

Solutions Bayesian Networks

Exercise 1

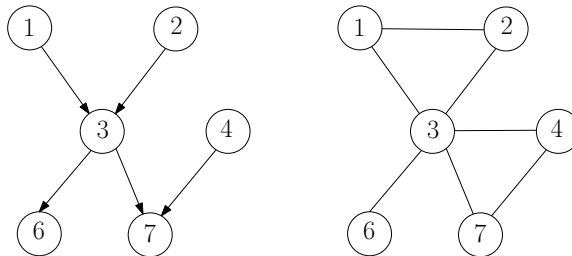
(a) Factorization:

$$\begin{aligned}
 P(X) &= \prod_{i=1}^9 P(X_i \mid X_{pa(i)}) \\
 &= P(X_1)P(X_2)P(X_3|X_1, X_2)P(X_4)P(X_5|X_2)P(X_6|X_3) \\
 &\quad P(X_7|X_3, X_4)P(X_8|X_6, X_7)P(X_9|X_5, X_7)
 \end{aligned}$$

(b) $6 \perp\!\!\!\perp 7$

To verify $X \perp\!\!\!\perp Y \mid Z$, take the directed independence graph on $\text{an}^+(X \cup Y \cup Z)$ and moralize this graph. Then you can verify the independence property in the resulting undirected graph using separation.

The directed independence graph on $\text{an}^+(\{6, 7\})$ is given left, the corresponding moral graph is given right:



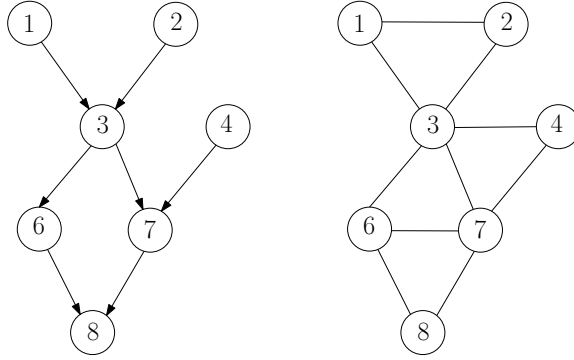
Since 6 and 7 are not separated by the empty set (there is a path between 6 and 7), they are not marginally independent.

(c) $6 \perp\!\!\!\perp 7 \mid 3$

For the graphs, see (b). Yes, every path between 6 and 7 must pass through 3.

(d) $6 \perp\!\!\!\perp 7 \mid 8$

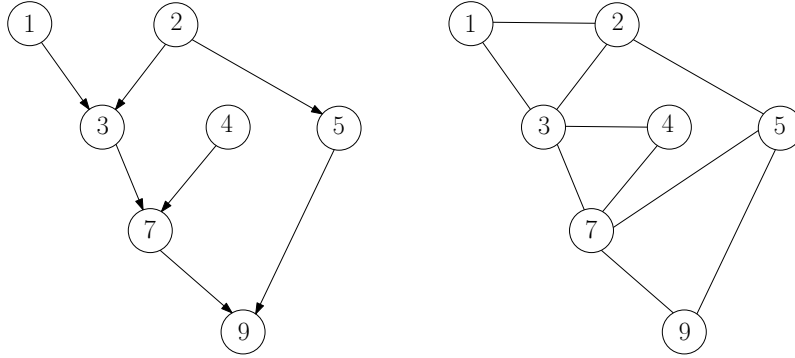
The directed independence graph on $\text{an}^+(\{6, 7, 8\})$ is given left, the corresponding moral graph is given right:



No, 8 does not separate 6 and 7 in the moral graph.

(e) $2 \perp\!\!\!\perp 9 \mid \{5, 7\}$

The directed independence graph on $\text{an}^+(\{2, 5, 7, 9\})$ is given left, the corresponding moral graph is given right:



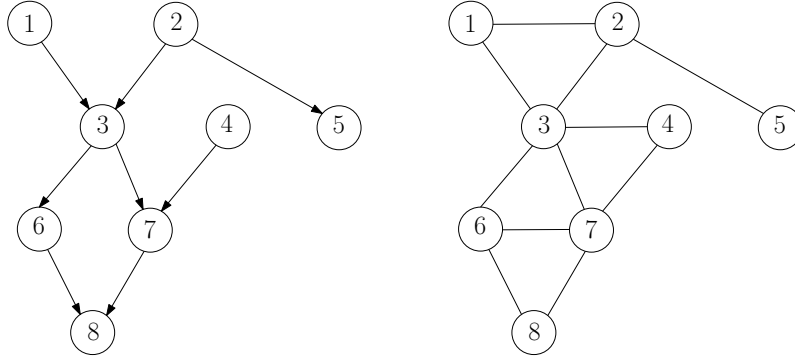
Yes: $\{5, 7\}$ separates 2 from 9, that is, every path from 2 to 9 must pass through a node in the set $\{5, 7\}$.

(f) $2 \perp\!\!\!\perp 9 \mid \{3, 5\}$

For the graphs, see (e). Yes: $\{3, 5\}$ separates 2 from 9.

(g) $5 \perp\!\!\!\perp 8$

The directed independence graph on $\text{an}^+(\{5, 8\})$ is given left, the corresponding moral graph is given right:



No, there is a path between 5 and 8.

(h) $5 \perp\!\!\!\perp 8 \mid 3$

For the graphs, see (g). Yes, 3 separates 5 from 8 in the moral graph.

Exercise 2

(a) The maximum likelihood estimates are:

$$\begin{aligned}\hat{p}_1(0) &= \frac{n_1(0)}{n} = \frac{5}{10} \\ \hat{p}_2(0) &= \frac{n_2(0)}{n} = \frac{6}{10} \\ \hat{p}_3(0) &= \frac{n_3(0)}{n} = \frac{5}{10}\end{aligned}$$

$$\begin{aligned}\hat{p}_1(1) &= \frac{n_1(1)}{n} = \frac{5}{10} \\ \hat{p}_2(1) &= \frac{n_2(1)}{n} = \frac{4}{10} \\ \hat{p}_3(1) &= \frac{n_3(1)}{n} = \frac{5}{10}\end{aligned}$$

where, for example, $\hat{p}_1(0)$ is shorthand for $\hat{p}(x_1 = 0)$.

(b) The contribution of each node (variable) to the loglikelihood score is:

Node 1: $5 \log \frac{5}{10} + 5 \log \frac{5}{10}$.

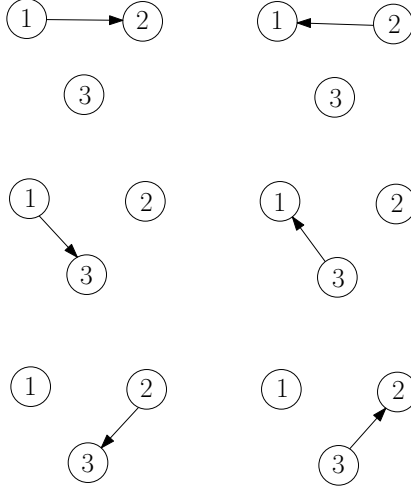
Node 2: $6 \log \frac{6}{10} + 4 \log \frac{4}{10}$.

Node 3: $5 \log \frac{5}{10} + 5 \log \frac{5}{10}$.

Hence, the total loglikelihood score is:

$$\begin{aligned}\mathcal{L} &= 5 \log \frac{5}{10} + 5 \log \frac{5}{10} + 6 \log \frac{6}{10} + 4 \log \frac{4}{10} \\ &\quad + 5 \log \frac{5}{10} + 5 \log \frac{5}{10} \approx -20.59\end{aligned}$$

(c) The neighbors are:



Pairs of models in the same row are equivalent, because moralisation does not require marrying parents, and the resulting undirected graphs are the same.

(d) No, X_1 and X_2 are independent in the data, that is, for all values x_1 of X_1 and x_2 of X_2 : $\hat{P}(x_2) = \hat{P}(x_2|x_1)$. This means that adding an edge from X_1 to X_2 does not improve the loglikelihood score. The BIC-score will go down because of the extra parameter.

(e) We compute

$$\hat{p}_{3|1}(0 | 0) = \frac{1}{5} \quad \hat{p}_{3|1}(1 | 0) = \frac{4}{5} \quad \hat{p}_{3|1}(0 | 1) = \frac{4}{5}, \quad \hat{p}_{3|1}(1 | 1) = \frac{1}{5}$$

where $\hat{p}_{3|1}(0 | 0)$ is shorthand for $\hat{p}(x_3 = 0 | x_1 = 0)$. Hence, the new contribution of node 3 to the loglikelihood score is:

$$\log \frac{1}{5} + 4 \log \frac{4}{5} + 4 \log \frac{4}{5} + \log \frac{1}{5}$$

The change in loglikelihood score is:

$$\Delta \mathcal{L} = \left(\log \frac{1}{5} + 4 \log \frac{4}{5} + 4 \log \frac{4}{5} + \log \frac{1}{5} \right) - \left(5 \log \frac{5}{10} + 5 \log \frac{5}{10} \right) \approx 1.93$$

The loglikelihood score improves by 1.93. This is at the cost of one extra parameter that costs $0.5 \log 10 = 1.15$. All in all adding an edge from X_1 to X_3 improves the BIC score by $1.93 - 1.15 = 0.78$.

Exercise 3

(a) The maximum likelihood estimates are:

$$\begin{aligned}
 \hat{p}_1(0) &= \frac{n_1(0)}{n} = \frac{4}{10} & \hat{p}_1(1) &= \frac{n_1(1)}{n} = \frac{6}{10} \\
 \hat{p}_2(0) &= \frac{n_2(0)}{n} = \frac{5}{10} & \hat{p}_2(1) &= \frac{n_2(1)}{n} = \frac{5}{10} \\
 \hat{p}_{3|12}(0|0,0) &= \frac{n_{123}(0,0,0)}{n_{12}(0,0)} = \frac{0}{2} = 0 & \hat{p}_{3|12}(1|0,0) &= \frac{n_{123}(0,0,1)}{n_{12}(0,0)} = \frac{2}{2} = 1 \\
 \hat{p}_{3|12}(0|0,1) &= \frac{n_{123}(0,1,0)}{n_{12}(0,1)} = \frac{1}{2} & \hat{p}_{3|12}(1|0,1) &= \frac{n_{123}(0,1,1)}{n_{12}(0,1)} = \frac{1}{2} \\
 \hat{p}_{3|12}(0|1,0) &= \frac{n_{123}(1,0,0)}{n_{12}(1,0)} = \frac{3}{3} = 1 & \hat{p}_{3|12}(1|1,0) &= \frac{n_{123}(1,0,1)}{n_{12}(1,0)} = \frac{0}{3} = 0 \\
 \hat{p}_{3|12}(0|1,1) &= \frac{n_{123}(1,1,0)}{n_{12}(1,1)} = \frac{1}{3} & \hat{p}_{3|12}(1|1,1) &= \frac{n_{123}(1,1,1)}{n_{12}(1,1)} = \frac{2}{3} \\
 \hat{p}_{4|3}(0|0) &= \frac{n_{34}(0,0)}{n_3(0)} = \frac{2}{5} & \hat{p}_{4|3}(1|0) &= \frac{n_{34}(0,1)}{n_3(0)} = \frac{3}{5} \\
 \hat{p}_{4|3}(0|1) &= \frac{n_{34}(1,0)}{n_3(1)} = \frac{4}{5} & \hat{p}_{4|3}(1|1) &= \frac{n_{34}(1,1)}{n_3(1)} = \frac{1}{5}
 \end{aligned}$$

where, for example, $\hat{p}_{3|12}(0|0,0)$ is shorthand for $\hat{p}(x_3 = 0|x_1 = 0, x_2 = 0)$.

(b) The loglikelihood score is:

$$\begin{aligned}
 \mathcal{L} &= \overbrace{4 \log \frac{4}{10} + 6 \log \frac{6}{10}}^{X_1} + \overbrace{5 \log \frac{5}{10} + 5 \log \frac{5}{10}}^{X_2} \\
 &\quad + 0 \log 0 + 2 \log 1 + \log \frac{1}{2} + \log \frac{1}{2} \\
 &\quad + 3 \log 1 + 0 \log 0 + \log \frac{1}{3} + 2 \log \frac{2}{3} \quad X_3 \\
 &\quad + \boxed{2 \log \frac{2}{5} + 3 \log \frac{3}{5} + 4 \log \frac{4}{5} + \log \frac{1}{5}} \quad X_4 \\
 &= -22.82450
 \end{aligned}$$

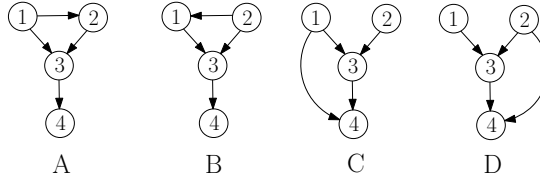
(c) Count the number of parameters per node (variable) as follows. Suppose a node has k different parent configurations (possible value assignments to its parents), and it can take on m different values itself. Then the number of parameters associated with that node is $k(m-1)$ because you have to estimate k different conditional distributions, and each conditional distribution requires the estimation of $m-1$ probabilities. If a node doesn't have any parents, then the number of parameters associated with it is $m-1$. Specified per node, the number of parameters is therefore:

- Node 1: 1.
- Node 2: 1.
- Node 3: $4 \times 1 = 4$.
- Node 4: $2 \times 1 = 2$.

Hence, the BIC score is:

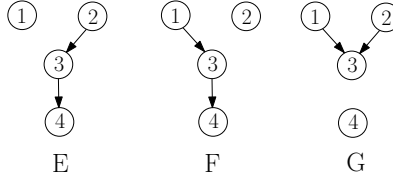
$$-22.82450 - 1.15 (1 + 1 + 4 + 2) = -32.02450$$

(d) Adding an arc:

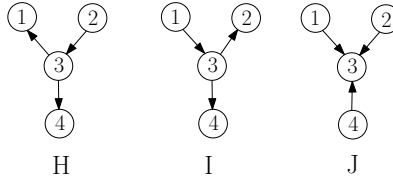


A and B are equivalent.

Removing an arc:



Reversing an arc:



H and I are equivalent.

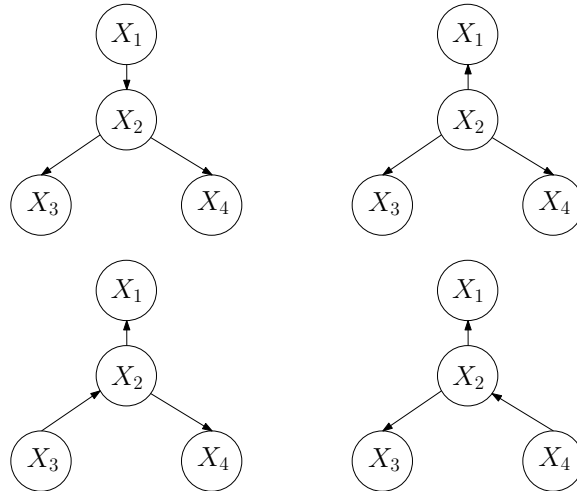
- (e) The parent set of X_4 changes so we have to recompute the part of the score corresponding to this node. The boxed part of the loglikelihood under (b) is replaced by

$$2 \log \frac{2}{4} + 2 \log \frac{2}{4} + \log \frac{1}{2} + \log \frac{1}{2} \approx -4.16,$$

where we left out all the terms that evaluate to zero. The boxed part under (b) evaluates to -5.86 so the loglikelihood increases by 1.7. This is however at the cost of two extra parameters, that cost 1.15 each, so all in all addition of an arc from X_1 to X_4 decreases the BIC score. Hence it is not preferred to the current model.

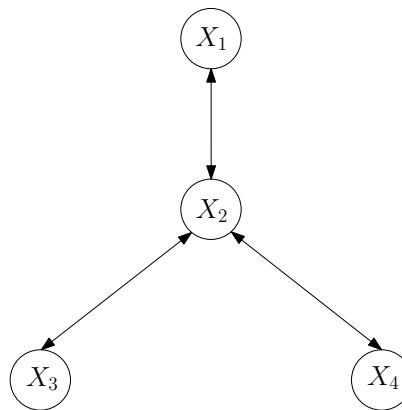
Exercise 4: Essential Graph

The equivalence class is:



These four graphs all have the same skeleton and the same set of v-structures (in this case: none).

The essential graph is:



All edges are bi-directional because each edge occurs in opposite directions in different members of the equivalence class.

Exercise 5: Structure Learning

We need to compute $\Delta \text{ Score}(\text{add}(B \rightarrow D))$, $\Delta \text{ Score}(\text{add}(C \rightarrow D))$, $\Delta \text{ Score}(\text{add}(E \rightarrow D))$. You may also mention $\Delta \text{ Score}(\text{remove}(A \rightarrow D))$, and $\Delta \text{ Score}(\text{reverse}(A \rightarrow D))$, although the first returns to the initial model, and the second leads to a model that is equivalent to the current model (and therefore has the same BIC-score).

In general: you need to compute the Δ scores for operations (addition, removal, reversal) that change the parent set of node D , because the parent set of node D has changed in the previous step. The other Δ scores are the same as in the previous step, and can therefore be retrieved from memory.

Exercise 6: Maximum Likelihood Estimation

The loglikelihood function is:

$$\mathcal{L} = \sum_{i=1}^k \left\{ \sum_{x_{pa(i)}} n(x_i = 0, x_{pa(i)}) \log p(x_i = 0 \mid x_{pa(i)}) + n(x_i = 1, x_{pa(i)}) \log(1 - p(x_i = 0 \mid x_{pa(i)})) \right\}$$

The partial derivative of \mathcal{L} with respect to $p(x_j = 0 \mid x_{pa(j)})$ is:

$$\frac{\partial \mathcal{L}}{\partial p(x_j = 0 \mid x_{pa(j)})} = \frac{n(x_j = 0, x_{pa(j)})}{p(x_j = 0 \mid x_{pa(j)})} - \frac{n(x_j = 1, x_{pa(j)})}{1 - p(x_j = 0 \mid x_{pa(j)})}$$

Let's introduce some shorthand to simplify notation. Let

- $n_0 = n(x_j = 0, x_{pa(j)})$
- $n_1 = n(x_j = 1, x_{pa(j)})$
- $p_0 = p(x_j = 0 \mid x_{pa(j)})$

The partial derivative of \mathcal{L} with respect to p_0 is:

$$\frac{\partial \mathcal{L}}{\partial p_0} = \frac{n_0}{p_0} - \frac{n_1}{1 - p_0}$$

Equate to zero and solve for p_0 to get

$$p_0 = \frac{n_0}{n_0 + n_1}$$