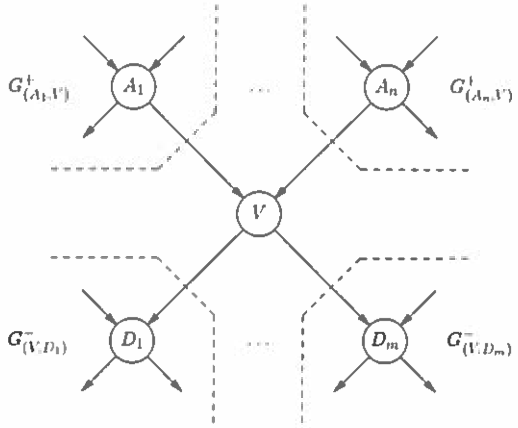
Formulas (Probabilistic Reasoning Exam Nov 2015)

Pearl in a singly connected digraph

Consider a Bayesian network $B = (G, \Gamma)$ and a node V in G with direct ancestors (parents) $\rho(V) = \{A_1, \ldots, A_n\}$ and direct descendants (children) $\sigma(V) = \{D_1, \ldots D_m\}$.



To compute its (prior or posterior) probability distribution with Pearl's algorithm, node V uses *data fusion*: $\alpha \cdot \pi(V) \cdot \lambda(V)$ and computes the following parameters for all c_V and c_{A_i}:

$$
\begin{aligned}
\pi(V) &= \sum_{c_{\rho(V)}} \Big(\gamma(V \mid c_{\rho(V)}) \cdot \prod_{i=1,\ldots,n} \pi_V^{A_i}(c_{A_i}) \Big) \\
\lambda(V) &= \prod_{j=1,\ldots,m} \lambda_{D_j}^V(V) \\
\pi_{D_j}^V(V) &= \alpha \cdot \pi(V) \cdot \prod_{\substack{k=1,\ldots,m \\ k \neq j}} \lambda_{D_k}^V(V) \\
\lambda_V^{A_i}(A_i) &= \alpha \cdot \sum_{c_V} \lambda(c_V) \cdot \sum_{c_{\rho(V)\setminus\{A_i\}}} \Big(\gamma(c_V \mid c_{\rho(V)\setminus\{A_i\}} \wedge A_i) \cdot \prod_{\substack{k=1,\ldots,n \\ k \neq i}} \pi_V^{A_k}(c_{A_k}) \Big)
\end{aligned}
$$

The MDL quality measure

Let $G = (\mathbf{V}_G, \mathbf{A}_G)$ be an acyclic digraph and let $\mathbf{D}$ be a dataset over N cases. Let $P(G)$ be a probability distribution over the set of acyclic graphs with node set $\mathbf{V}$. Then, the MDL quality measure for graph G is given by

$$
\begin{aligned}
Q_{MDL}(G, \mathbf{D}) &= \log P(G) - N \cdot H(G, \mathbf{D}) - \frac{1}{2} \log(N) \cdot \sum_{V_i \in V} 2^{|\rho(V_i)|} \\
&= \log P(G) + \sum_{V_i \in V} q(V_i, \rho(V_i), \mathbf{D})
\end{aligned}
$$

where $q(V_i, \rho(V_i), \mathbf{D})$ is the quality of node V_i and

$$-N \cdot H(G, \mathbf{D}) = \sum_{V_i \in V} \sum_{c_{V_i}} \sum_{c_{\rho(V_i)}} N(c_{V_i} \wedge c_{\rho(V_i)}) \cdot \log\Big(\frac{N(c_{V_i} \wedge c_{\rho(V_i)})}{N(c_{\rho(V_i)})} \Big) \qquad (0 \cdot \log \frac{0}{x} = 0, \text{ even if } x = 0)$$

Problem 3 (a: 10 pts., b: 10 pts., c: 15 pts., total: 35 pts.)

Suppose a dataset $\mathbf{D}$ is used for automated construction of a Bayesian network and that the learning algorithm implements the B search heuristic and the MDL quality measure (see attached formula sheet).

a. One of the assumptions underlying automated construction from a dataset is that the dataset does not contain *missing values*. In reality, databases typically contain many missing values. One approach to cope with this problem, as mentioned in class, is to use EM (expectation maximisation) to 'fill in the gaps'.

   Another approach to cope with missing values is to do a so-called *available case analysis*: use only those cases from the dataset that state values for all variables. Give a benefit and a drawback of this approach. Explain your answers.

Suppose we now have the following dataset $\mathbf{D}$ over the set $\mathbf{V} = \{V_1, V_2, V_3\}$ of binary statistical variables:

$\neg v_1 \wedge \neg v_2 \wedge \neg v_3$	$\neg v_1 \wedge \neg v_2 \wedge \neg v_3$	$\neg v_1 \wedge \neg v_2 \wedge v_3$
$v_1 \wedge \neg v_2 \wedge v_3$	$\neg v_1 \wedge v_2 \wedge v_3$	$\neg v_1 \wedge \neg v_2 \wedge \neg v_3$
$\neg v_1 \wedge \neg v_2 \wedge \neg v_3$	$\neg v_1 \wedge v_2 \wedge v_3$	$\neg v_1 \wedge \neg v_2 \wedge \neg v_3$
$\neg v_1 \wedge \neg v_2 \wedge \neg v_3$	$\neg v_1 \wedge v_2 \wedge \neg v_3$	$\neg v_1 \wedge v_2 \wedge \neg v_3$
$v_1 \wedge \neg v_2 \wedge \neg v_3$	$\neg v_1 \wedge \neg v_2 \wedge \neg v_3$	$\neg v_1 \wedge v_2 \wedge \neg v_3$

We will now consider the quality of Markov equivalent graphs. Two graphs are *Markov equivalent* if they both have the *same*

1. underlying structure (i.e. the same pairs of variables are connected), *and*
2. set of immoralities (i.e. exactly the same set of head-to-head connections $V_i \rightarrow V_j \leftarrow V_k$ where V_i and V_k are not directly connected ('unmarried'))

b. Suppose that during the learning process, a graph G_I is considered with (only) the following arcs:

$$V_1 \rightarrow V_2 \quad \text{and} \quad V_2 \rightarrow V_3$$

   Give a Markov equivalent graph G_{II} and show that the set of joint distributions $\Pr_I(V_1 \wedge V_2 \wedge V_3)$ that can be represented with G_I is identical to the set of joint distributions $\Pr_{II}(V_1 \wedge V_2 \wedge V_3)$ that can be represented with G_{II}.

   Hint: consider the factorisations of the different joint distributions.

c. Prove or provide a counter example for the following statement:

   Two Bayesian networks with Markov equivalent graphs, learned from the same dataset $\mathbf{D}$, have the same quality Q_{MDL}.