Bayesian networks – exercises

Collected by: Jiří Kléma, klema@labe.felk.cvut.cz

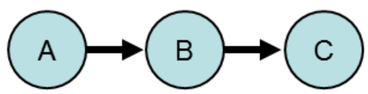
Fall 2014/2015

Note: The exercises 3b-e, 10 and 13 were not covered this term.

Goals: The text provides a pool of exercises to be solved during AE4M33RZN tutorials on graphical probabilistic models. The exercises illustrate topics of conditional independence, learning and inference in Bayesian networks. The identical material with the resolved exercises will be provided after the last Bayesian network tutorial.

1 Independence and conditional independence

Exercise 1. *Formally prove which (conditional) independence relationships are encoded by serial (linear) connection of three random variables.*



Only the relationship between A and C shall be studied (the variables connected by an edge are clearly dependent), let us concern $A \perp\!\!\!\perp C|\emptyset$ and $A \perp\!\!\!\perp C|B$:
$A \perp\!\!\!\perp C|\emptyset \Leftrightarrow Pr(A,C) = Pr(A)Pr(C) \Leftrightarrow Pr(A|C) = Pr(A) \wedge Pr(C|A) = Pr(C)$
$A \perp\!\!\!\perp C|B \Leftrightarrow Pr(A,C|B) = Pr(A|B)Pr(C|B) \Leftrightarrow Pr(A|B,C) = Pr(A|B) \wedge Pr(C|A,B) = Pr(C|B)$

It follows from BN definition: $Pr(A,B,C) = Pr(A)Pr(B|A)Pr(C|B)$

To decide on conditional independence between A and C, $Pr(A,C|B)$ can be expressed and factorized. It follows from both conditional independence and BN definition:

$$Pr(A,C|B) = \frac{Pr(A,B,C)}{Pr(B)} = \frac{Pr(A)Pr(B|A)}{Pr(B)}Pr(C|B) = Pr(A|B)Pr(C|B)$$

$Pr(A,C|B) = Pr(A|B)Pr(C|B)$ holds in the linear connection and $A \perp\!\!\!\perp C|B$ also holds.

Note 1: An alternative way to prove the same is to express $Pr(C|A,B)$ or $Pr(A|B,C)$:

$$Pr(C|A,B) = \frac{Pr(A,B,C)}{Pr(A,B)} = \frac{Pr(A)Pr(B|A)Pr(C|B)}{Pr(A)Pr(B|A)} = Pr(C|B) \text{ or}$$

$$Pr(A|B,C) = \frac{Pr(A,B,C)}{Pr(B,C)} = \frac{Pr(A)Pr(B|A)Pr(C|B)}{\sum_A Pr(A)Pr(B|A)Pr(C|B)} = \frac{Pr(A)Pr(B|A)Pr(C|B)}{Pr(C|B)\sum_A Pr(A)Pr(B|A)} =$$

$$= \frac{Pr(A)Pr(B|A)}{Pr(B)} = Pr(A|B)$$

Note 2: Even a more simple way to prove the same is to apply both the general and the BN specific definition of joint probability:

$$Pr(A)Pr(B|A)Pr(C|A,B) = Pr(A,B,C) = Pr(A)Pr(B|A)Pr(C|B) \Rightarrow Pr(C|A,B) = Pr(C|B)$$

To decide on independence of A from C, $Pr(A,C)$ needs to be expressed. Let us marginalize the BN definition:

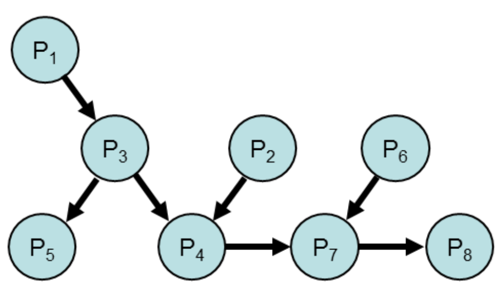
$$Pr(A,C) = \sum_B Pr(A,B,C) = \sum_B Pr(A)Pr(B|A)Pr(C|B) = Pr(A)\sum_B Pr(C|B)Pr(B|A) =$$

$$= Pr(A)\sum_B Pr(C|A,B)Pr(B|A) = Pr(A)\sum_B Pr(B,C|A) = Pr(A)Pr(C|A)$$

(the conditional independence expression proved earlier was used, it holds $Pr(C|B) = Pr(C|A,B)$). The independence expression $Pr(A,C) = Pr(A)Pr(C)$ does not follow from the linear connection and the relationship $A \perp\!\!\!\perp C|\emptyset$ does not hold in general.

Conclusion: D-separation pattern known for linear connection was proved on the basis of BN definition. Linear connection transmits information not given the intermediate node, it blocks the information otherwise. In other words, its terminal nodes are dependent, however, when knowing the middle node for sure, the dependence vanishes.

Exercise 2. *Having the network/graph shown in figure below, decide on the validity of following statements:*



a) $P_1, P_5 \perp\!\!\!\perp P_6|P_8$,

b) $P_2 \top\!\!\!\top P_6|\oslash$,

c) $P_1 \perp\!\!\!\perp P_2 | P_8$,

d) $P_1 \perp\!\!\!\perp P_2, P_5 | P_4$,

e) Markov equivalence class that contains the shown graph contains exactly three directed graphs.

Solution:

a) FALSE, the path through P_3, P_4 and P_7 is opened, neither the nodes P_1 and P_6 nor P_5 and P_6 are d-separated,

b) FALSE, the path is blocked, namely the node P_7,

c) FALSE, unobserved linear P_3 is opened, converging P_4 is opened due to P_8, the path is opened,

d) FALSE, information flows through unobserved linear P_3,

e) TRUE, $P_1 \to P_3$ direction can be changed (second graph) then $P_3 \to P_5$ can also be changed (third graph).

Exercise 3. *Let us have an arbitrary set of (conditional) independence relationships among N variables that is associated with a joint probability distribution.*

a) Can we always find a directed acyclic graph that perfectly maps this set (perfectly maps = preserves all the (conditional) independence relationships, it neither removes nor adds any)?

b) Can we always find an undirected graph that perfectly maps this set?

c) Can directed acyclic models represent the conditional independence relationships of all possible undirected models?

d) Can undirected models represent the conditional independence relationships of all possible directed acyclic models?

e) Can we always find a directed acyclic model or an undirected model?

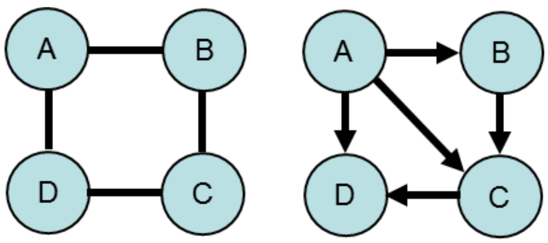
Solution:

a) No, we cannot. An example is $\{A \perp\!\!\!\perp C | B \cup D, B \perp\!\!\!\perp D | A \cup C\}$ in a four-variable problem. This pair of conditional independence relationships leads to a cyclic graph or converging connection which introduces additional independence relationships. In practice if the perfect map does not exist, we rather search for a graph that encodes only valid (conditional) independence relationships and it is minimal in such sense that removal of any of its edges would introduce an invalid (conditional) independence relationship.

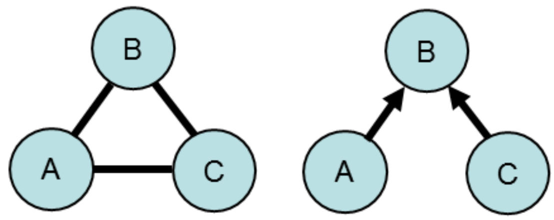
b) No, we cannot. An example is $\{A \perp\!\!\!\perp C | \emptyset\}$ in a three-variable problem (the complementary set of dependence relationships is $\{A \not\perp\!\!\!\perp B | \emptyset, A \not\perp\!\!\!\perp B | C, B \not\perp\!\!\!\perp C | \emptyset, B \not\perp\!\!\!\perp C | A, A \not\perp\!\!\!\perp C | B\}$). It follows that A and B must be directly connected as there is no other way to meet both $A \not\perp\!\!\!\perp B | \emptyset$ and $A \not\perp\!\!\!\perp B | C$. The same holds for B and C. Knowing $A \perp\!\!\!\perp C | \emptyset$, there can be no edge between A and C. Consequently, it necessarily holds $A \perp\!\!\!\perp C | B$ which contradicts the given set of independence relationships (the graph encodes an independence relationship that does not hold).

c) No, they cannot. An example is the set of relationships ad a) that can be encoded in a form of undirected graph (see the left figure below), but not as a directed graph. The best directed graph is the graph that encodes only one of the CI relationships (see the mid-left figure below) that stands for $\{B \perp\!\!\!\perp D | A \cup C\}$.

d) There are also directed graphs whose independence relationships cannot be captured by undirected models. Any directed graph with converging connection makes an example, see the graph on the right in the figure below which encodes the set of the relationships ad b). In the space of undirected graphs it needs to be represented as the complete graph (no independence assumptions). Any of two of the discussed graph classes is not strictly more expressive than the other.

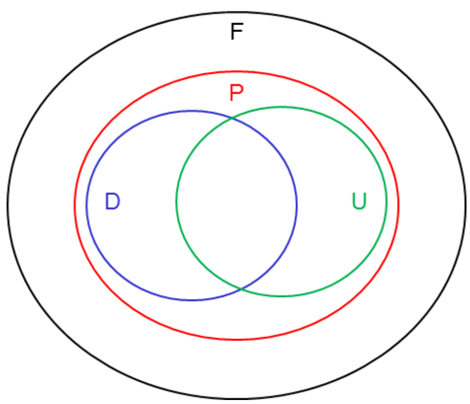


the undirected model ad a)
and its imperfect directed counterpart



the directed model ad b)
and its imperfect undirected counterpart

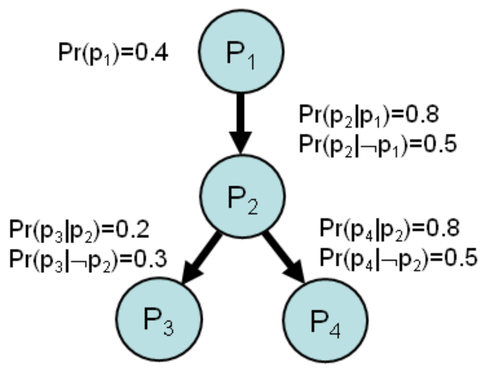
e) No, we cannot. Although the set of free CI relationship sets is remarkably restricted by the condition of existence of associated joint probability distribution (e.g., the set $\{A \perp\!\!\!\perp B, B \top\!\!\!\top A\}$ violates the trivial axiom of symmetry, there is no corresponding joint probability distribution), there are still sets of relationships that are meaningful (have their joint probability counterpart) but cannot be represented as any graph.



Venn diagram illustrating graph expressivity. F stands for free CI relationship sets, P stands for CI relationship sets with an associated probability distribution, D stands for distributions with the perfect directed map and U stands for distributions with the perfect undirected map.

2 Inference

Exercise 4. *Given the network below, calculate marginal and conditional probabilities* $Pr(\neg p_3)$, $Pr(p_2|\neg p_3)$, $Pr(p_1|p_2, \neg p_3)$ *a* $Pr(p_1|\neg p_3, p_4)$. *Apply the method of* **inference by enumeration**.



Inference by enumeration sums the joint probabilities of atomic events. They are calculated from the network model: $Pr(P_1, \ldots, P_n) = Pr(P_1|parents(P_1)) \times \cdots \times Pr(P_n|parents(P_n))$. The method does not take advantage of conditional independence to further simplify inference. It is a routine and easily formalized algorithm, but computationally expensive. Its complexity is exponential in the number of variables.

$$
\begin{aligned}
Pr(\neg p_3) &= \sum_{P_1,P_2,P_4} Pr(P_1, P_2, \neg p_3, P_4) = \sum_{P_1,P_2,P_4} Pr(P_1)Pr(P_2|P_1)Pr(\neg p_3|P_2)Pr(P_4|P_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2)+ \\
&+ Pr(p_1)Pr(\neg p_2|p_1)Pr(\neg p_3|\neg p_2)Pr(p_4|\neg p_2) + Pr(p_1)Pr(\neg p_2|p_1)Pr(\neg p_3|\neg p_2)Pr(\neg p_4|\neg p_2)+ \\
&+ Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2)+ \\
&+ Pr(\neg p_1)Pr(\neg p_2|\neg p_1)Pr(\neg p_3|\neg p_2)Pr(p_4|\neg p_2) + Pr(\neg p_1)Pr(\neg p_2|\neg p_1)Pr(\neg p_3|\neg p_2)Pr(\neg p_4|\neg p_2) = \\
&= .4 \times .8 \times .8 \times .8 + .4 \times .8 \times .8 \times .2 + .4 \times .2 \times .7 \times .5 + .4 \times .2 \times .7 \times .5+ \\
&+ .6 \times .5 \times .8 \times .8 + .6 \times .5 \times .8 \times .2 + .6 \times .5 \times .7 \times .5 + .6 \times .5 \times .7 \times .5 = \\
&= .2048 + .0512 + .028 + .028 + .192 + .048 + .105 + .105 = \mathbf{.762}
\end{aligned}
$$

$$
Pr(p_2|\neg p_3) = \frac{Pr(p_2, \neg p_3)}{Pr(\neg p_3)} = \frac{.496}{.762} = \mathbf{.6509}
$$

$$
\begin{aligned}
Pr(p_2, \neg p_3) &= \sum_{P_1,P_4} Pr(P_1, p_2, \neg p_3, P_4) = \sum_{P_1,P_4} Pr(P_1)Pr(p_2|P_1)Pr(\neg p_3|p_2)Pr(P_4|p_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2)+ \\
&+ Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2) = \\
&= .4 \times .8 \times .8 \times .8 + .4 \times .8 \times .8 \times .2 + .6 \times .5 \times .8 \times .8 + .6 \times .5 \times .8 \times .2 = \\
&= .2048 + .0512 + .192 + .048 = \mathbf{.496}
\end{aligned}
$$

$$
\begin{aligned}
Pr(p_1|p_2,\neg p_3) &= \frac{Pr(p_1,p_2,\neg p_3)}{Pr(p_2,\neg p_3)} = \frac{.256}{.496} = .\mathbf{5161} \\
Pr(p_1,p_2,\neg p_3) &= \sum_{P_4} Pr(p_1,p_2,\neg p_3,P_4) = \sum_{P_4} Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(P_4|p_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2) = \\
&= .4\times.8\times.8\times.8 + .4\times.8\times.8\times.2 = .2048 + .0512 = .\mathbf{256} \\
Pr(p_2,\neg p_3) &= Pr(p_1,p_2,\neg p_3) + Pr(\neg p_1,p_2,\neg p_3) = .256 + .24 = .\mathbf{496} \\
Pr(\neg p_1,p_2,\neg p_3) &= \sum_{P_4} Pr(\neg p_1,p_2,\neg p_3,P_4) = \sum_{P_4} Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|P_2)Pr(p_4|P_2) = \\
&= Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2) = \\
&= .6\times.5\times.8\times.8 + .6\times.5\times.7\times.2 = .192 + .048 = .\mathbf{24}
\end{aligned}
$$

$$
\begin{aligned}
Pr(p_1|\neg p_3,p_4) &= \frac{Pr(p_1,\neg p_3,p_4)}{Pr(\neg p_3,p_4)} = \frac{.2328}{.5298} = .\mathbf{4394} \\
Pr(p_1,\neg p_3,p_4) &= \sum_{P_2} Pr(p_1,P_2,\neg p_3,p_4) = \sum_{P_2} Pr(p_1)Pr(P_2|p_1)Pr(\neg p_3|P_2)Pr(p_4|P_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(p_1)Pr(\neg p_2|p_1)Pr(\neg p_3|\neg p_2)Pr(p_4|\neg p_2) = \\
&= .4\times.8\times.8\times.8 + .4\times.2\times.7\times.5 = .2048 + .028 = .\mathbf{2328} \\
Pr(\neg p_3,p_4) &= Pr(p_1,\neg p_3,p_4) + Pr(\neg p_1,\neg p_3,p_4) = .2328 + .297 = .\mathbf{5298} \\
Pr(\neg p_1,\neg p_3,p_4) &= \sum_{P_2} Pr(\neg p_1,P_2,\neg p_3,p_4) = \sum_{P_2} Pr(\neg p_1)Pr(P_2|\neg p_1)Pr(\neg p_3|P_2)Pr(p_4|P_2) = \\
&= Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(\neg p_1)Pr(\neg p_2|\neg p_1)Pr(\neg p_3|\neg p_2)Pr(p_4|\neg p_2) = \\
&= .6\times.5\times.8\times.8 + .6\times.5\times.7\times.5 = .192 + .105 = .\mathbf{297}
\end{aligned}
$$

Conclusion: $Pr(\neg p3) = 0.762$, $Pr(p2|\neg p3) = 0.6509$, $Pr(p_1|p_2,\neg p_3) = 0.5161$, $Pr(p1|\neg p3,p4) = 0.4394$.

Exercise 5. *For the same network calculate the same marginal and conditional probabilities again. Employ the properties of directed graphical model to* **manually simplify** *inference by enumeration carried out in the previous exercise.*

When calculating $Pr(\neg p3)$ (and $Pr(p2|\neg p3)$ analogically), P_4 is a leaf that is not a query nor evidence. It can be eliminated without changing the target probabilities.

$$
\begin{aligned}
Pr(\neg p_3) &= \sum_{P_1,P_2} Pr(P_1,P_2,\neg p_3) = \sum_{P_1,P_2} Pr(P_1)Pr(P_2|P_1)Pr(\neg p_3|P_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2) + Pr(p_1)Pr(\neg p_2|p_1)Pr(\neg p_3|\neg p_2) + \\
&+ Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2) + Pr(\neg p_1)Pr(\neg p_2|\neg p_1)Pr(\neg p_3|\neg p_2) = \\
&= .4\times.8\times.8 + .4\times.2\times.7 + .6\times.5\times.8 + .6\times.5\times.7 = \\
&= .256 + .056 + .24 + .21 = .\mathbf{762}
\end{aligned}
$$

The same result is reached when editing the following expression:

$$\begin{aligned}
Pr(\neg p_3) = \sum_{P_1,P_2,P_4} Pr(P_1,P_2,\neg p_3,P_4) = \sum_{P_1,P_2,P_4} Pr(P_1)Pr(P_2|P_1)Pr(\neg p_3|P_2)Pr(P_4|P_2) = \\
= \sum_{P_1,P_2} Pr(P_1)Pr(P_2|P_1)Pr(\neg p_3|P_2)\sum_{P_4} Pr(P_4|P_2) = \sum_{P_1,P_2} Pr(P_1)Pr(P_2|P_1)Pr(\neg p_3|P_2) \times 1
\end{aligned}$$

Analogically, $Pr(p2, \neg p3)$ and $Pr(p2|\neg p3)$ can be calculated:

$$\begin{aligned}
Pr(p_2,\neg p_3) &= \sum_{P_1} Pr(P_1,p_2,\neg p_3) = \sum_{P_1} Pr(P_1)Pr(p_2|P_1)Pr(\neg p_3|p_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2) + Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2) = \\
&= .4 \times .8 \times .8 + 6 \times .5 \times .8 = .256 + .24 = \mathbf{.496}
\end{aligned}$$

The computation of $Pr(p_1|p_2,\neg p_3)$ may take advantage of $P_1 \perp\!\!\!\perp P_3|P_2$ – P_2 makes a linear node between P_1 and P_3, when P_2 is given, the path is blocked and nodes P_1 and P_3 are d-separated. $Pr(p_1|p_2,\neg p_3)$ simplifies to $Pr(p_1|p_2)$ which is easier to compute (both P_3 and P_4 become unqueried and unobserved graph leaves, alternatively the expression could also be simplified by elimination of the tail probability that equals one):

$$\begin{aligned}
Pr(p_1|p_2) &= \frac{Pr(p_1,p_2)}{Pr(p_2)} = \frac{.32}{.62} = \mathbf{.5161} \\
Pr(p_1,p_2) &= Pr(p_1)Pr(p_2|p_1) = .4 \times .8 = \mathbf{.32} \\
Pr(p_2) &= Pr(p_1,p_2) + Pr(\neg p_1,p_2) = .32 + .3 = \mathbf{.62} \\
Pr(\neg p_1,p_2) &= Pr(\neg p_1)Pr(p_2|\neg p_1) = .6 \times .5 = \mathbf{.3}
\end{aligned}$$

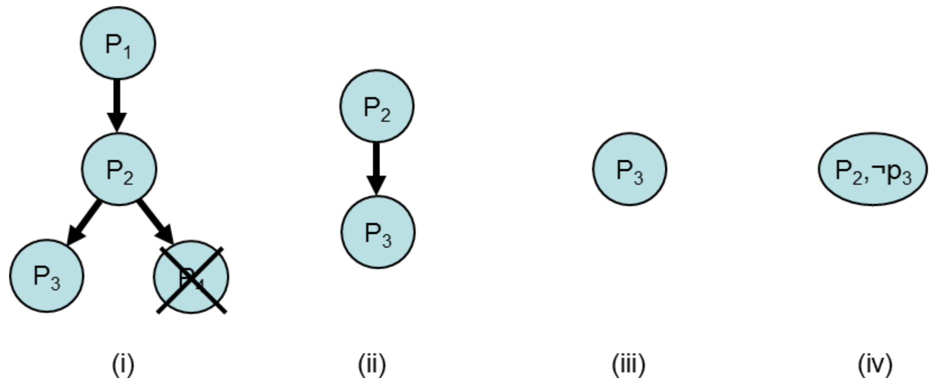
$Pr(p_1|\neg p_3, p_4)$ calculation cannot be simplified.

Conclusion: Consistent use of the properties of graphical models and precomputation of repetitive calculations greatly simplifies and accelerates inference.

Exercise 6. *For the same network calculate $Pr(\neg p_3)$ and $Pr(p_2|\neg p_3)$ again. Apply the method of* **variable elimination***.*

Variable elimination gradually simplifies the original network by removing hidden variables (those that are not query nor evidence). The hidden variables are summed out. The target network is the only node representing the joint probability $Pr(\mathbf{Q}, \mathbf{e})$. Eventually, this probability is used to answer the query $Pr(\mathbf{Q}|\mathbf{e}) = \frac{Pr(\mathbf{Q},\mathbf{e})}{\sum_{\mathbf{Q}} Pr(\mathbf{Q},\mathbf{e})}$.

The first two steps are the same for both the probabilities: (i) P_4 can simply be removed and (ii) P_1 is summed out. Then, (iii) P_2 gets summed out to obtain $Pr(\neg p_3)$ while (iv) the particular value $\neg p_3$ is taken to obtain $Pr(p_2|\neg p_3)$. See the figure below.



The elimination process is carried out by factors. The step (i) is trivial, the step (ii) corresponds to:

$$
\begin{aligned}
&f_{\bar{P}_1}(P_2) = \sum_{P_1} Pr(P_1, P_2) = \sum_{P_1} Pr(P_1)Pr(P_2|P_1) \\
&f_{\bar{P}_1}(p_2) = .4 \times .8 + .6 \times .5 = .62,\ f_{\bar{P}_1}(\neg p_2) = .4 \times .2 + .6 \times .5 = .38
\end{aligned}
$$

The step (iii) consists in:

$$
\begin{aligned}
&f_{\bar{P}_1,\bar{P}_2}(P_3) = \sum_{P_2} f_{\bar{P}_1}(P_2)Pr(P_3|P_2) \\
&f_{\bar{P}_1,\bar{P}_2}(p_3) = .62 \times .2 + .38 \times .3 = .238,\ f_{\bar{P}_1,\bar{P}_2}(\neg p_3) = .62 \times .8 + .38 \times .7 = .762
\end{aligned}
$$

The step (iv) consists in:

$$
\begin{aligned}
&f_{\bar{P}_1,\neg p_3}(P_2) = f_{\bar{P}_1}(P_2)Pr(\neg p_3|P_2) \\
&f_{\bar{P}_1,\neg p_3}(p_2) = .62 \times .8 = .496,\ f_{\bar{P}_1,\neg p_3}(\neg p_2) = .38 \times .7 = .266
\end{aligned}
$$

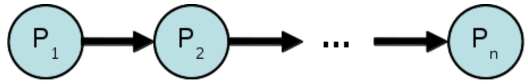
Eventually, the target probabilities can be computed:

$$Pr(\neg p_3) = f_{\bar{P}_1,\bar{P}_2}(\neg p_3) = \mathbf{.762}$$

$$Pr(p_2|\neg p_3) = \frac{f_{\bar{P}_1,\neg p_3}(p_2)}{f_{\bar{P}_1,\neg p_3}(p_2) + f_{\bar{P}_1,\neg p_3}(\neg p_2)} = \frac{.496}{.496 + .266} = \mathbf{.6509}$$

Conclusion: Variable elimination makes a building block for other exact and approximate inference algorithms. In general DAG it is NP-hard, nevertheless it is often "much more efficient" with a proper elimination order (it is difficult to find the best one, but heuristics exist). In our example, the enumeration approach takes 47 operations, the simplified method 17, while the variable elimination method needs 16 operations only.

Exercise 7. *Analyze the complexity of inference by enumeration and variable elimination on a chain of binary variables.*



The given network factorizes the joint probability as follows:

$$Pr(P_1,\dots,P_n) = \sum_{P_1,\dots,P_n} Pr(P_1)Pr(P_2|P_1)\dots Pr(P_n|P_{n-1})$$

The inference by **enumeration** works with up to 2^n atomic events. To get the probability of each event, $n-1$ multiplications must be carried out. To obtain $Pr(p_n)$, we need to enumerate and sum 2^{n-1} atomic events, which makes $(n-1)2^{n-1}$ multiplications and $2^{n-1}-1$ additions. The inference is apparently $\mathcal{O}(n2^n)$.

The inference by **variable elimination** deals with a trivial variable ordering $P_1 \prec P_2 \prec \dots \prec P_n$. In each step $i = 1,\dots,n-1$, the factor for P_i and P_{i+1} is computed and P_i is marginalized out:

$$Pr(P_{i+1}) = \sum_{P_i} Pr(P_i)Pr(P_{i+1}|P_i)$$

Each such step costs 4 multiplications and 2 additions, there are $n-1$ steps. Consequently, the inference is $\mathcal{O}(n)$. $Pr(p_n)$ (and other marginal and conditional probabilities even easier to be obtained) can be computed in linear time.

The linear chain is a graph whose largest clique does not grow with n and remains 2. That is why, variable elimination procedure is extremely efficient.

Exercise 8. *For the network from Exercise 4 calculate the conditional probability $Pr(p_1|p_2,\neg p_3)$ again. Apply a sampling approximate method. Discuss pros and cons of rejection sampling, likelihood weighting and Gibbs sampling. The table shown below gives an output of a uniform random number generator on the interval (0,1), use the table to generate samples.*

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
0.2551	*0.5060*	*0.6991*	*0.8909*	*0.9593*	*0.5472*	*0.1386*	*0.1493*	*0.2575*	*0.8407*
r_{11}	r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	r_{17}	r_{18}	r_{19}	r_{20}
0.0827	*0.9060*	*0.7612*	*0.1423*	*0.5888*	*0.6330*	*0.5030*	*0.8003*	*0.0155*	*0.6917*

Let us start with **rejection sampling**. The variables must be topologically sorted first (current notation meets the definition of topological ordering, $P_1 < P_2 < P_3 < P_4$). The individual samples will be generated as follows:

- $s^1:\ Pr(p_1) > r_1 \rightarrow p_1$,
- $s^1:\ Pr(p_2|p_1) > r_2 \rightarrow p_2$,
- $s^1:\ Pr(p_3|p_2) < r_3 \rightarrow \neg p_3$,
- $s^1:\ P_4$ is irrelevant for the given prob,
- $s^2:\ Pr(p_1) < r_4 \rightarrow \neg p_1$,
- $s^2:\ Pr(p_2|\neg p_1) < r_5 \rightarrow \neg p_2$, violates evidence, STOP.
- $s^3:\ Pr(p_1) < r_6 \rightarrow \neg p_1$,

- $s^3: \ Pr(p_2|\neg p_1) > r_7 \rightarrow p_2$,
- $s^3: \ Pr(p_3|p_2) > r_8 \rightarrow p_3$, violates evidence, STOP.
- ...

Using 20 random numbers we obtain 8 samples shown in the table below. The samples s^2, s^3, s^4, s^5 and s^7 that contradict evidence will be rejected. The rest of samples allows to estimate the target probability:

	s^1	s^2	s^3	s^4	s^5	s^6	s^7	s^8
P_1	T	F	F	T	T	F	F	F
P_2	**T**	F	T	F	F	**T**	F	**T**
P_3	**F**	?	T	?	?	**F**	?	**F**
P_4	?	?	?	?	?	?	?	?

$$Pr(p_1|p_2, \neg p_3) \approx \frac{N(p_1, p_2, \neg p_3)}{N(p_2, \neg p_3)} = \frac{1}{3} = 0.33$$

Likelihood weighting does not reject any sample, it weights the generated samples instead. The sample weight equals to the likelihood of the event given the evidence. The order of variables and the way of their generation will be kept the same as before, however, the evidence variables will be kept fixed (that is why random numbers will be matched with different probabilities):

- $s^1: \ Pr(p_1) > r_1 \rightarrow p_1$,
- $w^1: Pr(p_2|p_1)Pr(\neg p_3|p_2) = .8 \times .8 = 0.64$,
- $s^2: \ Pr(p_1) < r_2 \rightarrow \neg p_1$,
- $w^2: Pr(p_2|\neg p_1)Pr(\neg p_3|p_2) = .5 \times .8 = 0.4$,
- $s^3: \ Pr(p_1) < r_2 \rightarrow \neg p_1$,
- $w^3: Pr(p_2|\neg p_1)Pr(\neg p_3|p_2) = .5 \times .8 = 0.4$,
- ...

Using first 6 random numbers we obtain 6 samples shown in the table below. The target probability is estimated as the fraction of sample weights meeting the condition to their total sum:

	p^1	p^2	p^3	p^4	p^5	p^6
P_1	T	F	F	F	F	F
P_2	T	T	T	T	T	T
P_3	F	F	F	F	F	F
P_4	?	?	?	?	?	?
w^i	.64	.40	.40	.40	.40	.40

$$Pr(p_1|p_2, \neg p_3) \approx \frac{\sum_i w^i \delta(p_1^i, p_1)}{\sum_i w^i} = \frac{.64}{2.64} = 0.24$$

Conclusion: Both the sampling methods are consistent and shall converge to the target probability value .5161. The number of samples must be much larger anyway. Rejection sampling suffers from a large portion of generated and further unemployed samples (see s^2 and s^6). Their proportion grows for unlikely evidences with high topological indices. $Pr(p_1|\neg p_3, p_4)$ makes an example. For larger networks it becomes inefficient. Likelihood weighting shall deliver smoother estimates, nevertheless, it suffers from frequent insignificant sample weights under the conditions mentioned above.

Gibbs sampling removes the drawback of rejection sampling and likelihood weighting. On the other hand, in order to be able to generate samples, the probabilities $Pr(P_i|MB(P_i))$, where MB stands for Markov blanket of P_i node, must be computed. The blanket covers all P_i parents, children and their parents. Computation is done for all relevant hidden (unevidenced and unqueried) variables. In the given task, $Pr(P_1|P_2)$ must be computed to represent MB of P_1. The other MB probabilities

$Pr(P_2|P_1, P_3, P_4)$, $Pr(P_3|P_2)$ and $Pr(P_4|P_2)$ are not actually needed (P_2 and P_3 are evidences and thus fixed, P_4 is irrelevant), $Pr(P_3|P_2)$ and $Pr(P_4|P_2)$ are directly available anyway. However, finding $Pr(P_1|P_2)$ itself de facto solves the problem $Pr(p_1|p_2, \neg p_3)$. It follows that Gibbs sampling is advantageous for larger networks where it holds $\forall i = 1 \dots n \; |MB(P_i)| \ll n$.

Conclusion: Gibbs sampling makes sense in large networks with $\forall P_i : |MB(P_i)| \ll n$, where n stands for the number of variables in the network.

Exercise 9. *Let us have three tram lines – 6, 22 and 24 – regularly coming to the stop in front of the faculty building. Line 22 operates more frequently than line 24, 24 goes more often than line 6 (the ratio is 5:3:2, it is kept during all the hours of operation). Line 6 uses a single car setting in 9 out of 10 cases during the daytime, in the evening it always has the only car. Line 22 has one car rarely and only in evenings (1 out of 10 tramcars). Line 24 can be short whenever, however, it takes a long setting with 2 cars in 8 out of 10 cases. Albertov is available by line 24, lines 6 and 22 are headed in the direction of IP Pavlova. The line changes appear only when a tram goes to depot (let 24 have its depot in the direction of IP Pavlova, 6 and 22 have their depots in the direction of Albertov). Every tenth tram goes to the depot evenly throughout the operation. The evening regime is from 6pm to 24pm, the daytime regime is from 6am to 6pm.*

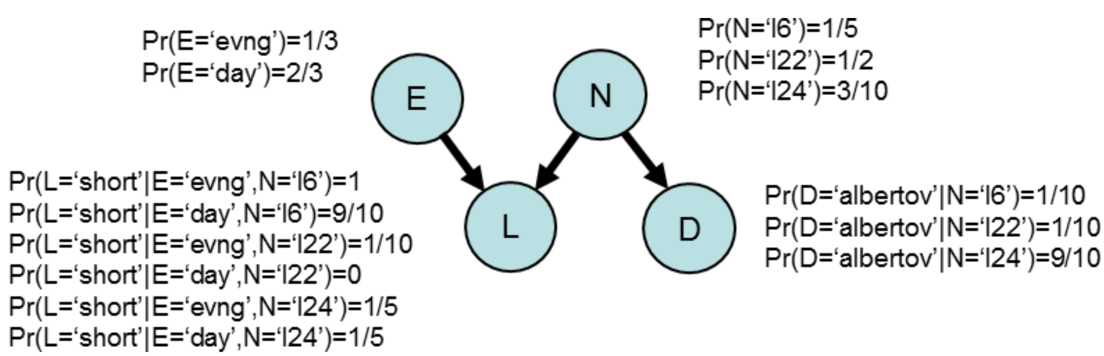
a) Draw a **correct, efficient and causal** *Bayesian network.*

b) Annotate the network with the conditional probability tables.

c) It is evening. A short tram is approaching the stop. What is the probability it will go to Albertov?

d) There is a tram 22 standing in the stop. How many cars does it have?

Ad a) and b)



Which conditional independence relationships truly hold?

- $E \perp\!\!\!\perp N|\emptyset$ – if not knowing the tram length then the tram number has nothing to do with time.
- $L \perp\!\!\!\perp D|N$ – if knowing the tram number then the tram length and its direction get independent.

- $E \perp\!\!\!\perp D|N$ – if knowing the tram number then time does not change the tram direction.
- $E \perp\!\!\!\perp D|\emptyset$ – if not knowing the tram length then time and the tram direction are independent.

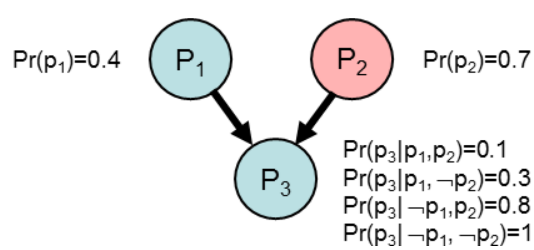
Ad c) We enumerate $Pr(D = albertov|E = evng, L = short)$, the path from E to D is opened (E is connected via the evidenced converging node L, L connects to unevidenced diverging node N, it holds $D \not\perp\!\!\!\perp E|L$, $D \not\perp\!\!\!\perp L|\emptyset$). The enumeration can be simplified by reordering of the variables and elimination of D in denominator:

$$Pr(D = albertov|E = evng, L = short) = \frac{Pr(D = albertov, E = evng, L = short)}{Pr(E = evng, L = short)} =$$
$$= \frac{\sum_N Pr(E = evng, N, L = short, D = albertov)}{\sum_{N,D} Pr(E = evng, N, L = short, D)} =$$
$$= \frac{Pr(E = evng) \sum_N Pr(N)Pr(L = short|E = evng, N)Pr(D = albertov|N)}{Pr(E = evng) \sum_N Pr(N)Pr(L = short|E = evng, N) \sum_D Pr(D|N))} =$$
$$= \frac{\sum_N Pr(N)Pr(L = short|E = evng, N)Pr(D = albertov|N)}{\sum_N Pr(N)Pr(L = short|E = evng, N)} =$$
$$= \frac{\frac{1}{5} \times 1 \times \frac{1}{10} + \frac{1}{2} \times \frac{1}{10} \times \frac{1}{10} + \frac{3}{10} \times \frac{1}{5} \times \frac{9}{10}}{\frac{1}{5} \times 1 + \frac{1}{2} \times \frac{1}{10} + \frac{3}{10} \times \frac{1}{5}} = \mathbf{.2548}$$

Ad d) In order to get $Pr(L = long|N = l22)$, it suffices the information available in nodes E and L, we will only sum out the variable E:

$$Pr(L = long|N = l22) = \sum_E Pr(L = long, E|N = l22) =$$
$$= \sum_E Pr(E|N = l22) \times Pr(L = long|E, N = l22) =$$
$$= \sum_E Pr(E) \times Pr(L = long|E, N = l22) =$$
$$= \frac{1}{3} \times \frac{9}{10} + \frac{2}{3} \times 1 = \mathbf{.9667}$$

Exercise 10. *Trace the algorithm of* **belief propagation** *in the network below knowing that* $\mathbf{e} = \{p_2\}$. *Show the individual steps, be as detailed as possible. Explain in which way the unevidenced converging node* P_3 *blocks the path between nodes* P_1 *and* P_2.



Obviously, the posterior probabilities are $Pr^*(P_1) = Pr(P_1|p_2) = Pr(P_1)$, $Pr^*(p_2) = 1$ and $Pr^*(P_3) = Pr(P_3|p_2) = \sum_{P_1} Pr(P_1)Pr(P_3|p_2, P_1)$ ($Pr^*(p_3) = 0.52$). The same values must be reached by belief propagation when message passing stops and the node probabilities are computed as follows: $Pr^*(P_i) = \alpha_i \times \pi(P_i) \times \lambda(P_i)$.

Belief propagation starts with the following list of initialization steps:

1. Unobserved root P_1 sets its compound causal $\pi(P_1)$, $\pi(p_1) = 0.4$, $\pi(\neg p_1) = 0.6$.
2. Observed root P_2 sets its compound causal $\pi(P_2)$, $\pi(p_2) = 1$, $\pi(\neg p_2) = 0$.
3. Unobserved leaf P_3 sets its compound diagnostic $\lambda(P_3)$, $\lambda(p_3) = 1$, $\lambda(\neg p_3) = 1$.

Then, iteration steps are carried out:

1. P_1 knows its compound π and misses one λ from its children only, it can send $\pi_{P_3}^{P_1}(P_1)$ to P_3:
$\pi_{P_3}^{P_1}(P_1) = \alpha_1 \pi(P_1) \rightarrow \alpha_1 = 1$, $\pi_{P_3}^{P_1}(p_1) = \pi(p_1) = 0.4$, $\pi_{P_3}^{P_1}(\neg p_1) = \pi(\neg p_1) = 0.6$
2. P_2 knows its compound π and misses one λ from its children only, it can send $\pi_{P_3}^{P_2}(P_2)$ to P_3:
$\pi_{P_3}^{P_2}(P_2) = \alpha_2 \pi(P_2) \rightarrow \alpha_2 = 1$, $\pi_{P_3}^{P_2}(p_2) = \pi(p_2) = 1$, $\pi_{P_3}^{P_2}(\neg p_2) = \pi(\neg p_2) = 0$
3. P_3 received all π messages from its parents, it can compute its compound $\pi(p_3)$:
$\pi(P_3) = \sum_{P_1,P_2} Pr(P_3|P_1,P_2) \prod_{j=1,2} \pi_{P_3}^{P_j}(P_j)$
$\pi(p_3) = .1 \times .4 \times 1 + .8 \times .6 \times 1 = 0.52$
$\pi(\neg p_3) = .9 \times .4 \times 1 + .2 \times .6 \times 1 = 1 - Pr(p3) = 0.48$
4. P_3 knows its compound λ, misses no π from its parents, it can send $\lambda_{P_3}^{P_1}(P_1)$ and $\lambda_{P_3}^{P_2}(P_2)$:
$\lambda_{P_3}^{P_j}(P_j) = \sum_{P_3} \lambda(P_3) \sum_{P_1,P_2} Pr(P_3|P_1,P_2) \prod_{k=1,2}^{k \neq j} \pi_{P_3}^{P_k}(P_k)$
$\lambda_{P_3}^{P_1}(p_1) = \sum_{P_3} \lambda(P_3) \sum_{p_1,P_2} Pr(P_3|p_1,P_2) \pi_{P_3}^{P_2}(P_2) = 1 \times 1(.1 + .9) = 1$
$\lambda_{P_3}^{P_1}(p_2) = \sum_{P_3} \lambda(P_3) \sum_{P_1,p_2} Pr(P_3|P_1,p_2) \pi_{P_3}^{P_1}(P_1) = 1(.4(.1 + .9) + .6(.8 + .2) = 1$
5. P_3 knows both its compound parameters, it can compute its posterior $Pr^*(P_3)$:
$Pr^*(P_3) = \alpha_3 \pi(P_3) \lambda(P_3)$
$Pr^*(p_3) = \alpha_3 \times .52 \times 1 = .52\alpha_3$, $Pr^*(\neg p_3) = \alpha_3 \times .48 \times 1 = .48\alpha_3$,
$Pr^*(p_3) + Pr^*(\neg p_3) = 1 \rightarrow \alpha_3 = 1$, $Pr^*(p_3) = .52$, $Pr^*(\neg p_3) = .48$
6. P_1 received all of its λ messages, the compound $\lambda(P_1)$ can be computed:
$\lambda(P_1) = \prod_{j=3} \lambda_{P_j}^{P_1}(P_1)$
$\lambda(p_1) = \lambda_{P_3}^{P_1}(p_1) = 1$, $\lambda(\neg p_1) = \lambda_{P_3}^{P_1}(\neg p_1) = 1$
7. P_1 knows both its compound parameters, it can compute its posterior $Pr^*(P_1)$: $Pr^*(P_1) = \alpha_4 \pi(P_1) \lambda(P_1)$
$Pr^*(p_1) = \alpha_4 \times .4 \times 1 = .4\alpha_4$, $Pr^*(\neg p1) = \alpha_4 \times .6 \times 1 = .6\alpha_4$,
$Pr^*(p_1) + Pr^*(\neg p_1) = 1 \rightarrow \alpha_4 = 1$, $Pr^*(p_1) = .4$, $Pr^*(\neg p_1) = .6$

Conclusion: Belief propagation reaches the correct posterior probabilities. The blocking effect of P_3 manifests in Step 4. Since P_3 is a unevidenced leaf, $\lambda(P_3)$ has a uniform distribution (i.e., $\lambda(p_3) = \lambda(\neg p_3) = 1$ as P_3 is a binary variable). It is easy to show that arbitrary normalized causal messages coming to P_3 cannot change this distribution (it holds $\sum_{P_j} \pi_{P_j}^{P_k}(P_j) = 1$). The reason is that it always holds $\sum_{P_3} Pr(P_3|P_1,P_2) = 1$. Step 4 can be skipped putting $\lambda_{P_3}^{P_j}(P_j) = 1$ automatically without waiting for the causal parameters.

3 (Conditional) independence tests, best network structure

Exercise 11. *Let us concern the frequency table shown below. Decide about independence relationships between A and B.*

	c		$\neg c$	
	b	$\neg b$	b	$\neg b$
a	*14*	*8*	*25*	*56*
$\neg a$	*54*	*25*	*7*	*11*

The relationships of independence ($A \perp\!\!\!\perp B|\emptyset$) and conditional independence ($A \perp\!\!\!\perp B|C$) represent two possibilities under consideration. We will present three different approaches to their analysis. The first one is based directly on the definition of independence and it is illustrative only. The other two approaches represent practically applicable methods.

Approach 1: Simple comparison of (conditional) probabilities.

Independence is equivalent with the following formulae:
$A \perp\!\!\!\perp B|\emptyset \Leftrightarrow Pr(A,B) = Pr(A)Pr(B) \Leftrightarrow Pr(A|B) = Pr(A) \wedge Pr(B|A) = Pr(B)$

The above-mentioned probabilities can be estimated from data by maximum likelihood estimation (MLE):
$Pr(a|b) = \frac{39}{100} = 0.39, Pr(a|\neg b) = \frac{64}{100} = 0.64, Pr(a) = \frac{103}{200} = 0.51$
$Pr(b|a) = \frac{39}{103} = 0.38, Pr(b|\neg a) = \frac{61}{97} = 0.63, Pr(b) = \frac{100}{200} = 0.5$

Conditional independence is equivalent with the following formulae:
$A \perp\!\!\!\perp B|C \Leftrightarrow Pr(A,B|C) = Pr(A|C)Pr(B|C) \Leftrightarrow Pr(A|B,C) = Pr(A|C) \wedge Pr(B|A,C) = Pr(B|C)$

Again, MLE can be applied:
$Pr(a|b,c) = \frac{14}{68} = 0.21, Pr(a|\neg b,c) = \frac{8}{33} = 0.24, Pr(a|c) = \frac{22}{101} = 0.22$
$Pr(a|b,\neg c) = \frac{25}{32} = 0.78, Pr(a|\neg b,\neg c) = \frac{56}{67} = 0.84, Pr(a|\neg c) = \frac{81}{99} = 0.82$
$Pr(b|a,c) = \frac{14}{22} = 0.64, Pr(b|\neg a,c) = \frac{54}{79} = 0.68, Pr(b|c) = \frac{68}{101} = 0.67$
$Pr(b|a,\neg c) = \frac{25}{81} = 0.31, Pr(b|\neg a,\neg c) = \frac{7}{18} = 0.39, Pr(b|\neg c) = \frac{32}{99} = 0.32$

In this particular case it is easy to see that the independence relationship is unlikely, the independence equalities do not hold. On the contrary, conditional independence rather holds as the definition equalities are roughly met. However, it is obvious that we need a more scientific tool to make clear decisions. Two of them will be demonstrated.

Approach 2: Statistical hypothesis testing.

Pearson's χ^2 independence test represents one of the most common options for independence testing. $A \perp\!\!\!\perp B|\emptyset$ is checked by application of the test on a contingency (frequency) table counting A and B co-occurrences (the left table):

O_{AB}	b	$\neg b$	sum
a	39	64	103
$\neg a$	61	36	97
sum	100	100	200

E_{AB}	b	$\neg b$	sum
a	51.5	51.5	103
$\neg a$	48.5	48.5	97
sum	100	100	200

The null hypothesis is independence of A and B. The test works with the frequencies expected under the null hypothesis (the table on right):

$$E_{AB} = \frac{N_A \times N_B}{N} \quad \rightarrow \quad E_{a\neg b} = \frac{N_a \times N_{\neg b}}{N} = \frac{103 \times 100}{200} = 51.5$$

The test compares these expected frequencies to the observed ones. The test statistic is:

$$\chi^2 = \sum_{A,B} \frac{(O_{AB} - E_{AB})^2}{E_{AB}} = 12.51 \gg \chi^2(\alpha = 0.05, df = 1) = 3.84$$

The null hypothesis is rejected in favor of the alternative hypothesis that A and B are actually dependent when the test statistic is larger than its tabular value. In our case, we took the tabular value for the common significance level $\alpha = 0.05$, df is derived from the size of contingency table ($df = (r-1)(c-1)$, where df stands for degrees of freedom, r and c stand for the number of rows and columns in the contingency table). Under the assumption that the null hypothesis holds, a frequency table with the observed and higher deviation from the expected counts can occur only with negligible probability $p = 0.0004 \ll \alpha$. **Variables A and B are dependent.**

$A \perp\!\!\!\perp B|C$ hypothesis can be tested analogically[1]. The χ^2 test statistic will be separately computed for the contingency tables corresponding to c and $\neg c$. The total value equals to sum of both the partial statistics, it has two degrees of freedom.

O_{AB}	c			$\neg c$		
	b	$\neg b$	sum	b	$\neg b$	sum
a	14	8	22	25	56	81
$\neg a$	54	25	79	7	11	18
sum	68	33	101	32	67	99

E_{AB}	c			$\neg c$		
	b	$\neg b$	sum	b	$\neg b$	sum
a	14.8	7.2	22	26.2	54.8	81
$\neg a$	53.2	25.8	79	5.8	12.2	18
sum	68	33	101	32	67	99

The null hypothesis is conditional independence, the alternative hypothesis is the full/saturated model with all parameters. The test statistic is:

$$\chi^2 = \sum_{A,B|C} \frac{(O_{AB|C} - E_{AB|C})^2}{E_{AB|C}} = 0.175 + 0.435 = 0.61 \ll \chi^2(\alpha = 0.05, df = 2) = 5.99$$

The null hypothesis cannot be rejected in favor of the alternative hypothesis based on the saturated model on the significance level $\alpha = 0.05$. A frequency table with the given or higher deviation from the expected values is likely to be observed when dealing with the conditional independence model – $p = 0.74 \gg \alpha$. **Variables A and B are conditionally independent given C.** Variable C explains dependence between A and B.

Approach 3: Model scoring.

Let us evaluate the null (A and B independent) and the alternative (A and B dependent) models of two variables, see the figure below.

[1] In practice, Pearson's χ^2 independence test is not used to test conditional independence for its low power. It can be replaced for example by the **likelihood-ratio test**. This test compares likelihood of the null model (AC,BC) with likelihood of the alternative modelu (AC,BC,AB). The null model assumes no interaction between A and B, it concerns only A and C interactions, resp. B and C interactions. The alternative model assumes a potential relationship between A and B as well.

null model alternative model

BIC (and Bayesian criterion) will be calculated for both models. The structure with higher score will be taken. At the same time, we will use the likelihood values enumerated in terms of BIC to perform the likelihood-ratio statistical test.

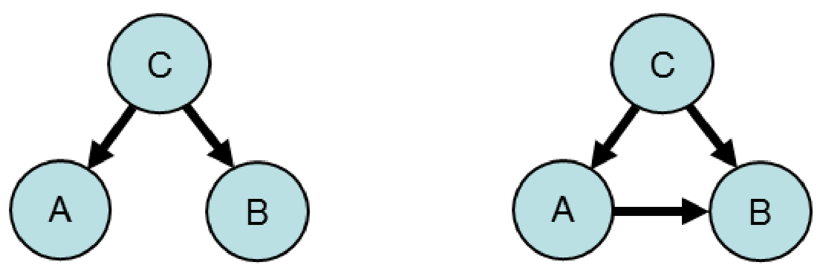
$\ln L_{null} = (39+64)\ln\frac{103}{200}\frac{100}{200} + (61+36)\ln\frac{97}{200}\frac{100}{200} = -277.2$
$\ln L_{alt} = 39\ln\frac{103}{200}\frac{39}{103} + 64\ln\frac{103}{200}\frac{64}{103} + 61\ln\frac{97}{200}\frac{61}{97} + 36\ln\frac{97}{200}\frac{36}{97} = -270.8$
$BIC(null) = -\frac{K}{2}\ln M + \ln L_{null} = -\frac{2}{2}\ln 200 - 277.2 = -282.5$
$BIC(alt) = -\frac{K}{2}\ln M + \ln L_{alt} = -\frac{3}{2}\ln 200 - 270.8 = -278.8$
$BIC(null) < BIC(alt) \Leftrightarrow$ the alternative model is more likely, the null hypothesis $A \perp\!\!\!\perp B|\emptyset$ does not hold.

Bayesian score is more difficult to compute, the evaluation was carried out in Matlab BNT (the function *score_dags*): $\ln Pr(D|null) = -282.9 < \ln Pr(D|alt) = -279.8 \Leftrightarrow$ the alternative model is more likely, the null hypothesis $A \perp\!\!\!\perp B|\emptyset$ does not hold.

The likelihood-ratio statistical test:
$D = -2(\ln L_{null} - \ln L_{alt}) = -2(-277.2 + 270.8) = 12.8$
D statistic follows χ^2 distribution with $3-2=1$ degrees of freedom. The null hypothesis has $p = 0.0003$ and we can reject it.

Conclusion: **Variables A and B are dependent.**

Analogically we will compare the null (A and B conditionally independent) and the alternative (A and B conditionally dependent) model of three variables, see the figure below.



null model alternative model

We will compare their scores, the structure with a higher score wins.

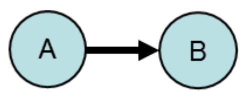
$\ln L_{null}$ and $\ln L_{alt}$ computed in Matlab BNT, the function *log_lik_complete*):
$BIC(null) = -\frac{K}{2}\ln M + \ln L_{null} = -\frac{5}{2}\ln 200 - 365.1 = -377.9$
$BIC(alt) = -\frac{K}{2}\ln M + \ln L_{alt} = -\frac{7}{2}\ln 200 - 364.3 = -382.9$
$BIC(null) > BIC(alt) \Leftrightarrow$ the null model has a higher score, the hypothesis $A \perp\!\!\!\perp B|C$ holds.

Bayesian score (carried out in Matlab BNT, the function *score_dags*): $\ln Pr(D|null) = -379.4 > \ln Pr(D|alt) = -385.5 \Leftrightarrow$ the alternative model has a lower posterior probability, the model assuming $A \perp\!\!\!\perp B|C$ will be used.

The likelihood-ratio statistical test:
$D = -2(\ln L_{null} - \ln L_{alt}) = -2(-365.1 + 364.3) = 1.6$
D statistic has χ^2 distribution with $7 - 5 = 2$ degrees of freedom. Assuming that the null hypothesis is true, the probability of observing a D value that is at least 1.6 is $p = 0.45$. As $p > \alpha$, the null hypothesis cannot be rejected.

Conclusion: **Variables A and B are conditionally independent given C.**

Exercise 12. *Let us consider the network structure shown in the figure below. Our goal is to calculate maximum likelihood (ML), maximum aposteriori (MAP) and Bayesian estimates of the parameter $\theta = Pr(b|a)$. 4 samples are available (see the table). We also know that the prior distribution of $Pr(b|a)$ is Beta(3,3).*



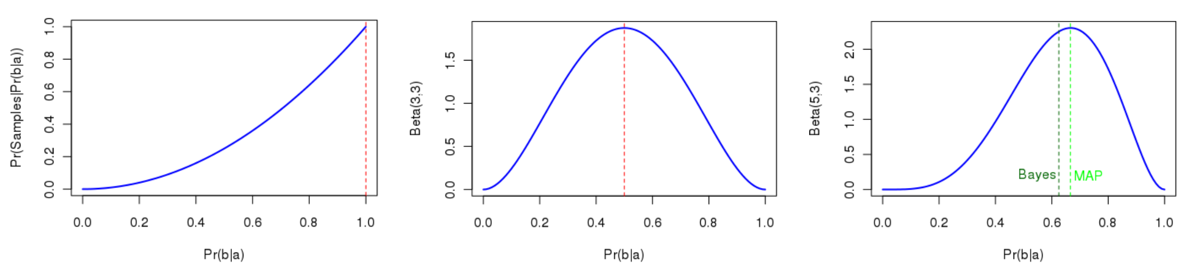
A	B
T	T
F	F
T	T
F	F

MLE of $Pr(b|a)$: $\widehat{\theta} = \arg\max_{\theta} L_B(\theta : D) = \frac{N(a,b)}{N(a)} = \frac{2}{2} = 1$
MLE sets $Pr(b|a)$ to maximize the probability of observations. It finds the maximum of function $Pr(b|a)^2(1 - Pr(b|a))^0$ shown in left graph.

MAP estimate maximizes posterior probability of parameter, it takes the prior distribution into consideration as well:
$Beta(\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$ where B plays a role of normalization constant.
The prior distribution Beta(3,3) is shown in middle graph. $Pr(b|a)$ is expected to be around 0.5, nevertheless the assumption is not strong (its strength corresponds to prior observation of four samples with positive A, two of them have positive B as well).

MAP of $Pr(b|a)$: $\widehat{\theta} = \arg\max_{\theta} Pr(\theta|D) = \frac{N(a,b)+\alpha-1}{N(a)+\alpha+\beta-2} = \frac{2}{3}$
The posterior distribution is proportional to $Pr(b|a)^4(1 - Pr(b|a))^2$, the estimated value was shifted towards the prior. See the graph to the right.

Similarly as MAP, Bayesian estimate deals with the posterior distribution $Pr(Pr(b|a)|D)$. Unlike MAP it takes its expected value. Bayesian estimation of $Pr(b|a)$:
$\widehat{\theta} = E\{\theta|D\} = \frac{N(a,b)+\alpha}{N(a)+\alpha+\beta} = \frac{5}{8}$
The expected value can be interpreted as the center of gravity of the posterior distribution shown in the graph to the right.



4 Dynamic Bayesian networks

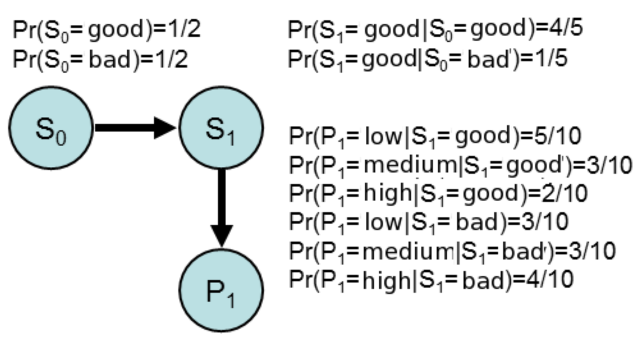
Exercise 13. *A patient has a disease N. Physicians measure the value of a parameter P to see the disease development. The parameter can take one of the following values {low, medium, high}. The value of P is a result of patient's unobservable condition/state S. S can be {good, poor}. The state changes between two consecutive days in one fifth of cases. If the patient is in good condition, the value for P is rather low (having 10 sample measurements, 5 of them are low, 3 medium and 2 high), while if the patient is in poor condition, the value is rather high (having 10 measurements, 3 are low, 3 medium and 4 high). On arrival to the hospital on day 0, the patient's condition was unknown, i.e., $Pr(S_0 = good) = 0.5$.*

a) Draw the transition and sensor model of the dynamic Bayesian network modeling the domain under consideration,

b) calculate probability that the patient is in good condition on day 2 given low P values on days 1 and 2,

c) can you determine the most likely patient state sequence in days 0, 1 and 2 without any additional computations?, justify.

ad a) The transition model describes causality between consecutive states, the sensor model describes relationship between the current state and the current evidence. See both the models in figure below:



ad b) $Pr(s_2|P_1 = low, P_2 = low)$ will be enumerated (the notation is: s good state, $\neg s$ poor state). It is a typical **filtering** task:

$$Pr(S_1|P_1 = low) = \alpha_1 Pr(P_1 = low|S_1) \sum_{S_0 \in \{s_0, \neg s_0\}} Pr(S_1|S_0)Pr(S_0)$$

$$Pr(s_1|P_1 = low) = \alpha_1 \times 0.5 \times 0.5 = 0.625$$

$$Pr(\neg s_1|P_1 = low) = \alpha_1 \times 0.3 \times 0.5 = 0.375$$

$$Pr(S_2|P_1 = low, P_2 = low) = \alpha_2 Pr(P_2 = low|S_2) \sum_{S_1 \in \{s_1, \neg s_1\}} Pr(S_2|S_1)Pr(S_1)$$

$$Pr(s_2|P_1 = low, P_2 = low) = \alpha_2 \times 0.5(0.8 \times 0.625 + 0.2 \times 0.375) = \alpha_2 \times 0.2875 = \mathbf{0.6928}$$

$$Pr(\neg s_2|P_1 = low, P_2 = low) = \alpha_2 \times 0.3(0.2 \times 0.625 + 0.8 \times 0.375) = \alpha_2 \times 0.1275 = 0.3072$$

The same task can be posed as a classical inference task:

$$
\begin{aligned}
&Pr(s_2|P_1 = low, P_2 = low) = \frac{Pr(s_2, P_1 = low, P_2 = low)}{Pr(P_1 = low, P_2 = low)} = \\
&= \frac{\sum_{S_0,S_1} Pr(S_0, S_1, s_2, P_1 = low, P_2 = low)}{\sum_{S_0,S_1,S_2} Pr(S_0, S_1, S_2, P_1 = low, P_2 = low)} = \\
&= \frac{Pr(s_0)Pr(s_1|s_0)Pr(s_2|s_1)Pr(P_1 = low|s_1)Pr(P_2 = low|s_2) + \ldots}{Pr(s_0)Pr(s_1|s_0)Pr(s_2|s_1)Pr(P_1 = low|s_1)Pr(P_2 = low|s_2) + \ldots} = \ldots
\end{aligned}
$$

ad c) No, we cannot. The most likely explanation task $Pr(S_{1:2}|P_{1:2})$ is a distinct task from filtering and smoothing. The states interact, moreover, at day 1 filtering computes $Pr(s_1|P_1 = low)$ instead of $Pr(s_1|P_1 = low, P_2 = low)$. Viterbi algorithm (a dynamic programming algorithm used in HMM) needs to be applied.