Bayesian Networks: An Introduction
Answers to Selected Exercises

John M. Noble,
Institute of Applied Mathematics and Mechanics
University of Warsaw
ul. Banacha 2
02-097 Warszawa, Poland

Contents

Chapter 1

1. (a)

$$\begin{aligned}\alpha(A,B) &= p(AB) - p(A)p(B) = p(A) - p(AB^c) - p(A)(1 - p(B^c)\\ &= -p(AB^c) + p(A)p(B^c) = -\alpha(A,B^c)\end{aligned}$$

   (b) Clearly $\alpha(A,B) = \alpha(B,A)$ for any A and B. It follows from the first part that

$$\alpha(A,B) = -\alpha(A,B^c) = -\alpha(B^c,A) = \alpha(B^c,A^c) = \alpha(A^c,B^c)$$

2. (a)

$$p(A|B) = \frac{p(AB)}{p(B)} = \frac{p(A) - p(AB^c)}{p(B)} = \frac{p(A) - p(A|B^c)p(B^c)}{p(B)}$$

   so $p(A|B^c) < p(A)$ implies

$$p(A|B) > \frac{p(A)(1 - p(B^c)}{p(B)} = P(A)$$

   The result is now clear.

   (b) If B is a prima facie cause then

$$\begin{aligned}p(A^c|B^c) &= \frac{1 - p(A \cup B)}{p(B^c)} = \frac{1 - p(A) - p(B) + p(AB)}{p(B^c)}\\ &= 1 - \frac{p(A) - p(A|B)p(B)}{p(B^c)} \geq 1 - \frac{p(A)(1 - p(B))}{p(B^c)} = p(A^c)\end{aligned}$$

   B^c is a prima facie cause of A^c.

   (c) If B is a prima facie cause then

$$\alpha(A,B) = p(AB) - p(A)p(B) = (p(A|B) - p(A))p(B) > (p(A) - p(A))p(B) = 0.$$

   Similarly, if it is a negative cause

$$\alpha(A,B) = p(AB) - p(A)p(B) = (p(A|B) - p(A))p(B) < (p(A) - p(A))p(B) = 0.$$

3.

$$\begin{aligned} W(A:E|B) &= \log \frac{O_p(A|BE)}{O_p(A|B)} = \log \frac{p(A|BE)}{p(A^c|BE)} \frac{p(A^c|B)}{p(A|B)} \\ &= \log \frac{p(ABE)p(BE)}{p(BE)p(A^cBE)} \frac{p(A^cB)p(B)}{p(B)p(AB)} = \log \frac{p(ABE)p(A^cB)}{p(A^cBE)p(AB)} = \log \frac{p(E|AB)}{p(E|A^cB)} \end{aligned}$$

4. Typo in question: should be G, not B.

$$\begin{aligned} W(H_1/H_2:E|G) &= \log \frac{O_p(H_1/H_2|GE)}{O_p(H_1/H_2|G)} = \log \frac{p(H_1|GE)p(H_2|G)}{p(H_2|GE)p(H_1|G)} \\ &= \log \frac{p(H_1GE)p(GE)p(H_2G)p(G)}{p(GE)p(H_2GE)p(G)p(H_1G)} = \log \frac{p(E|H_1G)}{p(E|H_2G)}. \end{aligned}$$

5. (a) Directly from equation (1.32) and (1.28),

$$Q_{\text{suff}}(E:F|U) = W(F^c:E^c|U) = \log \frac{p(E^c|F^cU)}{p(E^c|FU)}$$

as required.

(b) Directly from equation (1.33) and (1.28),

$$Q_{\text{nec}}(E:F|U) = W(F:E|U) = \log \frac{p(E|FU)}{p(E|F^cU)}$$

as required.

6. (a)

$$\begin{aligned} Q_{\text{suff}}(E:F|U) < 0 &\Leftrightarrow p(E^c|F^cU) < p(E^c|FU) \Leftrightarrow 1 - p(E|F^cU) < 1 - p(E|FU) \\ &\Leftrightarrow p(E|FU) < p(E|F^cU) \Leftrightarrow Q_{\text{necc}}(E:F|U) < 0. \end{aligned}$$

(b) Since $(E^c)^c = E$,

$$Q_{\text{suf}}(E^c:F_1/F_2|U) = W(F_1/F_2:E|U) = Q_{\text{nec}}(E:F_1/F_2|U)$$

(c)

$$Q_{\text{suff}}(E^c:F^c|U) = \log \frac{p(E|FU)}{p(E|F^cU)} = Q_{\text{nec}}(E:F|U).$$

7. (a) De-Finetti's result states that if the X_js are exchangeable, then there is a Q and π such that

$$p(\mathbf{X}^{(n)} = \mathbf{x}^{(n)}) = \int_\Theta \prod_{j=1}^n Q(X_i = x_i|\theta)\pi(\theta)d\theta.$$

For Bernoulli trials with parameter θ, $Q(X_i = 1|\theta) = \theta$ and $Q(X_i = 0|\theta) = 1 - \theta$, so that

$$p(\mathbf{X}^{(n)} = \mathbf{x}^{(n)}) = \int_\Theta \theta^t (1-\theta)^{n-t} \pi(\theta)d\theta,$$

where $t = \sum_{j=1}^{n} x_j$. It follows that

$$p(T = t) = \sum_{\mathbf{x}^{(n)} | \sum_{j=1}^{n} x_j = t} p(\mathbf{X}^{(n)} = \mathbf{x}^{(n)}) = \left(\begin{array}{c} n \\ t \end{array} \right) \int_0^1 \theta^t (1 - \theta)^{n-t} \pi(\theta) d\theta,$$

as required.

(b) Let $Y \sim Bi(n, \theta)$. Then clearly

$$E[T] = \int_0^1 E[Y|\theta] \pi(\theta) d\theta = n \int_0^1 \theta \pi(\theta) d\theta$$

as required.

8.

$$\pi_{\Theta|k}(\theta|k) = \frac{\pi_\Theta(\theta) p(X = k|\theta)}{p(X = k)}.$$

The prior is $\pi_\Theta(\theta) \equiv 1$, while $p(X = k|\theta) = \left(\begin{array}{c} n \\ k \end{array} \right) \theta^k (1 - \theta)^{n-k}$, so that

$$\pi_{\Theta|k}(\theta|k) = (\text{Const}) \theta^k (1 - \theta)^{n-k}.$$

Since $\int_0^1 \pi_{\Theta|k}(\theta|k) d\theta = 1$, it follows that

$$\text{Const} = \frac{1}{\int_0^1 \theta^k (1 - \theta)^{n-k} d\theta} = \frac{(n + 1)!}{k!(n - k)!}$$

The integral is the standard beta integral and is computed in the lecture notes.

9. Suppose $\mathbf{x}$ contains k heads. Then

$$\begin{aligned} p(X_{n+1} = H|\mathbf{x}) &= \int_0^1 p(X_{n+1} = H|\theta, \mathbf{x}) \pi(\theta|\mathbf{x}) d\theta \\ &= \frac{(n + 1)!}{k!(n - k)!} \int_0^1 \theta^{k+1} (1 - \theta)^{n-k} d\theta \\ &= \frac{(n + 1)!}{k!(n - k)!} \frac{(k + 1)!(n - k)!}{(n + 2)!} = \frac{k + 1}{n + 2}. \end{aligned}$$

10.

$$\pi_{\Theta|\mathbf{x}}(\theta|\mathbf{x}, \alpha) = (\text{Const}) \pi_\Theta(\theta|\alpha) p(\mathbf{X} = \mathbf{x}|\theta, \alpha)$$

and

$$p(\mathbf{X} = \mathbf{x}|\theta, \alpha) = (\text{Const}) \theta_1^{n_1} \ldots \theta_L^{n_L},$$

where the vector $\mathbf{x}$ has n_j outcomes in category j, $n_1 + \ldots + n_L = n$. It follows that

$$\pi_{\Theta|\mathbf{x}}(\theta|\mathbf{x}, \alpha) = (\text{Const}) \prod_{j=1}^{L} \theta_j^{n_j + \alpha q_j - 1}.$$

The constant may be obtained by noting that the expression in the exercise defines a probability density function.

11.

$$
\begin{aligned}
p(X_{n+1} = x_i|\mathbf{x}) &= \int_0^1 p(X_{n+1} = x_i|\theta, \mathbf{x})\pi_{\Theta|\mathbf{x}}(\theta|\mathbf{x})d\theta \\
&= \frac{n+\alpha}{\prod_{j=1}^L \Gamma(\alpha q_j + n_j)} \int_0^1 (\prod_{j\neq i} \theta_j^{n_j+\alpha q_j-1})\theta_i^{n_i+\alpha q_i} d\theta \\
&= \frac{\Gamma(n+\alpha)}{\prod_{j=1}^L \Gamma(n_j + \alpha q_j)} \frac{(\prod_{j\neq i} \Gamma(n_j + \alpha q_j))\Gamma(n_i + 1 + \alpha q_i)}{\Gamma(n+\alpha+1)} \\
&= \frac{\Gamma(n_i + 1 + \alpha q_i)\Gamma(n+\alpha)}{\Gamma(n_i + \alpha q_i)\Gamma(n+\alpha+1)} \\
&= \frac{n_i + \alpha q_i}{n+\alpha}.
\end{aligned}
$$

12. Let $\alpha = \alpha_1 + \ldots + \alpha_L$. The quantity $E[\theta_i]$ was calculated in previous exercise.

$$E[\theta_i^2] = \frac{\Gamma(n+\alpha)\Gamma(n_i + \alpha_i + 2)}{\Gamma(n+\alpha+2)\Gamma(n_i + \alpha_i)} = \frac{(n_i + \alpha_i + 2)(n_i + \alpha_i + 1)}{(n+\alpha+2)(n+\alpha+1)}$$

and

$$E[\theta_i] = \frac{n_i + \alpha_i + 1}{n+\alpha+1}.$$

This gives

$$\text{Var}(\theta_i) = \frac{n - n_i + \alpha - \alpha_i}{(n+\alpha+1)(n+\alpha+2)}.$$

13. Using the evaluation of the Beta integral,

$$
\begin{aligned}
&p(X_{n+1} = 1|X_1 + \ldots + X_n = s) = \int_0^1 p(X_{n+1} = 1|X_1 + \ldots + X_n = s, \theta)\pi(\theta)d\theta \\
&= \int_0^1 p(X_{n+1} = 1|X_1 + \ldots + X_n = s, \theta)\pi(\theta|X_1 + \ldots + X_n = s)d\theta \\
&= \int_0^1 \theta \frac{p(X_1 + \ldots + X_n = s|\theta)\pi(\theta)}{\int p(X_1 + \ldots + X_n = s|\theta)\pi(\theta)d\theta} d\theta \\
&= \frac{\begin{pmatrix} n \\ s \end{pmatrix} \int_0^1 \theta^{s+1}(1-\theta)^{n-s}d\theta}{\begin{pmatrix} n \\ s \end{pmatrix} \int_0^1 \theta^{s}(1-\theta)^{n-s}d\theta} \\
&= \frac{\frac{n!}{s!(n-s)!}\frac{(s+1)!(n-s)!}{(n+2)!}}{\frac{n!}{s!(n-s)!}\frac{s!(n-s)!}{(n+1)!}} \\
&= \frac{s+1}{n+2}
\end{aligned}
$$

14. The free variables are $(v_1, \dots, v_{K-1})$ with the constraint $v_K = 1 - \sum_{j=1}^{K-1} v_j$. Set

$$S = \sum_{j=1}^{K} \frac{v_j}{x_j} = \frac{1}{x_K} + \sum_{j=1}^{K-1} v_j \left(\frac{1}{x_j} - \frac{1}{x_K} \right)$$

then

$$u_j = \frac{v_j}{x_j S} \qquad j = 1, \dots, K \quad \text{and} \quad u_K = 1 - \sum_{j=1}^{K-1} u_j.$$

and, since $\sum_{j=1}^{K} v_j = 1$, it follows that $1 = \sum_{j=1}^{K} v_j = S \sum_{j=1}^{K} x_j u_j$ so that

$$S = \frac{1}{x_K + \sum_{j=1}^{K-1} (x_j - x_K) u_j} = \frac{1}{\sum_{j=1}^{K} x_j u_j}$$

The original density (in terms of the free variables) is

$$\frac{\Gamma\left(\sum_{j=1}^{k} a_j\right)}{\prod_{j=1}^{K} \Gamma(a_j)} \left(\prod_{j=1}^{K-1} v_j^{a_j - 1} \right) \left(1 - \sum_{j=1}^{K-1} v_j \right)^{a_K - 1}.$$

The Jacobian determinant for $\underline{v} \to \underline{u}$ may be computed by noting that $v_j = u_j x_j S$ and using

$$\frac{\partial S}{\partial u_\alpha} = -S^2 (x_\alpha - x_K)$$

so that

$$\frac{\partial v_i}{\partial u_\alpha} = \begin{cases} -S u_i x_i (x_\alpha - x_K) & \alpha \neq i \\ S x_i - S u_i x_i (x_i - x_K) & \alpha = i \end{cases}$$

The matrix of which the determinant is to be computed is therefore $S^{K-1} \prod_{i=1}^{K-1} x_i M$, where

$$M = I - S \begin{pmatrix} u_1 \\ \vdots \\ u_{K-1} \end{pmatrix} (x_1 - x_K, \dots, x_{K-1} - x_K).$$

Clearly 1 is an eigenvalue of multiplicity $K - 2$ for M. The remaining eigenvalue λ of M may be computed by noting that the vector $\underline{e}$ that satisfies

$$(M - \lambda) \underline{e} = 0$$

satisfies $\underline{e} = c \begin{pmatrix} u_1 \\ \vdots \\ u_{K-1} \end{pmatrix}$ and therefore λ satisfies

$$1 - \lambda = S \sum_{j=1}^{K-1} u_j (x_j - x_K) = S(\sum_{j=1}^{K-1} u_j x_j - x_K + x_K u_K) = S(\frac{1}{S} - x_K)$$

so that $\lambda = Sx_K$. It follows that the density in the new coordinates is

$$\frac{\Gamma\left(\sum_{j=1}^{k} a_j\right)}{\prod_{j=1}^{K}\Gamma(a_j)}\left(\prod_{j=1}^{K-1}(Sx_ju_j)^{a_j-1}\right)\left(1-S\sum_{j=1}^{K-1}x_ju_j\right)^{a_K-1} S^K\prod_{j=1}^{K}x_j.$$

Since $Sx_Ku_K = 1 - S\sum_{j=1}^{K-1}x_ju_j$, it follows that

$$\frac{\Gamma\left(\sum_{j=1}^{k} a_j\right)}{\prod_{j=1}^{K}\Gamma(a_j)}\prod_{j=1}^{K}x_j^{a_j}\left(\prod_{j=1}^{K}u_j^{a_j-1}\right)\left(\frac{1}{\sum_{j=1}^{K}x_ju_j}\right)^{\sum_{j=1}^{K}a_j}$$

as required.

15. The distribution was computed in the previous exercise; the remainder is a straightforward application of Bayes rule.

$$U_i = p(\{X=i\}|\text{data}) = \frac{p_X(i)l_i}{p(\text{data})} = \frac{V_il_i}{\sum_{i=1}^{K}V_il_i}$$

the denominator follows because $\sum_{i=1}^{K}U_i = 1$. The distribution of $\underline{U}$ now satisfies the definition of the $S(\underline{a}, \underline{l}^{-1})$ distribution of the previous exercise.

16. Again, assume data is obtained and the likelihood is $l_i = p(\text{data}|X=i)$ and the prior distribution is $S(\underline{a}, \underline{z})$. Then

$$U_i = p(\{X=i\}|\text{data}) = \frac{V_il_i}{p(\text{data})} = \frac{W_iz_i^{-1}l_i}{p(\text{data})\sum_{i=1}^{K}W_iz_i^{-1}},$$

where $\underline{W} \sim \text{Dir}(a_1,\ldots,a_K)$. Since $\sum_{i=1}^{K}U_i = 1$, it follows that

$$U_i = \frac{W_iz_i^{-1}l_i}{\sum_{i=1}^{K}W_iz_i^{-1}l_i}$$

so that the distribution of $\underline{U}$ satisfies the definition of a $S(\underline{a}, \underline{z}\times\underline{l}^{-1})$ distribution.

Chapter 2

1. Using characterisations of conditional independence:

   (a) $X \perp (Y, W)|Z$ means $p_{W,X,Y,Z}(w, x, y, z) = p_{X|Z}(x|z)p_{W,Y|Z}(w, y|z)p_Z(z)$ Summing over W gives $p_{X,Y,Z}(x, y, z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)p_Z(z)$; equivalent to $X \perp Y|Z$.

   Similarly, summing over Y gives $p_{W,X,Z}(w, x, z) = p_{X|Z}(x|z)p_{W|Z}(w|z)p_Z(z)$, equivalent to $X \perp W|Z$.

   (b) $X \perp Y|Z$ implies $p_{X,Y,Z}(x, y, z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)p_Z(z)$ and $X \perp W|(Y, Z)$ implies $p_{W,X,Y,Z}(w, x, y, z) = p_{X|Y,Z}(x|y, z)p_{W|Y,Z}(w|y, z)p_{Y,Z}(y, z)$. The first statement implies that for (x, y, z) such that $p_{X,Y,Z}(x, y, z) > 0$, $p_{X|Y,Z} = p_{X|Z}$, so, using $p_{Y,Z} = p_{Y|Z}p_Z$, it follows that

   $$\begin{aligned} p_{W,X,Y,Z}(w, x, y, z) &= p_{X|Z}(x|z)p_{W|Y,Z}(w|y, z)p_{Y|Z}(y|z)p_Z(z) \\ &= p_{X|Z}(x|z)p_{W,Y|Z}(w, y|z)p_Z(z), \end{aligned}$$

   so that $X \perp (W, Y)|Z$.

   (c) (Trick question) answer: the result stated is false. Counterexample: any distribution that factorises as

   $$p_Z p_{X|Z} p_{W|Z} p_{Y|W,Z}$$

   clearly satisfies $X \perp Y|(Z, W)$ and $X \perp W|(Y, Z)$, but there are distributions with such a factorisation that do not satisfy $X \perp Z|(Y, W)$.

2. Required factorisation is

   $$p_{A,B,C} = p_A p_{B|A} p_{C|B}$$

   From the joint probability table,

   $$p_A = (0.1, 0.9)$$

   $$p_{A,B} = \begin{array}{c|cc} A\backslash B & 0 & 1 \\ \hline 0 & 0.07 & 0.03 \\ 1 & 0.18 & 0.72 \end{array}$$

Division of tables $p_{B|A} = \frac{p_{A,B}}{p_A}$ yields

$$p_{B|A} = \begin{array}{c|cc} A\backslash B & 0 & 1 \\ \hline 0 & 0.7 & 0.3 \\ 1 & 0.2 & 0.8 \end{array}$$

$$p_B = (0.25, 0.75)$$

$$p_{B,C} = \begin{array}{c|cc} B\backslash C & 0 & 1 \\ \hline 0 & 0.1 & 0.15 \\ 1 & 0.00075 & 0.74925 \end{array}$$

$$p_{C|B} = \begin{array}{c|cc} B\backslash C & 0 & 1 \\ \hline 0 & 0.4 & 0.6 \\ 1 & 0.001 & 0.999 \end{array}$$

and now

$$p_A p_{B|A} p_{C|B} = \begin{array}{c|cc} A\backslash B & 0 & 1 \\ \hline 0 & (0.028, 0.042) & (0.00003, 0.02997) \\ 1 & (0.072, 0.108) & (0.00072, 0.71928) \end{array}$$

where the entries correspond to $C = (0, 1)$. This is the original distribution $p_{A,B,C}$, so the factorisation holds.

3.

$$p_A = (0.2, 0.4, 0.4)$$

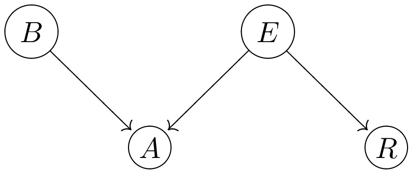
$$p_B = (0.17, 0.18, 0.65)$$

Division of tables:

$$p_{A|B} = \frac{p_{A,B}}{p_B}$$

$$p_{A|B} = \begin{array}{c|ccc} & a_1 & a_2 & a_3 \\ \hline b_1 & \frac{2}{17} & \frac{10}{17} & \frac{5}{17} \\ b_2 & \frac{1}{6} & 0 & \frac{5}{6} \\ b_3 & \frac{3}{13} & \frac{6}{13} & \frac{4}{13} \end{array}$$

$$p_{B|A} = \begin{array}{c|ccc} & b_1 & b_2 & b_3 \\ \hline a_1 & 0.10 & 0.15 & 0.75 \\ a_2 & 0.25 & 0.00 & 0.75 \\ a_3 & 0.125 & 0.375 & 0.50 \end{array}$$



4. (a) B - burglary yes/no; E - earth tremor yes / no; A - alarm yes / no. R - radio broadcast yes / no

   (b) i.

$$p_{B|A}(1|1) = \frac{p_{A|B}(1|1)p_B(1)}{p_A(1)}$$

$$p_{A|B}(1|1) = p_{A|B,E}(1|1,0)p_E(0) + p_{A|B,E}(1|1,1)p_E(1) = 0.97997$$

$$p(A=1|B=0) = 0.94908$$

$$p(A=1) = 0.94939$$

$$p(B=1|A=1) = .99967$$

   ii.

$$\begin{aligned} & p_{B|A,R}(1|1,1) \\ & = \frac{p_{B,A,R}(1,1,1)}{p_{A,R}(1,1)} = \frac{p_B(1)\sum_E p_E p_{A|B,E}(1|1,.)p_{R|E}(1|.)}{\sum_B p_B \sum_E p_E p_{A|B,E}(1|.,.)p_{R|E}(1|.)} \\ & = \frac{0.01(0.999 \times 0.98 \times 0.95 + 0.001 \times 0.95 \times 0.01)}{k} \end{aligned}$$

   where

$$\begin{aligned} k = & \; 0.01(0.999 \times 0.98 \times 0.95 + 0.001 \times 0.95 \times 0.01) \\ & +0.99(0.999 \times 0.95 \times 0.95 + 0.001 \times 0.03 \times 0.01) \end{aligned}$$

   so

$$p_{B|A,R}(1|1,1) = 0.01$$

   (This is a good exercise to try out HUGIN - the HUGIN LITE version works well).

5. (a) The variables D and G are d-separated from A instantiated chain connection at B; instantiated fork at J, uninstantiated collider at I - so no possibility to pass the Bayes ball through).

   The variables B, D and G are d-separated from F (uninstantiated collider at A prevents the ball from getting through to B - instantiated fork at J, uninstantiated collider at I also prevent the ball passing).

(b) C and F are d- separated from A. The ball passes through G; there is information on G; J is a descendant. The ball cannot pass through I which is an uninstantiated collider with no instantiated descendants.

All variables except for C and I are d separated from F. The Bayes Ball cannot pass through I, an uninstantiated collider node with no instantiated descendants.

6. Recall definition of *Markov blanket*; parents of A, children of A and any variables sharing a child with A. Consider the 'Bayes Ball' algorithm, started at A. The ball cannot travel through an instantiated chain or fork connection, nor can it travel through a collider, where none of the descendants are instantiated. Otherwise, it can travel through a node along the graph.

   Therefore: if all variables in the Markov blanket are instantiated, Bayes ball cannot pass through any of the parents (by definition, the connection is necessarily chain or fork). It cannot pass through a child to any offspring of the child (the connection necessarily chain). If it passes through an instantiated child to another parent of the instantiated child, it cannot pass further: connection at the point of the instantiated parent of the instantiated child is either chain or fork.

7. (a) Graph 1: no instantiation gives no separation. If A is instantiated, then B and C are separated.

   Graph 2: no instantiation gives no separation. If A is instantiated then B and C are separated.

   Graph 3: no instantiation gives no separation. If A is instantiated, then B and C are separated.

   Graph 4: no instantiation means that B and C are *separated*. If A is instantiated, then B and C are *connected*.

   All separations have been listed.

   Therefore: Graphs 1, 2 and 3 are I- equivalent to each other. Graph 4 is not I equivalent to 1,2 or 3. Neither is it a submap of 1,2 or 3. Neither are 1,2 or 3 a submap of 4.

   (b) Directly from first part, look at the graphs corresponding to the factorisations of p, q and r. It is straightforward to show that, in the case of three variables, these graphs give the *whole* independence structure. p corresponds to graph 1, q to graph 2, while r corresponds to graph 4. Therefore p and q have the same independence structure; the independence structure of r is different.

8. (a) The DAG gives $p_{A,B,C} = p_A p_{B|A} p_{C|A}$. Note that $p_{A,C} = p_A p_{C|A}$. From this,

$$p_{B|A,C} = \frac{p_{A,B,C}}{p_{A,C}} = \frac{p_A p_{B|A} p_{C|A}}{p_A p_{C|A}} = p_{B|A}.$$

   (b) $p_{A,B}$ is straightforward multiplication of tables.

$$p_{A,B} = p_A p_{B|A} = \begin{array}{c|cc} & b_1 & b_2 \\ \hline a_1 & 0.03 & 0.27 \\ a_2 & 0.56 & 0.14 \end{array}$$

p_B is obtained by summing over A

$$p_B = \begin{array}{cc} b_1 & b_2 \\ \hline 0.59 & 0.41 \end{array}$$

$p_{A,B,C}$ is obtained from multiplication of tables

$$p_{A,B,C} = p_A p_{B|A} p_{C|A} = p_{A,B} p_{C|A},$$

so

$$p_{A,B,C} = \begin{array}{c|cc} (c_1, c_2) & b_1 & b_2 \\ \hline a_1 & (0.018, 0.012) & (0.162, 0.108) \\ a_2 & (0.336, 0.224) & (0.084, 0.056) \end{array}$$

$p_{B,C}$ is obtained by summing over A

$$p_{B,C} = \begin{array}{c|cc} & b_1 & b_2 \\ \hline c_1 & 0.354 & 0.246 \\ c_2 & 0.236 & 0.164 \end{array}$$

$p_{C|B}$ is obtained by division of tables

$$p_{C|B} = \frac{p_{B,C}}{p_B} = \begin{array}{c|cc} B\backslash C & c_1 & c_2 \\ \hline b_1 & 0.6 & 0.4 \\ b_2 & 0.6 & 0.4 \end{array}$$

9. (a)

$$p_A = \begin{array}{cc} y & n \\ \hline 0.15 & 0.85 \end{array}$$

$$p_{A,T} = \begin{array}{c|cc} A\backslash T & y & n \\ \hline y & 0.1425 & 0.0075 \\ n & 0.00425 & 0.84575 \end{array}$$

$$p_T = \begin{array}{cc} y & n \\ \hline 0.14675 & 0.85325 \end{array}$$

$$p_{A|T} = \frac{p_{A,T}}{p_T} = \begin{array}{c|cc} T\backslash A & y & n \\ \hline y & 0.971... & 0.029... \\ n & 0.0088... & 0.9912... \end{array}$$

(b)

$$p_A = (0.0005, 0.9995)$$

$$p_{A,T} = p_A p_{T|A} = \begin{array}{c|cc} A\backslash T & y & n \\ \hline y & 0.000475 & 0.000025 \\ n & 0.0049975 & 0.9945025 \end{array}$$

$$p_T = \begin{array}{cc} y & n \\ \hline 0.0054725 & 0.9945275 \end{array}$$

$$p_{A|T} = \frac{p_{A,T}}{p_T} = \begin{array}{c|cc} T\backslash A & y & n \\ \hline y & 0.0868 & 0.9132 \\ n & 0.000025138 & 0.999974862 \end{array}$$

NOTE: This example should be used with HUGIN.

10.

$$X \perp Y|Z \Leftrightarrow p_{X,Y,Z}(x,y,z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)p_Z(z) \qquad \forall(x,y,z)$$

so that

$$\begin{aligned} p_{X,Y,Z}(x,y,z)p_{X,Y,Z}(x',y',z) &= p_{X|Z}(x|z)p_{Y|Z}(y|z)p_Z(z)p_{X|Z}(x'|z)p_{Y|Z}(y'|z)p_Z(z) \\ &= \left(p_{X|Z}(x'|z)p_{Y|Z}(y|z)p_Z(z)\right)\left(p_{X|Z}(x|z)p_{Y|Z}(y'|z)p_Z(z)\right) \\ &= p_{X,Y,Z}(x',y,z)p_{X,Y,Z}(x,y',z) \end{aligned}$$

for all (x,y,x',y',z) as required.

11. Yes: $H_2 - S_3 - H_3$ is a converging connection with middle variable instantiated, hence H_2 and H_3 are d-connected.

12. If $(X_i, X_j) \in E$ and $(X_j, X_r) \in E$ implies $(X_i, X_r) \in E$, then clearly any ancestor of a node X_ν is also a parent.

    If for each variable X_r any ancestor is also a parent, then if there are variables X_i and X_j such that $(X_i, X_j) \in E$ and $(X_j, X_r) \in E$, it follows that X_i is a parent of X_r and hence $(X_i, X_r) \in E$.

13. Question is badly phrased: here $X \subseteq V$ and $Y \subseteq V$, so it should say 'subset of' rather than 'in' and the conditioning set should be the *ancestors*.

    The question is asking for a proof of equivalence for transitive DAGS of LDMC and the statement that

    $$(X\backslash Y \cup \{an(X) \cap an(Y)\}) \perp (Y\backslash X \cup \{an(X) \cap an(Y)\})|\mathrm{an}(X) \cap \mathrm{an}(Y),$$

    where

    $$an(X) = \cup_{x \in X} an(x).$$

    for all sets X and Y such that both $X\backslash Y \cup \{an(X) \cap an(Y)\}$ and $Y\backslash X \cup \{an(X) \cap an(Y)\}$ are non empty.

    Short answer: Using small case letters for variables or nodes and large case letters for sets, the LDMC is that for each variable $x \in V$,

    $$x \perp V\backslash(V_x \cup \Pi_x)|\Pi_x$$

where V_x denotes descendants of x and Π_x denotes the parents.

For a transitive DAG, the local directed Markov condition means that a node x is conditionally independent, given the ancestor set (which is the parent set) of all nodes y which are neither ancestors nor descendants.

If the local directed Markov condition holds, then d separation implies conditional independence. Clearly, if all nodes in the set $an(X) \cap an(Y)$ are instantiated and only these nodes, then all chain and fork connections between $X\backslash(Y \cup \{an(X) \cap an(Y)\})$ and $Y\backslash(X \cup \{an(X) \cap an(Y)\})$ are instantiated and no collider connections are instantiated, neither are any descendants of collider connections on a trail between the two sets instantiated, hence they are d separated and hence the conditional independence statement holds.

If the conditional independence statement holds for all subsets X and Y of nodes for a transitive DAG, then consider $X = \{x\}$ for a node x. Then clearly $an(V\backslash(V_x \cup \Pi_x)) \supseteq \Pi_x$ so that $an(V\backslash(V_x \cup \Pi_x)) \cap \Pi_x = \Pi_x$ so that

$$x \perp V\backslash(V_x \cup \Pi_x)|\Pi_x,$$

hence LDMC holds.

14. This is a direct consequence of the definition. Let $\underline{x} = (x_1, \ldots, x_d)$ and $\underline{y} = (y_1, \ldots, y_d)$ where $y_j = x_j = w_j$ for $j \neq i$, $x_i = 1$, $y_i = 0$. Let $\pi_j(\underline{x})$ denote the parent configuration for variable j when $\underline{X} = \underline{x}$. Then, since

$$O_p(\{X_i = 1\}|\{\underline{X}_{V\backslash\{X_i\}} = \underline{w}\}) = \frac{p(\{X_i = 1\}, \{\underline{X}_{V\backslash\{X_i\}} = \underline{w}\})}{p(\{X_i = 1\}, \{\underline{X}_{V\backslash\{X_i\}} = \underline{w}\})}$$

$$\begin{aligned} O_p(\{X_i = 1\}|\{\underline{X}_{V\backslash\{X_i\}=\underline{w}}\}) &= \frac{\prod_{j=1}^d p_{X_j|\Pi_j}(x_j|\pi_j(\underline{x})}{\prod_{j=1}^d p_{X_j|\Pi_j}(y_j|\pi_j(\underline{y})} \\ &= \frac{p_{X_i|\Pi_i}(1|\pi_i(\underline{x}))\prod_{j|X_i\in\Pi_j} p_{X_j|\Pi_j}(x_j|\pi_j(\underline{x})}{p_{X_i|\Pi_i}(0|\pi_i(\underline{y}))\prod_{j|X_i\in\Pi_j} p_{X_j|\Pi_j}(y_j|\pi_j(\underline{y})} \end{aligned}$$

and, from the definition, this only involves the Markov blanket of X_i.

Chapter 3

1. Let A denote car type and C colour. Events to be updated: $p_C^*(\text{red}) = 0.4$, $p_C^*(\text{blue}) = 0.6$ Original *joint* table:

$$p_{A,C} = \begin{array}{c|cc} \text{car}\backslash\text{colour} & R & B \\ \hline V & 0.35 & 0.15 \\ S & 0.1 & 0.4 \end{array}$$

so

$$p_C(\text{red}) = 0.45 \qquad p_C(\text{blue}) = 0.55$$

and

$$p_{A|C} = \begin{array}{c|cc} \text{car}\backslash\text{colour} & R & B \\ \hline V & 7/9 & 3/11 \\ S & 2/9 & 8/11 \end{array}$$

Jeffrey's rule:

$$p_{A,C}^* = p_{A|C}p_C^* = \begin{array}{c|cc} \text{car}\backslash\text{colour} & R & B \\ \hline V & 14/45 & 9/55 \\ S & 4/45 & 24/55 \end{array}$$

$$p^*(\text{volvo}) = \frac{47}{99}$$

2. For any individual outcome, $x \in G_j$ for exactly one G_j and hence

$$p(x|A) = \frac{\lambda_j}{\sum_{k=1}^n \lambda_k p(G_k)} p(x) \qquad x \in G_j$$

as required.

3. Using Jeffreys update:

$$p^*(x) = p(x|G_j)p^*(G_j) = q_j p(x|G_j) \qquad x \in G_j.$$

Using Pearl's method, for $x \in G_j$

$$\begin{aligned} p^*(x) &= \frac{\lambda_j}{\sum_{k=1}^n \lambda_k p(G_k)} p(x) \\ &= \frac{q_j}{\sum_{k=1}^n p(G_k)} \frac{p(\{x\} \cap G_j)}{p(G_j)} = q_j p(x|G_j) \end{aligned}$$

so they both give the same.

4. By Jeffrey's rule,

$$\begin{aligned} p^*(C) &= \sum_j p(C|G_j)p^*(G_j) = \sum_j p(C|G_j)p(G_j|A) = \sum_j p(C|G_j)\frac{p(A|G_j)p(G_j)}{\sum_{k=1}^n p(A|G_k)p(G_k)} \\ &= \sum_j p(C\cap G_j)\frac{\lambda_j}{\sum_{k=1}^n \lambda_k p(G_k)} \end{aligned}$$

which is the Pearl update.

5. Virtual evidence on X_3 is treated as a node with a single parent X_3, so (using $p_{X_3}(1) = p_{X_3}(0) = \frac{1}{2}$),

$$\begin{aligned} \tilde{p}_{X_1,X_2,X_3}(x_1,x_2,1) &= p_{X_1,X_2,X_3}(x_1,x_2,1)\frac{p_{V|X_3}(1|1)}{p_{V|X_3}(1|1)p_{X_3}(1)+p_{V|X_3}(1|0)p_{X_3}(0)} \\ &= p(x_1,x_2,1)\frac{2\lambda}{\lambda+1} = \frac{\lambda}{4(\lambda+1)} \end{aligned}$$

$$\begin{aligned} \tilde{p}_{X_1,X_2,X_3}(x_1,x_2,0) &= p_{X_1,X_2,X_3}(x_1,x_2,0)\frac{p_{V|X_3}(1|0)}{p_{V|X_3}(1|1)p_{X_3}(1)+p_{V|X_3}(1|0)p_{X_3}(0)} \\ &= p(x_1,x_2,1)\frac{2}{\lambda+1} = \frac{1}{4(\lambda+1)} \end{aligned}$$

for each value of $(x_1,x_2) \in \{(0,0),(0,1),(1,0),(1,1)\}$.

6.

$$\pi(\psi|x_1,x_2) = \frac{p_{X_1,X_2}(x_1,x_2;\psi)}{p_{X_1,X_2}(x_1,x_2)}.$$

$\pi(\psi)$ is a function of ψ.

$$p_{X_1,X_2}(x_1,x_2;\psi) = \int_0^\psi p_{X_1,X_2}(x_1,x_2|\theta_1,\psi)\pi(\theta_1,\psi)d\theta_1.$$

Let $t = x_1 + x_2$, then

$$\begin{aligned} &\int_0^\psi p_{X_1,X_2}(x_1,x_2|\theta_1,\psi)\pi(\theta_1,\psi-\theta_1)d\theta_1 \\ &\quad = C(a_1,a_2,a_3,x_1,x_2)\int_0^\psi \theta_1^{x_1}(\psi-\theta_1)^{x_2}(1-\psi)^{n-(x_1+x_2)}\theta_1^{a_1-1}(\psi-\theta_1)^{a_2-1}(1-\psi)^{a_3-1}d\theta_1 \\ &\quad = C(a_1,a_2,a_3,x_1,x_2)(1-\psi)^{n+a_3-1-t}\psi^{t+a_1+a_2-2}\int_0^1 y^{x_1+a_1-1}(1-y)^{x_2+a_2-2}dy \\ &\quad = \tilde{C}(a_1,a_2,a_3,x_1,x_2)(1-\psi)^{n+a_3-1-t}\psi^{t+a_1+a_2-2}, \end{aligned}$$

where C and $\tilde{C}$ do note depend on ψ. Let $K(a_1,a_2,a_3,x_1,x_2) = \frac{\tilde{C}(a_1,a_2,a_3,x_1,x_2)}{p_{X_1,X_2}(x_1,x_2)}$. Then

$$\pi(\psi|x_1,x_2) = K(a_1,a_2,a_3,x_1,x_2)(1-\psi)^{n+a_3-1-t}\psi^{t+a_1+a_2-2}.$$

Then, since $\int \pi(\psi|x_1,x_2)d\psi = 1$, it follows that

$$\pi(\psi|x_1,x_2) = \frac{(1-\psi)^{n+a_3-1-t}\psi^{t+a_1+a_2-2}}{\int_0^1 (1-\psi)^{n+a_3-1-t}\psi^{t+a_1+a_2-2}d\psi},$$

which only depends on (x_1,x_2) through x_1+x_2. It follows that X_1+X_2 is Bayesian sufficient for ψ if the space of priors is the space specified.

7.
$$p(x \leq \min_j X_j \leq \max_j X_j \leq y) = \prod_{j=1}^n p(x \leq X_j \leq y) = \left(\frac{y-x}{\theta_2-\theta_1}\right)^n, \quad \theta_1 \leq x \leq y \leq \theta_2$$

so, for $n \geq 2$, joint density is

$$p(x,y|\theta_1,\theta_2) = \begin{cases} n(n-1)\frac{(y-x)^{n-2}}{(\theta_2-\theta_1)^n} & \theta_1 \leq x \leq y \leq \theta_2 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)\pi(\theta)$$

The density $p(\mathbf{x}|\theta)$ is given by

$$\begin{aligned} p(\mathbf{x}|\theta) &= \begin{cases} \frac{1}{(\theta_2-\theta_1)^n} & \theta_1 \leq x_j \leq \theta_2 \quad \forall j \in \{1,\dots,n\} \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{1}{(\theta_2-\theta_1)^n}\chi_{(\theta_1,\theta_2)}(\min_j X_j)\chi_{(\theta_1,\theta_2)}(\max_j X_j). \end{aligned}$$

It follows that $p(\mathbf{x}|\theta) = \phi(\theta, T(\mathbf{x}))$ for some function ϕ and hence T is sufficient.

8. **Bucket Elimination**

(a)
$$p(A,B,C,D,E,F,G) = p(A)p(B)p(C|A,B)p(D|C)p(E|D,A)p(G|D)p(F|G,D)$$

so
$$p(F=1|C=1) = \sum_{D,E,G} p(D|C=1)p(E|D,A)p(A)p(G|D)p(F|G,D)$$

Elimination order E, A, B, G, D gives

$$\sum_D p(D|C=1)\sum_G p(G|D)p(F=1|G,D)\sum_B p(B)\sum_A p(A)\sum_E p(E|A,D)$$

Table $p(E|A,D)$ has 4 free elements, because $p(E=0|A,D) = 1-p(E=1|A,D)$.
$\lambda_E \equiv \sum_E p(E|A,D) = 1$.
$\lambda_E p(A) = p(A)$ has one free element, $\lambda_A := \sum_A p(A) = 1$. $p(B)$ has one element.

$\sum_B \lambda_A p(B) = 1$.

$p(G|D)p(F=1|G,D)$ has 4 elements, $\lambda_G(D,F=1) := \sum_G p(G|D)p(F=1|G,D)$.

$\lambda_G(D,F=1)p(D|C=1)$ has 2 elements,

$$\sum_D \lambda_G(D,F=1)p(D|C=1) = p(F=1|C=1).$$

Maximum table size: 4. Sum of table sizes: $4+1+1+4+2=12$

If order is D,G,B,A,E, then sum is

$$\sum_E \sum_A p(A) \sum_B p(B)p(C=1|A,B) \sum_G \sum_D p(D|C=1)p(E|A,D)p(G|D)p(F=1|D,G)$$

$p(D|C=1)p(E|A,D)p(G|D)p(F=1|D,G)$ has 16 elements,

$\lambda_D(A,G,E,C=1,F=1) = \sum_D p(D|C=1)p(E|A,D)p(G|D)p(F=1|D,G)$ has 8 elements.

$\lambda_G(A,E,C=1,F=1) = \sum_G \lambda_D(A,D,E,C=1,F=1)$.

$p(B)p(C=1|A,B)\lambda_G(A,E,C=1,F=1)$ has 8 entries,

$\lambda_B(A,E,C=1,F=1) = \sum_B p(B)p(C=1|A,B)\lambda_G(A,E,C=1,F=1)$.

$p(A)\lambda_B(A,E,C=1,F=1)$ has 4 entries,

$\lambda_A(E,C=1,F=1) = \sum_A p(A)\lambda_B(A,E,C=1,F=1)$.

$\lambda_A(E,C=1,F=1)$ has 2 entries,

$$p(F=1|C=1) = \sum_E \lambda_A(E,C=1,F=1)$$

Maximum table: 16. Sum of table sizes: $16+8+8+4+2=38$

(b) (MATLAB exercise)

9. A age, T temperature, W weather, MH mental health, PH physical health, CF condition when found, H hypothermia, WC Wind chill CL condition when lost. Before divorcing is given by figure (3.1); after divorcing is given by figure (3.2).

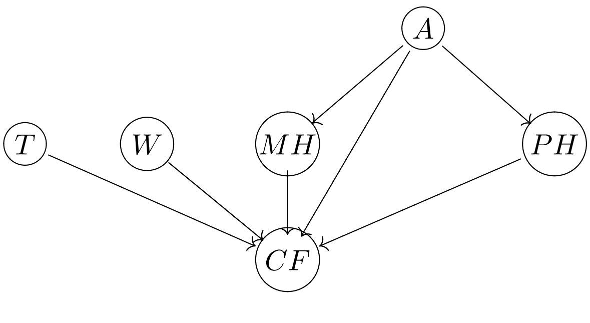


Figure 3.1: Before Divorcing

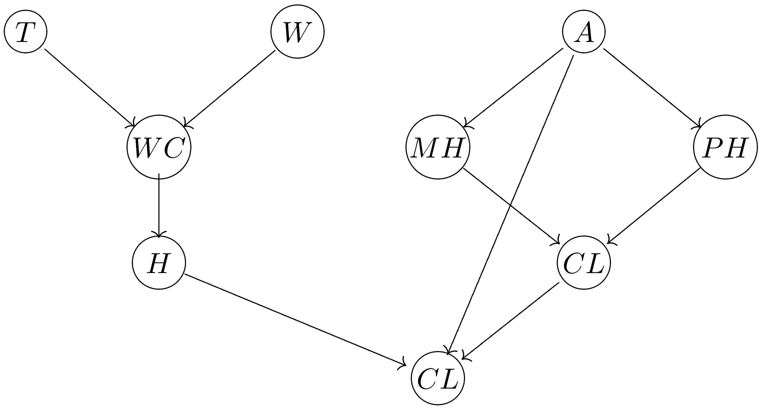


Figure 3.2: After Divorcing

10. The diagram is satisfied if and only if $I_1, \ldots, I_n$ are conditionally independent given H. It follows that the diagram is satisfied if and only if

$$p(H|(I_j)_{j=1}^n) = \frac{p(H, I_1, \ldots, I_n)}{p(I_1, \ldots, I_n)} = \frac{p(H)p(I_1|H)\ldots p(I_n|H)}{p(I_1, \ldots, I_n)}$$

so that the given causal structure implies the conditional probability structure.

11. This is a straightforward piece of MATLAB coding; the resulting chain is clearly *not* aperiodic; it has period 2.

12. (a) The notation $x_{\tilde{V}\backslash\tilde{U},l}$ denotes the lth component of $x_{\tilde{V}\backslash\tilde{U}}$. Here Z_n is irreducible (clearly for any two different states (a,b) there is a path with probability $P_{a,b}(k) = P_{b,a}(k) = \frac{1}{t^k}$ where k is the number of states that differ). It is also aperiodic. The original underlying chain is NOT aperiodic (it has period 2), but the chain (with the modification) is aperiodic because it can stay in the same state with positive probability (the rejection probability).

    (b)

$$p_{\underline{X}} = \prod_{j=1}^{d} p_{X_j|\Pi_j}$$

so

$$p_{\underline{X}_{\tilde{V}\backslash\tilde{U}}|\underline{X}_{\tilde{U}}}(\underline{x}_{\tilde{V}\backslash\tilde{U}}|\underline{y}) = \frac{1}{p_{\underline{X}_{\tilde{U}}}(\underline{y})} \prod_{j=1}^{d} p_{X_j|\Pi_j}(x_j|\pi_j(\underline{x}))\bigg|_{\underline{x}_{\tilde{U}}=\underline{y}}$$

and hence, with a transition of variable l, let $\tilde{C}_l$ denote the indexing set for the children of variable l, then

$$\frac{p_{\underline{X}_{\tilde{V}\backslash\tilde{U}}|\underline{X}_{\tilde{U}}}(\underline{x}^*_{\tilde{V}\backslash\tilde{U}}|\underline{y})}{p_{\underline{X}_{\tilde{V}\backslash\tilde{U}}|\underline{X}_{\tilde{U}}}(\underline{x}_{\tilde{V}\backslash\tilde{U}}|\underline{y})} = \frac{p_{X_l|\Pi_l}(x_l^*|\pi_l(\underline{x}))\prod_{j\in\tilde{C}_l} p_{X_j|\Pi_j}(x_j|\pi_j(\underline{x}^*))}{p_{X_l|\Pi_l}(x_l|\pi_l(\underline{x}))\prod_{j\in\tilde{C}_l} p_{X_j|\Pi_j}(x_j|\pi_j(\underline{x}))}$$

so take

$$\phi(x_{(\tilde{V}\backslash\tilde{U}),l}, M_l) = p_{X_l|\Pi_l}(x_l|\pi_l(\underline{x})) \prod_{j\in\tilde{C}_l} p_{X_j|\Pi_j}(x_j|\pi_j(\underline{x})).$$

This depends only on variable l together with its Markov blanket (its parents, its children and parents of its children).

(c) Since the transition matrix Q for the 'proposal' chain is symmetric, the result follows from the theorem stated and proved in the text.

The important point for this question is that at each step the method only requires the tables corresponding to the single variable that has been altered and its children.

13. This exercise may be carried out using HUGIN Lite; it gives a good introduction to the use of the software.

Chapter 4

1. α_1 and α_7 are simplicial. The graph is not triangulated: cycle $(\alpha_2, \alpha_3, \alpha_6, \alpha_5, \alpha_2)$ does not have a chord.

2. α_6, α_5 are simplicial. Graph is triangulated: consider sequence $(\alpha_6, \alpha_5, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$.

3. $C_1 = (\alpha_1, \alpha_2, \alpha_4, \alpha_5)$ with separator $S_1(\alpha_1, \alpha_2, \alpha_4)$, $C_2 = (\alpha_3, \alpha_4, \alpha_6)$, $S_2 = (\alpha_3, \alpha_4)$, $C_3 = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$

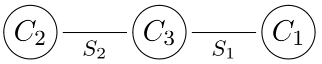


Figure 4.1: Junction Tree for Q3

4. No: it isn't triangulated - cycle $(\alpha_2, \alpha_3, \alpha_4, \alpha_5)$ does not have a chord. Decomposable and triangulated are equivalent.

5. (a) moral graph is shown in the figure.
   (b) $X_{12}, X_{11}, X_7, X_5, X_{13}, X_{10}, X_9, X_8, X_6, X_4, X_1, X_2, X_3$ is a perfect elimination sequence.
   (c) junction tree in the figure

6. Assume that the following hold for a singly connected DAG on n nodes:

   (a) The moral graph is triangulated.
   (b) The parent/children configurations form the cliques of the corresponding junction tree.
   (c) Each separator has exactly one node.

   This is clearly true for a graph with one node. Consider a singly connected DAG on $n+1$ nodes and consider a node which has no ancestors (this can clearly be done). The graph with this node removed is singly connected and is therefore (by the inductive hypothesis) the three assumptions listed above hold.

   Now consider the addition of this node to the DAG. By definition, it cannot be at the centre of a chain connection (since it is does not have parents) or a fork connection (since the DAG on n

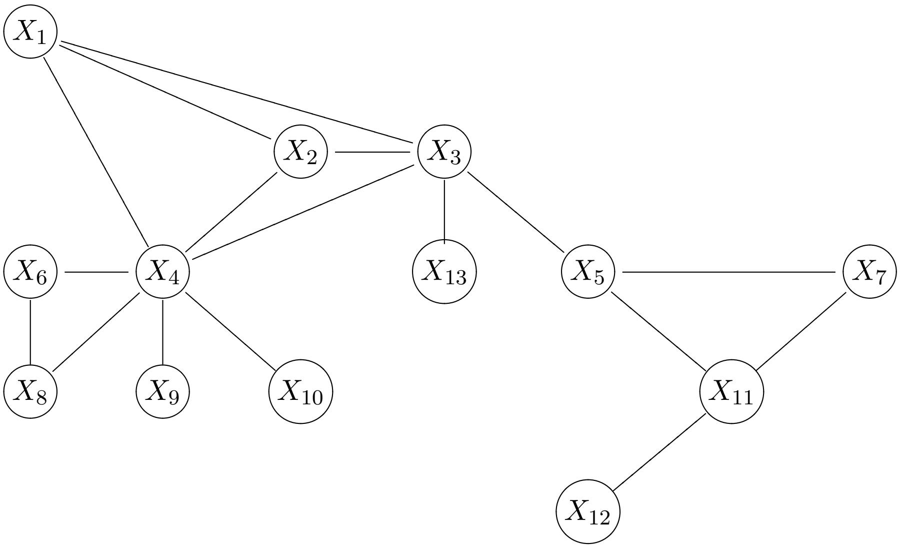


Figure 4.2: Moral graph for Q5

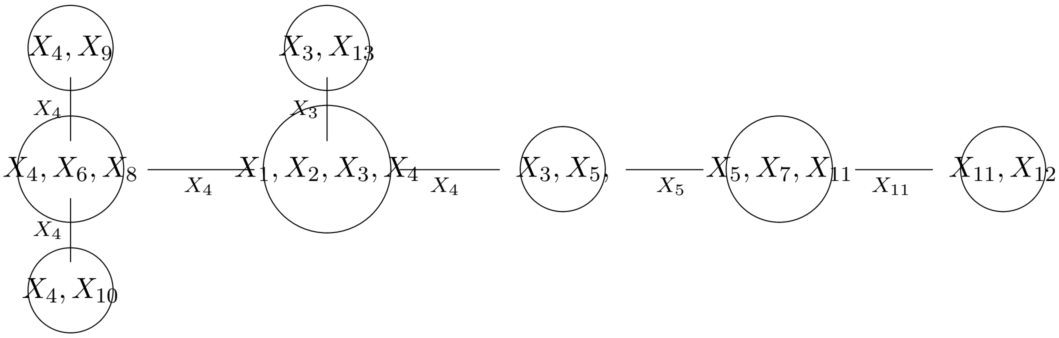


Figure 4.3: Junction Tree for Q5

variables was connected and singly connected). It is therefore linked to exactly one node, which is now a collider node. Clearly, this new edge does not produce a cycle of length ≥ 4.

When considering the moral graph, the new node is added to the clique of all the parents of the child of the new node. It is therefore a simplicial node; all its neighbours are parents of a common child together with the common child and therefore have edges between them). There is a perfect elimination sequence - start with the new node and then proceed with the perfect elimination sequence for for the graph on n variables. Hence the graph is triangulated.

Again, the child/parent configurations form the cliques in the junction tree for the graph on $n+1$ nodes, and the separators each contain exactly one node. The result follows by induction.

7. Consider a perfect elimination sequence for the original graph. Consider the nodes of A taken in the same order. The family of each node in A of the original graph formed a clique for the original elimination sequence. A clique in the original graph, restricted to A is clearly a clique in A; the edges between neighbours of a node in A have been preserved. Since no fill ins were required for the original sequence, no fill ins will be required for the graph restricted to A. Hence $\mathcal{G}_A$ is triangulated.

8. See diagrams

   (a) moralised graph figure 4.4

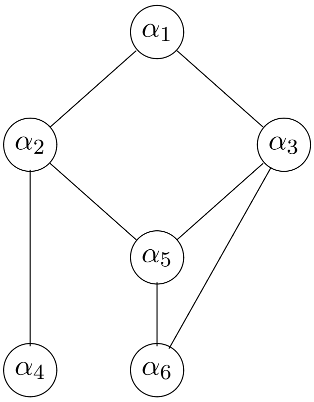


Figure 4.4: moralised graph Q8

   (b) A possible triangulation figure 4.5

   (c) A junction tree following the triangulation - figure 4.6

9. This simply follows from reading the definition and following the notations.

   Consider the expression in definition 4.30. For a directed graph, there are no undirected neighbours. From the definition of ancestral boundary, it follows that $\Lambda(\{X_j\}) = \Pi_{X_j}$. $D(X_j)$ is the set of descendants of X_j and, for a DAG, $\Phi_{\mathcal{G}}(X_j) = \{X_j\} \cup \Pi_{X_j}$ and hence, for a DAG, the

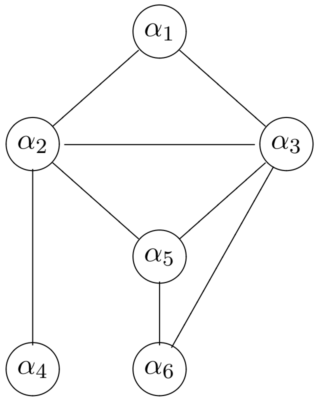


Figure 4.5: triangulated graph Q8

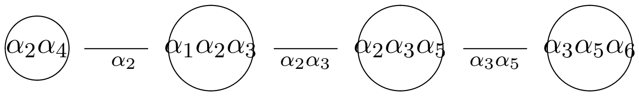

$\alpha_2\alpha_4$
α_2
$\alpha_1\alpha_2\alpha_3$
$\alpha_2\alpha_3$
$\alpha_2\alpha_3\alpha_5$
$\alpha_3\alpha_5$
$\alpha_3\alpha_5\alpha_6$

Figure 4.6: junction tree Q8

definition of local $\mathcal{G}$-Markovian given in definition 4.30 is equivalent to the local directed Markov condition given in definition 2.28.

10. The definition is that for any sets A, B, S such that S separates A and B in $\mathcal{G}^m_{An(A \cup B \cup S)}$, it holds that $A \perp B | S$.

    Here $An(A \cup B \cup S)$ denotes the smallest set containing all members of $A \cup B \cup S$, all parents and all undirected neighbours. In the case of a DAG, there are no undirected neighbours. The definition 4.31 for the moral graph of a chain graph clearly reduces to the definition 4.9 for the moral graph of a DAG when a DAG is being considered.

    The problem is therefore to show that, for a DAG, a set A is d-separated from B by S if and only if S separates A from B in $\mathcal{G}^m_{A \cup B \cup S \cup \Pi_A \cup \Pi_B \cup \Pi_S}$.

    S d-separates A from B if and only if for every trail from a node in A to a node in B, either

    - there is a fork or chain node in S or
    - there is a collider node not in S which does not have any descendant in S.

    Consider a Directed Acyclic Graph, and suppose that S separates A from B in

    $$\mathcal{G}^m_{A \cup B \cup S \cup \Pi_A \cup \Pi_B \cup \Pi_S}.$$

    Then any path from A to B in the restricted moralised graph contains a node in S.

    Consider a path from $\alpha \in A$ to $\beta \in B$ in the moralised restricted graph, where α is the only node in A and β is the only node in B, that contains a node $s \in S$. If s is a fork or chain in the

DAG, then the corresponding trail is blocked; if one of the edges to s is added in as a result of the moralisation then there is not a corresponding trail in the DAG.

Now suppose that in a path $\alpha, \gamma_1, \ldots, \gamma_k, \beta$ between $\alpha \in A$ and $\beta \in B$ in the moralised restricted graph, where $\gamma_1, \ldots, \gamma_k \notin A \cup B$ either there is node $\gamma_l \notin S$ which, in the DAG is a collider node with a descendant in s for some $l \in \{1, \ldots, k\}$, or else there is a node $\gamma_l \in S$ for some $l \in \{1, \ldots, k\}$ which is a collider in the DAG. For any node $\gamma_l \in S$, there is an edge between the two nodes which are *parents* in the original DAG, so it is possible to find a path that does not contain this node. Since the original path has a finite number of nodes, the new path without γ_1 has one fewer nodes, it is therefore possible to find a path from α to β that does not have any nodes in S.

Therefore no path between A and B in the moralised restricted graph contains a node γ that is a collider in the corresponding trail in the DAG such that either $\gamma \in S$ or γ has a descendant $s \in S$.

It follows that if A and B are separated by S in the restricted moralised graph, then A and B are d-separated by S in the DAG.

Suppose A and B are d-separated by S in the DAG. Consider $\mathcal{G}^m_{A \cup B \cup S \cup \Pi_A \cup \Pi_B \cup \Pi_S}$ and consider any undirected path in the moralised restricted graph $(\alpha, \gamma_1, \ldots, \gamma_k, \beta)$ where $\alpha \in A$, $\beta \in B$, $(\gamma_1, \ldots, \gamma_k) \notin A \cup B$.

For any edge $\langle \gamma_j, \gamma_{j+1} \rangle$ such that neither (γ_j, γ_{j+1}) nor (γ_{j+1}, γ_j) is present in the original DAG, there is a node c such that (γ_j, c) and (γ_{j+1}, c) are present in the original, where $c \in A \cup B \cup S$. Consider the trail formed from the directed version of the edges $\langle \gamma_j, \gamma_{j+1} \rangle$ in the DAG if it exists, or replacing it by $(\gamma_j, c), (\gamma_{j+1}, c)$ otherwise.

For this new trail, labelled $(\alpha, \gamma_1, \ldots, \gamma_n, \beta)$, consider colliders $(\gamma_{i-1}, \gamma_i, \gamma_{i+1})$ along the trail, starting from $i = 1$. If $\gamma_i \in S$, then the collider connection is open. If $\gamma_i \in \Pi_S$, then again the collider is open. If $\gamma_i \in \Pi_A$, consider the trail formed by $(\tilde{\alpha}, \gamma_i, \gamma_{i+1}, \ldots)$, where $\tilde{\alpha} \in A$. If $\gamma_i \in \Pi_B$, then there is a node $\beta^* \in B$ such that the trail $(\alpha, \gamma_1, \ldots, \gamma_i, \beta^*)$ is open. It follows that A is not d-separated from B by S.

It follows that if A and B are d-separated by S, then A is separated from B by S in the moralised restricted graph.

For a distribution p factorised along a DAG, it follows that the definitions of faithful and globally $\mathcal{G}$ Markovian are equivalent.

11. Simple consequence of the junction tree property; all the remaining nodes contain α. In the original tree, for any two nodes containing α, there is a path between them with α in each node along the path. Since the intervening nodes remain, the path remains (from the definition of eliminating a node).

Chapter 5

1.
$$D(\mathbf{f}|\mathbf{g}) = (1-p)\log\frac{1-p}{1-g} + p\log\frac{p}{g}.$$

2. Definition of convexity for a function ϕ: for any $\lambda \in [0,1]$ and any (x, y),

$$\phi(\lambda x + (1-\lambda)y) \leq \lambda\phi(x) + (1-\lambda)\phi(y).$$

Proof of result by induction: if $\mathcal{X} = \{x_1, x_2\}$, set $p_1 = \lambda$, $p_2 = 1-\lambda$, then $\mu = E[X] = p_1x_1+p_2x_2$ so definition of convexity gives

$$\phi(\mu) \leq p_1\phi(x_1) + p_2\phi(x_2) = E[\phi(X)].$$

Assume result is true for any probability distribution over $\{x_1, \ldots, x_n\}$. Consider a probability distribution $(p_1, \ldots, p_{n+1})$ over $(x_1, \ldots, x_{n+1})$. Then

$$\phi(\mu) = \phi\left(\sum_{j=1}^{n+1} p_jx_j\right) \leq p_{n+1}\phi(x_{n+1}) + (1-p_{n+1})\phi\left(\sum_{j=1}^{n}\frac{p_j}{1-p_{n+1}}x_j\right)$$

and, by the inductive hypothesis (since $\sum_{j=1}^{n}\frac{p_j}{1-p_{n+1}} = 1$) ,

$$\phi\left(\sum_{j=1}^{n}\frac{p_j}{1-p_{n+1}}x_j\right) \leq \sum_{j=1}^{n}\frac{p_j}{1-p_{n+1}}\phi(x_j)$$

so that

$$\phi(\mu) \leq \sum_{j=1}^{n+1}\phi(x_j)p_j$$

as required.

3.
$$k = D_{KL}(Be(1/2)|Be(g(k)) = \frac{1}{2}\log\frac{1}{2g(k)} + \frac{1}{2}\log\frac{1}{2(1-g(k))}$$

so that

$$-2k = 2\log 2 + \log g(k)(1-g(k)$$

$$g(k)(1-g(k)) = \frac{1}{4}e^{-2k}$$

$$g^2 - g + \frac{1}{4} = \frac{1}{4}(1-e^{-2k})$$
$$g(k) = \frac{1}{2} \pm \frac{1}{2}\sqrt{1-e^{-2k}}$$

4. The computation is

$$p(\{X_{j,n+1} = x_j^i\}|\{\Pi_j = \pi_j^l\}, \{\mathbf{X} = \mathbf{x}\}) = \int p_{X_{n+1}^j|\Theta,\Pi_j,\mathbf{X}}(x_j^i|\theta, \pi_j^l, \mathbf{x})\pi_{\Theta|\Pi_j,\mathbf{X}}(\theta|\pi_j^l, \mathbf{x})d\theta$$

and
$$p_{X_{n+1}^j|\Theta,\Pi_j,\mathbf{X}}(x_j^i|\theta, \pi_j^l, \mathbf{x}) = p_{X_{n+1}^j|\Pi_j,\Theta}(x_j^i|\pi_j^l, \theta) = \theta_{jil}.$$

$$\pi_{\Theta|\Pi_j,\mathbf{X}}(\theta|\pi_j^l, \mathbf{x}) = \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \text{Const} \times \pi_\Theta(\theta)p_{\mathbf{X}|\Theta}(\mathbf{x}|\theta).$$

Note that $\pi_\Theta(\theta) = \prod_{j,l}\pi_{\Theta^{j,l}}(\theta^{j.l}))$, so that

$$\pi_{\Theta|\mathbf{x}}(\theta) = \prod_{j,l}\pi_{\Theta^{j,l}|\mathbf{x}}(\theta^{j.l}))$$

Since these are probability densities, each integrates to 1 leaving

$$p_{X_{n+1}^j|\Pi_j,\mathbf{X}}\left(x_j^i|\pi_j^l, \mathbf{x}\right) = \int_{S_{j.l}} \theta_{jil}\pi_{\Theta|\mathbf{X}}\left(\theta^{j.l}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}; \alpha_{j,l}\right) d\theta^{j.l}.$$

Since

$$\pi_{\Theta^{j.l}|\mathbf{x}}(\theta^{j.l}) = \text{Dir}(n(x_j^1|\pi_j^l) + \alpha_{j1l}, \ldots n(x_j^{k_j}|\pi_j^l) + \alpha_{jk_jl}),$$

the result follows directly from the results stated in chapter 1 and the exercises at the end of chapter 1.

5. Firstly, note that

$$S^M(C) = -\sum_{c\in C}\log P^M(c) = -\sum_{v\in Sp(v)}(\#(v))\log P^M(v) = -\sum_{v\in Sp(v)} nP^C(v)\log P^M(v).$$

This is true for all M; in particular, take $M = C$, so that

$$S^C(C) = -\sum_{v\in Sp(v)} nP^C(v)\log P^C(v)$$

giving

$$S^M(C) - S^C(C) = n \sum_{v \in Sp(v)} P^C(v) \log \frac{P^C(v)}{P^M(v)} = nd_K(P^C|P^M).$$

6. (a) 5 instantiations with $A = 0$. These are $(0,0),(0,1),(0,0),(0,1),(0,1)$ so $\hat{\theta}_B = \frac{3}{5}$. Six instantiations with $(B,C) = (1,0)$. These are

$$(1,1,0,0),(1,1,0,1),(0,1,0,0),(0,1,0,1),(0,1,0,0),(1,1,0,1)$$

so $\hat{\theta}_D = \frac{3}{6} = \frac{1}{2}$

(b)

$$\pi(\theta_D|\mathbf{U}) \propto p(\mathbf{U}|\theta_D)\pi(\theta_D) \propto \theta_D^3(1-\theta_D)^3\theta_D^{-0.5} = \theta_D^{2.5}(1-\theta_D)^3.$$

$$\int_0^1 \theta_D^{2.5}(1-\theta_D)^3 d\theta_D = \frac{\Gamma(3.5)\Gamma(4)}{\Gamma(7.5)} = \frac{3!}{(6.5).(5.5).(4.5).(3.5)}$$

using

$$\Gamma(z+1) = z\Gamma(z)$$

and $\Gamma(n+1) = n!$. The result follows directly.

7. (a) For Θ_a: 14y / 6n, so the updated distribution is

$$\pi_{\Theta_a|\mathbf{X}}(\theta_a|\mathbf{x}) = \frac{31!}{21!9!}\theta_a^{21}(1-\theta_a)^9.$$

9 (y,y) and 5 (y,n) so

$$\pi_{\Theta_{b|y}|\mathbf{X}}(\theta_{b|y}|\mathbf{x}) = \frac{23!}{15!7!}\theta_{b|y}^{15}(1-\theta_{b|y})^7.$$

4 (n,n) and 2 (n,y) so

$$\pi_{\Theta_{b|n}|\mathbf{X}}(\theta_{b|n}|\mathbf{x}) = \frac{12!}{4!7!}\theta_{b|n}^{4}(1-\theta_{b|n})^7.$$

(b) With fading, posteriors are

$$\pi_{\Theta_a|\mathbf{X}}(\theta_a|\mathbf{x}) = \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)}\theta_a^{\alpha}(1-\theta_a)^{\beta},$$

where

$$\alpha = 7 \times 0.9^{20} + \sum_{j=16}^{19} 0.9^j + (0.9^{12} + 0.9^{11}) + \sum_{j=5}^{9} 0.9^j + \sum_{j=0}^{2} 0.9^j \simeq 6.90$$

$$\beta = 3 \times 0.9^{20} + 0.9^{15} + 0.9^{14} + 0.9^{13} + 0.9^{10} + 0.9^4 + 0.9^3 = 2.79$$

For $\Theta_{b|y}$,

$$\pi_{\Theta_{b|y}|\mathbf{X}}(\theta_{b|y}|\mathbf{x}) = \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)}\theta_{b|y}^{\alpha}(1-\theta_{b|y})^{\beta},$$

where

$$\alpha = 6 \times 0.9^{20} + 0.9^{18} + 0.9^{16} + 0.9^{12} + 0.9^{8} + 0.9^{7} + 0.9^{6} + 0.9^{5} + 0.9 + 1 = 5.71$$

$$\beta = 2 \times 0.9^{20} + 0.9^{19} + 0.9^{17} + 0.9^{11} + 0.9^{9} + 0.9^{2} = 2.06$$

For $\Theta_{b|n}$,

$$\pi_{\Theta_{b|n}|\mathbf{X}}(\theta_{b|n}|\mathbf{x}) = \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)}\theta_{b|n}^{\alpha}(1-\theta_{b|n})^{\beta},$$

where

$$\alpha = 2 \times 0.9^{20} + 0.9^{13} + 0.9^{4} = 1.15$$

$$\beta = 3 \times 0.9^{20} + 0.9^{15} + 0.9^{14} + 0.9^{10} + 0.9^{3} = 1.88$$

8. Use

$$n^*(x_j^i|\pi_j^l) = n(x_j^i|\pi_j^l) + p(x_j^i, \pi_j^l|e^*)$$

where e^* denotes the incomplete piece of evidence. Note that

$$p_A(1) = E_a[\theta] = 3\int_0^1 \theta^3 d\theta = \frac{3}{4}$$

$$p_A(0) = \frac{1}{4}$$

$$p_{B|A}(1|1) = E_{b|y}[\theta] = 12\int_0^1 \theta^3(1-\theta)d\theta = \frac{3}{5}$$

$$p_{B|A}(0|1) = \frac{2}{5}$$

$$p_{B|A}(1|0) = 12\int_0^1 \theta^2(1-\theta)^2 = \frac{2}{5}$$

$$p_{B|A}(0|0) = \frac{3}{5}$$

so

$$p_B(1) = \frac{2}{5} \times \frac{1}{4} + \frac{3}{5} \times \frac{3}{4} = \frac{11}{20}$$

$$p_B(0) = \frac{9}{20}$$

$$p_{A|E^*}(1|e^*) = p_{A|B}(1|1) = \frac{p_A(1)p_{B|A}(1|1)}{p_B(1)} = \frac{9}{11}$$

$$p_{A|E^*}(0|e^*) = \frac{2}{11}$$

$$p_{A,B|B}((0,1)|1) = p_{A|B}(0|1) = \frac{2}{11}$$

$$p_{A,B|B}((1,1)|1) = p_{A|B}(1|1) = \frac{9}{11}$$

$$p_{A,B|B}((0,0)|1) = p_{A,B|B}((1,0)|1) = 0.$$

So updating is

$$\pi_{a|e^*}(\theta) = \frac{\Gamma(5)}{\Gamma(3+\frac{9}{11})\Gamma(1+\frac{2}{11})}\theta^{2+\frac{9}{11}}(1-\theta)^{\frac{2}{11}}, \quad \theta \in [0,1]$$

$$\pi_{b|y,e^*}(\theta) = \frac{\Gamma(5+\frac{9}{11}))}{\Gamma(3+\frac{9}{11})\Gamma(2)}\theta^{2+\frac{9}{11}}(1-\theta), \quad \theta \in [0,1]$$

$$\pi_{b|n,e^*}(\theta) = \frac{\Gamma(5+\frac{2}{11}))}{\Gamma(2+\frac{2}{11})\Gamma(3)}\theta^{1+\frac{2}{11}}(1-\theta)^2, \quad \theta \in [0,1]$$

9. Firstly, a part that is not asked for in the question. If the graph *structure* is given, and it is required to estimate the conditional probability potentials from the data (no prior distribution has been stated), then the maximum likelihood estimates for the CPPs, from the data, are $\hat{p}_{X_1}(1) = \frac{39}{100}$,

$$\hat{p}_{X_2|X_1} = \begin{array}{c|cc} X_2\backslash X_1 & 1 & 0 \\ \hline 1 & \frac{12}{39} & \frac{50}{61} \\ 0 & \frac{27}{39} & \frac{11}{61} \end{array}$$

$$\hat{p}_{X_3|X_1} = \begin{array}{c|cc} X_3\backslash X_1 & 1 & 0 \\ \hline 1 & \frac{27}{39} & \frac{24}{61} \\ 0 & \frac{12}{39} & \frac{37}{61} \end{array}$$

$$\hat{p}_{X_4|X_2} = \begin{array}{c|cc} X_4\backslash X_2 & 1 & 0 \\ \hline 1 & \frac{29}{62} & \frac{23}{38} \\ 0 & \frac{33}{62} & \frac{15}{38} \end{array}$$

$$\hat{p}_{X_5|X_3,X_4}(1|.,.) = \begin{array}{c|cc} X_4\backslash X_3 & 1 & 0 \\ \hline 1 & \frac{15}{16} & 1 \\ 0 & 1 & 1 \end{array}$$

Upon inspection, these values appear reasonably close and it may be verified, using (for example) χ^2 tests, that the null hypothesis, that the stated probability tables are correct, is not rejected. For example,

$$p_{X_5|X_3,X_4}(.|1,1) = (0.9, 0.1)$$

There are 16 observations with $(X_3, X_4) = (1,1)$ of which 15 give $X_5 = 1$ and 1 gives $X_5 = 0$.

$$G = 2 \times (16 \sum_{x_5} 16\hat{p}_{X_5|X_3,X_4}(x_5|1,1) \log \frac{\hat{p}_{X_5|X_3,X_4}(x_5|1,1)}{p_{X_5|X_3,X_4}(x_5|1,1)} \overset{approx}{\sim} \chi^2_{2-1}.$$

When some events happen with very small probability, one technique (to prevent the problem of $\log 0$ appearing in the computation) is to add 1 to the total number of observations, which is divided evenly between the cells. Using this, under the null hypothesis,

$$G = 32\left(\frac{15.5}{17}\log\frac{15.5}{17\times 0.9} + \frac{1.5}{17}\log\frac{1.5}{17\times 0.1}\right) = 1.967 < 2.706 = \chi_1^2(0.9),$$

so the null hypothesis is not rejected at the $\alpha = 0.1$ significance level. Of course, the nominal significance level is greater than the true significance level, since multiple tests are being carried out.

But the question with such a data set is whether the decomposition

$$p_{X_1,X_2,X_3,X_4,X_5} = p_{X_1}p_{X_2|X_1}p_{X_3|X_1}p_{X_4|X_2}p_{X_5|X_3,X_4}$$

is valid. The computation by hands is tedious and should be carried out using software.

If the test were made without reference to the given probability tables and was simply to check whether such a decomposition held. Then, for each $(x_1,\ldots,x_5)\in\{0,1\}^5$, set

$$\hat{p}_{X_1,X_2,X_3,X_4,X_5}(x_1,x_2,x_3,x_4,x_5) = \frac{1}{101}\left(\frac{1}{32}+\right.$$
$$\left.\hat{p}_{X_1}(x_1)\hat{p}_{X_2|X_1}(x_2|x_1)\hat{p}_{X_3|X_1}(x_3|x_1)\hat{p}_{X_4|X_2}(x_4|x_2)\hat{p}_{X_5|X_3,X_4}(x_5|x_3,x_4)\right)$$

and

$$\tilde{p}_{X_1,X_2,X_3,X_4,X_5} = \frac{1}{101}\left(\frac{1}{32}+n(x_1,\ldots,x_5)\right)$$

where $n(x_1,\ldots,x_5)$ denotes the number of observations for instantiation $(x_1,\ldots,x_5)$ given in the table. The addition of the $\frac{1}{32}$ to each cell is to prevent $\log 0$ appearing in the computation. The test statistic is

$$G = 200\sum_{x_1,x_2,x_3,x_4,x_5}\hat{p}_{X_1,X_2,X_3,X_4,X_5}(x_1,x_2,x_3,x_4,x_5)\log\frac{\hat{p}_{X_1,X_2,X_3,X_4,X_5}(x_1,x_2,x_3,x_4,x_5)}{\tilde{p}_{X_1,X_2,X_3,X_4,X_5}(x_1,x_2,x_3,x_4,x_5)}$$

and, under the null hypothesis that the decomposition holds, $G \overset{approx}{\sim} \chi^2_{31-11}$. This is the difference in d.f. between the full model and the fitted model.

To test whether the observed data comes from the distribution given by

$$p_{X_1,\ldots,X_5} = p_{X_1}p_{X_2|X_1}p_{X_3|X_1}p_{X_4|X_2}p_{X_5|X_3,X_4},$$

where the 'theoretical' probabilities are given by the CPPs listed, firstly compute $p_{X_1,\ldots,X_5}$ by multiplication of tables. The test statistic is then

$$G = 200\sum_{x_1,\ldots,x_5}\tilde{p}_{X_1,\ldots,X_5}(x_1,\ldots,x_5)\log\frac{\tilde{p}_{X_1,\ldots,X_5}(x_1,\ldots,x_5)}{p_{X_1,\ldots,X_5}(x_1,\ldots,x_5)}$$

which, under the hypothesis that the data fits the model, has χ_{31} distribution. The fitted model is simply the theoretical probability $p_{X_1,\ldots,X_5}$ which has no degrees of freedom.

These numerical computations should be carried out using appropriate software - MATLAB coding is straightforward.

10. First note that

$$\begin{aligned}
\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) &= (\text{const})\pi_{\Theta}(\theta)p_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) \\
&= (\text{const})\sum_i \lambda_i \text{Dir}(\alpha^{(i)}q_1^{(i)} + n_1, \ldots, \alpha^{(i)}q_L^{(i)} + n_L) \\
&= \sum \lambda_i \frac{\Gamma(\sum_j \alpha^{(i)}q_j^{(i)} + n_j)}{\prod_j \Gamma(\alpha^{(i)}q_j^{(i)} + n_j)} \prod_{j=1}^{L} \theta_j^{\alpha^{(i)}q_j^{(i)}+n_j-1}.
\end{aligned}$$

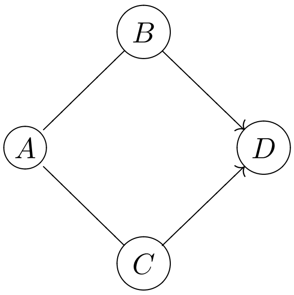
By standard Dirichlet integral calculations (see chapter 1),

$$E[\theta_j] = \sum_i \lambda_i \frac{\alpha^{(i)}q_j^{(i)} + n_j + 1}{n + \alpha^{(i)} + 1}$$

where n is the total sample size.

Chapter 6

1. DAGs are Markov equivalent if and only if they have the same skeletons and same immoralities. The first three are equivalent to each other, the fourth is not equivalent to any of the first three; it has an additional immorality at A. The essential graph for the first three is the graph



The fourth is already an essential graph.

2. (a) Yes: $A \to B \leftarrow C$ is an essential graph.

   (b) Recall that the essential graph is the graph where directions are retained on and only on those edges that retain the same direction in every graph in the Markov equivalence class. Hence $A - B \leftarrow C$ is not an essential graph since $A \to B \leftarrow C$ and $A \leftarrow B \leftarrow C$ are not Markov equivalent; if $B \leftarrow C$ is present and (A, B, C) is not an immorality, this forces $A \leftarrow B$.

   With this in mind, the essential graphs are:

   The three graphs $A \to B \leftarrow C$, $B \to A \leftarrow C$, $A \to C \leftarrow B$, the three graphs with one (undirected) edge between two of the nodes and the third node unconnected, the graph with no edges between any of the nodes, the three graphs with two undirected edges $A - B - C$, $A - C - B$, $C - A - B$. The graph with three undirected edges between A, B and C.

   (c) $A, B \leftarrow C$; $A - B \leftarrow C$; $A \leftarrow B \leftarrow C$; $A \to B, C$; $A \to B - C$; $A \to B \to C$. None of them are essential graphs.

   The motivation for this problem: this shows that to a Monte Carlo technique that runs through the space of essential graphs, it is necessary to change at least two edges for each transition.

3. (a) $p_{\underline{X}_k|\Theta}(\underline{x}_{(k)}|\underline{\theta}) = \prod_{j=1}^{d} \theta_{jil}^{n_k(j,i,l)}$ where $n_k(jil) = 1$ if the parent child configuration $(x_j^{(i)}, \pi_j^l)$ appears and 0 otherwise. It follows that $\prod_{k=1}^{n} p_{\underline{X}_k|\Theta}(\underline{x}_{(k)}|\underline{\theta}) = \prod_{j=1}^{n} \theta_{jil}^{n(jil)}$ where $n(jil)$

denotes the total number of times the parent child configuration $(x_j^{(i)}, \pi_j^l)$ appears in the n instantiations. Let $n(j.l) = \sum_{i=1}^{k_j}$ denote the number of times that the parent configuration π_j^l appears in the n instantiations. It follows that

$$\begin{aligned} p_{\mathbf{X}|\mathcal{E}}(\mathbf{x}|E) &= \prod_{j=1}^{d}\prod_{l=1}^{q_j}\int\left(\prod_{i=1}^{k_j}\theta_{jil}^{n(jil)}\right)\phi(\underline{\theta}_{j.l}, \underline{\alpha}_{j.l})d\underline{\theta}_{j.l} \\ &= \prod_{j=1}^{d}\prod_{l=1}^{q_j}\frac{\Gamma(\sum_{i=1}^{k_j}\alpha_{jil})}{\prod_{i=1}^{k_j}\Gamma(\alpha_{jil})}\frac{\prod_{i=1}^{k_j}\Gamma(n(jil)+\alpha_{jil})}{\Gamma(n(j.l)+\sum_{i=1}^{k_j}\alpha_{jil})} \end{aligned}$$

which, after rearrangement, gives the advertised result.

(b) To get uniform prior, clearly $\alpha_{jil} = 1$ for each jil. The likelihood function is then

$$p_{\mathbf{X}|\mathcal{E}}(\mathbf{x}|E) = \prod_{j=1}^{d}\prod_{l=1}^{q_j}\frac{(k_j-1)!}{(n(\pi_j^l)+k_j-1)!}\prod_{j=1}^{k_j} n(x_j^i|\pi_j^l)!.$$

4. This is a straightforward MATLAB exercise. and should produce edges

   (a) $\langle C_4, C_5\rangle$

   (b) $\langle C_4, C_1\rangle$

   (c) $\langle C_4, C_2\rangle$

   (d) $\langle C_5, C_3\rangle$.

The second and third may arise in either order. A Chow- Liu tree assumes that there is a decomposition

$$p_{X_1,\ldots,X_n} = p_{X_{\sigma(1)}}\prod_{j=2}^{n} p_{X_{\sigma(j)}|X_{\sigma(j-1)}}.$$

for some permuation σ. Therefore, there are no immoralities in the tree. The first edge can take either direction. Say (C_5, C_4) is chosen. The next two edge necessarily have the directions (C_4, C_1) and (C_4, C_2). The remaining edge (C_3, C_5) can take either direction.

The dependencies between $C_1 \leftrightarrow C_5$ and $C_2 \leftrightarrow C_5$ have been omitted. It is useful to check if the Chow Liu tree fits the data. The Chow Liu tree suggested above corresponds to a decomposition

$$p_{C_3}p_{C_5|C_3}p_{C_4|C_5}p_{C_1|C_4}p_{C_2|C_4}$$

It is straightforward to construct a MATLAB code to compute estimates of the potentials corresponding to the edges chosen; in this case the potentials to be estimated are: $\hat{p}_{C_3}$, $\hat{p}_{C_5|C_3}$, $\hat{p}_{C_4|C_5}$, $\hat{p}_{C_1|C_4}$ and $\hat{p}_{C_2|C_4}$.

Any test based on a χ^2 statistic will indicate that the Chow - Liu tree does not fit the data here. The distribution factorises according to

$$p_{C_1}p_{C_2}p_{C_3}p_{C_4|C_1,C_2}p_{C_5|C_3,C_4}$$

Although it has the same skeleton, it contains immoralities that are not present in a Chow - Liu decomposition.

5. MATLAB exercise - investigate the BN toolbox. (The final column is the category and therefore is *not* one of the variables to be used in the construction of the classifier.)

   The naive classifier (where it is assumed that the variables are independent) works very well with this data set; check the web site for other data sets where the naive classifier works less well and where the Chow-Liu tree may offer some improvement.

   Try a χ^2 test, both for the 'independence' model and Chow-Liu tree to see if either of these models fit the data. The 'independence' model does not fit. Appropriate adjustments should be made to accommodate cell probability estimates of zero.

6. MATLAB exercise.

7. MATLAB exercise. The code for MMHC is available in the Bayes nets toolbox.

8. MATLAB exercise. Compare with HUGIN.

9. MATLAB The algorithm is for finding a faithful DAG if there is one; success means that there is a DAG for which d-separation implies conditional independence; failure means that there does not. For a reasonable nominal significance level, it should be clear that there does not exist a DAG corresponding to the set of CI statements produced.

10. MATLAB

11. MATLAB The outputs from an McMC algorithm and the constraint based MMPC algorithm are likely to be different; the algorithms are based on different principles, with different critera for scoring the output.

Chapter 7

1. Assume that the variables are ordered so that $\Pi_j \subseteq \{X_1, \ldots, X_{j-1}\}$. Then

$$\begin{aligned}
D_{KL}(p,q) &= \sum_x p(x) \ln \frac{p(x)}{q(x)} \\
&= \sum_{\underline{x}=(x_1^{(i_1)},\ldots,x_d^{(i_d)})} \prod_{k=1}^{d} p_{X_k|\Pi_k}(x_k^{(i_k)}|\pi_k(\underline{x})) \ln \frac{\prod_{k=1}^{d} p_{X_k|\Pi_k}(x_k^{(i_k)}|\pi_k(\underline{x}))}{\prod_{k=1}^{d} q_{X_k|\Pi_k}(x_k^{(i_k)}|\pi_k(\underline{x}))} \\
&= \sum_{\underline{x}|\pi_j(\underline{x})=\pi_j^l} \prod_{k=1}^{d} p_{X_k|\Pi_k}(x_k^{(i_k)}|\pi_k(\underline{x})) \ln \frac{\theta_{ji_jl}}{\tilde{\theta}_{ji_jl}} \\
&= \sum_{i_j} \theta_{ji_jl} \ln \frac{\theta_{ji_jl}}{\tilde{\theta}_{ji_jl}} \sum_{(x_1^{(i_1)},\ldots,x_{j-1}^{(i_{j-1})})|\pi_j(\underline{x})=\pi_j^l} \prod_{k=1}^{j-1} p_{X_k|\Pi_k}(x_k^{(i_k)}|\pi_k(\underline{x}) \\
&= d_{KL}(\theta_{j.l}, \tilde{\theta}_{j.l}) p_{\Pi_j}(\pi_j^l).
\end{aligned}$$

2.

$$D_{CD}(p,q) = \max_x \ln \frac{p(x)}{q(x)} - \min_x \ln \frac{p(x)}{q(x)}.$$

Clearly, for any function f, $\max_x f(x) \geq \min_x f(x)$ so the distance is non negative. If $D_{CD}(p,q) = 0$, it follows that $\frac{p(x)}{q(x)} = \alpha$, a constant, for all $x \in \mathcal{X}$. It follows that $p(x) = \alpha q(x)$ so that

$$1 = \sum_x p(x) = \alpha \sum_x q(x) = \alpha$$

and hence $\alpha = 1$, so that $p(x) = q(x)$ for all $x \in \mathcal{X}$.

For the second point, if $p(x) > 0$ for a point where $q(x) = 0$, or $p(x) = 0$ for a point where $q(x) > 0$, then $D_{CD}(p,q) = +\infty$.

Now consider $p(x) > 0 \Leftrightarrow q(x) > 0$. For a strictly positive function f, $\max_x f(x) = \frac{1}{\min_x (1/f(x))}$ since the point where the maximum of $f(x)$ is attained is the point where the minimum of $1/f(x)$

is attained. It follows that

$$\begin{aligned} D_{CD}(p,q) &= \max \ln \frac{p(x)}{q(x)} - \min \ln \frac{p(x)}{q(x)} \\ &= \ln \frac{1}{\min_x \frac{q(x)}{p(x)}} - \ln \frac{1}{\max_x \frac{q(x)}{p(x)}} \\ &= -\min_x \ln \frac{q(x)}{p(x)} + \max_x \ln \frac{q(x)}{p(x)} \\ &= D_{CD}(p,q). \end{aligned}$$

3. Take any point $x \in \mathcal{X}$, then $x \in G_j$ for exactly one j. It follows that $q(x) = \mu_j p(x|G_j) = \mu_j \frac{p(x)}{\lambda_j}$ for j such that $x \in G_j$. Therefore

$$\begin{aligned} D_{CD}(p,q) &= \max_x \ln \frac{p(x)}{q(x)} - \min_x \ln \frac{p(x)}{q(x)} \\ &= \max_j \ln \frac{\lambda_j}{\mu_j} - \min_j \ln \frac{\lambda_j}{\mu_j} \end{aligned}$$

as required.

4.

$$\begin{aligned} S(p)(t) &= \max_{x\in\{0,1\}} \frac{d}{dt} \ln p_X(x|t) - \min_{x\in\{0,1\}} \frac{d}{dt} \ln p_X(x|t) \\ &= \max_{x\in\{0,1\}} \left(x\frac{d}{dt} \ln t + (1-x)\frac{d}{dt} \ln(1-t) \right) \\ &\quad - \min_{x\in\{0,1\}} \left(x\frac{d}{dt} \ln t + (1-x)\frac{d}{dt} \ln(1-t) \right) \\ &= \max_{x\in\{0,1\}} \left(\frac{x}{t} - \frac{1-x}{1-t} \right) - \min_{x\in\{0,1\}} \left(\frac{x}{t} - \frac{1-x}{1-t} \right) \\ &= \frac{1}{t} + \frac{1}{1-t} = \frac{1}{t(1-t)}. \end{aligned}$$

5. Let $\mathbb{P} = (p_0, p_1)$ be a Bernoulli trial with success probability $p_1 = \frac{1}{2}$ and $\mathbb{Q}$ a Bernoulli trial with success probability $q_1 = \theta > \frac{1}{2}$. Then

$$D_{CD}(\mathbb{Q},\mathbb{P}) = \ln 2\theta - \ln 2(1-\theta) = \ln \frac{\theta}{1-\theta}.$$

Hence, let $\theta(k)$ denote the value of θ such that $D_{CD}(\mathbb{Q},\mathbb{P}) = k$, then

$$e^k = \frac{\theta}{1-\theta}$$

giving

$$\theta(k) = \frac{e^k}{1+e^k}.$$

Considering $0 \leq \theta \leq \frac{1}{2}$ gives, for a Chan Darwiche distance k,

$$\theta(k) = \frac{e^{-k}}{1+e^{-k}}.$$

6. Typo: question should ask for sensitivity of $p_{A|C}(.|1)$. Let

$$\begin{aligned}p_C(1) = {} & p_A(1)(p_B(0)p_{C|A,B}(1|1,0) \\ & +p_B(1)p_{C|A,B}(1|1,1)) + p_A(0)(p_B(0)p_{C|A,B}(0|1,0) + p_B(1)p_{C|A,B}(0|1,1))\end{aligned}$$

$$\begin{aligned} p_{A|C}(1|1) &= \frac{p_A(1)(p_B(0)p_{C|A,B}(1|1,0) + p_B(1)p_{C|A,B}(1|1,1))}{p_C(1)} \\ &= \frac{s(0.8 \times (1-t) + t)}{s(0.8 \times (1-t) + t) + (1-s)((1-t) \times 0.2 + t \times 0)} \\ &= \frac{0.2st + 0.8s}{0.2st + 0.8s + 0.2 - 0.2s - 0.2t + 0.2st} \\ &= \frac{st + 4s}{2st + 3s - t + 1} \end{aligned}$$

$$p_{A|C}(0|1) = \frac{st - s - t + 1}{2st + 3s - t + 1} = \frac{(1-t)(1-s)}{2st + 3s - t + 1}$$

$$\frac{\partial}{\partial s} \ln p_{A|C}(1|1) = \frac{1}{s} - \frac{2t-5}{2st - 5s - t + 1} \qquad \frac{\partial}{\partial s} \ln p_{A|C}(0|1) = -\frac{1}{1-s} - \frac{2t-5}{2st - 5s - t + 1}$$

$$S(p_{A|C}(.|1))(s) = \frac{1}{s} + \frac{1}{1-s} = \frac{1}{s(1-s)}.$$

$$\frac{\partial}{\partial t} \ln p_{A|C}(1|1) - \frac{\partial}{\partial t} \ln p_{A|C}(0|1) = \frac{1}{t+4} + \frac{1}{1-t} = \frac{5}{(t+4)(1-t)} = S(p_{A|C}(.|1))(t).$$

7. D for disease; 1 yes, 0 no. R for result, 1 positive 0 negative

$$p_D(1) = t \qquad p_{R|D} = \begin{array}{c|cc} D\backslash R & 1 & 0 \\ \hline 1 & 0.99 & 0.01 \\ 0 & 0.02 & 0.98 \end{array}$$

$$p_R(1) = 0.99t + 0.02(1-t), \qquad p_{R,D} = \begin{array}{c|cc} D\backslash R & 1 & 0 \\ \hline 1 & 0.99t & 0.01t \\ 0 & 0.02(1-t) & 0.98(1-t) \end{array}$$

$$S(p_{D|R})(.|1) = \left| \frac{\partial}{\partial t} \ln p_{D,R}(1,1) - \frac{\partial}{\partial t} \ln p_{D,R}(0,1) \right| = \frac{1}{t} + \frac{1}{1-t} = \frac{1}{t(1-t)}.$$

$$p_{D|R}(1|1)(t = 0.01) = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.02 \times 0.99} = \frac{1}{3}$$

$$p_{D|R}(1|1)(t = 0.02) = \frac{0.99 \times 0.02}{0.99 \times 0.02 + 0.02 \times 0.98} = \frac{0.99}{1.97} \simeq 0.50$$

8. (solution omitted)

9. Question should ask: $p_{A|D}(1|1) = 0.8$.

$$\begin{aligned}
p_{A,D}(1,1) &= 0.25\sum_{b,c} p_{C|A,B}(c|1,b)p_{D|C}(1|c) \\
&= 0.25(p_{C|A,B}(1|1,1) + p_{C|A,B}(1|1,0))p_{D|C}(1|1) \\
&\qquad +0.25(p_{C|A,B}(0|1,1) + p_{C|A,B}(0|1,0))p_{D|C}(1|0) \\
&= 0.35
\end{aligned}$$

$$\begin{aligned}
p_{A,D}(0,1) &= 0.25\sum_{b,c} p_{C|A,B}(c|0,b)p_{D|C}(1|c) \\
&= 0.25(p_{C|A,B}(1|0,1) + p_{C|A,B}(1|0,0))p_{D|C}(1|1) \\
&\quad +0.25(p_{C|A,B}(0|0,1) + p_{C|A,B}(0|0,0))p_{D|C}(1|0) \\
&= 0.175t + 0.075 \times (2-t) = 0.1t + 0.15
\end{aligned}$$

$$p_D(1) = 0.1t + 0.5$$

$$p_{A|D}(1|1) = \frac{0.35}{0.1t + 0.15}$$

$$0.8 = \frac{7}{2t+3} \Rightarrow 1.6t + 2.4 = 7 \Rightarrow t = \frac{23}{8}$$

Not a solution; any solution must lie in the range $[0,1]$ hence no solution.

10. Gradient descent problem - solution omitted.

Chapter 8

1. (a) no

   (b) yes

$$f(x;\mu) = \exp\left\{-\frac{1}{2}(\mu + \ln \mu) - \frac{1}{2\mu}x^2\right\}\frac{e^x}{\sqrt{2\pi}}$$

   minimal sufficient statistic: X^2.

   (c) Yes:

$$\mathbb{P}(X = x | X \geq 1) = \frac{\binom{n}{x}p^x(1-p)^{n-x}}{1-(1-p)^n} \qquad x = 1, \ldots, n$$

$$p_X(x) = \binom{n}{x}\exp\left\{x\ln\left(\frac{p}{1-p}\right) + \ln\frac{(1-p)^n}{1-(1-p)^n}\right\}.$$

   minimal sufficient statistic: X

2.

3.

$$M(\lambda) = \mathbb{E}[e^{\lambda\phi(X)}] = \int \exp\{(\lambda+\theta)\phi(x) - A(\theta)\}h(x)dx = e^{A(\lambda+\theta)-A(\theta)}$$

$$M'(\lambda) = A'(\lambda+\theta)M(\lambda) \qquad M''(\lambda) = (A''(\lambda+\theta) + (A'(\lambda+\theta))^2)M(\lambda)$$

$$M'(0) = A'(\theta) \qquad M''(0) = (A''(\theta) + (A'(\theta))^2)$$

$$\mathbb{E}[\phi(X)] = M'(0) = A'(\theta)$$

$$\mathbb{E}[\phi^2(X)] = M''(0) = A''(\theta) + (A'(\theta))^2$$

$$\text{Var}(\phi(X)) = A''(\theta).$$

4. (a) Yes - using $x_3 = n - (x_1 + x_2)$, the probability function can be rewritten as:

$$p(x_1, x_2, x_3|\eta) = \frac{2^{x_2}n!}{x_1!x_2!x_3!}\exp\left\{(2x_1 + x_2)\ln\left(\frac{\eta}{1-\eta}\right) + 2n\ln(1-\eta)\right\}$$

   This is in the form required for an exponential family. The canonical parameter is:

$$\theta = \ln\left(\frac{\eta}{1-\eta}\right)$$

so that

$$\frac{\eta}{1-\eta} = e^{\theta} \Rightarrow \eta(1+e^{\theta}) = e^{\theta} \Rightarrow \eta = \frac{1}{1+e^{-\theta}}.$$

and the log partition function is

$$A(\theta) = -2n\ln(1-\eta) = -2n\ln\left(\frac{1}{1+e^{\theta}}\right) = 2n\ln(1+e^{\theta}).$$

(b) The minimal sufficient statistic is $2X_1 + X_2$.

(c) The mean parameter is:

$$\mu = \mathbb{E}_{\eta}[2X_1 + X_2] = 2n\eta^2 + 2n\eta(1-\eta) = 2n\eta = \frac{2n}{1+e^{-\theta}}.$$

(d)

$$A^*(\mu) = \sup_{\theta}\{\theta\mu - A(\theta)\} = \sup_{\theta}\{\theta\mu - 2n\ln(1+e^{\theta})\}$$

Maximum:

$$\mu - \frac{2ne^{\theta}}{1+e^{\theta}} = 0 \Rightarrow \mu = \frac{2n}{1+e^{-\theta}} \Rightarrow \theta = \ln\left(\frac{\mu}{2n-\mu}\right)$$

$$A^*(\mu) = \mu\ln\left(\frac{\mu}{2n-\mu}\right) - 2n\ln\left(\frac{2n}{2n-\mu}\right) = (2n-\mu)\ln(2n-\mu) + \mu\ln\mu - (2n)\ln(2n).$$

(e) Canonical parameters $\theta = \ln\frac{\eta}{1-\eta}$

$$\begin{aligned} D(\theta_1|\theta_2) &= \sum_{x_1,x_2,x_3} p(x_1,x_2,x_3|\eta_1)\ln\frac{p(x_1,x_2,x_3|\eta_1)}{p(x_1,x_2,x_3|\eta_2)} \\ &= \left(\ln\frac{\eta_1}{1-\eta_1} - \ln\frac{\eta_2}{1-\eta_2}\right)\mathbb{E}_{\eta_1}[2X_1+X_2] + (2n\ln(1-\eta_1) - 2n\ln(1-\eta_2)) \\ &= (\theta_1-\theta_2)\mu_1 + A(\theta_2) - A(\theta_1). \end{aligned}$$

so the result is verified in this example. For the next part,

$$\mu = \frac{2n}{1+e^{-\theta}} \Rightarrow \theta = \ln\left(\frac{\mu}{2n-\mu}\right).$$

$$A(\theta) = 2n\ln(1+e^{\theta}) = 2n\ln\left(\frac{2n}{2n-\mu}\right) = -A^*(\mu) + \mu\ln\left(\frac{\mu}{2n-\mu}\right)$$

Using this,

$$\begin{aligned} \tilde{D}(\mu_1|\theta_2) &= D(\theta_1|\theta_2) = (\theta_1-\theta_2)\mu_1 + A(\theta_2) - A(\theta_1) \\ &= A(\theta_2) - \mu_1\theta_2 + A^*(\mu_1) - \mu_1\ln\left(\frac{\mu_1}{2n-\mu_1}\right) + \mu_1\ln\left(\frac{\mu_1}{2n-\mu_1}\right) \\ &= A(\theta_2) + A^*(\mu_1) - \mu_1\theta_2 \end{aligned}$$

so the result for the mixed form of the Kullback Leibler divergence is verified in this case. Finally,

$$\begin{aligned} \tilde{\tilde{D}}(\mu_1|\mu_2) &= A^*(\mu_1) - A^*(\mu_2) - \mu_2\ln\left(\frac{\mu_2}{2n-\mu_2}\right) - \mu_1\theta_2 \\ &= A^*(\mu_1) - A^*(\mu_2) - \theta_2(\mu_1-\mu_2). \end{aligned}$$

(f) The primal form is given by

$$D(\theta_1|\theta_2) = (\theta_1 - \theta_2)\frac{2n}{1+e^{-\theta_1}} + 2n \ln\left(\frac{1+e^{\theta_2}}{1+e^{\theta_1}}\right)$$

The dual form by:

$$\begin{aligned}\tilde{\tilde{D}}(\mu_1|\mu_2) &= (2n-\mu_1)\ln(2n-\mu_1) \\ &-(2n-\mu_2)\ln(2n-\mu_2) + \mu_1 \ln \mu_1 - \mu_2 \ln \mu_2 - (\mu_1 - \mu_2)\ln\left(\frac{\mu_2}{2n-\mu_2}\right).\end{aligned}$$

5. This is a 'noisy logic gate' shown in figure 8.1

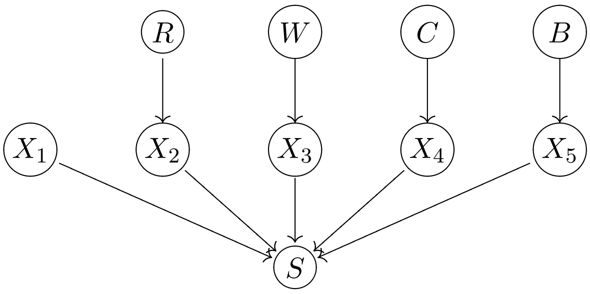


Figure 8.1: Noisy logic gate

Here X_1 denotes 'normal' circumstances; the treatment to fails with probability 0.1, hence p_{X_1} has probability table

$$p_{X_1}(0) = 0.9 \qquad p_{X_1}(1) = 0.1$$

Throughout 1 will denote that the cause for the treatment to fail is activated. The variable S denotes success or failure; 1 for success, 0 for failure.

$p_{X_2|R} =$

$R\backslash X_2$	1	0
1	$\frac{1}{7}$	$\frac{6}{7}$
0	0	1

$p_{X_3|W} =$

$W\backslash X_3$	1	0
1	$\frac{1}{4}$	$\frac{3}{4}$
0	0	1

$p_{X_4|C} =$

$C\backslash X_4$	1	0
1	$\frac{1}{3}$	$\frac{2}{3}$
0	0	1

$p_{X_5|B} =$

$B\backslash X_5$	1	0
1	$\frac{1}{2}$	$\frac{1}{2}$
0	0	1

$$p_{S|X_1,X_2,X_3,X_4,X_5}(1|x_1,x_2,x_3,x_4,x_5) = \begin{cases} 1 & x_1 = \ldots = x_5 = 0 \\ 0 & \text{otherwise} \end{cases}$$

There are 16 cases; take (for example) $(R,W,C,B) = (1,1,0,0)$

$$p_{S|R,W,C,B}(1|1,1,0,0) = 0.9 \times \frac{6}{7} \times \frac{3}{4}$$

the others are similar.

6. (a) Misprint in question: should read:

$$A_H^*(\underline{\mu}) = \sum_{j=1}^n \{\mu(j) \ln(1-\mu(j)) + (1-\mu(j)) \ln \mu(j)\}$$

q is the independence model;

$$q_{\underline{X}}(\underline{x}|\underline{\theta}) = \prod_{j=1}^n q_{X_j}(x(j)|\theta(j))$$

where

$$q_{X_j}(1|\theta(j)) = \frac{e^{\theta(j)}}{1+e^{\theta(j)}} \qquad q_{X_j}(0|\theta(j)) = \frac{1}{1+e^{\theta(j)}}$$

We have a precise expression for the partition function:

$$A_H(\underline{\theta}) = \sum_{j=1}^n \ln(1+e^{\theta(j)})$$

$$\frac{\partial}{\partial\theta(i)} A_H(\underline{\theta}) = \frac{1}{1+e^{\theta(i)}}$$

so $\theta(i)$ which maximises expression for $A^*(\mu)$ satisfies:

$$\mu(i) = \frac{1}{1+e^{\theta(i)}} \Rightarrow \theta(i) = \ln\left(\frac{1}{\mu(i)} - 1\right)$$

$$A_H(\underline{\theta}) = \sum_{j=1}^n \ln \frac{1}{\mu(j)}$$

which gives:

$$A_H^*(\underline{\mu}) = \sum_{j=1}^n \mu(j) \ln\left(\frac{1}{\mu(j)} - 1\right) - \sum_{j=1}^n \ln\frac{1}{\mu(j)} = \sum_{j=1}^n \{\mu(j)\ln(1-\mu(j)) + (1-\mu(j))\ln\mu(j)\}$$

as required.

(b) The partition function is given by:

$$\begin{aligned}
A(\underline{\theta}) &= \sup_{\mu} \{\langle \mu, \theta\rangle - A^*(\underline{\mu})\} \\
&= \sup_{\mu} \left\{ \sum_{j=1}^n \theta(j)\mu(j) + \sum_{\langle j,k\rangle \in E} \theta(j,k)\mu(j,k) - A^*(\underline{\mu}) \right\} \\
&\geq \sup_{\underline{\mu}|\mu(i,j)=\mu(i)\mu(j)\ \forall \langle i,j\rangle \in E} \left\{ \sum_{j=1}^n \theta(j)\mu(j) + \sum_{\langle j,k\rangle \in E} \theta(j,k)\mu(j,k) - A^*(\underline{\mu}) \right\} \\
&= \sup_{\underline{\mu}} \left\{ \sum_{j=1}^n \theta(j)\mu(j) + \sum_{\langle j,k\rangle \in E} \theta(j,k)\mu(j)\mu(k) - A_H^*(\underline{\mu}) \right\}
\end{aligned}$$

(c) Expressions are incorrect, they should be:

$$\begin{cases} \theta(1)+\theta(1,2)\mu(2)+\theta(1,3)\mu(3)+\ln\left(\frac{\mu(1)}{1-\mu(1)}\right)+\frac{2\mu(1)-1}{\mu(1)(1-\mu(1))}=0 \\ \theta(2)+\theta(1,2)\mu(1)+\ln\left(\frac{\mu(2)}{1-\mu(2)}\right)+\frac{2\mu(2)-1}{\mu(2)(1-\mu(2))}=0 \\ \theta(3)+\theta(1,3)\mu(1)+\ln\left(\frac{\mu(3)}{1-\mu(3)}\right)+\frac{2\mu(3)-1}{\mu(3)(1-\mu(3))}=0 \end{cases}$$

The expression to be maximised is:

$$\sum_{j=1}^{3}\theta(j)\mu(j)+\theta(1,2)\mu(1)\mu(2)+\theta(1,3)\mu(1)\mu(3)-\sum_{j=1}^{3}\mu(j)\ln(1-\mu(j))-\sum_{j=1}^{3}(1-\mu(j))\ln\mu(j)$$

Maximise: take derivatives with respect to $\mu(1)$, $\mu(2)$ and $\mu(3)$ respectively: For $\mu(1)$

$$\theta(1)+\theta(1,2)\mu(2)+\theta(1,3)\mu(3)-\ln(1-\mu(1))+\frac{\mu(1)}{1-\mu(1)}+\ln\mu(1)-\frac{1}{\mu(1)}+1=0$$

giving:

$$\theta(1)+\theta(1,2)\mu(2)+\theta(1,3)\mu(3)+\ln\left(\frac{\mu(1)}{1-\mu(1)}\right)+\frac{2\mu(1)-1}{\mu(1)(1-\mu(1))}=0$$

For $\mu(2)$:

$$\theta(2)+\theta(1,2)\mu(1)+\ln\left(\frac{\mu(2)}{1-\mu(2)}\right)+\frac{2\mu(2)-1}{\mu(2)(1-\mu(2))}=0$$

For $\mu(3)$:

$$\theta(3)+\theta(1,3)\mu(1)+\ln\left(\frac{\mu(3)}{1-\mu(3)}\right)+\frac{2\mu(3)-1}{\mu(3)(1-\mu(3))}=0$$

7. Notation implies one variable X_Γ which, conditioned on the random vector $\underline{X}_\Delta$ has Gaussian distribution; $X_\Gamma|\underline{X}_\Delta=\underline{i}\sim N(\mu(\underline{i}),\sigma^2(\underline{i}))$. The discrete random vector has probability function

$$p_{\underline{X}_\Delta}(\underline{i})=p(\underline{i}).$$

$$E[X_\Gamma]=E[E[X_\Gamma|\underline{X}_\Delta]]=\sum_{\underline{i}}p(\underline{i})E[X_\Gamma|\underline{X}_\Delta=\underline{i}]=\sum_{\underline{i}}p(\underline{i})\mu(\underline{i}).$$

$$E[X_\Gamma^2]=\sum_{\underline{i}}p(\underline{i})E[X_\Gamma^2|\underline{X}_\Delta=\underline{i}]=\sum_{\underline{i}}p(\underline{i})(\sigma^2(\underline{i})+\mu(\underline{i})^2)$$

$$\mathrm{Var}(X_\Gamma)=E[X_\Gamma^2]-E[X_\Gamma]^2=\sum_{\underline{i}}p(\underline{i})\sigma^2(\underline{i})+\sum_{\underline{i}}p(\underline{i})(\mu(i)-E[X_\Gamma])^2$$

since $E[X_\Gamma]^2=\sum_{\underline{i}}p(\underline{i})E[X_\Gamma]^2$ and

$$\begin{aligned}\sum_{\underline{i}}p(\underline{i})\left(\mu(\underline{i})^2-E[X_\Gamma]^2\right) &= \sum_{\underline{i}}p(\underline{i})\left((\mu(\underline{i})-E[X_\Gamma])^2+2\mu(\underline{i})E[X_\Gamma]-2E[X_\Gamma]^2\right) \\ &= \sum_{\underline{i}}p(\underline{i})(\mu(i)-E[X_\Gamma])^2\end{aligned}$$

8. I_1 and I_2 are binary variables with joint distribution $p_{I_1,I_2}(i_1,i_2)$ for $(i_1,i_2) \in \{0,1\}^2$. The continuous variables (X_1,X_2) have conditional distribution

$$(X_1,X_2)|(I_1,I_2) = \underline{i} \sim N\left(\underline{\mu}(\underline{i}), C(\underline{i})\right)$$

and the canonical parameters are (g,h,K) where

$$K(\underline{i}) = C(\underline{i})^{-1}, \qquad h(\underline{i}) = \underline{\mu}(\underline{i})C(\underline{i})^{-1}$$

$$g(\underline{i}) = \log p_{I_1,I_2}(\underline{i}) + \frac{1}{2}(\log \det K(\underline{i}) - 2\log 2\pi - \underline{\mu}(\underline{i})K(\underline{i})\underline{\mu}(\underline{i})^t).$$

(vectors are taken as row vectors)

9. Let $\underline{X} = (X_1,\ldots,X_d)$ be a random vector. The result is is a straightforward consequence of the fact that if $\underline{X} \sim N(\underline{\mu},C)$, then any subvector $\underline{X}^r = (X_{j_1},\ldots,X_{j_r})$ is obtained by marginalising over the components not in $\underline{X}$ and has distribution $\underline{X}^r \sim N(\underline{\mu}^r, C^r)$ where $\underline{\mu}^r = (\mu_{j_1},\ldots,\mu_{j_r})$ and $C^r_{kl} = C_{j_k,j_l}$. The canonical characteristics of this marginalised distribution are

$$K^r(\underline{i}) = C^r(\underline{i})^{-1}, \qquad h^r(\underline{i}) = \mu^r(\underline{i})C^r(\underline{i})^{-1}$$

$$g^r(\underline{i}) = g(\underline{i}) + \frac{1}{2}(\log\frac{\det K^r(\underline{i})}{\det K(\underline{i})} + (n-r)\log(2\pi) + \underline{\mu}(\underline{i})K(\underline{i})\underline{\mu}(\underline{i})^t - \underline{\mu}^r(\underline{i})K^r(\underline{i})\underline{\mu}^r(\underline{i})^t).$$

10. Conditioned on the discrete variables, the distribution is multivariate Gaussian. Let $\underline{X}_\Gamma = (\underline{X}_1, \underline{X}_2)$, where the conditioning is on $\underline{X}_1 = \underline{x}_1$. Then, conditioned on the discrete variables, $\underline{X}_\Delta = \underline{i}$, let

$$C(\underline{i}) = \left(\begin{array}{cc} C_{11}(\underline{i}) & C_{12}(\underline{i}) \\ C_{21}(\underline{i}) & C_{22}(\underline{i}) \end{array}\right).$$

The covariance (conditioned on $\underline{X}_\Delta = \underline{i}$) for $\underline{X}_2 - C_{21}C_{11}^{-1}\underline{X}_1$ and $\underline{X}_1$ is

$$C_{21} - C_{21}C_{11}^{-1}C_{11} = 0$$

hence $\underline{X}_2 - C_{21}C_{11}^{-1}\underline{X}_1 \perp \underline{X}_1 | \underline{X}_\Delta$ and

$$\underline{X}_2 - C_{21}C_{11}^{-1}\underline{X}_1 | \underline{X}_\Delta = \underline{i} \sim N(\underline{\mu}_2(\underline{i}) - C_{21}C_{11}^{-1}\underline{\mu}_1(\underline{i}), C_{22}(\underline{i}) - C_{21}(\underline{i})C_{11}^{-1}(\underline{i})C_{12}(\underline{i}))$$

It follows that

$$\underline{X}_2|\{\underline{X}_1 = \underline{x}_1, \underline{X}_\Delta = \underline{i}\} \sim N\left(\underline{\mu}_2(\underline{i}) + C_{21}(\underline{i})C_{11}^{-1}(\underline{i})(\underline{x}_1 - \underline{\mu}_1(\underline{i})), C_{22}(\underline{i}) - C_{21}(\underline{i})C_{11}^{-1}(\underline{i})C_{12}(\underline{i})\right).$$

The parameters $p(\underline{i})$ remain unaltered.

Chapter 9

1. (a) $p_{B\|C}(1\|1) = p_B(1) = \frac{1}{2}$

   (b) $p_{A\|C}(1\|1) = p_A(1) = \frac{1}{2}$

   (c) $p_{A\|B}(1\|1) = p_A(1) = \frac{1}{2}$

   (d) $p_{C\|B}(1\|1) = p_{C|B}(1|1) = \frac{1}{8}$

2.

$$p_{A,B} = \begin{array}{c|cc} A\backslash B & 1 & 0 \\ \hline 1 & 0.37 & 0.13 \\ 0 & 0.13 & 0.37 \end{array}$$

$$p_{C|A,B}(1|.,.) = \begin{array}{c|cc} A\backslash B & 1 & 0 \\ \hline 1 & 0.41 & 0.31 \\ 0 & 0.77 & 0.70 \end{array}$$

$$p_{B,C} = \begin{array}{c|cc} B\backslash C & 1 & 0 \\ \hline 1 & 0.25 & 0.25 \\ 0 & 0.3 & 0.2 \end{array}$$

$$p_B(1) = 0.5$$

$$p_{C|B} = \begin{array}{c|cc} B\backslash C & 1 & 0 \\ \hline 1 & 0.5 & 0.5 \\ 0 & 0.6 & 0.4 \end{array}$$

To show that this is an example of Simpson's paradox:

$$p_{C|A,B}(1|1,1) = 0.41 > 0.31 = p_{C|A,B}(1|1,0)$$

$$p_{C|A,B}(1|0,1) = 0.77 > 0.74 = p_{C|A,B}(1|0,0)$$

but

$$p_{C|B}(1|1) = 0.5 < 0.6 = p_{C|B}(1|0)$$

Hence this gives an example of Simpson's paradox.

(a) p_B, p_C, $p_{A|B,C}$

(b) $p_A(1) = 0.5$, $p_B(1) = 0.5$, $p_C(1) = 0.55$

$$p_{A|B,C}(1|.,.) = \begin{array}{c|cc} B\backslash C & 1 & 0 \\ \hline 1 & 0.6 & 0.88 \\ 0 & 0.13 & 0.2 \end{array}$$

so estimated probabilities obtained using

$$\hat{p}_{A,B,C} = p_B p_C p_{A|B,C} = \begin{array}{cc|cc} & & C & \\ A & B & 1 & 0 \\ \hline 1 & 1 & 0.165 & 0.198 \\ & 0 & 0.0358 & 0.101 \\ \hline 0 & 1 & 0.11 & 0.033 \\ & 0 & 0.239 & 0.124 \end{array}$$

(c) 6 parameters; $p_C(1)$, $p_B(1)$, 4 from $p_{A|B,C}(1|.,.)$.

(d)

$$\begin{aligned} d_{KL}(\mathbf{x}|\mathbf{y}) &= -\sum_j x_j \log \frac{y_j}{x_j} \\ &= -0.15 \log \frac{0.165}{0.15} - 0.22 \log \frac{0.198}{0.22} - 0.04 \log \frac{0.0358}{0.04} - 0.09 \log \frac{0.101}{0.09} \\ &\quad -0.1 \log \frac{0.11}{0.1} - 0.03 \log \frac{0.033}{0.03} - 0.26 \log \frac{0.239}{0.26} - 0.11 \log \frac{0.124}{0.11} \\ &= 0.0088 \end{aligned}$$

If the model fits, then $2nd_K(\mathbf{x}|\mathbf{y}) \sim \chi_1$. Test statistic $800 \times 0.0088 = 7.04 > \chi_1^2(95\%) = 3.84$, so the model does not fit the data (testing with a 5% significance level).

(e) Need appropriate potentials;

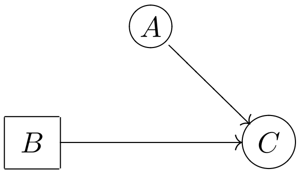
$$p_A = (0.5, 0.5)$$

$$p_{A,B} = \begin{array}{c|cc} A\backslash B & 1 & 0 \\ \hline 1 & 0.37 & 0.13 \\ 0 & 0.13 & 0.37 \end{array}$$

$$p_{B|A} = \begin{array}{c|cc} A\backslash B & 1 & 0 \\ \hline 1 & 0.74 & 0.26 \\ 0 & 0.26 & 0.74 \end{array}$$

$$p_{C|A,B}(1|.,.) = \begin{array}{c|cc} A\backslash B & 1 & 0 \\ \hline 1 & 0.41 & 0.31 \\ 0 & 0.77 & 0.70 \end{array}$$

Forcing B results in the network



so $p_{C|A\|B}(1|1\|1) = p_{C|B,A}(1|1,1)$ as before, but

$$p_{C\|B}(1\|1) = p_A(1)p_{C|A,B}(1|1,1) + p_A(0)p_{C|A,B}(1|0,1) = 0.59$$

$$p_{C\|B}(1\|0) = 0.505$$

so the inequality is in the same direction for the marginal and both the conditional distributions.

3. Let V denote the variable set and let $\tilde{p}_{V\backslash C} = p_{V\backslash C\|C}(.\|x_C)$. Then $\tilde{p}$ factorises along the graph $\mathcal{G}^{-C}_{V\backslash C}$ (the subgraph of $\mathcal{G}^{-C}$ with the nodes C removed) and, for $X \notin C$, conditional probability potentials $\tilde{p}_{X|\tilde{\Pi}_X} = p_{X|\Pi_X}$ where $\tilde{\Pi}_X = \Pi_X \backslash C$, Π_X denotes the original neighbour set, and the variables in $\Pi_X \cap C$ instantiated with the appropriate values.

   If $A \perp B\|_{\mathcal{G}^{-C}} C \cup W$ then any trail from A to B either has a fork or chain node in $C \cup W$ or a collider node that is not in $C \cup W$ with no descendants in $C \cup W$. It follows that, on the graph $\mathcal{G}^{-C}_{V\backslash C}$, any trail from A to B either has a fork or chain node in W or a collider node that is not in W with no descendants in W; edges are deleted, but not added, by taking the subgraph restricted to the variables of $V\backslash C$ and hence no new trails are added by removing the nodes in C. It follows that $A \perp B\|_{\mathcal{G}^{-C}_{V\backslash C}} W$ and hence that

$$p_{A|W,B\|C}(x_A|x_W, x_B\|x_C) = \tilde{p}_{A|W,B}(x_A|x_W, x_B) = \tilde{p}_{A|W}(x_A|x_W) = p_{A|W\|C}(x_A|x_W\|x_C)$$

   which is the result.

4. Let $\tilde{p}_{V\backslash\{X\}} = p_{V\backslash\{X\}\|X}(.\|x)$. Then $\tilde{p}$ factorises along $\mathcal{G}^{-X}_{V\backslash\{X\}}$, the subgraph of $\mathcal{G}^{-X}$ over the variables $V\backslash\{X\}$. The probability tables are, for $Y \neq X$ and parent sets $\tilde{\Pi}_Y = \Pi_Y\backslash\{X\}$ ($\tilde{\Pi}_Y$ is the original parent set of Π_Y with X removed) $\tilde{p}_{Y|\tilde{\Pi}_Y} = p_{Y|\Pi_Y}$ with the instantiation $X \leftarrow x$ for every appearance of X in Π_Y. If $Y \perp Z\|_{\mathcal{G}^{-X}} X$, then all trails between Y and Z in $\mathcal{G}^{-X}$ have either X as a fork or chain node, or else have a collider node that is not X and which does not

have X as a descendant. It follows that all trails between Y and Z in $\mathcal{G}_{V\backslash\{X\}}^{-X}$ have at least one collider node and hence that $Y \perp Z\|_{\mathcal{G}_{V\backslash\{X\}}^{-X}} \phi$ (d separated when none of the other variables are instantiated. It follows that, under probability distribution $\tilde{p}$, $Y \perp Z$, so that

$$p_{Y|Z\|X}(.|.\|x) = \tilde{p}_{Y|Z} = \tilde{p}_Y = p_{Y\|X}(.\|x).$$

5. $C_1 = \{X_1, X_2\}$ does not satisfy the back door criterion; $Y - X_4 - Z$ is a trail between Y and Z with an edge pointing to Y which is not blocked by C_1.

   $C_2 = \{X_4, X_5\}$ satisfies the back door criterion; trail $Y - X_6 - Z$ does not have an edge pointing towards Y. The other trails pass through X_4. For the trails $Y - X_4 - Z$ and $Y - X_3 - X_1 - X_4 - Z$, X_4 is an instantiated fork or chain respectively, hence C_2 blocks the trail. For $Y - X_1 - X_4 - X_2 - X_5 - Z$, X_5 is an instantiated chain and hence the trail is blocked. All trails between Y and Z have been considered.

   For the backdoor criterion with respect to (Z, Y), the sets have to block all trails with an arrow pointing towards Z. This means that any set that contains X_6, X_4 and any node from $\{X_3, X_1, X_2, X_5\}$ will satisfy the backdoor criterion with respect to (Z, Y); any set that does not will not.

6. It is clear that
$$p_{Y\|X}(y\|x) = \sum_c p_{Y|C\|X}(y|c\|x)p_{C\|X}(c\|x).$$
   Since C blocks all trails between Y and X that have an edge pointing towards X, it follows that $Y \perp (\Pi_X\backslash C)\|_{\mathcal{G}}C$. It follows, with notation that should be clear, using the proposition 9.6 (adjustment for direct causes) the text that

$$\begin{aligned}
p_{Y|C\|X}(y|c\|x) &= \sum_{\pi\backslash c} p_{Y|C,\Pi_X\|X}(y|c,\pi\backslash c\|x)p_{\Pi_X\backslash C\|X}(\pi\backslash c\|x) \\
&= \sum_{\pi\backslash c} p_{Y|C,\Pi_X,X}(y|c,\pi\backslash c,x)p_{\Pi_X\backslash C}(\pi\backslash c) \\
&= \sum_{\pi\backslash c} p_{Y|C,X}(y|c,x)p_{\Pi_X\backslash C}(\pi\backslash c) \\
&= p_{Y|C,X}(y|c,x).
\end{aligned}$$

   Furthermore, since none of the variables in C are descendants of X, it follows (again, using proposition 9.6 (adjustment for direct causes) in the text) that

$$p_{C\|X}(c\|x) = p_C(c)$$

   and the result follows. The fact that $p_{C\|X}(c\|x) = p_C(c)$,

$$p_{Y|C,\Pi_X\|X}(y|c,\pi\backslash c\|x) = p_{Y|C,\Pi_X,X}(y|c,\pi\backslash c,x)$$

and $p_{\Pi_X\backslash C\|X}(\pi\backslash c\|x) = p_{\Pi_X\backslash C}(\pi\backslash c)$ is clear by comparing the original DAG and the mutilated graph. A formal algebraic proof that $p_{C\|X}(c\|x) = p_C(c)$ is given in the next exercise.

7. The variables may be ordered as $V = \{Y_1, \ldots, Y_n, X, Y_{n+1}, \ldots, Y_{n+m}\}$ where the ordering is chosen such that $\Pi(Y_j) \subseteq \{Y_1, \ldots, Y_{j-1}\}$ for $j \leq n$, $\Pi(X) \subseteq \{Y_1, \ldots, Y_n\}$,

$$\Pi(Y_j) \subseteq \{Y_1, \ldots, Y_n, X, Y_{n+1}, \ldots, Y_{j-1}\}$$

for $j \in \{m+1, \ldots, n+m\}$ and where $C \subseteq \{Y_1, \ldots, Y_n\}$. From the intervention formula,

$$p_{V\backslash X\|X}(y_1, \ldots, y_{m+n}\|x) = \prod_{j=1}^{m+n} p_{Y_j|\Pi_j}(y_j|\pi_j)$$

while

$$p_V(y_1, \ldots, y_{m+n}, x) = p_{X|\Pi(X)}(x|\pi_X) \prod_{j=1}^{m+n} p_{Y_j|\Pi_j}(y_j|\pi_j).$$

Now, sum over variables $Y_{n+1}, \ldots, Y_{n+m}$ in both expressions, then sum over X in the second expression. Then sum over all remaining variables not in C. The same answer obtains for both expressions, so that

$$p_{C\|X} = p_C.$$

8. Firstly,

$$p_{V\backslash\{X_j\}\|\{X_j\}} = p_{V\backslash(\{X_j\}\cup\Pi_j)|X_j,\Pi_j}p_{\Pi_j}.$$

Now let $\Pi_U = \Pi_j \cap U$, U_a denote ancestors of X_j in U, U_b ancestors of $Z\backslash(\{X_j\} \cup \Pi_j)$ in U and $U_c = U\backslash(\Pi_U \cup U_a \cup U_b)$.

Sum over the variables in $V\backslash(Z\cup\Pi_U)$, then, from the condition that there are no trails between X_j and other variables in Z that contain only fork or chain connections, $Z\backslash\{X_j\} \cup \Pi_j$ is d-separated from Π_U by $\{X_j\}$. It follows that

$$\begin{aligned} p_{Z\cup\Pi_U\backslash\{X_j\}\|\{X_j\}} &= p_{Z\backslash(\{X_j\}\cup(\Pi_j\cap Z))|X_j,(\Pi_j\cap Z),\Pi_U}p_{\Pi_j} \\ &= p_{Z\backslash(\{X_j\}\cup(\Pi_j\cap Z))|X_j,(\Pi_j\cap Z)}p_{\Pi_j} \end{aligned}$$

so that

$$p_{Z\backslash\{X_j\}\|\{X_j\}} = p_{Z\backslash(\{X_j\}\cup(\Pi_j\cap Z)\cup\Pi_U)|X_j,(\Pi_j\cap Z)}p_{\Pi_j\cap Z}.$$

Chapter 10

1. (a) The domain graph is given in figure 10.1.

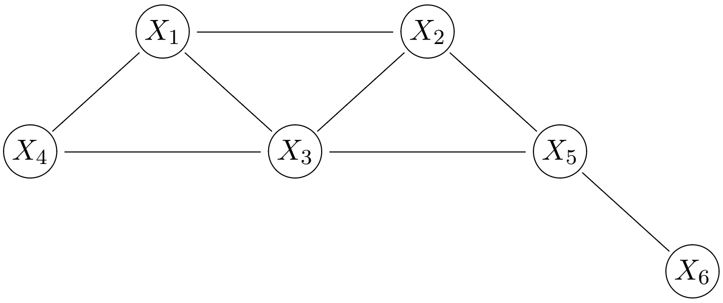


Figure 10.1: Domain Graph

   (b) The domain graph with X_3 eliminated given in figure 10.2. The corresponding potentials, after elimination, are

$$\tilde{\phi}_{1;X_1,X_2,X_4,X_5} = \sum_{x \in \mathcal{X}_3} \phi_{1;X_1,X_2,X_3}(.,.,x)\phi_{2;X_2,X_3,X_5}(.,x,.)\phi_{3;X_1,X_3,X_5}(.,x,.)$$

$$\tilde{\phi}_{2;X_5,X_6} = \phi_{4;X_5,X_6}.$$

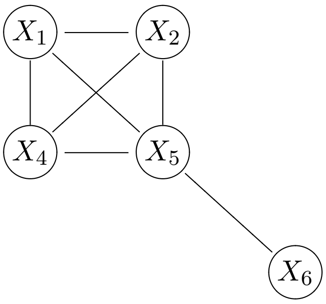


Figure 10.2: Domain Graph, X_3 eliminated

(c) For example: the sequence $X_6, X_5, X_2, X_4, X_3, X_1$ does not require any fill-ins and hence is a perfect elimination sequence.

2. (a) The domain graph given in figure 10.3.

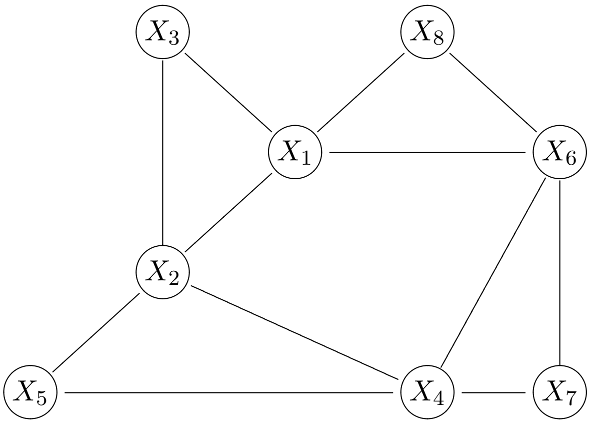


Figure 10.3: Domain Graph

(b) The domain graph, with X_1 eliminated, is given in figure 10.4. Potentials are

$$\tilde{\phi}_{1;X_2,X_3,X_6,X_8} = \sum_{x \in \mathcal{X}_1} \phi_{1;X_1,X_2,X_3}(x,.,.)\phi_{4;X_1,X_6,X_8}(x,.,.)$$

$$\tilde{\phi}_{2;X_2,X_4,X_5} = \phi_{2;X_2,X_4,X_5}$$

$$\tilde{\phi}_{3;X_4,X_6,X_7} = \phi_{3;X_4,X_6,X_7}.$$

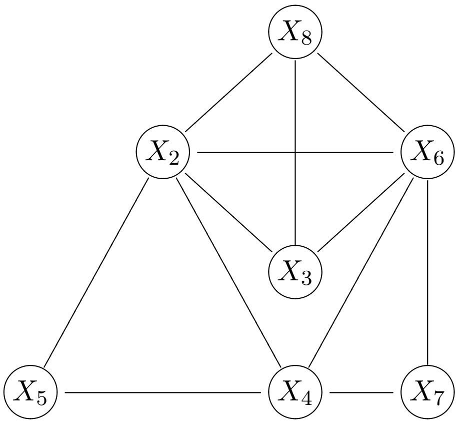


Figure 10.4: Domain Graph after elimination of X_1.

(c) No: X_1, X_2, X_4, X_6, X_1 is a cycle of length 4 without a chord. The graph is not triangulated, hence there is no perfect elimination sequence.

3. (a) The potentials, where the variables have been put in the subscript after the label, are

$$\phi_{1;X_1} = p_{X_1}, \quad \phi_{2;X_6} = p_{X_6}, \quad \phi_{3;X_1,X_2} = p_{X_2|X_1},$$

$$\phi_{4;X_1,X_3} = p_{X_3|X_1}, \quad \phi_{5;X_3,X_5} = p_{X_5|X_3}, \quad \phi_{6;X_2,X_4,X_6} = p_{X_4|X_2,X_6},$$

$$\phi_{7;X_5,X_8} = p_{X_8|X_5}, \quad \phi_{8;X_4,X_5,X_7} = p_{X_7|X_4,X_5}.$$

Adding in the edges between the parents and making all edges undirected gives the moral graph, given in figure 10.5.

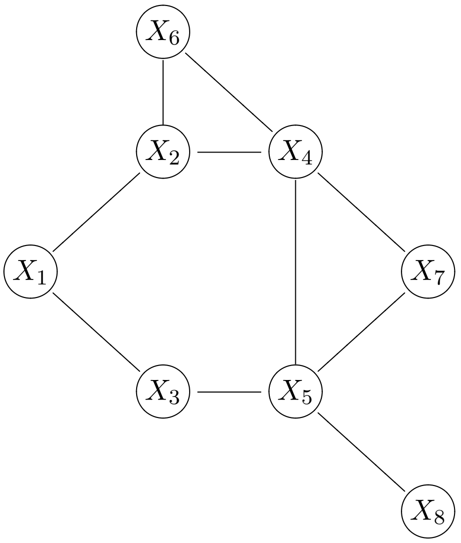


Figure 10.5: Moral Graph

(b) No: there is clearly a cycle length 5 ($X_1, X_2, X_4, X_5, X_3, X_1$) hence the graph is not triangulated.

(c) At least two new chords are needed. An example is given in the graph in figure 10.6 where the edges X_2, X_3 and X_4, X_5 have been added.

4. (a) A possible elimination sequence, with no fill ins, associated with the triangulated graph in figure 10.6: $X_8, X_6, X_7, X_1, X_5, X_2, X_3, X_4$

(b) The factorisation according to the DAG in the statement of the question is

$$p_{X_1,\ldots,X_8} = p_{X_1}p_{X_6}p_{X_2|X_1}p_{X_3|X_1}p_{X_4|X_2,X_6}p_{X_5|X_3}p_{X_7|X_4,X_5}p_{X_8|X_5}.$$

This requires the potentials p_{X_1}, p_{X_6}, $p_{X_2|X_1}$, $p_{X_3|X_1}$, $p_{X_4|X_2,X_6}$, $p_{X_5|X_3}$, $p_{X_7|X_4,X_5}$, $p_{X_8|X_5}$. Note that (for example) potential p_{X_1} needs only 2 entries, since $\sum_x p_{X_1}(x) = 1$. Let $n(.)$ denote the number of entries required for a potential. It follows that

$$n(p_{X_1}) = 2, \; n(p_{X_2|X_1}) = 6, \; n(p_{X_3|X_1}) = 6, \; n(p_{X_6}) = 2, \; n(p_{X_5|X_3}) = 6,$$

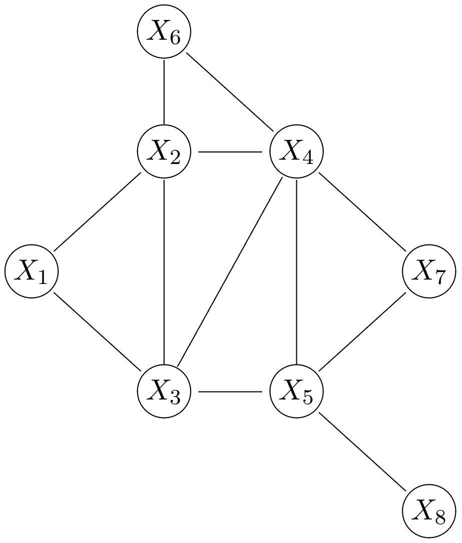


Figure 10.6: triangulated graph

$$n(p_{X_4|X_2.X_6}) = 18,\ n(p_{X_7|X_4,X_5}(x_7^{(1)}|.,.)) = 9,\ n(p_{X_8|X_5}(x_8^{(1)}|.)) = 3$$
$$\text{total} = 52$$

(c) The factorisation is

$$p_{X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8} = p_{X_1}p_{X_2|X_1}p_{X_3|X_1}p_{X_4|X_2,X_6}p_{X_5|X_3}p_{X_6}p_{X_7|X_4,X_5}p_{X_8|X_5}$$

$$p_{X_1|X_7,X_8}(.|x_7^{(1)}, x_8^{(1)}) = \frac{p_{X_1,X_7,X_8}(., x_7^{(1)}, x_8^{(1)})}{p_{X_7,X_8}(x_7^{(1)}, x_8^{(1)})}$$

and

$$p_{X_7,X_8}(x_7^{(1)}, x_8^{(1)}) = \sum_{j=1}^{3} p_{X_1,X_7,X_8}(x_1^{(j)}, x_7^{(1)}, x_8^{(1)}),$$

so only necessary to compute $p_{X_1,X_7,X_8}(., x_7^{(1)}, x_8^{(1)})$ and then marginalise.

The bucket elimination procedure according to the elimination order X_2, X_6, X_4, X_5, X_3 to obtain $p_{X_1,X_7,X_8}(., x_7^{(1)}, x_8^{(1)})$

$$\begin{aligned}
&p_{X_1,X_7,X_8}(., x_7^{(1)}, x_8^{(1)}) \\
&\quad = p_{X_1}(.) \sum_{j_3=1}^{3} p_{X_3|X_1}(x_3^{(j_3)}|.) \sum_{j_5=1}^{3} p_{X_8|X_5}(x_8^{(1)}|x_5^{(j_5)}) p_{X_5|X_3}(x_5^{(j_5)}|x_3^{(j_3)}) \\
&\quad \times \sum_{j_4=1}^{3} p_{X_7|X_4,X_5}(x_7^{(1)}|x_4^{(j_4)}, x_5^{(j_5)}) \sum_{j_6=1}^{3} p_{X_6}(x_6^{(j_6)}) \\
&\quad \times \sum_{j_2=1}^{3} p_{X_2|X_1}(x_2^{(j_2)}|.) p_{X_4|X_2,X_6}(x_4^{(j_4)}|x_2^{(j_2)}, x_6^{(j_6)}),
\end{aligned}$$

where the summations are performed from right to left (i.e. first $\sum_{j_2=1}^{3}$ then $\sum_{j_6=1}^{3}$, etc ....) The size of the potentials required are as follows:

The potential

$$\phi_{2;X_1,X_2,X_4,X_6} := p_{X_2|X_1} p_{X_4|X_2,X_6}$$

needs 3^4 entries. Marginalise over X_2 and then multiply by p_{X_6}

$$\phi_{6;X_1,X_4,X_6} := p_{X_6} \sum_{X_2} \phi_{2;X_1,X_2,X_4,X_6}$$

needs 3^3 entries. Marginalise over X_6 and then multiply by $p_{X_7|X_4,X_5}(x_7^{(1)}|.,.)$

$$\phi_{4;X_1,X_4,X_5} = p_{X_7|X_4,X_5}(x_7^{(1)}|.,.) \sum_{X_6} \phi_{6;X_1,X_4,X_6}$$

needs 3^3 entries. Marginalise over X_4 and then multiply by $p_{X_8|X_5}(x_8^{(1)}|.)p_{X_5|X_3}(.|.)$ gives

$$\phi_{5;X_1,X_3,X_5} = p_{X_8|X_5}(x_8^{(1)}|.)p_{X_5|X_3}(.|.) \sum_{X_4} \phi_{4;X_1,X_4,X_5}$$

needs 3^3 entries. Marginalise over X_5 and multiply by $p_{X_3|X_1}$ gives

$$\phi_{3;X_1,X_3} = p_{X_3|X_1} \sum_{X_5} \phi_{5;X_1,X_3,X_5}$$

needs 3^2 entries. Marginalise over X_3 and multiply by p_{X_1} gives

$$p_{X_1,X_7,X_8}(.,x_7^{(1)},x_8^{(1)}) = p_{X_1} \sum_{X_3} \phi_{3;X_1,X_3}$$

needs 3 entries.

(d) Since all variables have the same number of states (i.e. 3), try an elimination sequence that requires as few fill-ins as possible at each stage. X_6, X_2, X_4, X_5, X_3

Potential $\phi_{6;X_2,X_4,X_6} = p_{X_6} p_{X_4|X_2,X_6}$ requires 3^3 entries. Marginalise over X_6 and then multiply by $p_{X_2|X_1}$

$$\phi_{2;X_1,X_2,X_4} = p_{X_2|X_1} \sum_{X_6} \phi_{6;X_2,X_4,X_6}$$

Need 3^3 entries. Marginalise over X_2, then multiply by $p_{X_7|X_4}(x_7^{(1)}|.)$

$$\phi_{4;X_1,X_4,X_5} = p_{X_7|X_4,X_5}(x_7^{(1)}|.,.) \sum_{X_2} \phi_{2;X_1,X_2,X_4,}$$

Need 3^3 entries. Marginalise over X_4 and multiply by $p_{X_8|X_5}(x_8^{(1)}|.)p_{X_5|X_3}(.|.)$

$$\phi_{5;X_1,X_3,X_5} = p_{X_8|X_5}(x_8^{(1)}|.)p_{X_5|X_3}(.|.) \sum_{X_4} \phi_{4;X_1,X_4,X_5}$$

Need 3^3 entries. Marginalise over X_5 and multiply by $p_{X_3|X_1}$ to get

$$\phi_{3;X_1,X_3} = p_{X_3|X_1} \sum_{X_5} \phi_{5;X_1,X_3,X_5}$$

Marginalise over X_3 and multiply by p_{X_1} to get

$$p_{X_1,X_7,X_8}(.,x_7^{(1)},x_8^{(1)}) = p_{X_1}(.) \sum_{X_3} \phi_{3;X_1,X_3}$$

a potential that requires 3 elements. The size of the potentials is most easily checked from elimination in the domain graph, corresponding to an elimination sequence where X_7 and X_8 are considered in the appropriate place. The sequence corresponding to the procedure described above is:

Eliminate X_6; no fill ins.
Eliminate X_2; one fill in $\langle X_1, X_4\rangle$
Eliminate X_7: no fill ins
Eliminate X_4; one fill in $\langle X_1, X_5\rangle$
Eliminate X_8: no fill ins
Eliminate X_5; no fill ins
Eliminate X_3; no fill ins

5. (a) The question should state 'Let $\mathcal{G}$ be the domain graph for a charge Φ'. The result then follows trivially from the definition. When eliminating a node, all potentials in Φ containing the node α_v are multiply them together and this product is marginalised over α_v. These potentials that have been multiplied together are then replaced with this new potential.

   From this, it follows that the domain graph is modified precisely by adding edges between all neighbours of α_v and only these edges, and then removing α_v. This is the definition of removing a node. The domain graph, for $\Phi^{-\alpha_v}$ is therefore, by definition of removing a node, the graph $\mathcal{G}^{-\alpha_v}$.

   (b) This, again, follows almost trivially from the definition. When a node α is eliminated, the current potentials containing α are taken and multiplied together to construct a new potential. From the definition, the domain of the new potential is N_α in the graph just before the elimination.

6. Moralise the network by undirecting the edges and adding in $\langle X_1, X_2\rangle$, $\langle X_4, X_5\rangle$, $\langle X_5, X_6\rangle$. The moral graph is already triangulated; no fill ins. The junction tree given in figure 10.7.

   (a) fully active schedule:
   $X_2X_9 \to X_1X_2X_3$, $X_1X_2X_3 \to X_3X_4X_5$, $X_4X_5X_7 \to X_3X_4X_5$, $X_5X_6X_8 \to X_3X_5X_6$, $X_3X_5X_6 \to X_3X_4X_5$
   $X_3X_4X_5 \to X_3X_5X_6$, $X_3X_5X_6 \to X_5X_6X_8$, $X_3X_4X_5 \to X_4X_5X_7$, $X_3X_4X_5 \to X_1X_2X_3$, $X_1X_2X_3 \to X_2X_9$.

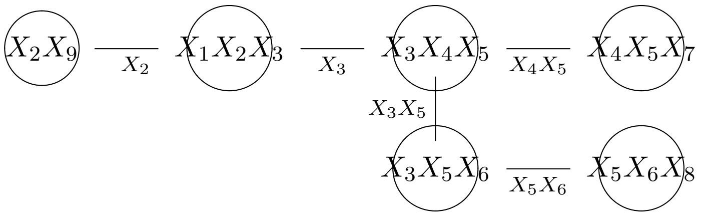


Figure 10.7: Junction Tree

(b) Since the evidence does not affect $X_1X_2X_3$ or X_2X_9, the initial messages $X_2X_9 \rightarrow X_1X_2X_3$ and $X_1X_2X_3 \rightarrow X_3X_4X_9$ are unnecessary. All other messages (including $X_3X_4X_5 \rightarrow X_1X_2X_3$ and $X_1X_2X_3 \rightarrow X_2X_9$) are to be included.

The implementation is now discussed. The algorithm proceeds as follows:

$$\phi_{X_4,X_5,X_7}(.,x_4^{(1)},x_5^{(3)}) = p_{X_4,X_5,X_7}(.,x_4^{(1)},x_5^{(3)})$$
$$\phi^*_{X_4,X_5}(x_4^{(1)},x_5^{(3)}) = \sum_y \phi_{X_4,X_5,X_7}(.,x_4^{(1)},x_5^{(3)})\phi_{X_4,X_5}(x_4^{(1)},x_5^{(3)})$$

(in this case, the updated potential over the cliques is the same as the original - nothing happens in the transmission)

$$\phi^*_{X_3,X_4,X_5}(.,x_4^{(1)},x_5^{(1)}) = \phi_{X_3,X_4,X_5}(.,x_4^{(1)},x_5^{(1)})\frac{\phi^*_{X_4X_5}(x_4^{(1)},x_5^{(1)})}{\phi_{X_4X_5}(x_4^{(1)},x_5^{(1)})} = \phi_{X_3,X_4,X_5}(.,x_4^{(1)},x_5^{(1)})$$

(nothing is altered by the transmission)

$$\phi^*_{X_5X_6}(x_5^{(3)},.) = \phi_{X_5X_6X_8}(x_5^{(3)},.,x_8^{(2)})$$

(again nothing is altered by the transmission)

$$\phi^*_{X_3X_5X_6}(.,x_5^{(3)},.) = \phi_{X_3X_5X_6}(.,x_5^{(3)},.)\frac{\phi^*_{X_5X_6}(x_5^{(3)},.)}{\phi_{X_5X_6}(x_5^{(3)},.)} = \phi_{X_3X_5X_6}(.,x_5^{(3)},.)$$

(again, no alteration)

$$\phi^*_{X_3X_5}(.,x_5^{(3)}) = \sum_y \phi_{X_3X_5X_6}(.,x_5^{(3)},y)$$
$$\phi^*_{X_3X_4X_5}(.,x_4^{(1)},x_5^{(3)}) = \phi_{X_3X_4X_5}(.,x_4^{(1)},x_5^{(3)})\frac{\phi^*_{X_3X_5}(.,x_5^{(3)})}{\phi_{X_3X_5}(.,x_5^{(3)})}.$$

This is the message passing into the clique chosen as a root; $X_3X_4X_5$. Now all the transmissions out to the leaves are necessary.

After this has been carried out, the cliques and separators will not be probabilities; they will all differ from probabilities by the same factor. Dividing through the potentials on all the cliques and separators by the marginalisation of any of them over all the uninstantiated variables (for example, in this case simply taking the updated value of $\phi_{X_4,X_5}(x_4^{(1)},x_5^{(3)})$, the potential over the separator $\{X_4, X_5\}$) will give the resulting conditional probabilities.

7. (a) The moral graph is given in the figure 10.8. The elimination sequence is: $X_{13}(8)$, $X_{10}(30)$, $X_{12}(30)$, $X_6(45)$ (these are simplicial), $X_2(12)$, $X_4(18)$, $X_1(24)$, $X_5(24)$, $X_7(24)$, $X_{11}(48)$ $X_8(12)$ $X_3(4)$, $X_{10}(2)$.

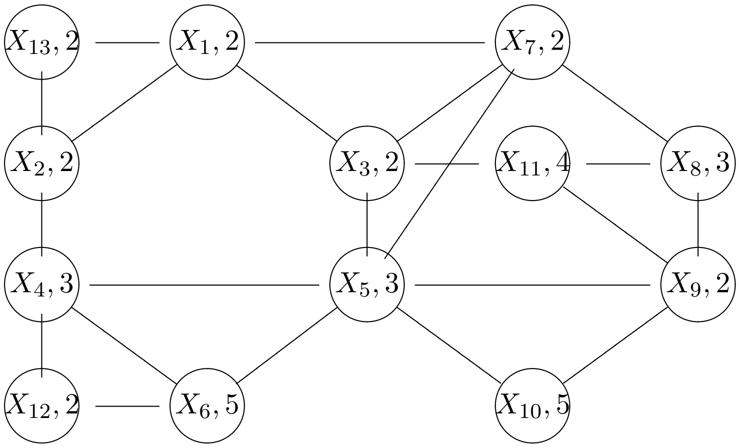


Figure 10.8: Moral Graph

The junction tree is given in figure 10.9.

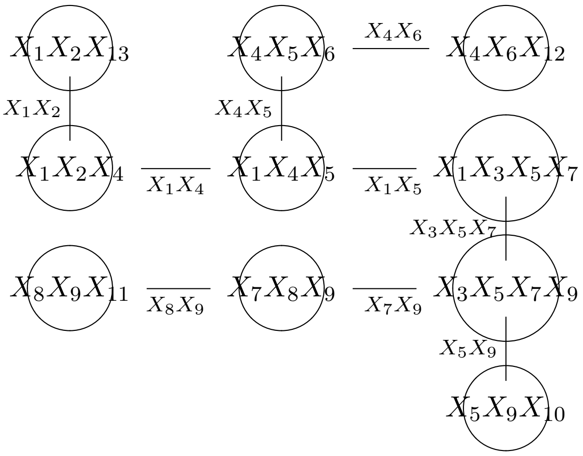


Figure 10.9: Junction Tree

(b) Solution omitted.

8. A junction tree has already been constructed. It is found in figure 4.3. With $X_5 = x_5^{(1)}$, $X_6 = x_6^{(3)}$, $X_9 = x_9^{(2)}$, there is no evidence on the clique $\{X_4, X_{10}\}$ or $\{X_3, X_13\}$ or $\{X_{11}, X_{12}\}$. A suitable propagation is therefore

$$\{X_4, X_9\} \to \{X_4, X_6, X_8\}, \{X_4, X_6, X_8\} \to \{X_1, X_2, X_3, X_4\}$$

$$\{X_5, X_7, X_{11}\} \to \{X_3, X_5\} \to \{X_1, X_2, X_3, X_4\}$$

$$\{X_1, X_2, X_3, X_4\} \to \{X_3, X_5\} \to \{X_5, X_7, X_{11}\} \to \{X_{11}, X_{12}\}$$

$$\{X_1, X_2, X_3, X_4\} \to \{X_3, X_{13}\}$$

$$\{X_1, X_2, X_3, X_4\} \to \{X_4, X_6, X_8\}, \{X_4, X_6, X_8\} \to \{X_4, X_9\}, \{X_4, X_6, X_8\} \to \{X_4, X_{10}\}.$$

To obtain the conditional probabilities, the potentials over the cliques and separators have to be divided through by a normalising constant, obtained by taking the easiest potential to marginalise. Here, the value of the separator over X_5 (after the fully active schedule) gives the normalising constant; no marginalisation is necessary.

9. (a) From the definition, clearly a junction tree is a pre-I-tree. Proof that it is of maximal weight: assume that the result is true for n cliques and consider a junction tree on $n+1$ cliques. Let $\mathcal{T}$ denote the tree with greatest weight that can be formed on n nodes by removing one of the leaf nodes. Label the leaf node C_{n+1} and assume that it is attached to C_n. Then $\mathcal{T}$ (from the construction) is a junction tree, since the tree on $n+1$ nodes was a junction tree - it clearly satisfies the required property.

   Assume that the junction tree on $n+1$ nodes is not of maximal weight. Since the choice of edges between $C_1, \ldots, C_n$ forms a pre-I-tree of maximal weight between these cliques, and since the weight of this tree is greater than any tree on n cliques that includes C_{n+1}, it follows that a tree of greater weight may be obtained by attaching C_{n+1} as a leaf to a node to C_i where $i \in \{1, \ldots, n-1\}$. But, by the junction tree property, $C_i \cap C_{n+1} \subseteq C_n \subset C_{n+1}$ hence a contradiction. Hence a junction tree is a pre-I-tree of maximal weight.

   (b) This may be proved by induction. A pre-I-tree on two cliques is necessarily of maximal weight. Assume that any pre-I-tree of maximal weight on n or fewer cliques is a junction tree. Consider a pre-I-tree of maximal weight on $n+1$ cliques, formed by taking the pre-I-tree of maximal weight that may be constructed from n of the cliques (which by the inductive hypothesis is a junction tree) and then adding C_{n+1} as a leaf node, attaching it to the node that gives maximal weight. Label this node C_n. Then $|C_{n+1} \cap C_n| \geq |C_{n+1} \cap C_i|$ for $i = 1, \ldots, n-1$.

   If this is a not a junction tree, then it holds for a tree S, one of the connected components obtained after removing C_{n+1} and C_n, that $C_{n+1} \cap S \not\subseteq C_n \cap S$. Since $|C_{n+1} \cap C_n| \geq |C_{n+1} \cap C_i|$ for all $i \in \{1, \ldots, n-1\}$, it follows that there exists a node $W \in C_{n+1} \cap S \backslash C_n \cap S$. Take $X \in C_{n+1} \cap C_n$, $P \in C_n \backslash (C_{n+1} \cup S)$, $Q \in C_n \cap S$. Then in the original graph (W, X, P, Q, W) forms a cycle of length 4 without a chord. Hence the cliques do not come from a triagulated graph.

   The proof is complete.

10. (a) Conditioning on a chain or fork connection blocks the 'Bayes ball'. It is not possible to obtain a singly connected graph by conditioning on one variable. Conditioning on X_5 and X_7 gives a resulting DAG that is singly connected.

(b) If X_5 and X_7 are instantiated, there are 10 different singly connected DAGS (one for each possible instantiation of (X_5, X_7)). Condition on X_1 (and note that $X_4 \rightarrow X_2 \rightarrow X_5$ is a chain connection), X_2 and X_7. This gives 8 singly connected DAGS.

11. This is dealt with in the text.

Chapter 11

The factor graph is shown in figure 11.1.

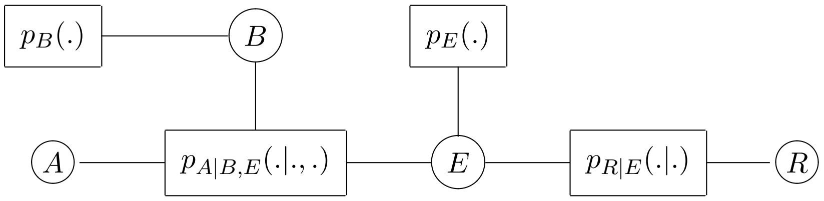


Figure 11.1: Factor Graph

1. Computation of $p_{B|A}(.|1)$. This can be treated as a 'one - eye - problem' where $A = 1$ only is considered. A root has to be chosen; E is a suitable root.

   From definition 11.3 for variable to function (recall that initialisations are all identically 1),

$$\mu_{p_B \to B}(b) = p_B(b) = \begin{array}{ccc} b & 0 & 1 \\ \hline & 0.99 & 0.01 \end{array}$$

$$\mu_{B \to p_{A|B,E}}(b) = \mu_{p_B \to B}(b) = \begin{array}{ccc} b & 0 & 1 \\ \hline & 0.99 & 0.01 \end{array}$$

   The variable A is instantiated with value 1, so

$$\mu_{A \to p_{A|B,E}}(a) = \begin{array}{ccc} a & 0 & 1 \\ \hline & 0 & 1 \end{array}.$$

   This is how definition 11.3 is to be understood. It states that $\mu_{A \to p_{A|B,E}}(a)$ takes the value 1; it should state 1 for all possible values of A and 0 otherwise.

   Following the algorithm, the message $\mu_{p_{A|B,E} \to E}(e)$ needs a marginalisation

$$\begin{aligned}\mu_{p_{A|B,E}\to E}(e) &= \sum_{a,b} p_{A|B,E}(a|b,e)\mu_{B\to p_{A|B,E}}(b)\mu_{A\to p_{A|B,E}}(a)\\ &= \sum_{b} p_{A|B,E}(1|b,e)\mu_{B\to p_{A|B,E}}(b)\\ &= (0.03\times 0.99+0.95\times 0.01, 0.95\times 0.99+0.98\times 0.01) = (0.0392, 0.9503)\end{aligned}$$

$$\mu_{p_E\to E} = p_E$$

The messages $\mu_{R\to p_{R|E}}$, $\mu_{p_{R|E}\to E}$, $\mu_{E\to p_{R|E}}$, $\mu_{p_{R|E}\to R}$ are clearly unnecessary for the problem.

The message $\mu_{E\to p_E}$ does not influence the distribution over E and B; it would only be useful for computing $p_{E|A}(.|1)$.

$$\mu_{E\to p_E} = \mu_{p_{A|B,E}\to E} = (0.0392, 0.9503)$$

$$\mu_{E\to p_{A|B,E}} = \mu_{p_E\to E} = (0.001, 0.999)$$

Message $\mu_{p_{A|B,E}\to A}$ is not needed for the termination at B.

$$\begin{aligned}\mu_{p_{A|B,E}\to B}(b) &= \sum_{e,a} p_{A|B,E}(a|b,e)\mu_{E\to p_{A|B,E}}(e)\mu_{A\to p_{A|B,E}}(a)\\ &= \sum_{e} p_{A|B,E}(1|b,e)\mu_{E\to p_{A|B,E}}(e)\\ &= (0.03\times 0.001+0.95\times 0.999, 0.95\times 0.001+0.98\times 0.999)\\ &= (0.94908, 0.97997)\end{aligned}$$

Message $\mu_{B\to p_B}$ is not needed for the termination at B.

From the formula for the termination at B, it follows that

$$\begin{aligned}(p_{A,B}(1,0), p_{A,B}(1,1)) &= (\mu_{p_B\to B}(0)\mu_{p_{A|B,E}\to B}(0), \mu_{p_B\to B}(1)\mu_{p_{A|B,E}\to B}(1))\\ &= (0.037203936, 0.93165691)\end{aligned}$$

and hence that

$$\begin{aligned}(p_{B|A}(0|1), p_{B|A}(1|1))) &= \frac{1}{0.968469627}(0.037203936, 0.93165691)\\ &= (0.038415181, 0.961584819)\end{aligned}$$

2. For this one, the messages are slightly different. Take E as root.

$$\mu_{p_B\to B} = (0.99, 0.01)$$

$$\mu_{B\to p_{A|B,E}} = (0.99, 0.01)$$

$$\mu_{A\to p_{A|B,E}} = (0,1)$$

$$\mu_{p_{A|B,E}\to E} = (0.0392, 0.9503)$$

$$\mu_{p_E\to E} = p_E = (0.001, 0.999)$$

$$\mu_{R\to p_{R|E}} = (0,1)$$

$$\mu_{p_{R|E}\to E} = p_{R|E}(1|.) = (0.01, 0.95)$$

$$\mu_{E\to p_E} = \mu_{p_{R|E}\to E}\mu_{p_{A|B,E}\to E} = (0.000392, 0.90535)$$

$$\mu_{E\to p_{A|B,E}} = \mu_{p_E\to E}\mu_{p_{R|E}\to E} = (0.00001, 0.94905)$$

$$\begin{aligned}\mu_{p_{A|B,E}\to A}(1) = p_{A|B,E}(1|b,e)\mu_{B\to p_B}(b)\mu_{E\to p_{A|B,E}}(e) = 0.03 \times 0.99 \times 0.00001 \\ +0.95 \times 0.00001 \times 0.01 + 0.95 \times 0.94905 \times 0.99 + 0.98 \times 0.01 \times 0.94905 = 0.90188\end{aligned}$$

$$\begin{aligned}\mu_{p_{A|B,E}\to B} &= \sum_e p_{A|B,E}(1|.,e)\mu_{E\to p_{A|B,E}}(e) \\ &= (0.03 \times 0.00001 + 0.95 \times 0.94905, 0.95 \times 0.00001 + 0.98 \times 0.94905) \\ &= (0.90160, 0.93052)\end{aligned}$$

$$\mu_{B\to p_B} = \mu_{p_{A|B,E}\to B} = (0.90160, 0.93052)$$

From the termination formula, $p_{A,B,R}(1,b,1) = \mu_{p_B\to B}(b)\mu_{p_{A|B,E}\to B}(b)$ so that

$$(p_{A,B,R}(1,0,1), p_{A,B,R}(1,1,1)) = (0.99 \times 0.90160, 0.01 \times 0.93052) = (0.89259, 0.0093052)$$

so that

$$p_{B|A,R}(.|1,1) = (0.98968, 0.01032)$$