures 7, 8 and 9 of Topics in Statistical Sciences 1. During the oral exam you will have 20 min to present the exercise. You decide what topics to cover and how to present them, however, we will ask questions to any part of the exercise and presentation.

1 Check that a directed graph is acyclic

A graph $G = (V, E)$ with nodes V and edges E is directed if and only if all edges are directed. A cycle in a directed graph is a sequence of edges such that if you walk along the direction of the edges then you will come back to where you started. A directed graph is acyclic if and only if the graph contains no cycles. A directed acyclic graph is called a DAG.

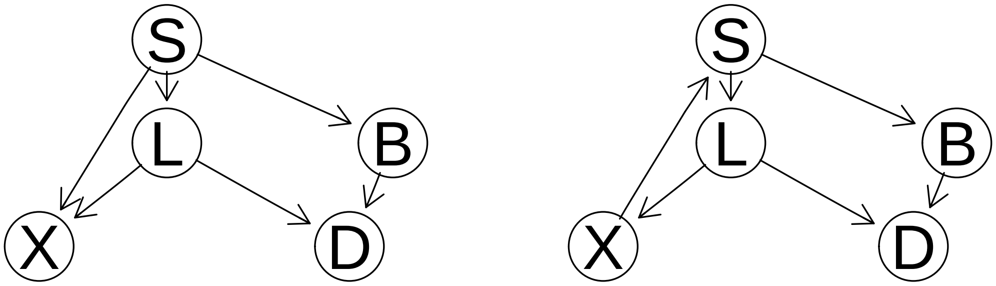
We have seen in the course that given a DAG $G = (V, E)$ a joint density can be specified as

$$p_{X_V}(x_V) = \prod_{v \in V} p_{X_v|X_{pa(v)}}(x_v|x_{pa(v)})$$

where $pa(v)$ denotes the parents of v in G. If the graph has cycles, then the product on the right hand side does not in general define a joint density. Therefore it is of interest to be able to check if a directed graph has cycles.

Consider these examples:

```r
library(gRbase)
library(Rgraphviz)
dg1 <- dag(~ S + L|S + X|L:S + B|S + D|L:B)
dg2 <- dag(~ S|X + L|S + X|L + B|S + D|L:B)
par(mfrow=c(1,2))
plot(dg1)
plot(dg2)
```



A graph can be represented on a computer as an adjacency matrix, for example

```
as(dg1, "matrix")
```

```
##   S L X B D
## S 0 1 1 1 0
## L 0 0 1 0 1
## X 0 0 0 0 0
## B 0 0 0 0 1
## D 0 0 0 0 0
```

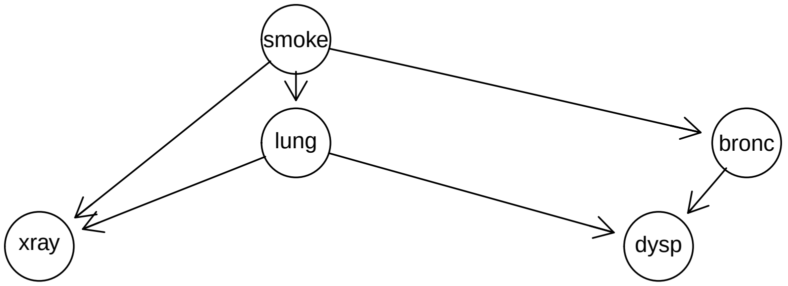
in which entry (i, j) is non-zero if and only if there is as a directed edge from node i to node j. Here we make the convention, that a node can not point to itself! (Bonus information: An undirected edge between node i and node j can be thought of as a bidirected edge and can be represented as entry (i, j) and entry (j, i) being non-zero.)

1. Propose an algorithm which as input takes a directed graph and as output returns `TRUE` if the graph is acyclic and `FALSE` otherwise.
2. Implement the algorithm in R where you base the implementation on that the graph is represented as an adjacency matrix; demonstrate that it works.

2 An excerpt of the chest clinic example

Consider the following excerpt from the chest clinic examples:

```
dg <- dag(~ smoke + lung|smoke + xray|lung:smoke + bronc|smoke + dysp|bronc:lung )
plot(dg)
```



The dataset

```
data(chestSim1000, package="gRbase")
head(chestSim1000)
```

```
##   asia tub smoke lung bronc either xray dysp
## 1   no  no    no   no   yes     no   no  yes
## 2   no  no   yes   no   yes     no   no  yes
## 3   no  no   yes   no    no     no   no   no
## 4   no  no    no   no    no     no   no   no
## 5   no  no   yes   no   yes     no   no  yes
## 6   no  no   yes  yes   yes    yes  yes  yes
```

contains “data” from which the conditional probability tables (CPTs) can be estimated.

1. Extract the necessary CPTs from data, and construct the Bayesian network.
2. What does information about `dysp` tell is about `smoke`, i.e. what is the conditional distribution `smoke` given `dysp`?

3. If we know `smoke`, what does additional information about `bronc` tell us about `lung`? That is, what is the conditional distribution of `lung` given `smoke`, and what is the conditional distribution of `lung` given `smoke` *and* `bronc`?

4. If we know `smoke` and `dysp`, what does additional information about `bronc` tell us about `lung`?

5. Sketch the message passing algorithm (`CollectEvidence` and `DistributeEvidence`) for this specific example.