Some Topics on Graphical Models in Statistics

Mark John Brewer

submitted for the degree of Doctor of Philosophy

The University of Edinburgh

1994

Abstract

This thesis considers graphical models that are represented by families of probability distributions having sets of conditional independence constraints specified by an influence diagram.

Chapter 1 introduces the notion of a directed acyclic graph, a particular type of independence graph, which is used to define the influence diagram. Examples of such structures are given, and of how they are used in building a graphical model. Models may contain discrete or continuous variables, or both. Local computational schemes using exact probabilistic methods on these models are then reviewed.

Chapter 2 presents a review of the use of graphical models in legal reasoning literature. The use of likelihood ratios to propagate probabilities through an influence diagram is investigated in this chapter, and a method for calculating LRs in graphical models is presented.

The notion of recovering the structure of a graphical model from observed data is studied in Chapter 3. An established method on discrete data is described, and extended to include continuous variables. Kernel methods are introduced and applied to the probability estimation needed in these methods.

Chapters 4 and 5 describe the use of stochastic simulation on mixed graphical association models. Simulation methods, in particular the Gibbs sampler, can be used on a wider range of models than exact probabilistic methods. Also, estimates of marginal density functions of continuous variables can be obtained by using kernel estimates on the simulated values; exact methods generally only provide the marginal means and variances of continuous variables.

A *standard* mixed graphical association model is introduced in Chapter 4—this has Normal conditional density functions defined for continuous variables, where the mean is a linear function of certain other continuous variables. Gibbs sampling applied here is straightforward. For non-standard models (Chapter 5) this is not so, and other Markov Chain Monte Carlo (MCMC) methods must be used in addition to the Gibbs sampler. It is shown that the use of MCMC methods enables a very wide choice of model.

To hell and back

Acknowledgments

Warmest thanks are extended to the following:

Colin Aitken For massive amounts of guidance, and knowing exactly when to push.

Alex Gammerman and Zhiyuan Luo For many interesting discussions, and donations of software.

Table of Contents

Summary

This thesis considers graphical models that are represented by families of probability distributions having sets of conditional independence constraints specified by an influence diagram. Graphical models are gaining more and more importance in statistics, and this work will hopefully be a useful addition to the growing literature on the subject.

Chapter 1 introduces the notion of a directed acyclic graph (DAG), a particular type of independence graph, which is used to define the influence diagram. Examples of such structures are given, and of how they are used in building a graphical model. Models may contain discrete or continuous variables, or both. Local computational schemes using exact probabilistic methods on these models are then described.

Chapter 2 presents a review of the use of graphical models in legal reasoning literature. The use of likelihood ratios (LRs) to propagate probabilities through an influence diagram is also investigated in that chapter, and an algorithm for calculating LRs in graphical models is presented.

The problem of estimating the structure of an influence diagram from observed data is described in chapter 3; a standard method for recovering skeleton tree structure and (partial) directionality using sample frequencies is compared with another using kernel functions. The method is extended to include continuous variables, and finally some ideas are presented for recovering networks that contain (undirected) loops.

Chapters 4 and 5 describe the use of stochastic simulation on mixed graphical association models. Simulation methods, in particular the Gibbs sampler, can be used on a wider range of models than exact probabilistic methods. Also, estimates of marginal density functions of continuous variables can be obtained by using kernel estimates on the simulated values. Exact methods generally only provide the marginal means and variances of continuous variables.

A *standard* mixed graphical association model is introduced—this has Normal conditional density functions defined for continuous variables, where the mean is a linear function of certain other continuous variables—in Chapter 4. Gibbs sampling applied here is straightforward. For non-standard models, which appear in Chapter 5, this is not so, and other Markov Chain Monte Carlo (MCMC) methods must be used in addition to the Gibbs sampler. It is shown that the use of MCMC methods enables a very wide choice of model.

Finally some conclusions and suggestions for future research are presented.

Chapter 1

Introduction to Graphical Models

1.1 Purpose

The objective of this thesis is to develop the theory of graphical models that form a basis for expert systems. Lauritzen and Spiegelhalter (1988) broadly define an expert system as "a computer program intended to make reasoned judgments or give assistance in a complex area in which human skills are fallible or scarce". In addition to reviewing the recent literature on the development of computational schemes for expert systems, a wide range of examples is presented, with special focus in Chapter 2 on the use of graphical models in legal reasoning.

The use of graphical representations for probabilistic information goes back as far as Sewal Wright (1921), a geneticist who developed path analysis. His work was however shunned by statisticians—see Niles (1922), for example. This outlook changed in the 1960s, beginning with the early work of Birch (1963,1964) on contingency tables. The likes of Goodman (1970) and Haberman (1974) realised that some conditional independence properties of log-linear models could be illustrated by a graph. More references can be found in Kiiveri and Speed (1982).

Whilst there has been a considerable amount of work concentrating on manipulating a given model, the construction of the model itself has been somewhat neglected. A method for recovering structure from observed data is thus described.

Another main concern of this thesis is to extend the range of models that can be defined. Models which have discrete variables only have been catered for well enough—see Lauritzen and Spiegelhalter (1988) or Pearl (1988) for example. Lauritzen and Wermuth (1984,1989), Wermuth and Lauritzen (1990), Lauritzen (1992) include continuous variables by introducing the *conditional Gaussian* (CG) family of distributions. It is shown here that stochastic simulation, using a Gibbs sampler, can provide more information about a CG model than exact probabilistic methods. Furthermore, simulation methods remove the necessity to use CG distributions, so that, theoretically, *any* conditional distribution, as the model might require, can be defined. There is also no compulsion to have simple linear relationships between variables and parameters—again *any* relationship should be feasible.

This chapter first introduces some graph theory and related concepts. The computational schemes of Lauritzen and Spiegelhalter (1988), Pearl (1988) and Lauritzen (1992) are then sketched, and illustrated with examples. These schemes can therefore be contrasted with the simulation procedures of Chapters 4 and 5.

1.2 Terminology

A graphical model is described by Whittaker (1990) as "a family of probability density functions that incorporates a specific set of conditional independence constraints listed in an independence graph". In this work, the independence graph takes the form of a *directed acyclic graph* (see below).

The explanations that follow are largely taken from Whittaker (1990) and Lauritzen (1992).

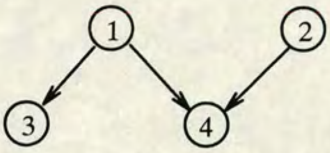
A *graph* G is a mathematical object consisting of two sets—a set of *nodes* (or *vertices*) V and a set of *edges* E. The set E consists of distinct pairs of elements of V. A *directed edge* exists between nodes i and j in V if the pair (i, j) occurs in the set E. Node i is then defined to be a *parent* of j, and j is similarly a *child* of i. A *root* node has no parents; a *leaf* node has no children. An *undirected edge* occurs when both (i, j) and (j, i) reside in the set E.

A graph can have two kinds of node—nodes which represent discrete variables in the model (and are termed *discrete nodes*), and nodes which represent continuous variables (*continuous nodes*). The set of nodes V is partitioned into two groups:

$$V = \Delta \mathbin{\dot\cup} \Gamma.$$

The set Δ represents discrete nodes, and the set Γ continuous nodes. A graph is *pure* if it has only one type of node.

Nodes are represented pictorially as circles or ellipses; in non-pure graphs, discrete nodes are shaded, while continuous nodes are not. An edge is shown by a line from one node to another, with an arrow pointing from i to j if i is a parent of j. Thus the graph with $V = \{1,2,3,4\}$ and $E = \{(1,3),(1,4),(2,4)\}$ has the diagram



If there is an edge between i and j then they are said to be *neighbours*, and are *adjacent* to each other. A *path* of length m is a sequence of distinct nodes $i_1, i_2, \ldots, i_m$ for which (i_l, i_{l+1}) is in E for each $l = 1, 2, \ldots, m-1$. A path is a *cycle* if $i_1 = i_m$. A cycle is *chordless* if no other than successive pairs of nodes in the cycle have an edge between them. An undirected graph is *triangulated* if and only if every chordless cycle in the graph contains no more than 3 nodes.

A subset of nodes C *separates* two nodes i and j if every path joining i and j has at least one node in C.

A *tree* is an undirected graph with a unique path between any two nodes. A *polytree* is a tree that has some directed edges.

A graph is *acyclic* if it has no directed cycles. This thesis is mainly concerned with *directed acyclic graphs.*

A graph is *complete* if each node has an edge with each other node. The induced subgraph G_C of G (where $C \subseteq V$) is the graph obtained by deleting all the nodes in $V \setminus C$ and the associated edges from E. The subgraph G_C is a *clique* if it is complete and the addition of a node in $V \setminus C$ (and edges) would render the new graph incomplete; that is, a clique is maximally complete.

The *boundary* of a subset C, written $\text{bd}(C)$, is defined as the set of nodes in $V \setminus C$ that have an edge with a node in C.

A *pure linear* structure is a sequence of nodes $X_1, X_2, \ldots X_n$ where an edge joins X_{i-1} to X_i for $i = 2, 3, \ldots n$, and there exist no other edges in the graph. Note the distinction with a "pure graph" above.

1.3 Independence Graphs

Let $X = (X_1, X_2, \ldots, X_k)$ denote the random variables in a model, and $V = (1, 2, \ldots, k)$ the associated set of nodes. The graph for the model is an *independence graph* (or more correctly a *conditional independence graph*) if there is no edge between two nodes when the corresponding variables are independent given the remaining variables. For two such variables X_i and X_j, the shorthand notation $i \perp\!\!\!\perp j \,|\, V \setminus \{i,j\}$ is used for $X_i \perp\!\!\!\perp X_j \,|\, X \setminus \{X_i, X_j\}$.

1.3.1 Markov Properties of Undirected Graphs

Consider an undirected graph $G^u = (V, E^u)$. This graph has several Markov properties:

- The *pairwise Markov* property, that for all nodes i and j where no edge (i,j) or (j,i) exists in E^u,

$$X_i \perp\!\!\!\perp X_j \,|\, X_C, \qquad \text{where} \quad C = V \setminus \{i,j\}.$$

- The *global Markov* property, that, for all disjoint subsets A, B and C of V, whenever A and B are separated by C in the graph, then

$$X_B \perp\!\!\!\perp X_A \,|\, X_C.$$

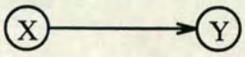
- The *local Markov* property, that, for every node i, if $C = \text{bd}(i)$ is its boundary set and B the set of remaining nodes, then

$$X_i \perp\!\!\!\perp X_B \,|\, X_C, \qquad \text{where} \quad B = V \setminus (\{i\} \cup C).$$

Whittaker (1990) shows that these properties are equivalent.

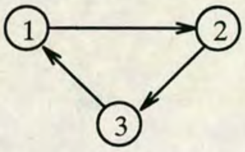
1.3.2 Directed Acyclic Independence Graphs

A directed graph $G = (V, E)$ can be used to display a notion such as "X effects Y", giving the diagram



which together with a conditional probability distribution $f_{Y|X}$ is a natural object of study for the statistical modeller.

If directed cycles were allowed, for example as in



then the desired joint density function would look something like $f_{3|2}f_{2|1}f_{1|3}$, but this is generally not a well defined probability density function. Hence, directed cycles are forbidden.

A consequence of excluding directed cycles is that the nodes can be *completely ordered*, that is there exists a relation $\prec$ on the elements of V such that for all i and j in V, (i) either $i \prec j$ or $j \prec i$, (ii) $\prec$ is irreflexive, and (iii) $\prec$ is transitive, such that if $i \prec j$ and $j \prec l$, then $i \prec l$. It will be helpful to assume that $1 \prec 2 \prec \ldots \prec k$ in V, and to consider that each variable has a "past" and a "future"; for example X_1 precedes X_2.

Definition 1 *The* directed independence graph *of X is the directed graph $G = (V, E)$, where $V = \{1, 2, \ldots, k\}$, $V(j) = \{1, 2, \ldots, j\}$ and the edge (i, j), with $i \prec j$, is* not *in the edge set E if and only if $j \perp\!\!\!\perp i \,|\, V(j) \setminus \{i, j\}$.*

The conditional densities $f_{i|V(i)\setminus\{i\}}$ (for $i = 1, 2, \ldots, k$) contain enough information to define the joint distribution as a consequence of the *recursive factorisation identity*:

$$f_{12\ldots k} = f_{k|V(k)\setminus\{k\}} f_{k-1|V(k-1)\setminus\{k-1\}} \cdots f_{2|1} f_1.$$

In fact, due to the conditional independencies inherent in the graph, for a variable X_i, where node i has the set of parents $\mathrm{pa}(i)$,

$$f_{i|V(i)\setminus\{i\}} = f_{i|\mathrm{pa}(i)}.$$

It is true that $\mathrm{pa}(i) \subseteq V(i) \setminus \{i\}$ since the nodes are assumed to be completely ordered.

To be able to comment on the Markov properties of directed graphs, the following definition is required:

Definition 2 *The* moral graph *of the directed graph $G = (V, E)$ is the undirected graph $G^m = (V, E^m)$ on the same set of nodes, and (i) directed edges replaced by undirected edges, (ii) an edge added between any two nodes (not already joined) with a common child in G.*

Whittaker (1990) proves the following theorem:

Theorem 1 (Markov Theorem for Directed Graphs) *The directed independence graph G possesses the Markov properties of its associated moral graph, G^m.*

Thus it is now possible to define a pure discrete graphical model. The probability distributions $f_{i|\mathrm{pa}(i)}$ are specified simply as numerical probabilities in this case. There is, however, a need to be able to manipulate defined models; if a variable has its value observed (or hypothesised) the effect on the remaining variables is of interest. Two *propagation* schemes for updating a model are presented in the next section. A similar procedure for mixed models follows in section 1.6.

1.4 Propagation in Pure Discrete Models

The two schemes that follow are due to Lauritzen and Spiegelhalter (1988) and Pearl (1988). They use exact probabilistic methods to distribute the information supplied by an observed variable around a graph. This is known as *propagation.*

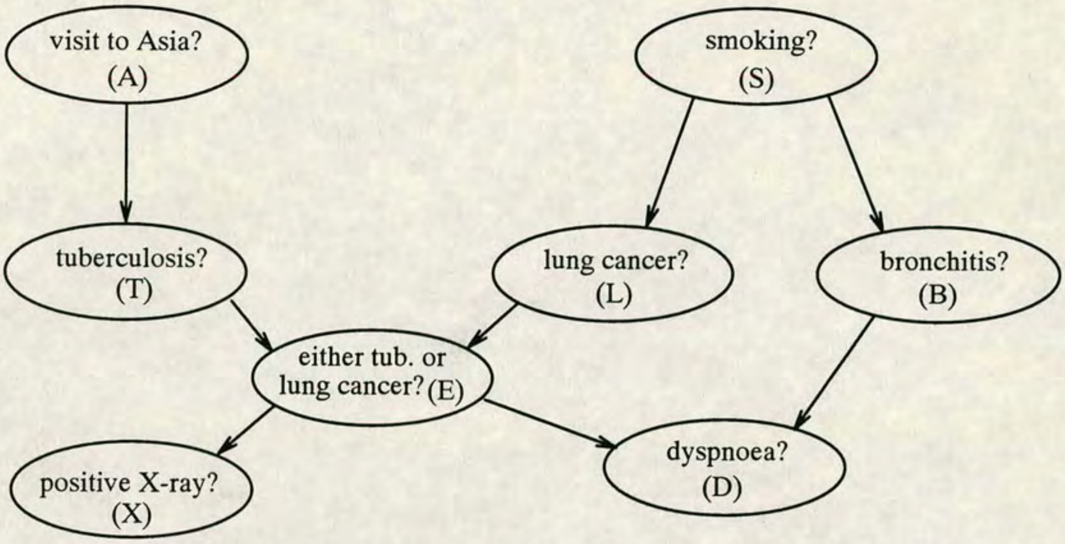


Figure 1–1: *Dyspnoea example—the capital letters in brackets are the abbreviated node names.*

1.4.1 Lauritzen and Spiegelhalter's Method

A brief sketch of this method is presented. The example from Lauritzen and Spiegelhalter (1988) is first described and then worked through.

> Shortness-of-breath (dyspnoea) may be due to tuberculosis, lung cancer or bronchitis, or none of them, or more than one of them. A recent visit to Asia increases the chance of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea.

This situation is represented by eight binary discrete variables with graph as in Figure 1–1. For this example, the variables are denoted by capital letters—e.g. A represents the answer to the question "visit to Asia?". The yes/no responses for each variable are denoted by lower-case letters: $A = a$ means "yes" (the patient has visited Asia), $A = \overline{a}$ means "no"; and so on for the remaining variables. Note here that a node is given the same name as its corresponding variable. The

A:	$p(a)$	$= .01$	E:	$p(e \mid l, t)$	$= 1$
				$p(e \mid l, \overline{t})$	$= 1$
T:	$p(t \mid a)$	$= .05$		$p(e \mid \overline{l}, t)$	$= 1$
	$p(t \mid \overline{a})$	$= .01$		$p(e \mid \overline{l}, \overline{t})$	$= 0$
S:	$p(s)$	$= .50$	X:	$p(x \mid e)$	$= .98$
				$p(x \mid \overline{e})$	$= .05$
L:	$p(l \mid s)$	$= .10$			
	$p(l \mid \overline{s})$	$= .01$	D:	$p(d \mid e, b)$	$= .90$
				$p(d \mid e, \overline{b})$	$= .70$
B:	$p(b \mid s)$	$= .60$		$p(d \mid \overline{e}, b)$	$= .80$
	$p(b \mid \overline{s})$	$= .30$		$p(d \mid \overline{e}, \overline{b})$	$= .10$

Table 1–1: *Conditional probabilities for dyspnoea example.*

shorthand form $p(a)$ is used in place of $\Pr(A = a)$. Assessments of the relevant conditional probabilities are given in Table 1–1. These probabilities are, however, fictitious.

A doctor, equipped with this model, might want to consider such questions as: given that a patient has dyspnoea and has recently visited Asia, what are the chances of each of the diseases being present?

From the graph, it is clear to see that the belief is that the joint distribution $p(A, T, X, E, D, L, B, S)$ can be expressed as

$$p(A)p(T \mid A)p(X \mid E)p(E \mid T, L)p(D \mid E, B)p(L \mid S)p(B \mid S)p(S). \tag{1.1}$$

To calculate the probability $p(x \mid a, d)$ of a positive X-ray for the patient with dyspnoea who has been to Asia, the simple way would be to use expression (1.1), get the $2^8 = 256$ joint probabilities and perform summations to get $p(x, a, d)/p(a, d)$. This is clearly very inefficient; a better method of obtaining $p(x, a, d)$ is to use the summation

$$p(A)\sum_T p(T|a)\left[\sum_E p(x|E)\left[\sum_L p(E|T,L)\left[\sum_B p(d|E,B)\left[\sum_S p(L|S)p(B|S)p(S)\right]\right]\right]\right]. \tag{1.2}$$

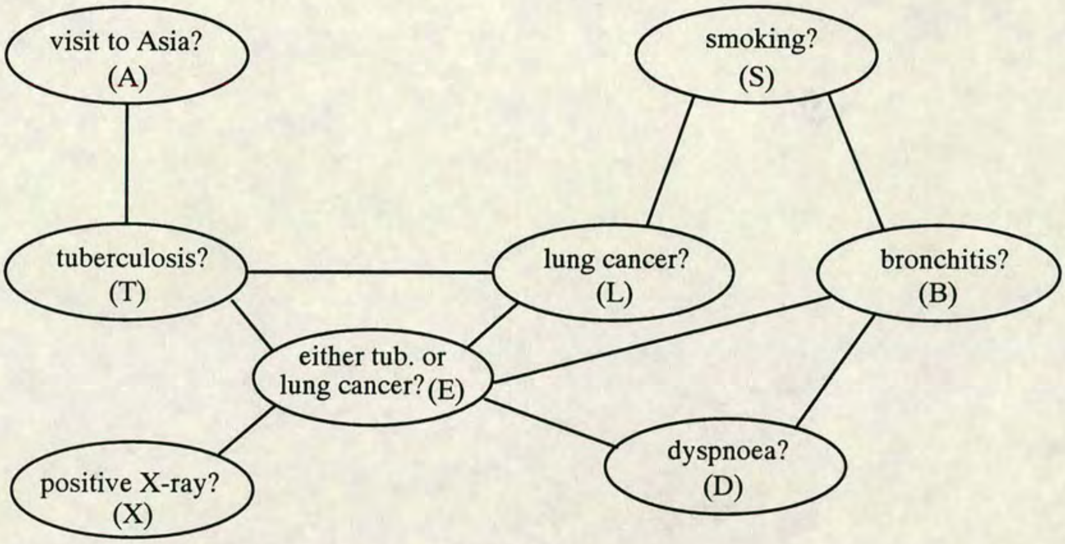


Figure 1–2: *Moral graph of the dyspnoea example.*

The following approach exploits an adapted topology of the graph to perform expressions such as (1.2).

It is possible to work with proportionality; the calculation of normalising constants can be left until needed. Here, when the states of particular nodes are revealed, the observed states can be inserted into (1.1) to give an expression (correct up to a normalising factor) for the conditional probability of states at remaining nodes given the observed. Terms in (1.1) can no longer be thought of as probabilities, however. Instead, it is regarded as a less structured expression by introducing *evidence potentials*, denoted by ψ. Hence (1.1) becomes

$$\psi(A)\psi(T,A)\psi(X,E)\psi(E,T,L)\psi(D,E,B)\psi(L,S)\psi(B,S)\psi(S) \qquad (1.3)$$

where, initially, $\psi(A) = p(A)$, $\psi(T,A) = p(T \mid A)$, and so on.

The graph of Figure 1–1 is now converted to its corresponding moral graph—see Figure 1–2. This means that (1.3) now involves functions of sets of nodes which are complete (in the moral graph).

Note that summing out S in (1.2) would result in the creation of a function of L and B, which are not joined in the graph—thus we would have a function of

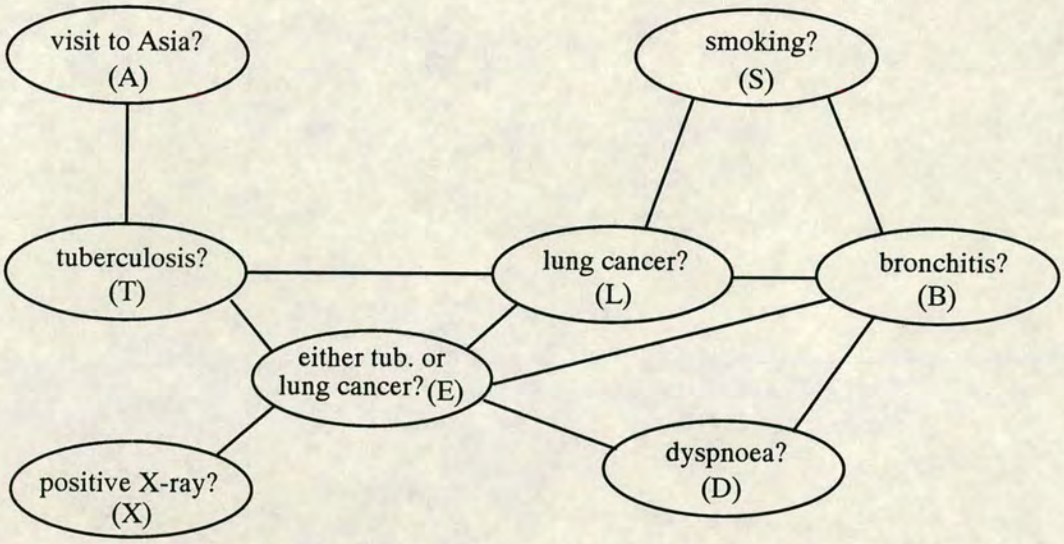


Figure 1–3: *Triangulated moral graph of the dyspnoea example.*

a set of nodes which is *not* complete. To be able to perform all such summations (after finding a suitable ordering) *without* creating functions of nodes not joined in the graph, the graph is "filled in" so that it becomes triangulated. In this example, an edge between L and B is added to cut the cycle (S, L, E, B)[1]—see Figure 1–3.

For convenience, a representation of the joint distribution p is adopted whose ψ functions are defined on the cliques of the filled-in graph:

$$p \propto \psi(A,T)\psi(T,L,E)\psi(L,E,B)\psi(L,B,S)\psi(E,B,D)\psi(E,X) \qquad (1.4)$$

The potentials are obtained by matching with terms in (1.1): $\psi(E,X)=p(X \mid E)$, $\psi(A,T)=p(A)p(T \mid A)$, $\psi(T,L,E)=p(E \mid T,L)$, $\psi(L,B,S)=p(L \mid S)p(B \mid S)p(S)$, $\psi(E,B,D)=p(D \mid E,B)$. $\psi(L,E,B)$ is not yet defined, and may be assumed to take any non-zero constant value—1, say. The other potentials can be calculated from Table 1–1, and are displayed in Table 1–2, third column.

[1] An edge between E and S could have been added instead.

Original clique order	Configuration	Potentials from conditional probability tables	Potentials from set chain	Clique marginals	Potentials after absorbing a, d
			$\{p(A,T)\}$		
$C_1 = \{A,T\}$	a,t	.0005	.0005	.0005	
	$a,\overline{t}$	.0095	.0095	.0095	
	$\overline{a},t$	.0099	.0099	.0099	
	$\overline{a},\overline{t}$	.9801	.9801	.9801	
			$\{p(L,E \mid T)\}$		
$C_2 = \{T,L,E\}$	t,l,e	1	.0550	.00057	.000028
	$t,l,\overline{e}$	0	.0	.0	.0
	$t,\overline{l},e$	1	.9450	.00983	.000473
	$t,\overline{l},\overline{e}$	0	.0	.0	.0
	$\overline{t},l,e$	1	.0550	.05443	.000523
	$\overline{t},l,\overline{e}$	0	.0	.0	.0
	$\overline{t},\overline{l},e$	0	.0	.0	.0
	$\overline{t},\overline{l},\overline{e}$	1	.9450	.9352	.008978
			$\{p(B \mid L,E)\}$		
$C_3 = \{L,E,B\}$	l,e,b	1	.5727	.03150	.5154
	$l,e,\overline{b}$	1	.4273	.02350	.2991
	$l,\overline{e},b$	1	.5727	.0	.4582
	$l,\overline{e},\overline{b}$	1	.4273	.0	.0427
	$\overline{l},e,b$	1	.4429	.00435	.3986
	$\overline{l},e,\overline{b}$	1	.5571	.00548	.3899
	$\overline{l},\overline{e},b$	1	.4429	.4142	.3543
	$\overline{l},\overline{e},\overline{b}$	1	.5571	.5210	.0557
			$\{p(S \mid L,B)\}$		
$C_4 = \{L,B,S\}$	l,b,s	.0300	.9524	.0300	.9524
	$l,b,\overline{s}$	.0015	.0476	.0015	.0476
	$l,\overline{b},s$	.0200	.8511	.0200	.8511
	$l,\overline{b},\overline{s}$	.0035	.1489	.0035	.1489
	$\overline{l},b,s$	.2700	.6452	.2700	.6452
	$\overline{l},b,\overline{s}$	.1485	.3548	.1486	.3548
	$\overline{l},\overline{b},s$	.1800	.3419	.1800	.3419
	$\overline{l},\overline{b},\overline{s}$	.3465	.6581	.3464	.6581
			$\{p(D \mid E,B)\}$		
$C_5 = \{E,B,D\}$	e,b,d	.9	.9	.03227	
	$e,b,\overline{d}$	.1	.1	.00359	
	$e,\overline{b},d$	.7	.7	.02029	
	$e,\overline{b},\overline{d}$	.3	.3	.00869	
	$\overline{e},b,d$	.8	.8	.3314	
	$\overline{e},b,\overline{d}$	.2	.2	.08284	
	$\overline{e},\overline{b},d$	.1	.1	.05210	
	$\overline{e},\overline{b},\overline{d}$	.9	.9	.4689	
			$\{p(X \mid E)\}$		
$C_6 = \{E,X\}$	e,x	.98	.98	.06354	.98
	$e,\overline{x}$	.02	.02	.00130	.02
	$\overline{e},x$	.05	.05	.04676	.05
	$\overline{e},\overline{x}$	.95	.95	.8884	.95

Table 1–2: *Stages in calculating clique marginals before and after absorption of evidence.*

A bonus of using cliques on a triangulated graph in this way is that the joint distribution can be expressed as a simple function of the individual marginal distributions on the cliques. The joint probability can, in fact, be written as

$$\frac{p(A,T)p(T,L,E)p(L,E,B)p(L,B,S)p(E,B,D)p(E,X)}{p(T)p(L,E)p(L,B)p(E,B)p(E)}. \qquad (1.5)$$

Probabilities of single nodes can be calculated easily from stored clique marginals, simply by summing out the other variables in a particular clique. The expressions (1.1), (1.4) and (1.5) are different *local representations* of the joint distribution—that is, they separate the set of nodes into "local" (defined by the topology of the graph) subsets. The method of Lauritzen and Spiegelhalter (1988) is based upon moving between these representations.

To obtain the clique marginals, the system must first be *initialised*, where the potentials take the form of conditional probability tables on a chain of sets of nodes. This *set chain* exploits the properties of triangulated graphs.

The nodes are ordered using *maximum cardinality search* (see Lauritzen and Spiegelhalter, 1988). One such ordering is shown in Figure 1–4. Next, the cliques are ordered too, ranking according to the earliest labelled node in each clique (see Table 1–3), and this forms the set chain. Obtaining the set chain in this way ensures the cliques have the *running intersection property*, that is the nodes of a clique C_i also contained in previous cliques $C_1, \ldots, C_{i-1}$ are all members of *one* previous clique, called a *parent* clique. For example, from Table 1–3, $C_4 \cap (C_1 \cup C_2 \cup C_3) = \{L, B\}$ is contained in C_3. These *separating* nodes are denoted as $S_i = C_i \cap (C_1 \cup \ldots \cup C_{i-1})$, and the residual $C_i \setminus S_i$ as R_i.

The running intersection property ensures that the joint probability can be written as

$$p(A,T)p(L,E\,|\,T)p(B\,|\,L,E)p(S\,|\,L,B)p(D\,|\,E,B)p(X,E) \qquad (1.6)$$

directly from (1.5). Expression (1.6) is yet another potential representation of p, but it allows clique marginals to be obtained more easily. The problem is to get

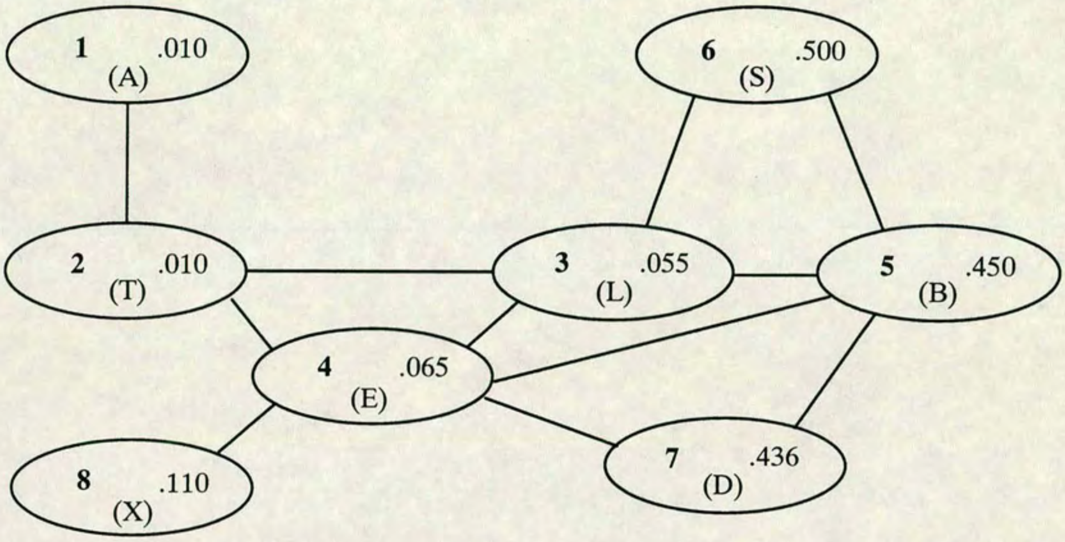


Figure 1–4: *Triangulated moral graph, showing possible ordering of nodes and initial node marginals.*

i	Cliques C_i	Residuals R_i	Separators S_i	Possible parent cliques
1	A, T	A, T	$\emptyset$	
2	T, L, E	L, E	T	1
3	L, E, B	B	L, E	2
4	L, B, S	S	L, B	3
5	E, B, D	D	E, B	3
6	E, X	X	E	2,3,5

Table 1–3: *Initial set chain of triangulated moral graph.*

from (1.4) to (1.6). The running intersection property allows a simple algorithm to obtain (1.6) term by term. For example,

$$p(X \mid E) = p(R_6 \mid S_6) = \psi(X, E) \Big/ \sum_X \psi(X, E)$$

so that the last term in (1.6) is obtained directly from the potentials on C_6. Furthermore, the potential representation on all nodes except X is unchanged except that $\psi(C_5) = \psi(E, B, D)$ becomes $\psi(E, B, D) \sum_X \psi(X, E)$. In this case, $\sum_X \psi(X, E) = 1$, but in general, when i cliques remain, $\psi(C_i)$ is transformed to

$$p(R_i \mid S_i) = \psi(C_i) \Big/ \sum_{R_i} \psi(C_i),$$

then the potentials of a parent clique of C_i are multiplied by $\sum_{R_i} \psi(C_i)$—see Table 1–2, fourth column.

Having obtained (1.6), the clique marginals can be derived. From $p(C_1) = p(A, T)$, $p(S_2) = p(T)$ is obtained by marginalisation. Then, multiplication gives $p(C_2) = p(R_2 \mid S_2) = p(L, E \mid T)p(T)$. Working back through the chain in this way gives the clique marginals displayed in Table 1–2, column five. The marginals on individual nodes are shown in Figure 1–4. The system has now been initialised.

Suppose now evidence has been gathered; namely, a patient has dyspnoea and has recently visited Asia, i.e. $A = a$ and $D = d$. Thus the required conditional distribution is, from (1.4),

$$\begin{aligned} p(T, L, E, B, S, X \mid a, d) &\propto p(a, T, L, E, B, S, d, X) \\ &\propto \psi(a, T)\psi(T, L, E)\psi(L, E, B)\psi(L, B, S)\psi(E, B, d)\psi(E, X). \qquad (1.7) \end{aligned}$$

Figure 1–5 shows the updated graph. A potential representation of the cliques of this new graph,

$$\psi^*(T, L, E)\psi^*(L, E, B)\psi^*(L, B, S)\psi^*(E, X), \qquad (1.8)$$

is obtained by matching terms in (1.7):

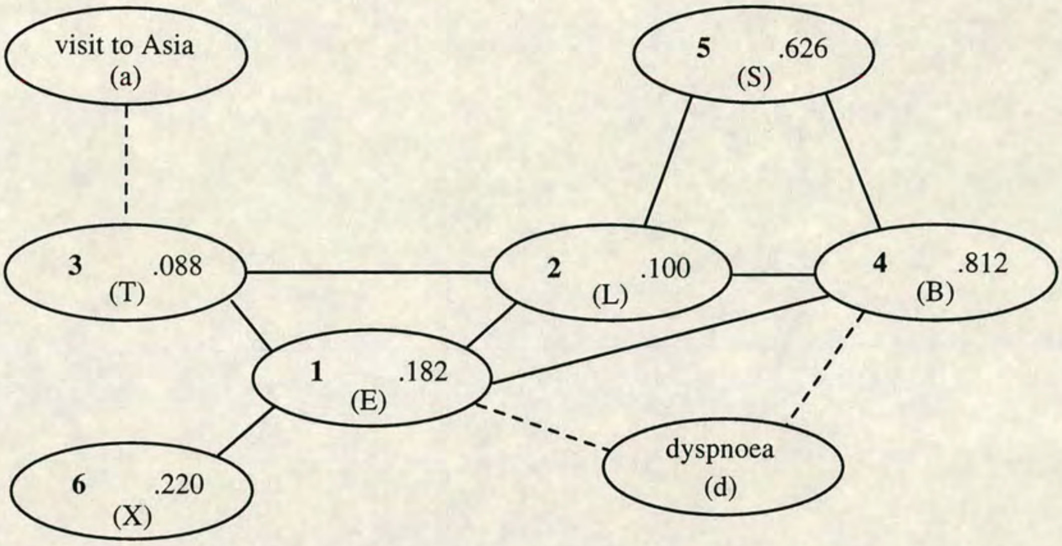


Figure 1–5: *Updated graph after absorbing evidence.*

$$\psi^*(T,L,E) = \psi(T,L,E)\psi(a,T), \qquad \psi^*(L,B,S) = \psi(L,B,S),$$
$$\psi^*(L,E,B) = \psi(L,E,B)\psi(E,B,d), \qquad \psi^*(E,X) = \psi(E,X).$$

Essentially, evidence is absorbed by projecting potentials either onto a new, reduced clique, or, if a clique can be removed, onto another clique. Using the set chain potentials of Table 1–2, fourth column, the conditional potentials of (1.8) are shown in Table 1–2, sixth column. For example, $\psi^*(l,e,b) = \psi(l,e,b)\psi(e,b,d) = 0.5727 \times 0.9 = 0.5154$.

The new marginal distributions on the the cliques can now be found in the same manner as the original system was initialised earlier. Maximum cardinality search again provides a node ordering (see Figure 1–5), with clique ordering $C_1 = \{T,L,E\}$, $C_2 = \{L,E,B\}$, $C_3 = \{L,B,S\}$, $C_4 = \{E,X\}$.

The conditionals

$$p(R_4 \mid S_4) = p(X \mid E), \qquad p(R_3 \mid S_3) = p(S \mid L,B)$$

in the set chain are the same as before, but $p(R_2 \mid S_2) = p(B \mid L,E)$ is affected by the revised potential on $\{L,E,B\}$. When the clique marginals have been found, the updated node marginals can be obtained—see Figure 1–5. It can be seen that the evidence has increased the disease probabilities, nearly nine times for tuberculosis. The dyspnoea has resulted in a raised expectation of the patient being a smoker, and the chance of a positive X-ray has doubled.

1.4.2 Pearl's Method

Pearl's method of propagating evidence (Pearl, 1988) begins by developing an algorithm for polytrees, and then using *conditioning* to include general directed acyclic graphs. The exposition that follows is due to Luo (1992). The notation here is different from that in §1.4.1 for consistency with Pearl (1988) and Luo (1992).

Consider a node j in a polytree, representing a variable X_j. The posterior probability of X_j given evidence E (a set of observed variables) can be written as $\Pr(X_j \mid E)$; this probability is called a *belief function* by Pearl, and is denoted $\text{Bel}(X_j)$. The evidence E can be divided into two parts:

E_j^- : The subset of variables of E in the polytree rooted at j; and

E_j^+ : The subset $E \setminus E_j^-$.

The probability $\Pr(X_j \mid E)$ can then be written

$$\begin{aligned}\Pr(X_j \mid E) &= \Pr(X_j \mid E_j^-, E_j^+) = \frac{\Pr(X_j, E_j^-, E_j^+)}{\Pr(E_j^-, E_j^+)} \\ &= \frac{\Pr(E_j^+)\Pr(X_j \mid E_j^+)\Pr(E_j^- \mid X_j, E_j^+)}{\Pr(E_j^-, E_j^+)}.\end{aligned}$$

Using the conditional independence inherent in the structure of the polytree, $\Pr(E_j^- \mid X_j, E_j^+) = \Pr(E_j^- \mid X_j)$, so that

$$\Pr(X_j \mid E) = \frac{\Pr(E_j^+)}{\Pr(E_j^-, E_j^+)}\Pr(X_j \mid E_j^+)\Pr(E_j^- \mid X_j),$$

where $\dfrac{\Pr(E_j^+)}{\Pr(E_j^-, E_j^+)}$ is a normalising constant and is denoted by α. Next two *messages* for X_j are defined: $\pi(X_j) = \Pr(X_j \,|\, E_j^+)$ and $\lambda(X_j) = \Pr(E_j^- \,|\, X_j)$, so that

$$\Pr(X_j \,|\, E) = \mathrm{Bel}(X_j) = \alpha\pi(X_j)\lambda(X_j). \tag{1.9}$$

Thus $\pi(X_j)$ represents a message from the parents of j, and $\lambda(X_j)$ a message from the children of j.

Two assumptions are made at this stage: firstly, for any root node k, having E_k^+ empty, $\pi(X_k) = \Pr(X_k \,|\, \emptyset) = \Pr(X_k)$; and secondly, for any leaf node l, having E_l^- empty $\lambda(X_l) = \Pr(\emptyset \,|\, X_l) = 1$.

Pearl's algorithm follows from this. Assume now, for illustrative purposes, that node j has two parents f and g, and two children c and d. If X_j has not yet been observed, the set E_j^- can be partitioned further into E_c^- and E_d^-, whereby $E_j^- = E_c^- \cup E_d^-$.

Thus the message $\lambda(X_j)$ can be expressed

$$\begin{aligned}\lambda(X_j) &= \Pr(E_j^- \,|\, X_j) \;=\; \Pr(E_c^- \cup E_d^- \,|\, X_j) \\ &= \Pr(E_c^- \,|\, X_j)\Pr(E_d^- \,|\, X_j, E_c^-) \\ &= \Pr(E_c^- \,|\, X_j)\Pr(E_d^- \,|\, X_j).\end{aligned}$$

Letting $\lambda_c(X_j) = \Pr(E_c^- \,|\, X_j)$ and $\lambda_d(X_j) = \Pr(E_d^- \,|\, X_j)$, it is seen that

$$\lambda(X_j) = \lambda_c(X_j)\lambda_d(X_j). \tag{1.10}$$

Consider now the calculation of, for example, $\lambda_c(X_j)$. Using the conditional independence properties as before and conditioning on the possible states of X_c,

$$\begin{aligned}\lambda_c(X_j) &= \Pr(E_c^- \,|\, X_j) \\ &= \sum_{X_c}\Pr(E_c^- \,|\, X_j, X_c)\Pr(X_c \,|\, X_j) \\ &= \sum_{X_c}\Pr(E_c^- \,|\, X_c)\Pr(X_c \,|\, X_j) \\ &= \sum_{X_c}\lambda(X_c)\Pr(X_c \,|\, X_j).\end{aligned} \tag{1.11}$$

Thus the message $\lambda_c(X_j)$ is calculated using c's λ message and the conditional probability of X_c given X_j.

In a similar way, the message $\pi(X_j)$ can be found from its parents:

$$\begin{aligned}
\pi(X_j) &= \Pr(X_j \,|\, E_j^+) \\
&= \sum_{X_f, X_g} \Pr(X_j \,|\, X_f, X_g, E_j^+) \Pr(X_f, X_g \,|\, E_j^+) \\
&= \sum_{X_f, X_g} \Pr(X_j \,|\, X_f, X_g) \Pr(X_f, X_g \,|\, E_{j(f)}^+, E_{j(g)}^+) \\
&= \sum_{X_f, X_g} \Pr(X_j \,|\, X_f, X_g) \Pr(X_f \,|\, E_{j(f)}^+) \Pr(X_g \,|\, E_{j(g)}^+),
\end{aligned}$$

where $E_{j(g)}^+$ (for example) stands for the evidence contained in the subgraph on the *tail* side of the directed edge from g to j—i.e. in the opposite direction of the arrow.

The message $\pi_f(X_j)$ is defined as $\pi_f(X_j) = \Pr(X_f \,|\, E_{j(f)}^+)$, so that

$$\pi(X_j) = \sum_{X_f, X_g} \Pr(X_j \,|\, X_f, X_g) \pi_f(X_j) \pi_g(X_j), \tag{1.12}$$

(with $\pi_g(X_j)$ defined similarly) where the message sent by j's parent f is calculated using

$$\pi_f(X_j) = \Pr(X_f \,|\, E_{j(f)}^+) = \alpha \Pr(X_f \,|\, E_f^+) \Pr(\{E_f^- \setminus E_{f(j)}^-\} \,|\, X_f),$$

where $E_{f(j)}^-$ represents the evidence contained in the subgraph on the *head* side of the directed edge from f to j—i.e. in the direction of the arrow. So the set $\{E_f^- \setminus E_{f(j)}^-\}$ refers to the children of f other than j. Denoting these r children of f by i, $i = 1, 2, \ldots, r$, then

$$\begin{aligned}
\pi_f(X_j) &= \alpha \Pr(X_f \,|\, E_f^+) \Pr(\{\cup_i E_i^-\} \,|\, X_f) \\
&= \alpha \Pr(X_f \,|\, E_f^+) \prod_i \Pr(E_i^- \,|\, X_f) \qquad \textit{(by conditional independence)} \\
&= \alpha \pi(X_f) \prod_i \lambda_i(X_f). \qquad (1.13)
\end{aligned}$$

The calculations required for this method are local in that messages are passed between neighbouring nodes. It can be shown that initially, before any variables have been observed, the λ message for each node is the unit vector. The π messages alone determine the initial marginal probability of each variable.

When information is acquired, the effects of observed evidence are propagated through the polytree. Messages on nodes will change when the corresponding variables are observed; the following algorithm is used to absorb evidence—note that when expressions (1.9) through (1.13) are referred to, they may have to be replaced by the obvious generalisations[2]. So, if a variable Y (taking values $\{y^1, y^2, \dots, y^n\}$) is found to have value y^ϕ, then this procedure is carried out:

```
BEGIN
    set $\mathrm{Bel}(y^\phi) = \lambda(y^\phi) = \pi(y^\phi) = 1$,
                        and for $\theta \neq \phi$ set $\mathrm{Bel}(y^\theta) = \lambda(y^\theta) = \pi(y^\theta) = 0$
    send new $\lambda$ messages to $Y$'s parents by expression (1.11)
    send new $\pi$ messages to $Y$'s children by expression (1.13)
END
```

Having entered the evidence into the system, the following algorithm performs the propagation:

```
BEGIN
    WHILE not all nodes are updated DO
        BEGIN
            IF a variable $B$ receives a new $\lambda$ message from a child
                AND $B$ is not already observed
            THEN
```

[2] These expressions were described using the two parent/two children case.

```
                BEGIN
                        compute new value of $\lambda(B)$ by (1.10)
                        compute new value of Bel($B$) by (1.9)
                        send new $\lambda$ messages to $B$'s parents by (1.11)
                        send new $\pi$ messages to $B$'s other children by (1.13)
                END
        IF a variable $B$ receives a new $\pi$ message from a parent
                AND $B$ is not already observed
        THEN
                BEGIN
                        compute new value of $\pi(B)$ by (1.12)
                        compute new value of Bel($B$) by (1.9)
                        send new $\lambda$ messages to $B$'s other parents by (1.11)
                        send new $\pi$ messages to $B$'s children by (1.13)
                END
        END
END
```

This algorithm however only applies to polytrees. In order to extend the method to general directed acyclic graphs, i.e. graphs where there may be more than one directed path between two nodes, conditioning is employed. The idea is to split the graph into a number of polytrees upon which the algorithm can be applied.

The task is to find a set of *loop cut nodes*, so that when the nodes in this set are assumed observed and removed from the graph, the resulting structure is a polytree. If all possible combinations of values of the loop cut nodes are considered then the results of independent applications of the algorithm can be combined using the conditional probability of the loop cut set nodes. So, given a loop cut set L and evidence E,

$$\mathrm{Bel}(X_j) = \Pr(X_j \mid E) = \sum_L \Pr(X_j \mid E, L) \Pr(L \mid E)$$

where each $\Pr(X_j \mid E, L)$ is obtained from the algorithm, and

$$\Pr(L \mid E) = \alpha \Pr(E \mid L) \Pr(L),$$

α being a normalising constant. $\Pr(L \mid E)$ is termed a *mixing weight.* $\Pr(E \mid L)$ and $\Pr(L)$ can be found using the algorithm on the respective polytrees.

Clearly this method relies on finding a loop cut set small enough so that the number of polytrees to be considered does not become unmanageable.

Pearl (1988) presents the following example, originally due to Cooper (1984):

> Metastatic cancer is a possible cause of a brain tumour and is also an explanation for increased total serum calcium. In turn, either of these could explain a patient falling into a coma. Severe headaches are also possibly associated with a brain tumour.

There are five variables:

Metastatic cancer. Denoted by A; $A = a$ implies the presence, and $A = \overline{a}$ the absence, of the cancer.

Increased total serum calcium. Denoted by B; $B = b$ implies an increase in calcium, while $B = \overline{b}$ implies no increase.

Brain tumour. Denoted by C; $C = c$ implies the presence, and $C = \overline{c}$ the absence, of a brain tumour.

Coma. Denoted by D; $D = d$ implies that a patient "occasionally lapses into a coma", while $D = \overline{d}$ implies not.

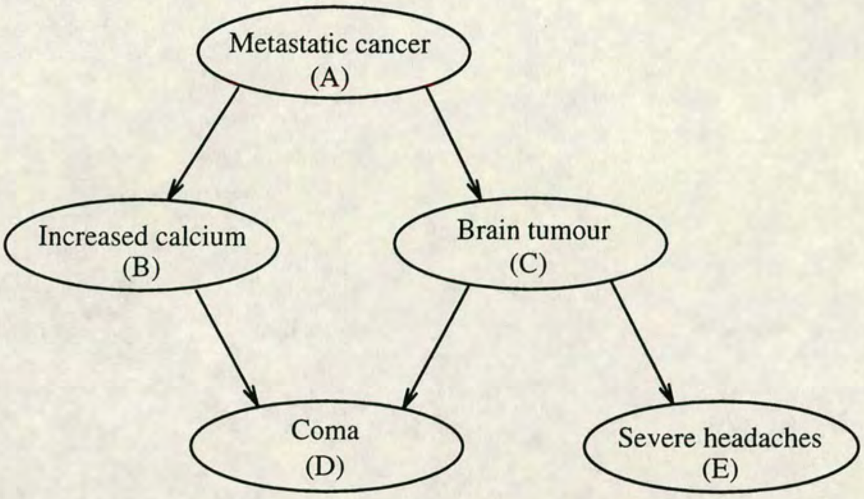


Figure 1–6: *Graph of Pearl's metastatic cancer example.*

$\Pr(A)$:	$\Pr(a) = 0.20$	
$\Pr(B \mid A)$:	$\Pr(b \mid a) = 0.80$	$\Pr(b \mid \overline{a}) = 0.20$
$\Pr(C \mid A)$:	$\Pr(c \mid a) = 0.20$	$\Pr(c \mid \overline{a}) = 0.05$
$\Pr(D \mid B, C)$:	$\Pr(d \mid b, c) = 0.80$	$\Pr(d \mid \overline{b}, c) = 0.80$
	$\Pr(d \mid b, \overline{c}) = 0.80$	$\Pr(d \mid \overline{b}, \overline{c}) = 0.05$
$\Pr(E \mid C)$:	$\Pr(e \mid c) = 0.80$	$\Pr(e \mid \overline{c}) = 0.60$

Table 1–4: *Conditional probabilities for Pearl's metastatic cancer example.*

Severe headaches. Denoted by E; $E = e$ implies that a patient suffers from severe headaches, while $E = \overline{e}$ implies the absence of such headaches.

The structure of this model is represented by Figure 1–6, and the numerical assessments of the conditional probabilities are shown in Table 1–4.

Assume now that evidence has been declared, namely a patient is suffering from severe headaches ($E = e$) but has not fallen into a coma ($D = \overline{d}$). The task is to compute the updated posterior probabilities of that patient having the cancer or a brain tumour.

Firstly, the initial belief distributions for B, C, D and E are calculated

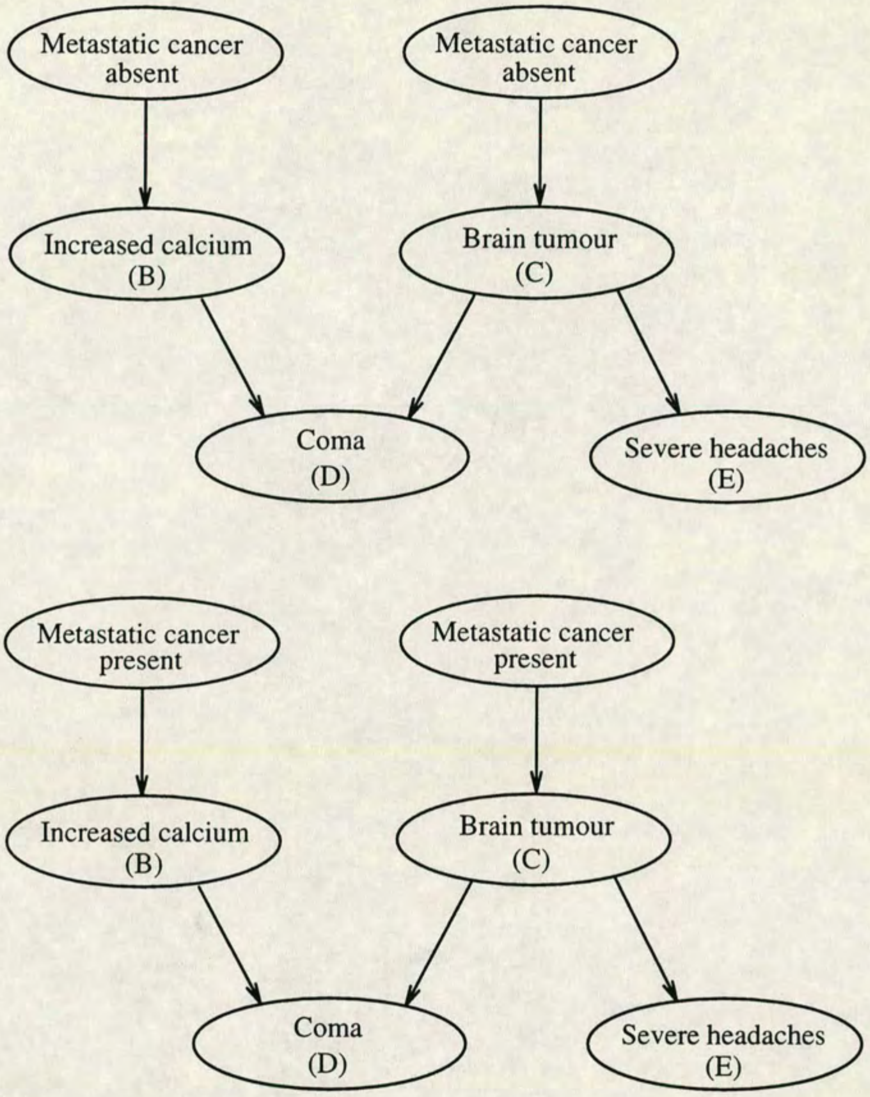


Figure 1–7: *The graph of Pearl's example becomes two polytrees.*

under the two assumptions $A = a$ and $A = \overline{a}$—the loop cut set here is thus $\{A\}$. The graph now becomes two polytrees, displayed in Figure 1–7. Messages, belief functions and normalising constants associated with each polytree are given superscripts: "1" for $A = a$, and "0" for $A = \overline{a}$.

From Table 1–4, for $A = a$,

$$\begin{aligned} \pi^1(b) &= \Pr(b \,|\, a) &= 0.80 \\ \pi^1(c) &= \Pr(c \,|\, a) &= 0.20, \end{aligned}$$

so that $\mathrm{Bel}^1(b) = \alpha\pi^1(b)\lambda^1(b)$, and since the λ messages are all unit vectors at this stage, $\mathrm{Bel}^1(b) = 0.80$. Similarly $\mathrm{Bel}^1(c) = 0.20$.

From equation (1.12)

$$\begin{aligned}\mathrm{Bel}^1(d) &= \pi^1(d) = \sum_{B,C} \Pr(d\,|\,B,C)\pi^1(B)\pi^1(C)\\ &= [0.80 \times 0.80 \times 0.20 + 0.80 \times 0.20 \times 0.20 +\\ &\qquad 0.80 \times 0.80 \times 0.80 + 0.05 \times 0.20 \times 0.20]\\ &= 0.68,\end{aligned}$$

and $\mathrm{Bel}^1(e) = \pi^1(e) = \sum_C \Pr(e\,|\,C)\pi^1(C) = 0.80 \times 0.20 + 0.60 \times 0.80 = 0.64$.

Correspondingly, for $A = \overline{a}$,

$$\begin{aligned}\pi^0(b) &= \Pr(b\,|\,\overline{a}) = 0.20\\ \pi^0(c) &= \Pr(c\,|\,\overline{a}) = 0.05.\end{aligned}$$

As with Bel^1,

$$\begin{aligned}\mathrm{Bel}^0(d) &= \pi^0(d) = \sum_{B,C} \Pr(d\,|\,B,C)\pi^0(B)\pi^0(C)\\ &= [0.80 \times 0.20 \times 0.05 + 0.80 \times 0.80 \times 0.05 +\\ &\qquad 0.80 \times 0.20 \times 0.95 + 0.05 \times 0.80 \times 0.95]\\ &= 0.23,\end{aligned}$$

and $\mathrm{Bel}^0(e) = \pi^0(e) = \sum_C \Pr(e\,|\,C)\pi^0(C) = 0.80 \times 0.05 + 0.60 \times 0.95 = 0.61$.

Each node stores its two belief functions Bel^1 and Bel^0, calculated above, along with the initial mixing weights: $w^1 = \Pr(a) = 0.20$ and $w^0 = \Pr(\overline{a}) = 0.80$.

The initial marginal probabilities can be calculated at this stage. For example $\Pr(b) = \mathrm{Bel}(b) = w^1\mathrm{Bel}^1(b) + w^0\mathrm{Bel}^0(b) = 0.20 \times 0.80 + 0.80 \times 0.20 = 0.32$. Also, $\Pr(c) = \mathrm{Bel}(c) = 0.08$.

Evidence $E = e$ is observed at this point. Node E sends new mixing weights to the other nodes:

$$\begin{aligned} w_E^1 &= \Pr(a \,|\, e) \;=\; \alpha \Pr(e \,|\, a) \Pr(a) \\ &= \alpha \mathrm{Bel}^1(e) \times w^1 \\ &= \alpha \times 0.64 \times 0.20 \end{aligned}$$

so that when w_E^0 is considered the same way, α can be found, giving $w_E^1 = 0.208$ and $w_E^0 = 0.792$.

Simultaneously, node E sends λ messages to C, namely (since $\Pr(E \,|\, C, A) = \Pr(E \,|\, C)$),

$$\begin{aligned} \lambda_E^1(c) &= \lambda_E^0(c) = \Pr(e \,|\, c) = 0.80 \\ \lambda_E^1(\overline{c}) &= \lambda_E^0(\overline{c}) = \Pr(e \,|\, \overline{c}) = 0.60. \end{aligned}$$

Next, node C sends π messages to D, for example:

$$\pi_D^1(c) = \mathrm{Bel}^1(c) = \alpha^1 \pi^1(c) \lambda^1(c) = \alpha^1 \times 0.20 \times 0.80,$$

and in fact $\pi_D^1(c) = 0.25$, $\pi_D^1(\overline{c}) = 0.75$, $\pi_D^0(c) = 0.066$ and $\pi_D^0(\overline{c}) = 0.934$. The belief functions for D thus become

$$\begin{aligned} \mathrm{Bel}^1(d) &= \sum_{B,C} \Pr(d \,|\, B, C) \pi^1(B) \pi^1(C) = 0.6875 \\ \mathrm{Bel}^0(d) &= \sum_{B,C} \Pr(d \,|\, B, C) \pi^0(B) \pi^0(C) = 0.24 \end{aligned}$$

with $\mathrm{Bel}^1(\overline{d}) = 0.3125$ and $\mathrm{Bel}^0(\overline{d}) = 0.76$.

The next piece of evidence, $D = \overline{d}$, is now introduced. Again, new mixing weights are computed:

$$\begin{aligned} w_{E,D}^1 &= \Pr(a \,|\, e, \overline{d}) \;=\; \alpha \Pr(\overline{d} \,|\, a, e) \Pr(a \,|\, e) \\ &= \alpha \mathrm{Bel}^1(\overline{d}) \times w_E^1 \;=\; \alpha \times 0.3125 \times 0.208 \end{aligned}$$

so that $w_{E,D}^1 = 0.0975$ and $w_{E,D}^0 = 0.9025$.

Next, node D sends λ messages to B and C:

$$\begin{aligned} \lambda^1_D(c) &= \sum_B \Pr(\overline{d} \,|\, B, c)\pi^1_D(B) \\ &= (0.2 \times 0.8 + 0.2 \times 0.2) \;=\; 0.20, \end{aligned}$$

together with $\lambda^1_D(\overline{c}) = 0.35$, $\lambda^0_D(c) = 0.20$, $\lambda^0_D(\overline{c}) = 0.80$, $\lambda^1_D(b) = 0.20$, $\lambda^1_D(\overline{b}) = 0.76$, $\lambda^0_D(b) = 0.20$ and $\lambda^0_D(\overline{b}) = 0.90$.

Now the belief functions can be computed for B and C, using expressions (1.9) and (1.10):

$$\mathrm{Bel}^1(b) = \alpha^1 \pi^1(b) \lambda^1_d(b) = \alpha^1(0.8 \times 0.2)$$

so that $\mathrm{Bel}^1(b) = 0.512$. It can also be found that $\mathrm{Bel}^0(b) = 0.053$, $\mathrm{Bel}^1(c) = 0.16$ and $\mathrm{Bel}^0(c) = 0.017$.

Finally, the combined belief functions can be recovered using the mixing weights:

$$\begin{aligned} \mathrm{Bel}(b) &= w^1_{E,D}\mathrm{Bel}^1(b) + w^0_{E,D}\mathrm{Bel}^0(b) \;=\; 0.096 \\ \mathrm{Bel}(c) &= w^1_{E,D}\mathrm{Bel}^1(c) + w^0_{E,D}\mathrm{Bel}^0(c) \;=\; 0.031. \end{aligned}$$

$\mathrm{Bel}(a)$ is, of course, equal to $w^1_{E,D}$, i.e. $\mathrm{Bel}(a) = 0.0975$.

It is seen that the declared evidence $D = \overline{d}$ and $E = e$ has reduced the marginal probabilities of nodes A, B and C; that is, the evidence causes us to infer reduced probabilities of the patient having metastatic cancer (from 0.20 initially down to 0.0975), increased total serum calcium (from 0.32 to 0.096) or a brain tumour (from 0.08 to 0.031).

Pearl's method relies on being able to find a suitable loop cut set within a graph. The speed of the algorithm is thus affected by the number of separate polytrees formed by the observation of loop cut nodes. The speed of the Lauritzen and Spiegelhalter method however is mainly due to the clique sizes; larger cliques will slow down the algorithm, so this method is expected to perform well on sparse graphs, i.e. graphs where the edge-to-node ratio is low. Such is the efficiency of both methods, the problem of choosing between them only becomes a serious matter when dealing with very large graphs. For a fuller comparison see Luo (1992).

In situations where neither work well, stochastic simulation may provide a solution—see Chapters 4 and 5.

Having presented methods for pure discrete models, mixed models are now introduced.

1.5 Mixed Graphical Association Models

In this section, models containing both discrete and continuous variables are considered. These models are based upon directed Markov fields of conditional Gaussian (CG) distributions. The CG distributions, introduced in Lauritzen and Wermuth (1989), have the property that the conditional distribution of the continuous variables, given the discrete variables, is multivariate Gaussian. What follows is taken from Lauritzen (1992).

The set of variables V is partitioned as $V = \Delta \cup \Gamma$, where Δ is the set of discrete variables, and Γ the continuous. An element of the state space $\mathcal{X} = \mathcal{I} \times \mathcal{Y}$ is denoted as x or as (i, y), where i represents the values of the discrete variables, and y the values of the continuous variables. The joint distribution of all variables has density f such that

$$f(x) = f(i, y) = \chi(i) \exp\{g(i) + h(i)'y - y'K(i)y/2\},$$

where $\chi(i) \in \{0, 1\}$ indicates whether f is positive at i. The variable X is then said to have a CG distribution, i.e.

$$\mathcal{L}(X_\Gamma \mid X_\Delta = i) = N_{|\Gamma|}(\xi(i), \Sigma(i)) \qquad \text{whenever} \qquad p(i) = \Pr(X_\Delta = i) > 0$$

where X_Γ and X_Δ represent the continuous and discrete variables respectively, and $\xi(i) = K(i)^{-1}h(i)$, $\Sigma(i) = K(i)^{-1}$, the latter being positive definite.

The triple (g, h, K) (defined only for $\chi(i) > 0$) constitutes the *canonical* characteristics of the model, and (p, ξ, Σ) the *moment* characteristics. The notion of a CG distribution is extended to that of a CG *potential*, which is a function ψ of the form

$$\psi(x) = \psi(i, y) = \chi(i) \exp\{g(i) + h(i)'y - y'K(i)y/2\},$$

where $K(i)$ is now only assumed to be a symmetric matrix, so that ψ may not be a density. The triple (g, h, K) is used here also, with $\psi \approx (g, h, K)$. The moment characteristics are well-defined only when K is positive definite for all i with $\chi(i) > 0$. In this case, Σ and ξ are defined as before, but

$$p(i) \propto \{\det \Sigma(i)\}^{\frac{1}{2}} \exp\{g(i) + h(i)'\Sigma(i)h(i)/2\}.$$

If the moment characteristics (p, ξ, Σ) are given, the canonical characteristics can be calculated:

$$K(i) = \Sigma(i)^{-1}, \qquad h(i) = K(i)\xi(i)$$

and

$$g(i) = \log p(i) + \{\log \det K(i) - |\Gamma| \log(2\pi) - \xi(i)'K(i)\xi(i)\}/2.$$

The joint distribution of x is assumed to satisfy the directed Markov property, i.e. the density is is equal to the product of the conditional densities of the variables corresponding to each node, given the values of parent nodes—see Kiiveri *et al.* (1984).

In order for the properties of CG distributions to be exploited, it is necessary to assume that no continuous nodes have discrete children. Thus discrete nodes can only have discrete parents, and conditional probabilities are defined for them as usual. For a continuous node A, with associated variable X_A, the conditional distribution is of the type

$$\mathcal{L}(X_A \mid \mathrm{pa}(A)) = N(\alpha(i) + \beta(i)'z, \gamma(i))$$

and so a separate Gaussian distribution emerges for each i, that is for each combination of values of A's discrete parents (represented by i), there is a different Gaussian distribution for X_A. Here: $\mathrm{pa}(A)$ is the set of parents of A; $\gamma(i)$ is the variance of X_A when A's discrete parents take the set of values i; the $\alpha(i)$ are real-valued constants; z is the vector of values of A's continuous parents; and $\beta(i)$ is a vector of multiplicative constants for z.

Thus the mean of $\mathcal{L}$ is a linear function of the values of A's continuous parents, with the form of the linear function chosen by the values of discrete parents. The variance depends on discrete parents, but not continuous ones.

So $\mathcal{L}$ corresponds to a CG potential $\phi_A \approx (g_A, h_A, K_A)$ with

$$g_A(i) = -\frac{\alpha(i)^2}{2\gamma(i)} - \{\log(2\pi\gamma(i))\}/2, \qquad h_A(i) = \frac{\alpha(i)}{\gamma(i)} \begin{pmatrix} 1 \\ -\beta(i) \end{pmatrix}$$

and

$$K_A(i) = \frac{1}{\gamma(i)} \begin{pmatrix} 1 & -\beta(i)' \\ -\beta(i) & \beta(i)\beta(i)' \end{pmatrix}.$$

This set-up is illustrated by the following example, taken from Lauritzen (1992), which is concerned with control of emissions of heavy metals from a waste incinerator.

> The emission from a waste incinerator differs because of compositional differences in incoming waste. Another important factor is the waste burning regime which can be monitored by measuring the concentration of CO_2 in the emission. The filter efficiency depends on the technical state of the electrofilter and the composition of waste. The emission of heavy metal depends both on the concentration of metal in the incoming waste and the emission of dust in general. The emission of dust is monitored through measuring the penetrability of light.

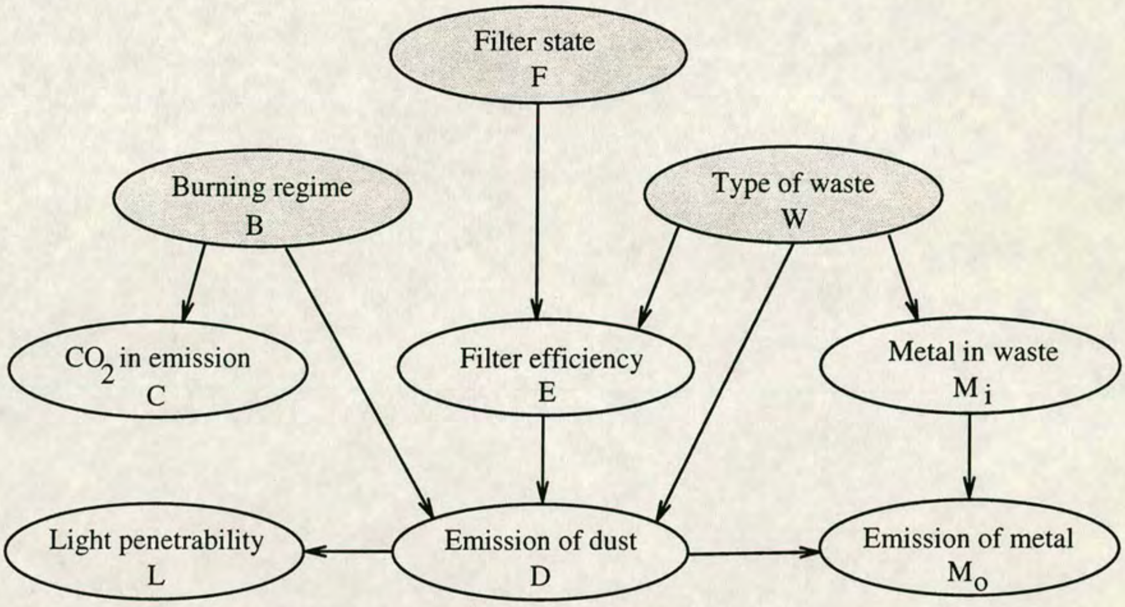


Figure 1–8: *Lauritzen's waste incinerator example.*

Note that the time aspect of this problem is ignored.

The graph relating to this description is shown in Figure 1–8; as it represents a mixed model, the discrete nodes are shaded. The distributions of the variables are defined thus:

Burning regime (Discrete) Denoted by B. Let

$$\Pr(B = b = \textit{stable}) = 0.85 = 1 - \Pr(B = \overline{b} = \textit{unstable}).$$

Filter state (Discrete) Denoted by F. Let

$$\Pr(F = f = \textit{intact}) = 0.95 = 1 - \Pr(F = \overline{f} = \textit{defective}).$$

Type of waste (Discrete) Denoted by W. Let

$$\Pr(W = w = \textit{industrial}) = 2/7 = 1 - \Pr(W = \overline{w} = \textit{household}).$$

Filter efficiency (Continuous) Denoted on a logarithmic scale by E. Assume the relation

$$\text{waste}_{\text{out}} = \text{waste}_{\text{in}} \times \rho$$

and so

$$\log \text{waste}_{\text{out}} = \log \text{waste}_{\text{in}} + \log \rho.$$

Then let $E = \log \rho$ (admitting filter *inefficiency* might be a better name for E) and specify

$$\begin{aligned}
\pounds(E \mid f, \overline{w}) &= N(-3.2, 0.00002)\\
\pounds(E \mid \overline{f}, \overline{w}) &= N(-0.5, 0.0001)\\
\pounds(E \mid f, w) &= N(-3.9, 0.00002)\\
\pounds(E \mid \overline{f}, w) &= N(-0.4, 0.0001)
\end{aligned}$$

which correspond to filter efficiencies $1 - \rho$ of about 96%, 39%, 98% and 33% respectively.

Emission of dust (Continuous) Denoted on a logarithmic scale by D. Let

$$\begin{aligned}
\pounds(D \mid b, w, e) &= N(6.5 + e, 0.03)\\
\pounds(D \mid b, \overline{w}, e) &= N(6.0 + e, 0.04)\\
\pounds(D \mid \overline{b}, w, e) &= N(7.5 + e, 0.1)\\
\pounds(D \mid \overline{b}, \overline{w}, e) &= N(7.0 + e, 0.1)
\end{aligned}$$

So on a day when household waste is burned under a stable regime with an intact filter, the mean concentration will be $\exp(6.0 - 3.2) = 16.4\text{mg/Nm}^3$.

Concentration of $\mathbf{CO_2}$ (Continuous) Denoted on a logarithmic scale by C. Let

$$\pounds(C \mid b) = N(-2, 0.1) \qquad \text{and} \qquad \pounds(C \mid \overline{b}) = N(-1, 0.3).$$

Thus concentration of CO_2 is usually around 14% under a stable regime, and 37% when things are unstable.

Penetrability of light (Continuous) Denoted on a logarithmic scale by L. Let

$$\pounds(L \mid d) = N(3 - d/2, 0.25)$$

which corresponds to the penetrability being roughly inversely proportional to the square root of the dust concentration.

Metal in waste (Continuous) Denoted on a logarithmic scale by M_i. Let

$$\mathcal{L}(M_i \mid w) = N(0.5, 0.01) \qquad \text{and} \qquad \mathcal{L}(M_i \mid \overline{w}) = N(-0.5, 0.005)$$

which correspond to industrial waste concentration of heavy metal being about three times that of household waste.

Emission of metal (Continuous) Denoted on a logarithmic scale by M_o. Let

$$\mathcal{L}(M_o \mid d, m_i) = N(d + m_i, 0.002)$$

which simply assumes that the concentration of emitted heavy metal is about the same as in the original waste.

This example will be analysed in the next section by Lauritzen's exact method, and in Chapter 4 by stochastic simulation.

1.6 Propagation in Mixed Models

Lauritzen (1992) presents a propagation scheme for mixed graphical association models. For continuous variables, only means and variances are computed.

As in section 1.4.1, the directed graph of Figure 1–8 is converted to a triangulated moral graph. However, one further step is necessary—the graph must be made *decomposable.*

Definition 3 *An undirected graph is decomposable if and only if it is triangulated and does not contain any path $(j, i_1, i_2, \ldots, i_n, k)$ between two non-adjacent discrete nodes passing through continuous nodes only, i.e. with $i_x \in \Gamma$ for $1 < x < n$ and no edge between j and k.*

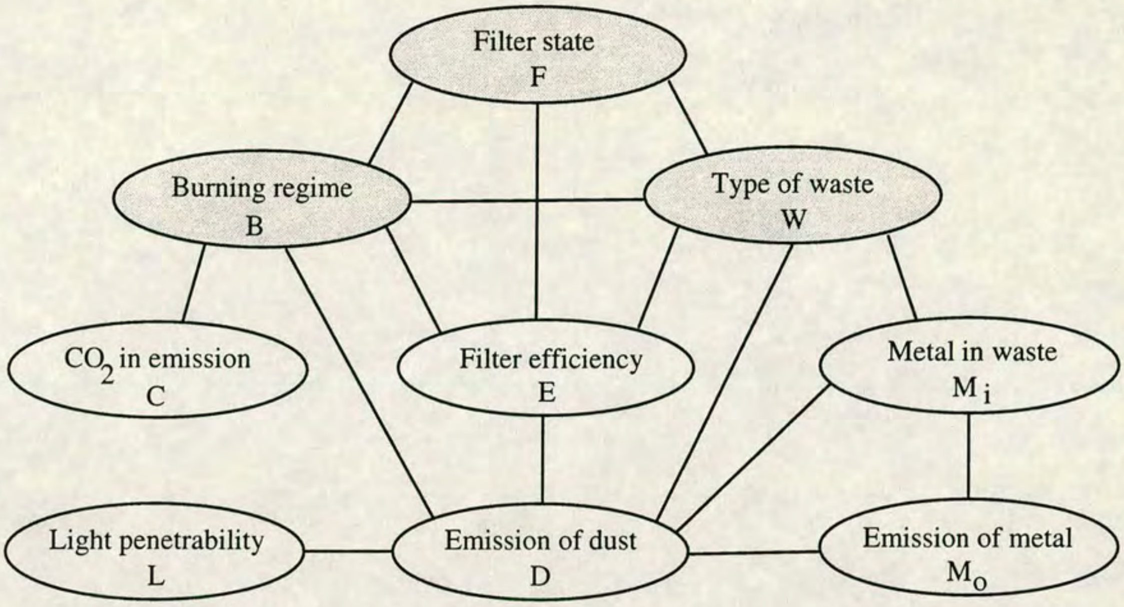


Figure 1–9: *Decomposable graph for waste incinerator example.*

The decomposable graph for the current example is shown in Figure 1–9. Note that an edge between B and F has been added; there would otherwise have been a forbidden path (B, E, F) with B and F non-adjacent.

Lauritzen's mixed algorithm works on cliques in much the same way as the method of section 1.4.1. Here, the cliques are placed in a *junction tree*—a tree of subsets of V satisfying the condition that if A and B are subsets in the tree, $A \cap B$ is a subset of all sets on the path between A and B (see Jensen *et al.* (1990)).

A final condition is required for the propagation scheme; a *strong root* is required.

Definition 4 *A subset R in a junction tree is a strong root if any pair A, B of neighbours in the tree with A closer to R than B satisfies*

$$(B \setminus A) \subseteq \Gamma \qquad or \qquad (B \cap A) \subseteq \Delta.$$

Statement iii)' of Theorem 2' of Leimer (1989) ensures that a junction tree with a strong root can be formed from the cliques of a decomposable graph. Figure 1–10

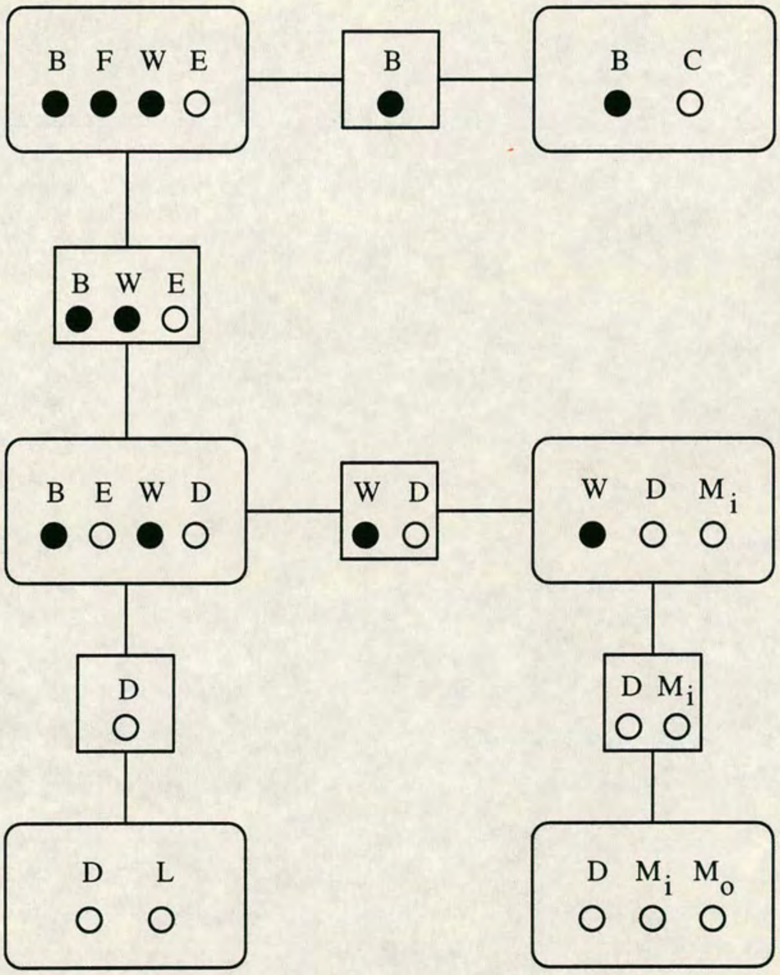


Figure 1–10: *Junction tree for waste incinerator example; • for discrete nodes, ◦ for continuous nodes.*

shows a junction tree for the waste incinerator example. The clique $\{B, F, E, W\}$ is a strong root, since whenever a separator set is not purely discrete, the clique furthest away has only continuous vertices beyond those in the separator. For example $\{W, D, M_i\}$ has only M_i beyond separator $\{W, D\}$.

The propagation scheme relies on the following set of basic operations on CG potentials.

Extension: If (g, h, K) are the characteristics of a CG potential ϕ defined on $\mathcal{X} = \mathcal{I} \times \mathcal{Y}$, the extension $\bar{\phi}$ defined on $\mathcal{W} = (\mathcal{I} \times \mathcal{J}) \times (\mathcal{Y} \times \mathcal{Z})$ is

$$\bar{\phi}(i, j, y, z) = \phi(i, y)$$

with characteristics

$$\bar{b}(i, j) = g(i), \qquad \bar{h}(i, j) = \begin{pmatrix} h(i) \\ 0 \end{pmatrix}, \qquad \bar{K}(i, j) \begin{pmatrix} K(i) & 0 \\ 0 & 0 \end{pmatrix}.$$

Multiplication and division: Multiplication is defined in the obvious way:

$$(\phi_1 \times \phi_2)(x) = \phi_1(x)\phi_2(x)$$

after extension has been carried out as above. The canonical characteristics become

$$(g_1, h_1, K_1) \times (g_2, h_2, K_2) = (g_1 + g_2, h_1 + h_2, K_1 + K_2).$$

For division, care has to be taken in case of dividing by zero:

$$(\phi_1/\phi_2)(x) = \begin{cases} 0 & \text{if } \phi_1(x) = 0 \\ (\phi_1(x)/\phi_2(x)) & \text{if } \phi_2(x) \neq 0 \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Marginalisation: Addition is not defined in general, since adding two CG potentials typically results in a function with a different structure. There are several cases to consider; firstly marginalisation over continuous variables. Let

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \qquad h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \qquad K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$$

with y_1 having dimension p and y_2 dimension q. The integral $\int \phi(i, y_1, y_2)dy_1$ is then finite if and only if K_{11} is positive definite, in which case the new CG potential $\tilde{\phi}$ has canonical characteristics

$$\begin{aligned}
\tilde{g}(i) &= g(i) + \left\{p\log(2\pi) - \log\det K_{11}(i) + h_1(i)'K_{11}(i)^{-1}h_1(i)\right\}/2 \\
\tilde{h}(i) &= h_2(i) - K_{21}(i)K_{11}(i)^{-1}h_1(i) \\
\tilde{K}(i) &= K_{22}(i) - K_{21}(i)K_{11}(i)^{-1}K_{12}(i).
\end{aligned}$$

Secondly, there is marginalisation over discrete variables; if $h(i, j)$ and $K(i, j)$ do not depend on j,

$$\tilde{g}(i) = \log \sum_{j:\chi(i,j)=1} \exp\{g(i, j)\}, \qquad \tilde{h}(i) = h(i, j), \qquad \tilde{K}(i) = K(i, j).$$

If however there is dependence on j, the following procedure is followed, although it is only well-defined for $K(i, j)$ positive definite. It is best to use moment characteristics here; the marginal $\tilde{\phi}$ has characteristics $(\tilde{p}, \tilde{\xi}, \tilde{\Sigma})$ where

$$\tilde{p}(i) = \sum_j p(i, j), \qquad \tilde{\xi}(i) = \sum_j \xi(i, j)p(i, j)/\tilde{p}(i),$$

and

$$\tilde{\Sigma}(i) = \sum_j \Sigma(i, j)p(i, j)/\tilde{p}(i) + \sum_j (\xi(i, j) - \tilde{\xi}(i))'(\xi(i, j) - \tilde{\xi}(i))p(i, j)/\tilde{p}(i).$$

Finally, there is marginalisation over both continuous and discrete variables; in this case, the above procedures are carried out, marginalising first over the continuous variables and then over the discrete.

Once a model has been specified, the algorithm acts on the junction tree representation. The set of cliques is denoted $\mathcal{C}$, and the intersections of neighbouring cliques in the tree are called *separators*; the set of these is denoted $\mathcal{S}$. Both cliques and separators have CG potentials attached to them. The joint system belief is written as

$$\phi_U = \frac{\prod_{V\in\mathcal{C}} \phi_V}{\prod_{S\in\mathcal{S}} \phi_S}, \tag{1.14}$$

and is proportional to the joint density of all the variables. Since the potentials involved are CG potentials, the joint density will be a CG density itself.

The junction tree with its strong root must be initialised. Firstly each node A is assigned to a clique V in the tree, such that the union of A and A's parents is a subset of V. Then ϕ_V is declared to be the product of all the (extensions of) potentials ϕ_A for nodes assigned to each clique V. For the separators S, $\phi_S \equiv 1$, i.e. the potential with canonical characteristics $(0, 0, 0)$. This same potential is given to cliques not assigned nodes. With this initialisation, expression (1.14) will be a correct representation of the joint system belief.

For the waste incinerator example, the nodes could be assigned thus: B, C to clique $\{B, C\}$; F, W, E to $\{B, F, W, E\}$; D to $\{B, W, E, D\}$; L to $\{L, D\}$; M_i to $\{W, D, M_i\}$; and M_o to $\{D, M_i, M_o\}$. Example potentials include

$$\begin{aligned} g_{\{B,C\}}(stable) &= -19.930 \\ h_{\{B,C\}}(stable) &= -20 \\ K_{\{B,C\}}(stable) &= 10 \end{aligned}$$

and

$$\begin{aligned} g_{\{B,C\}}(unstable) &= -3.881 \\ h_{\{B,C\}}(unstable) &= -3.333 \\ K_{\{B,C\}}(unstable) &= 3.333 \end{aligned}$$

as well as

$$h_{\{L,D\}} = \begin{pmatrix} 12 \\ 6 \end{pmatrix}, \qquad K_{\{L,D\}} = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}.$$

Now the structure has been established, the propagation procedure is described. The cliques carry information in the form of potentials and the separators act as communication channels between the cliques. Incoming evidence is divided up into *items of evidence*, which can be either of the following:

- a function $\chi_W(i_W) \in \{0,1\}$, where W is a set of discrete variables that is a subset of some clique V in the junction tree; or

- a statement that $Y_A = y_A^*$ for a continuous node A.

Discrete evidence is entered simply by multiplying χ_W onto the potential ϕ_V. Continuous evidence is entered by modifying the potentials of all cliques and separators containing A. If a potential ϕ has canonical characteristics (g, h, K) with

$$h(i) = \begin{pmatrix} h_1(i) \\ h_A(i) \end{pmatrix}, \qquad K(i) = \begin{pmatrix} K_{11}(i) & K_{1A}(i) \\ K_{A1}(i) & K_{AA}(i) \end{pmatrix},$$

then the modified potential ϕ^* will have characteristics (g^*, h^*, K^*) as

$$\begin{aligned} K^*(i) &= K_{11}(i) \\ h^*(i) &= h_1(i) - y_A^* K_{A1}(i) \\ g^*(i) &= g(i) + h_A(i) y_A^* - K_{AA}(i)(y_A^*)^2/2. \end{aligned}$$

In the example, assume now that industrial waste has been burned, the light penetration has been found to be 1.1, and the CO_2 concentration −0.9 (on logarithmic scales). The type of waste information is entered as function χ_W:

$$\chi_W(\mathit{industrial}) = 1, \qquad \chi_W(\mathit{household}) = 0.$$

The potentials for the continuous evidence become, for example,

$$\begin{aligned} g^*_{\{B\}}(\mathit{stable}) &= -19.930 + 18 - 4.050 &= -5.980 \\ g^*_{\{B\}}(\mathit{unstable}) &= -3.881 + 3 - 1.350 &= -2.231 \end{aligned}$$

as well as

$$h^*_{\{D\}} = 6 - 1.1 \times 12 = -7.2, \qquad K^*_{\{D\}} = 1.$$

The most important stage now follows: the effects of evidence are propagated through the junction tree. Consider such a tree with cliques $\mathcal{C}$ and separators $\mathcal{S}$.

Let $V \in \mathcal{C}$ and $W_1, W_2, \ldots, W_m$ be neighbours of V with separators $S_1, S_2, \ldots, S_m$ respectively. The clique V is said to ***absorb*** from $W_1, W_2, \ldots, W_m$ if the following calculations are performed:

$$\begin{aligned} \phi'_{S_i} &= \sum_{W_i \setminus V} \phi_{W_i} \qquad \textit{(for } i = 1, 2, \ldots, m) \\ \phi'_V &= \phi_V \times (\phi'_{S_1}/\phi_{S_1}) \times \cdots \times (\phi'_{S_m}/\phi_{S_m}). \end{aligned} \tag{1.15}$$

Dividing both sides of expression (1.15) by $\phi_{S_1} \times \cdots \times \phi_{S_m}$) shows that the joint system belief is not changed by absorption.

The propagation scheme is then based on this operation. Each $V \in \mathcal{C}$ is given the action CollectEvidence: when CollectEvidence in V is called from a neighbour W, then V calls CollectEvidence in all its other neighbours. When they have finished their CollectEvidence, V absorbs from them.

Next, each $V \in \mathcal{C}$ is given the action DistributeEvidence: when DistributeEvidence is called in V from a neighbour W, V absorbs from W and calls DistributeEvidence in all its other neighbours.

Note that the joint system belief is unchanged under both CollectEvidence and DistributeEvidence. In practice, once evidence has been entered, evoking first CollectEvidence and then DistributeEvidence from a strong root performs the propagation correctly. Marginals for each node can then be obtained by marginalising within the cliques of the junction tree. This scheme thus supplies the correct updated probabilities of states at discrete nodes, and the correct means and variances at continuous nodes.

Figure 1–11 shows the initial and updated marginal probabilities, means and variances at each node. The logarithmic value for node C of -0.9 implies a concentration of CO_2 of 41% in the emission, which as expected considerably increases the probability of an unstable burning regime. The evidence assumed has resulted in the expected emission of heavy metal increasing by a factor of $\exp(1.3) \approx 3.7$.

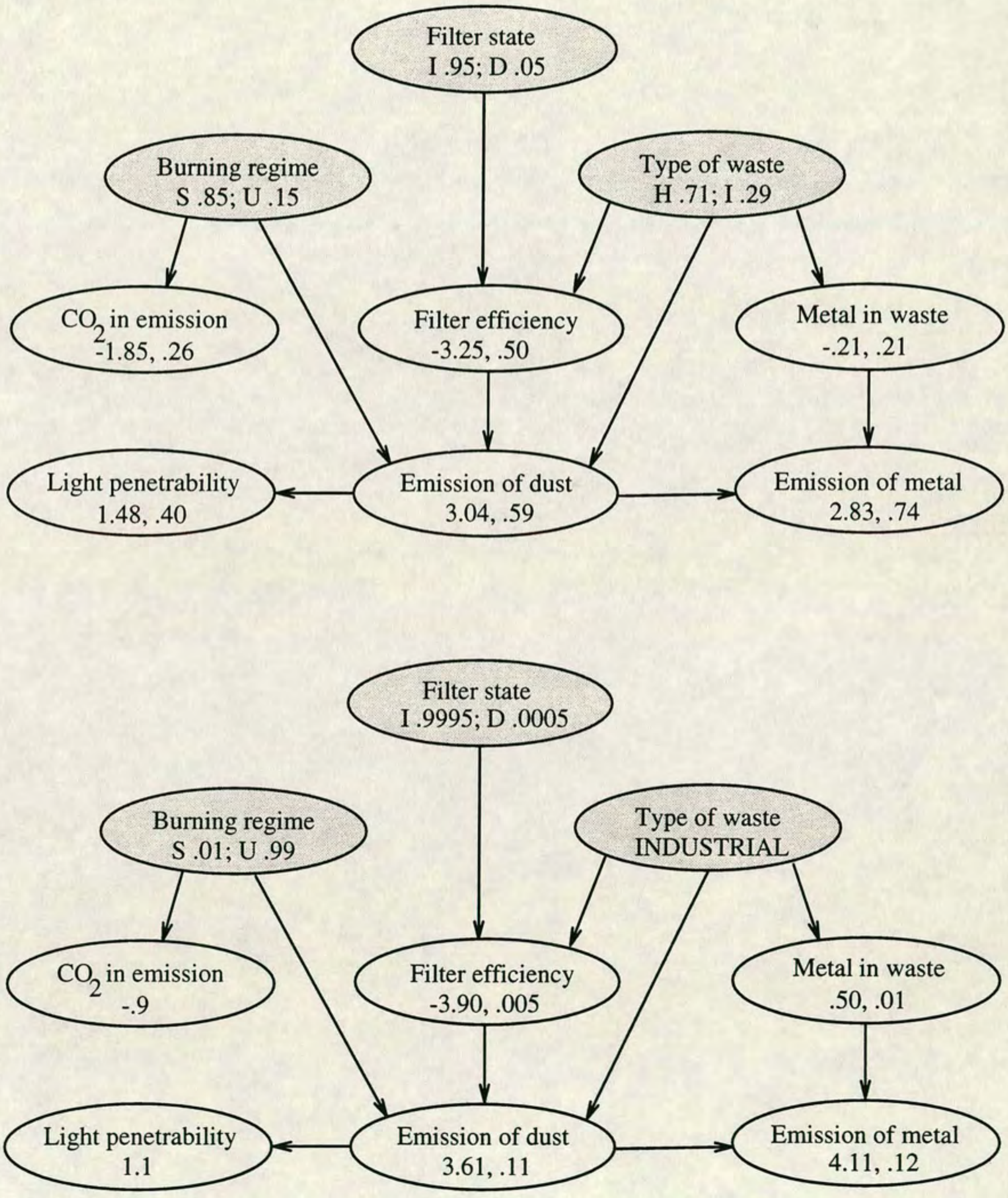


Figure 1–11: *Initial (top) and updated marginal probabilities, means and variances for each node of the waste incinerator example.*

Lauritzen (1992) states that the computation of the marginal densities can be "forbiddingly complex". In Chapter 4 a method of estimating these densities is discussed.

In this chapter graphical models have been introduced and propagation schemes for analysing them have been described. The next chapter considers a particular application of graphical models: legal networks.

Chapter 2

Legal Networks

2.1 Introduction

The use of graphs in a legal context can be traced as far back as Wigmore (1913). He introduced a method of organising evidence and facts arising in any particular legal case into charts. This method was expanded in Wigmore (1937), and it is clear that his charts can be seen as directed acyclic graphs, where nodes represent propositions, and directed edges represent probative (inferential) force.

Following an appearance of probabilistic arguments[1] in the "Collins case" (see, for example, Edwards, 1991), Tribe (1971) launched a full-scale attack on what he described as "trial by mathematics." Finkelstein and Fairley (1970) had proposed that Bayes' Theorem be used in criminal trials, making the simplifying assumption that "factual questions such as the identity of the person who is the source of pertinent evidence (such as fingerprints) are decisive on the issue of the guilt or innocence of a given defendant." Tribe criticised this assumption and claimed that Bayesian analysis becomes unduly complicated without it; he

[1] Which turned out to have serious flaws—see section 2.3.1.

argued that even taking into account the possibility of a defendant having been framed "strain[s] the [Bayesian] system beyond its breaking point."

Tribe had good reason to criticise; the logical structure of criminal cases can be very complex. Wigmore's Chart Method had shown that it *is* feasible to break down a case into its constituent parts for analysis; Morgan (1961) also used a graphical representation to illustrate the "stages" of inference. As early as 1971, David Schum was working on the "Bayesian analysis of criminal trials", resulting in Schum (1987), a massive two-volume study of the structure and theory of evidence—the crucial point here being that Schum incorporated probabilities into the evidence structure. Koehler (1991) provided an excellent counterattack to Tribe's arguments.

This chapter starts by looking at Wigmore's charts in section 2.2. The ideas of Morgan (1961) and Schum (1987) follow, along with an example of a probabilistic legal network from Edwards (1991), in §2.3.

Forensic science problems are closely related to the legal cases, and one such example (from Aitken and Gammerman, 1989) is described in section §2.4.

The remainder of the chapter is concerned with likelihood ratios, which may prove easier for juries to understand than probabilities. Martin (1980) developed an algorithm for efficient calculation of likelihood ratios in a network; in a legal context it is useful to discover the relative value of a particular piece of evidence. Finally, the notion of using likelihood ratios (instead of probabilities) as input to an expert system is explored.

2.2 Wigmore's Chart Method

Wigmore (1937) proposed a method for analysing the myriad of facts and evidence which can arise during preparation of a case for trial. A lawyer must construct persuasive arguments without leaving weaknesses to be exploited by the opposite side. Wigmore's aim was to introduce a logical, or scientific, approach for doing this.

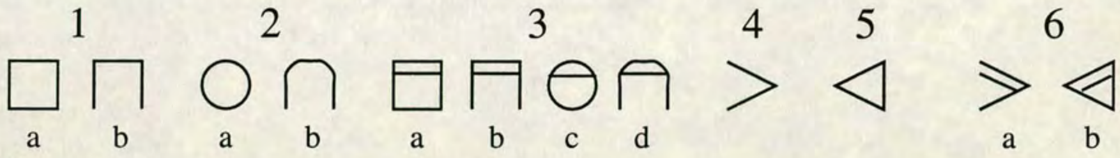
In Wigmore (1937) it is stated that the chart method is a "logical (or psychological) process" for combining many ideas into rather fewer "... until the number and kind is such that the mind can consciously juxtapose them with due attention to each."

The process involves designing a *key-list*, that is a list of evidence items, propositions, and so on. The entries in the key-list are represented pictorially by a chart resembling a directed acyclic graph. This chart is intended to enable facts and evidence, and the links between them, to be exhibited concisely and transparently. Keeping track of all information in the mind is a very difficult task in other than the simplest cases.

Rather than simply use a circle or ellipse for each node, Wigmore uses a number of different symbols to represent different types of evidence—see Figure 2–1. The edges between the "nodes" mostly have arrows on them, but other marks are made too—see Figure 2–2. The "nodes" are numbered to tally with the entries in the key-list.

It is important to note at this point that Wigmore does *not* make numerical assessments of probabilities in his charts[2]; these are merely hinted at by the use of ciphers. For example, placing a circle (○) inside a node signifies that the evidence

[2] For example, of the probability of one piece of evidence given another.



Key:

(1) Testimonial evidence : (a) *affirmatory* (e.g. testimony that the defendant had the knife); (b) *negatory* (e.g. the defendant did *not* have the knife).

(2) Circumstantial evidence : (a) *affirmatory* (e.g. knife was found near defendant, hence defendant had it); (b) *negatory* (e.g. knife was found elsewhere, hence defendant did *not* have it).

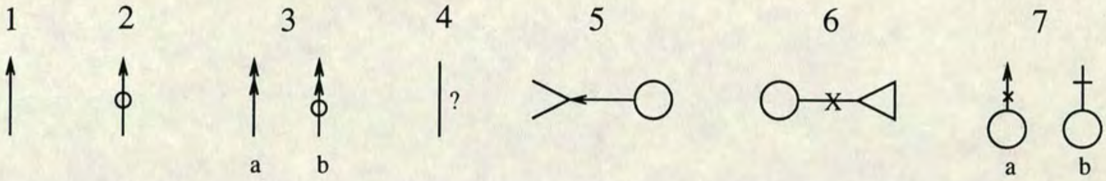
(3) Same four types of evidence as (1) and (2) when offered by the *defendant* in a case.

(4) Explanatory evidence : for circumstantial evidence, explaining away the effect (e.g. knife may have been dropped by a third person); for testimonial evidence, discrediting it (e.g. witness was too excited to see who picked up the knife).

(5) Corroborative evidence : for circumstantial evidence, closing up possible explanations (e.g. no third person around when knife was found); for testimonial evidence, supporting testimony (e.g. witness was stood close by and was calm).

(6) Same types of evidence as (4) and (5) when offered by the defendant.

Figure 2–1: *Wigmore's symbols for "nodes".*



Key:

(1) Provisional force given to an inference from *affirmatory* evidence is shown by adding an arrow-head to the edge.

(2) Provisional force given to *negatory* evidence is shown by adding an arrow-head and a small circle to the edge.

(3) Stronger force given to evidence is represented by doubling the arrow-head. For example, several witnesses making the same testimony might result in stronger force.

(4) A small question mark by an edge signifies doubt as to the probative effect of the evidence.

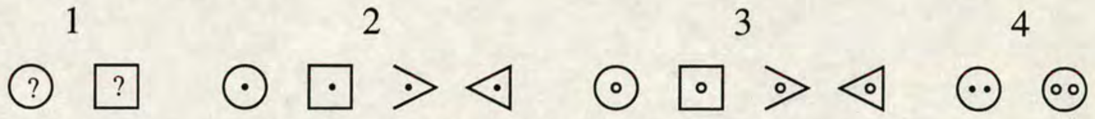
(5) If a single *explanatory* fact detracts from the force of the desired inference (e.g. if it discredits the assertion of a witness), this is signified by an arrow-head as shown.

(6) If a single *corroborative* fact is believed, a cross is placed on the edge.

(Note that doubling an arrow-head or a cross from (5) or (6) also increases the intensity of the effect;)

(7) Determining the overall effect of facts on one particular fact; if this fact is (a) *corroborated* then a cross is placed on the edge above it; and (b) *explained* then a short horizontal line is placed on the edge above it.

Figure 2–2: *Wigmore's symbols on edges.*



Key:

(1) A question mark inside a symbol signifies a mental balance; the item is neither believed nor disbelieved.

(2) A dot inside a symbol signifies belief in the item.

(3) A small circle inside a symbol signifies disbelief in the item.

(4) Doubling the mark inside a symbol increases the effect; for example two circles signifies *strong* disbelief in the evidence or fact.

Figure 2–3: *Wigmore's ciphers representing belief or otherwise.*

or fact is *disbelieved*; a dot $(\cdot)$ implies *belief* for that node—see Figure 2–3 for more examples.

The positioning of the symbols in relation to each other is important. A supposed fact tending to prove or disprove another is placed *below* it. A supposed explanatory or corroborative fact tending to lessen or increase the force of another is placed to the *left* or *right* of it, respectively. When a fact is judicially admitted or observed by the court (such as the presentation of "Exhibit A") the symbols $(\P, \infty)$ respectively are placed below the "node".

Wigmore's treatise describes how a chart and key-list should be constructed. Each piece of evidence must be analysed; it must be classified as to its type, and its inferences considered. For a human act, motives must be made distinct, and a separate node reserved for each. Similarly, explanatory facts should be separated as far as possible into individual items. The ultimate leaf node, represented in Wigmore's charts as the top-most node, generally refers to the matter to be proven, e.g. the guilt or innocence of the defendant.

Consider then the example of Figure 2–4, which is a small section of a chart. (The reader is referred to Wigmore (1937) for fuller examples[3].)

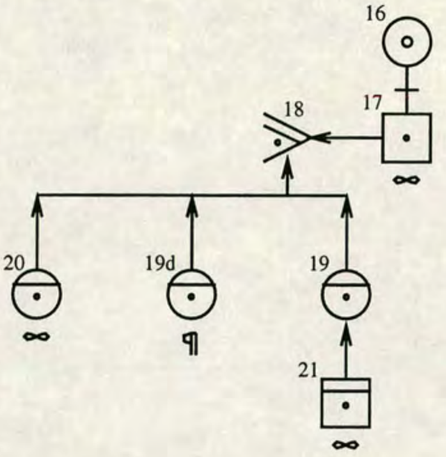


Figure 2–4: *A small section of a Wigmore chart.*

A witness has offered affirmatory circumstantial evidence in a trial. Node 17 represents the witness's testimony of the evidence (node 16). Suppose the witness is thought to have some bias against the defendant. Let node 18 be the supposed general fact of bias, with nodes 19 and 20 signifying the two circumstances that might cause it. Node 19 represents the witness's relation to the defendant as a discharged employee; a second witness's testimony to this is shown by node 21. Note that node 19 here is supported by node 19*d*, the supposed *general truth* that discharged employees have some hostility towards their erstwhile employers. Node 20 displays the first witness's strong air of bias while on the stand.

Figure 2–4 also shows the total probative effect of this small group of nodes; the witness's evidence (16) has been rejected since the fact of bias is believed.

This small example illustrates how a Wigmore chart can be used to represent the structure of argument in law, and to consider the net effect of all the evidence. Wigmore's work can aid the design of an influence diagram for a legal case; the following section is concerned with incorporating probabilities into a chart.

[3] A full example of a chart and key-list for a real case would take up too much room in this thesis. Indeed, Schum (1989a) refers to a chart discussed in Twining (1984) that measures *37 feet* in length...

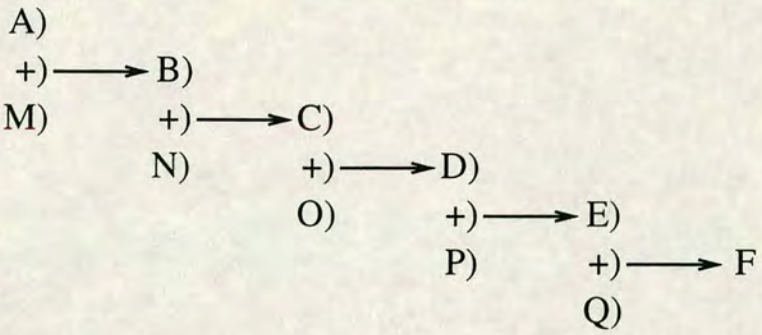


Figure 2–5: *Morgan's multistage inference diagram of a murder case.*

2.3 Probabilistic Legal Networks

This section considers the development of expert systems for legal cases.

The first point to make is that in probabilistic legal networks, nodes and edges are represented as in §1.2; while Wigmore's multitude of symbols should prove useful to a legal analyst, they are not required here. The dot (·) and the circle (∘) are not needed since they refer to degrees of belief or disbelief in a proposition—something perhaps better quantified by probabilities. The symbols ¶ and ∞ imply acceptance of a proposition; in a graphical model the relevant variable would have its state or value fixed, the probabilities would be updated and the node removed from the graph.

Morgan (1961) portrayed the inferences involved in a murder case as in Figure 2–5. The evidence is a love letter written by the defendant to the murder victim's wife and the question is whether the person who wrote the letter killed the husband of the female addressee. Apart from A, which represents the love letter itself, the letters in the chart stand for various factual inferences and certain supporting generalisations. For example, B represents the defendant's love of the victim's wife; C, the defendant's desire for exclusive possession (sic) of the victim's wife; and M, the generalisation "A man who loves a woman probably desires her for himself alone." Also D stands for the defendant's desire to get rid of the victim and O the generalisation "a man who loves a married woman and desires her for himself alone desires to get rid of her husband."

Morgan's diagrams of inference are always *chains*; that is, they consist of a single thread running between single pieces of evidence and single facts in issue.

This attempt to create a formal logical inferential structure for a legal case is appealing; Professor David Schum, as well as working on the theory of evidence (see, for example, Schum, 1987), has also investigated such inferences. Schum favoured more a Wigmore-style network, believing that "such charts are rich enough to reconstruct and mimic the sort of thinking that people are actually inclined to use when they face complex real-world problems." Tillers and Schum (1988) concentrated on making Wigmore's methods easier to implement.

Figure 2–6 is a simple example of the kind of chart Schum designed. Here, *two* pieces of evidence (represented by the filled boxes) lead (eventually) to the same factual statement (statements represented by unfilled boxes). Note that Schum's networks, since they are derived from Wigmore charts, have the arrows on some of the edges pointing in the opposite way to what might be expected; for example, although box A points to H, the probabilities $\Pr(A \mid H)$ must be defined. Of course, $\Pr(H \mid A)$ is ultimately of interest. Eyewitness and testimony boxes however (C and D in this example) point to their relevant "facts" in the expected way. Here, box C points to box B, and so $\Pr(B \mid C)$ must be defined.

The "H" at the top stands for the point to be proven. For example, box C might represent a report that a defendant escaped from jail; box B would then represent "defendant escaped from jail." Also, box A is interpreted as the defendant's belief in his own guilt, with D then as the defendant's evidence "I did it." Finally, H would represent the defendant's guilt or otherwise.

Schum (1989a) contained descriptions of different types and combinations of evidence, and of how they might be incorporated into the graph of a case: for example, hearsay evidence, corroborative and contradictory evidence.

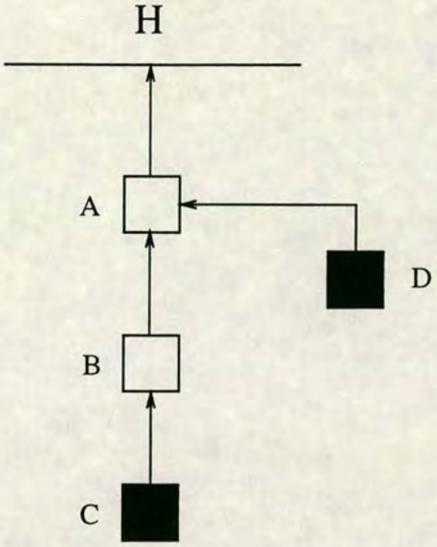


Figure 2–6: *A Schum legal network.*

The next important step is to introduce the "probabilistic" element into the models; consider again Wigmore's symbols. Note that on edges between symbols representing evidence, double arrow-heads are allowed, implying "stronger force"; this might relate to a higher (or lower, if the evidence is negatory) conditional probability between two items. Similarly, a question mark placed by an edge (see Figure 2–2) would mean that one fact or piece of evidence has little effect on the outcome of the other.

In Wigmore charts, ciphers are placed inside evidence symbols to represent belief or otherwise in the evidence or fact; the scale (see Figure 2–3) running from ∘∘ to ·· relates to strong disbelief through strong belief. A question mark implies a "mental balance" and thus perhaps a probability of 0.5 on the evidence. Normal belief and disbelief could be represented by probabilities of 0.6 and 0.4 respectively; then strong belief and strong disbelief by 0.8 and 0.2. In this way, what are essentially quantitative ideas from Wigmore can be converted into numerical assessments.

For example, in Figure 2–6, consider boxes B and C. Given a report that the defendant escaped from jail ($C = c$), this would naturally affect our belief in B, the fact of the defendant having escaped from jail. So we might let $\Pr(b \mid c) = 0.8$ (i.e. strong belief that the defendant escaped from jail given a report stating so) and $\Pr(b \mid \overline{c}) = 0.4$ (i.e. normal (only) disbelief that the defendant escaped given a lack of a report stating so).

The Collins case (Edwards, 1991) is now presented, first with the prosecution argument, and then with Edwards' analysis.

2.3.1 The Collins Case—The Prosecution Argument

This case concerned Janet and Malcolm Collins, who were convicted in Los Angeles of second-degree robbery. The details that follow are taken from Edwards (1991).

Malcolm Collins appealed to the Supreme Court of California, his main complaint being "that the introduction of evidence pertaining to the mathematical theory of probability and the use of the same by the prosecution during the trial was error prejudicial to the defendant." The Supreme Court agreed and the conviction was reversed. The court's description of the case was as follows:

> On June 18, 1964, about 11:30 AM Mrs. Juanita Brooks, who had been shopping, was walking home along an alley in the San Pedro area of the City of Los Angeles... As she stooped down to pick up an empty carton, she was suddenly pushed to the ground by a person whom she neither saw nor heard approach... She managed to look up and saw a young woman running from the scene. According to Mrs. Brooks the latter appeared to weigh about 145 pounds, was wearing "something dark," and had her hair "between a dark blond and a light blond,"... [H]er purse, containing between \$35 and \$40, was missing.
>
> About the same time..., John Bass, who lived on the street at the end of the alley,... [heard] "a lot of crying and screaming" coming from the alley. As he looked in that direction, he saw a woman run

Characteristic	Probability
Partly yellow car	0.1
Man with moustache	0.25
Girl with ponytail	0.1
Girl with blonde hair	0.333
Black man with beard	0.1
Interracial couple in car	0.001

Table 2–1: *Probabilities used by the prosecution in the Collins case.*

> out of the alley and enter a yellow automobile parked across the street from him... The car... passed within six feet of Bass... [I]t was being driven by a male negro, wearing a mustache and beard...
>
> ... Bass described the woman who ran from the alley as a Caucasian, slightly over five feet tall, of ordinary build, with her hair in a dark blond ponytail, and wearing dark clothing.

The defense did not challenge greatly the descriptions of the couple. The main complaint was about expert testimony on the product rule for independent events. The prosecutor had introduced hypothesised probabilities of the reported characteristics of the perpetrators if the crime were committed by some couple other than Janet and Malcolm Collins. These probabilities are shown in Table 2–1. The prosecutor then multiplied these numbers together and came up with a probability of $1/12,000,000$ of another couple having the same characteristics.

There has been much criticism of Table 2–1 and the prosecutor's argument in legal literature—see Finkelstein and Fairley (1970) for example. Both the court and Tribe (1971) complained that no evidence was presented to support the probability estimates. Further, the events are *not* independent; a blonde girl and a black man seem very likely to be an interracial couple.

Fairley and Mosteller (1974) have since corrected the details of the probabilities for the features of a couple such as the Collinses. Edwards (1991) presents an analysis of the case using an influence diagram, and this analysis follows.

2.3.2 The Collins Case—Edwards' Analysis

A major error in the prosecution's analysis was that all the events (in Table 2–1) were assumed independent. Edwards (1991) structures his events more carefully, using an influence diagram to display the conditional independence structure of his model. His description of how he set about creating the graph is similar to Wigmore's ideas.

One device Edwards used was the notion of *imputed stipulations.* The idea is that if neither side in a case disputed a particular proposition, then it can be assumed that both sides accepted it; that is the proposition is regarded as a truth. The imputed stipulations that Edwards' used for the Collins case were:

1. The mugging occurred.
2. The mugging was committed by a Caucasian woman.
3. The getaway vehicle was a car that waited for the woman outside the alley.
4. The driver of that car, also waiting outside the alley, was male.
5. The Collinses were together at the time of the crime.
6. The Collinses had no alibi. (They testified to being elsewhere, but no other evidence was offered.)

Admittedly, stipulations such as these become clearer *after* a trial, but a legal analyst could certainly use such a method to organise his thoughts.

Figure 2–7 shows the network Edwards designed for the Collins case. The network illustrates, for example, the assumed conditional independence of skin

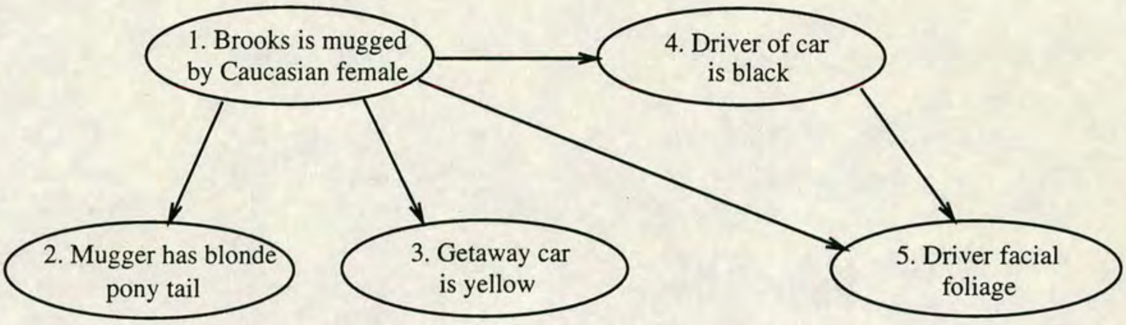

1. Brooks is mugged by Caucasian female
4. Driver of car is black
2. Mugger has blonde pony tail
3. Getaway car is yellow
5. Driver facial foliage

Figure 2–7: *Edwards' influence diagram for the Collins case.*

colour of the driver and car colour, and the lack of such independence of skin colour and facial foliage.

Additional evidence would be incorporated into Figure 2–7 by including more nodes; if, for example, evidence was presented on the credibility or otherwise of the eyewitness, nodes would be added and linked to the nodes representing issues upon which the eyewitness testified. (Schum (1989b) presents an analysis of witness credibility issues.)

The related conditional probabilities are displayed in Table 2–2, and represent subjective judgements by the prosecution in the Collins case and by Edwards himself. Those coming from Edwards reflect such judgements as if the mugger was Janet Collins and the driver of the getaway car was black, then he was virtually certain to have a moustache and beard, since he was virtually certain to be Malcolm Collins.

Now that the model has been specified, it can be analysed. The target node is number 1, the identity of the mugger. The prior probability of Janet Collins being the attacker is 0.0001, from Table 2–2. Now consider that evidence has been presented in court, and that this evidence relates to one or more of the variables in the model. A propagation scheme, such as that by Lauritzen and Spiegelhalter (1988) described in section 1.4.1, can then be used to calculate the posterior probability of Janet Collins being guilty. Some possible testimonies,

<table>
<tr><td colspan="5"><b>Node 1:</b> Brooks is mugged by a Caucasian female</td></tr>
<tr><td>Outcome</td><td colspan="4">Probability</td></tr>
<tr><td>Yes, Janet Collins</td><td colspan="4">0.0001</td></tr>
<tr><td>Other Caucasian female</td><td colspan="4">0.9999</td></tr>
<tr><td colspan="5"><b>Node 2:</b> Mugger has blonde ponytail</td></tr>
<tr><td></td><td colspan="4">Outcomes</td></tr>
<tr><td><i>Condition of node 1</i></td><td>Yes</td><td>Blonde, no ponytail</td><td>Ponytail, not blonde</td><td>Neither</td></tr>
<tr><td>Yes, Janet Collins</td><td>1.000</td><td>0.000</td><td>0.000</td><td>0.000</td></tr>
<tr><td>Other Caucasian female</td><td>0.033</td><td>0.500</td><td>0.150</td><td>0.317</td></tr>
<tr><td colspan="5"><b>Node 3:</b> Getaway car is yellow</td></tr>
<tr><td></td><td colspan="4">Outcomes</td></tr>
<tr><td><i>Condition of node 1</i></td><td colspan="2">Yes</td><td colspan="2">No</td></tr>
<tr><td>Yes, Janet Collins</td><td colspan="2">0.999</td><td colspan="2">0.001</td></tr>
<tr><td>Other Caucasian female</td><td colspan="2">0.100</td><td colspan="2">0.900</td></tr>
<tr><td colspan="5"><b>Node 4:</b> Driver of car is black</td></tr>
<tr><td></td><td colspan="4">Outcomes</td></tr>
<tr><td><i>Condition of node 1</i></td><td colspan="2">Yes</td><td colspan="2">No</td></tr>
<tr><td>Yes, Janet Collins</td><td colspan="2">0.999</td><td colspan="2">0.001</td></tr>
<tr><td>Other Caucasian female</td><td colspan="2">0.001</td><td colspan="2">0.999</td></tr>
<tr><td colspan="5"><b>Node 5:</b> Driver facial foliage</td></tr>
<tr><td></td><td colspan="4">Outcomes</td></tr>
<tr><td><i>If node 1 is "Yes, Janet Collins" and node 4 is...</i></td><td>Beard and moustache</td><td>Moustache no beard</td><td>Beard, no moustache</td><td>No beard or moustache</td></tr>
<tr><td>Yes (driver black)</td><td>0.999</td><td>0.000</td><td>0.000</td><td>0.001</td></tr>
<tr><td>No (driver not black)</td><td>0.025</td><td>0.300</td><td>0.010</td><td>0.665</td></tr>
<tr><td colspan="5"><i>If node 1 is "Other Caucasian female" and node 4 is...</i></td></tr>
<tr><td>Yes (driver black)</td><td>0.025</td><td>0.600</td><td>0.010</td><td>0.365</td></tr>
<tr><td>No (driver not black)</td><td>0.025</td><td>0.300</td><td>0.010</td><td>0.665</td></tr>
</table>

Table 2–2: *Conditional probabilities for Edwards' network of the Collins case.*

Evidence	Posterior probability
Blonde with ponytail	0.0030
Yellow car	0.0010
Black with moustache and beard	0.7997
All three together	*0.9992*
Blackness alone	0.0908
Facial foliage alone	0.0040

Table 2–3: *Posterior probabilities of the mugger being Janet Collins for various possible testimonies.*

such as eyewitness evidence that the getaway car was yellow, and their effect on node 1 are shown in Table 2–3. For example, the posterior probability of Janet Collins being the mugger given that the mugger had a blonde ponytail is 0.003. If the variables 2 to 5 are assumed to take values indicating that the mugger had a blonde ponytail, the getaway car was yellow, and the driver of the car was black and had a moustache and beard, then the posterior probability of Janet's guilt becomes 0.9992.

Once the posterior probabilities of guilt given evidence have been obtained, the problem remains of what to do with them; should a suspect be imprisoned for assault and robbery on a 0.9992 probability? The conditional probabilities defining the model are mostly subjective, after all. Certainly, however, the prosecution in the Collins case should be able to argue that Janet Collins seems highly likely to be the mugger, given the evidence, although whether the jury would appreciate the statistics is another matter. Likelihood ratios may present a more intuitive and acceptable alternative to probabilities—see section 2.5.

2.4 A Forensic Science Application

Forensic science has a very important rôle in criminal legal cases. The methods of the previous section work well with forensic data, since this data may supply reliable estimates for conditional probabilities required for the model, reducing the need for subjective judgements to be made.

Aitken and Gammerman (1989) present a forensic science example of an expert system. The (fictional) case is described thus:

> A murder has been committed. There are two suspects, X and Y, who are associates and who say they met the victim V some time before the commission of the crime. If there were an eyewitness to this meeting it would be interesting to know from that witness if the meeting had been cordial or not and, in particular, if there had been a fight. The reliability of the eyewitness is also of interest. Since X and Y are associates, it is feasible that Y may pick up something from X and then deposit it at the scene of the crime. For example, fibres from a jacket of X may be picked up by some garment of Y and then be left at the crime scene by Y, thus incriminating X who may, in fact, be perfectly innocent. Such transfer from X to Y may take place if, say, Y drives X's car frequently.

The graph of Figure 2–8 shows the structure linking the following variables:

- A: X committed the murder;
- B: Y committed the murder;
- E: Eyewitness evidence given of a fight between X, Y and the victim sometime before the commission of the crime;

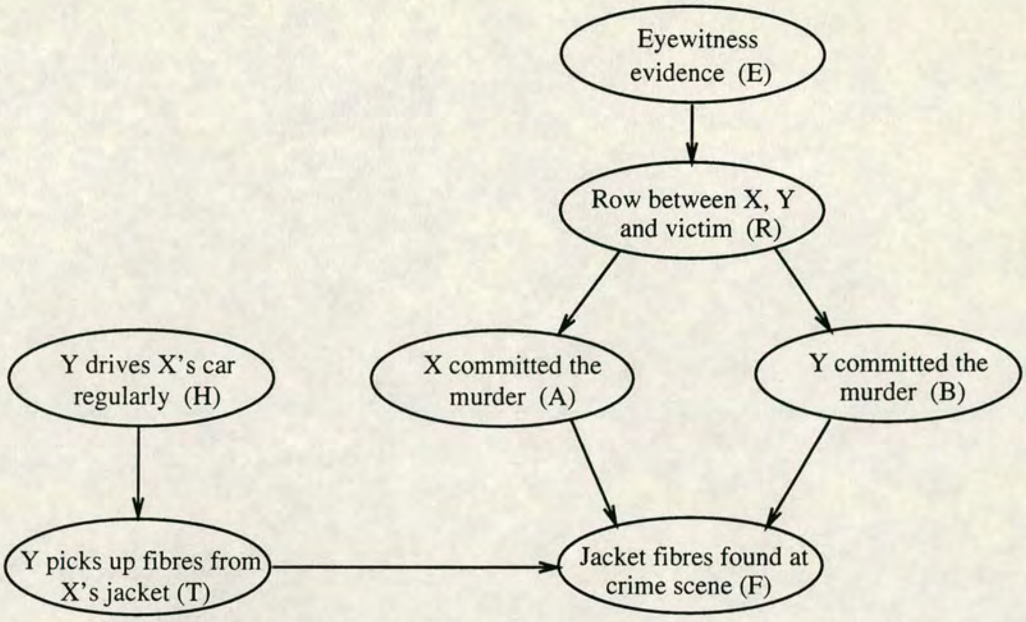

Eyewitness evidence (E)
Row between X, Y and victim (R)
Y drives X's car regularly (H)
X committed the murder (A)
Y committed the murder (B)
Y picks up fibres from X's jacket (T)
Jacket fibres found at crime scene (F)

Figure 2–8: *Influence diagram for forensic science example.*

- R: A fight occurred between X, Y and the victim;
- F: Fibres from a jacket similar to the one found in the possession of X are found at the crime scene;
- H: Y drives X's car regularly; and
- T: Y picks up fibres from X's jacket.

These variables are binary, and take the values "true" or "false". For example, if $A = a$ then X did commit the murder, but if $A = \overline{a}$ then X did not kill the victim. Table 2–4 shows the suggested conditional probabilities for the model, along with an interpretation of the numbers.

Given the model, it is possible to apply an algorithm from section 1.4 to obtain posterior probabilities of variables after accepting pieces of evidence. The important variables here are A and B, the guilt or innocence of X and Y.

Initially, $\Pr(A = a) = \Pr(B = b) = 0.053$, i.e. for example the probability of X committing the murder is 0.053, before acceptance of evidence. If the fibre

Probability	Interpretation in influence diagram
$\Pr(e) = 0.05$	Eyewitness evidence of row between X and Y is unlikely.
$\Pr(r \mid e) = 0.80$	Row likely given eyewitness account;
$\Pr(r \mid \overline{e}) = 0.05$	$\Pr(r \mid e) \neq 1$ since eyewitness may be lying/mistaken.
$\Pr(a \mid r) = 0.50$	X and Y are unlikely to commit the crime a priori
$\Pr(a \mid \overline{r}) = 0.01$	but possibly might have, given row.
$\Pr(b \mid r) = 0.50$	
$\Pr(b \mid \overline{r}) = 0.01$	
$\Pr(h) = 0.70$	It is quite likely that Y drives X's car.
$\Pr(t \mid h) = 0.20$	If Y often drives X's car, there is a possibility that Y will
$\Pr(t \mid \overline{h}) = 0.01$	pick up fibres from X's jacket; otherwise it is unlikely.
$\Pr(f \mid a, b, t) = 0.80$	Fibres from X's jacket are quite likely to be found if X
$\Pr(f \mid a, b, \overline{t}) = 0.70$	committed the crime, regardless of whether Y was
$\Pr(f \mid a, \overline{b}, t) = 0.70$	involved or not. If Y did not pick up fibres from X's
$\Pr(f \mid a, \overline{b}, \overline{t}) = 0.70$	jacket and X did not commit the crime, it is unlikely
$\Pr(f \mid \overline{a}, b, t) = 0.20$	that fibres from X's jacket will be found at the scene.
$\Pr(f \mid \overline{a}, b, \overline{t}) = 0.03$	
$\Pr(f \mid \overline{a}, \overline{b}, t) = 0.01$	
$\Pr(f \mid \overline{a}, \overline{b}, \overline{t}) = 0.01$	

Table 2–4: *Suggested probability values for forensic science example.*

evidence is accepted ($F = f$) then the posterior probabilities are $\Pr(A = a \mid f) = 0.888$ and $\Pr(B = b \mid f) = 0.090$; thus the chance of X being involved in the murder has increased to over three-quarters, which makes sense since it was his jacket fibres that were found at the scene of the crime. The probability of Y being involved has also increased, though not as much as for X, since there is a *possibility* of fibre transfer.

2.4.1 Uncertain evidence

In a forensic or legal context, there may be a variable in a model relating to evidence which is not totally reliable for a number of reasons; the witness might be lying, or mistaken. In this case, evidence will need to be accepted (if it is presented) but with an element of uncertainty—this is achieved in the following way.

Assume that evidence has been presented to the court relating to whether or not Y drives X's car regularly, but that the witness is not wholly reliable, and that there is a 20% chance that the witness is lying. A node H' is added to the influence diagram to give Figure 2–9. Letting, for example,

$$\Pr(H = h, H' = h') = 0.28, \qquad \Pr(h, \overline{h}') = 0.42,$$

$$\Pr(\overline{h}, h') = 0.07 \qquad \text{and} \qquad \Pr(\overline{h}, \overline{h}') = 0.20$$

ensures that $\Pr(h) = 0.70$ and $\Pr(h \mid h') = 0.80$. So in order to accept uncertain evidence, H' is set to take value h' and the effect is propagated through the graph. The posterior probability for $H = h$ is obviously 0.80. The probability of fibre transfer ($T = t$) becomes $\Pr(t \mid h') = 0.162$, whereas before accepting the "uncertain" evidence it was $\Pr(t) = 0.143$.

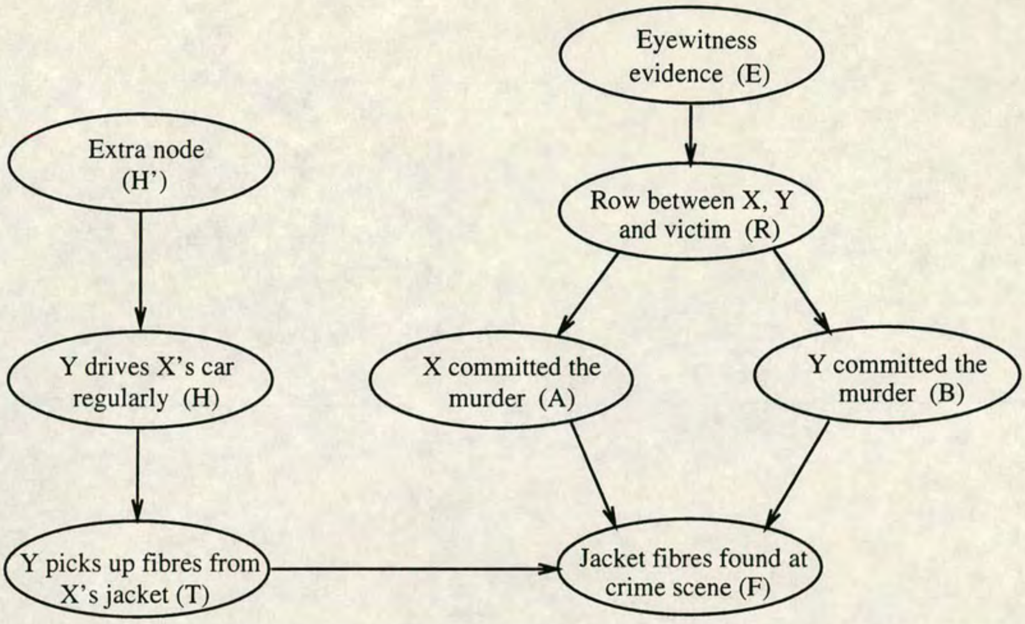

Eyewitness evidence (E)
Extra node (H')
Row between X, Y and victim (R)
Y drives X's car regularly (H)
X committed the murder (A)
Y committed the murder (B)
Y picks up fibres from X's jacket (T)
Jacket fibres found at crime scene (F)

Figure 2–9: *Forensic science example with extra node for uncertain evidence.*

2.5 Likelihood Ratios

In section 2.3 it was mentioned that likelihood ratios may be useful in legal inference of the kind discussed in this chapter.

Witnesses and investigators in criminal trials might well feel happier providing statements of the form

> Event X is q times more likely to occur if event Y were true than if Y were false,

than of the form

> The probability of event X, given event Y is true, is θ, while the probability of X, given Y is false, is ϕ.

Edwards *et al.* (1990) claimed that "[a]s several decades of research in psychophysics has shown, people are much better at making relative judgements than at making absolute judgements."

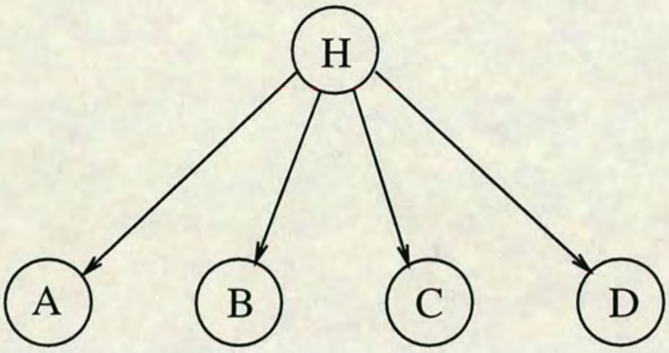


Figure 2–10: *Typical structure of model analysed by PIP*

This section then is concerned with the use of likelihood ratios in expert systems.

2.5.1 PIP and CASPRO

In the 1960's, Ward Edwards, David Schum and others worked on a computational reasoning system known as Probabilistic Information Processing (PIP)—see, for example, Edwards (1962), Edwards *et al.* (1968).

In PIP, people were required to assess probabilities and likelihood ratios to enable a computer program to calculate posterior probabilities or odds. The initial idea was for people to assess prior odds on certain hypotheses of interest, and then to judge probabilities or likelihood ratios for each new evidence item the system would be required to process. The hypotheses, in "expert systems" terms, were represented by a node connected to child nodes denoting the evidence items. These child nodes had no children of their own, nor were they joined to any node other than the hypotheses node; thus the model had a single-stage, non-hierarchical structure. Figure 2–10 shows a possible influence diagram for such a model with four evidence nodes.

Schum, after discovering Wigmore's work, decided that the single-stage structure of the models used by PIP did not faithfully represent human inference—see Schum (1989a). PIP did not achieve great success, partly because of the burden

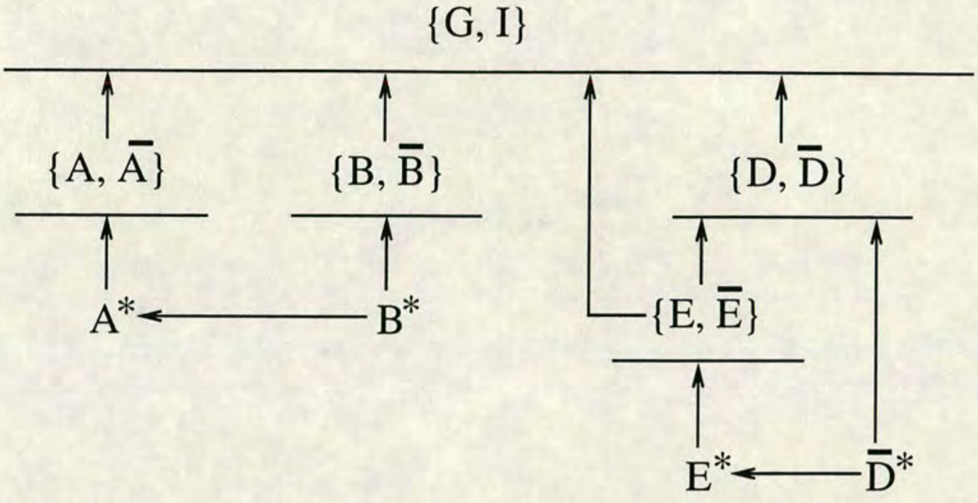


Figure 2–11: *Example influence diagram for CASPRO.*

of having to give assessments of conditional probabilities (as opposed to purely likelihood ratios—see Edwards *et al.*, 1990), and partly because work on the hierarchical[4] (multistage) nature of inference was in its infancy in this context.

Such work was carried out by David Schum in the years following, and this led to the program CASPRO[5], developed by Martin (1980). This program resembled many of the expert system programs today, in that the structure of a network and conditional probabilities are entered into the system. The program was quite a basic one however; it did not calculate marginal probabilities for the variables in a model, but likelihood ratios. The type of likelihood ratios output by CASPRO are illustrated by the following example from Martin (1980).

The influence diagram for this example is shown in Figure 2–11. The node A^* represents a report given by a witness in a case; the report is the testimony

[4] Schum, Martin, Edwards *et al.* use the term *cascaded.*

[5] Oddly, no reference is made in Martin (1980) to the phrase for which CASPRO is an acronym; "CAS" clearly refers to "cascaded", but "PRO" could be any of "probability", "propagation", or "prototype"—the latter word is used frequently in the paper.

that event A occurred while $\overline{A}$ did not. Similarly B^* is the report of a second witness that event B has occurred, and again for E^* and event E. The node $\overline{D}^*$ represents a report that event $\overline{D}$ occurred. The top node contains the two hypotheses of interest, the guilt (G) or innocence (I) of a suspect.

CASPRO concentrates on the reports that events occurred; it calculates how much more likely a report is if the suspect is guilty than if he/she is innocent—that is, it calculates for example,

$$\Lambda_{A^*} = \frac{\Pr(A^* \mid G)}{\Pr(A^* \mid I)}$$

in the case of report A^*. Since the report B^* depends on A^* the likelihood ratio[6] will also depend on A^*, so that

$$\Lambda_{B^*|A^*} = \frac{\Pr(B^* \mid A^*, G)}{\Pr(B^* \mid A^*, I)}.$$

The conditional probabilities given for this example by Martin (1980) are shown in Table 2–5. The resulting likelihood ratios (which can be checked with a routine from section 1.4) are:

$$\begin{aligned} \Lambda_{A^*} &= 8.349 \\ \Lambda_{B^*|A^*} &= 1.267 \\ \Lambda_{E^*} &= 1.090 \\ \Lambda_{\overline{D}^*|E^*} &= 0.858. \end{aligned}$$

Martin (1980) also considered *pure linear* inference, as displayed in Figure 2–12 with the case where there are three intermediate steps between the report (D^*) and the ultimate hypotheses (D_0 and $\overline{D_0}$). In Figure 5 of that paper, the expression for the likelihood ratio for this pure linear case was shown—see Figure

[6] The arrows are the "wrong" way round on this network; probabilities such as $\Pr(B^* \mid A^*, B)$ must be defined.

$\Pr(A \mid G) = 0.50$	$\Pr(A^* \mid A) = 0.90$
$\Pr(A \mid I) = 0.05$	$\Pr(A^* \mid \overline{A}) = 0.01$
$\Pr(B \mid G, A) = 0.70$	$\Pr(B \mid I, A) = 0.60$
$\Pr(B \mid G, \overline{A}) = 0.40$	$\Pr(B \mid I, \overline{A}) = 0.30$
$\Pr(B^* \mid A^*, B) = 0.90$	$\Pr(D \mid G) = 1.00$
$\Pr(B^* \mid A^*, \overline{B}) = 0.01$	$\Pr(D \mid I) = 0.01$
$\Pr(E \mid G, D) = 0.80$	$\Pr(E \mid I, D) = 0.75$
$\Pr(E \mid G, \overline{D}) = 0.00$	$\Pr(E \mid I, \overline{D}) = 0.40$
$\Pr(E^* \mid E) = 0.50$	$\Pr(\overline{D}^* \mid E^*, D) = 0.60$
$\Pr(E^* \mid \overline{E}) = 0.40$	$\Pr(\overline{D}^* \mid E^*, \overline{D}) = 0.70$

Table 2–5: *Conditional probability values for CASPRO example.*

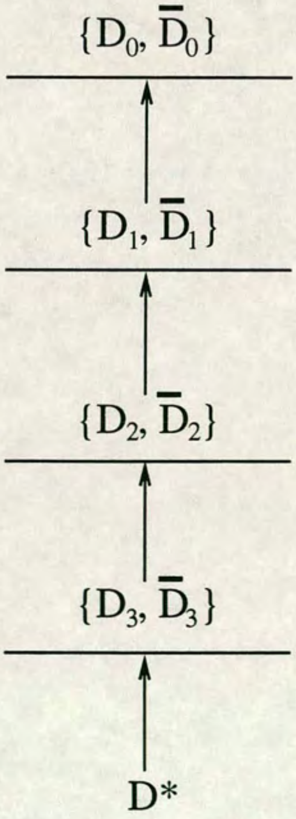


Figure 2–12: *Pure linear inference—four stages.*

$$\frac{\Pr(D^*|D_0)}{\Pr(D^*|\overline{D_0})} = \frac{\Pr(D_1|D_0) + \left[\dfrac{\Pr(D_2|D_1)+\left[\dfrac{\Pr(D_3|D_2)+\left[\frac{\Pr(D^*|D_3)}{\Pr(D^*|\overline{D_3})}-1\right]^{-1}}{\Pr(D_3|\overline{D_2})+\left[\frac{\Pr(D^*|D_3)}{\Pr(D^*|\overline{D_3})}-1\right]^{-1}}-1\right]^{-1}}{\Pr(D_2|\overline{D_1})+\left[\dfrac{\Pr(D_3|D_2)+\left[\frac{\Pr(D^*|D_3)}{\Pr(D^*|\overline{D_3})}-1\right]^{-1}}{\Pr(D_3|\overline{D_2})+\left[\frac{\Pr(D^*|D_3)}{\Pr(D^*|\overline{D_3})}-1\right]^{-1}}-1\right]^{-1}} - 1\right]^{-1}}{\Pr(D_1|\overline{D_0}) + \left[\dfrac{\Pr(D_2|D_1)+\left[\dfrac{\Pr(D_3|D_2)+\left[\frac{\Pr(D^*|D_3)}{\Pr(D^*|\overline{D_3})}-1\right]^{-1}}{\Pr(D_3|\overline{D_2})+\left[\frac{\Pr(D^*|D_3)}{\Pr(D^*|\overline{D_3})}-1\right]^{-1}}-1\right]^{-1}}{\Pr(D_2|\overline{D_1})+\left[\dfrac{\Pr(D_3|D_2)+\left[\frac{\Pr(D^*|D_3)}{\Pr(D^*|\overline{D_3})}-1\right]^{-1}}{\Pr(D_3|\overline{D_2})+\left[\frac{\Pr(D^*|D_3)}{\Pr(D^*|\overline{D_3})}-1\right]^{-1}}-1\right]^{-1}} - 1\right]^{-1}}$$

Figure 2–13: *Martin's expression for a pure linear likelihood ratio.*

2–13 here. Schum (1989a) described the expression as "picturesque"; however, a more compact formula using likelihood ratios (such as $\Pr(D_3 \mid D_2)/\Pr(D_3 \mid \overline{D_2})$) is derived in section 2.5.2.

2.5.2 Likelihood Ratios as Input to Expert Systems

Edwards *et al.* (1990) claimed that "[when analysing a probabilistic network] we can no longer get by with just likelihoods or likelihood ratios; we need exact conditional probabilities."

My comment about this statement is that it depends what likelihood ratios you are prepared to define.

Pure Linear Structure

Consider the following segment of the network of Figure 2–8. Take the three (binary) nodes, H, T, and F:

- H: Y drives X's car regularly;

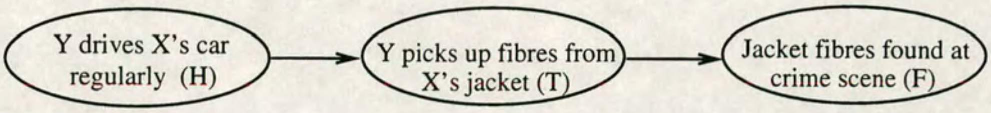

Y drives X's car regularly (H)
Y picks up fibres from X's jacket (T)
Jacket fibres found at crime scene (F)

Figure 2–14: *Segment of forensic science example—pure linear.*

- T: Y picks up fibres from X's jacket; and
- F: Fibres from a jacket similar to the one found in the possession of X are found at the crime scene.

These three nodes in isolation have a pure linear structure, as displayed in Figure 2–14. Usually the four probabilities $\Pr(T = t \,|\, H = h)$, $\Pr(t \,|\, \overline{h})$, $\Pr(f \,|\, t)$ and $\Pr(f \,|\, \overline{t})$ would need to be defined.

Suppose the information given is in the form of likelihood ratios p_1, q_1, p_2 and q_2 such that

$$p_1 = \frac{\Pr(t \,|\, h)}{\Pr(t \,|\, \overline{h})}, \qquad q_1 = \frac{\Pr(\overline{t} \,|\, h)}{\Pr(\overline{t} \,|\, \overline{h})}, \tag{2.1}$$

$$p_2 = \frac{\Pr(f \,|\, t)}{\Pr(f \,|\, \overline{t})} \qquad \text{and} \qquad q_2 = \frac{\Pr(\overline{f} \,|\, t)}{\Pr(\overline{f} \,|\, \overline{t})}. \tag{2.2}$$

From (2.1) it is straightforward to show that

$$\Pr(t \,|\, h) = \frac{p_1(q_1 - 1)}{(q_1 - p_1)} \qquad \text{and} \qquad \Pr(t \,|\, \overline{h}) = \frac{(q_1 - 1)}{(q_1 - p_1)}, \tag{2.3}$$

so that either $p_1 < 1 < q_1$ or $p_1 > 1 > q_1$ in order to satisfy the axioms of probability. The case $p_1 = q_1 = 1$ from (2.1) corresponds to independence of H and T, and need not concern us.

Similarly, from (2.2) it can be shown that

$$\Pr(f \,|\, t) = \frac{p_2(q_2 - 1)}{(q_2 - p_2)} \qquad \text{and} \qquad \Pr(f \,|\, \overline{t}) = \frac{(q_2 - 1)}{(q_2 - p_2)}. \tag{2.4}$$

The likelihood ratio $\Pr(f \,|\, h) / \Pr(f \,|\, \overline{h})$ is of interest here. Now,

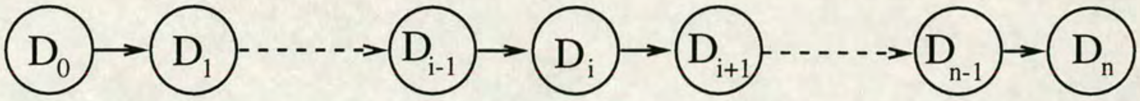


Figure 2–15: *Pure linear structure for* $(n+1)$ *events.*

$$\begin{aligned}
\Pr(f \mid h) &= \Pr(t \mid h)\Pr(f \mid t, h) + \Pr(\bar{t} \mid h)\Pr(f \mid \bar{t}, h) \\
&= \Pr(t \mid h)\Pr(f \mid t) + \Pr(\bar{t} \mid h)\Pr(f \mid \bar{t}) \qquad \textit{(conditional independence)} \\
&= \frac{p_1 p_2 (q_1 - 1)(q_2 - 1) + q_1 (q_2 - 1)(1 - p_1)}{(q_1 - p_1)(q_2 - p_2)},
\end{aligned}$$

and similarly

$$\Pr(f \mid \overline{h}) = \frac{p_2 (q_1 - 1)(q_2 - 1) + (q_2 - 1)(1 - p_1)}{(q_1 - p_1)(q_2 - p_2)},$$

so that the likelihood ratio is

$$\frac{\Pr(f \mid h)}{\Pr(f \mid \overline{h})} = \frac{p_1 p_2 (q_1 - 1)(q_2 - 1) + q_1 (q_2 - 1)(1 - p_1)}{p_2 (q_1 - 1)(q_2 - 1) + (q_2 - 1)(1 - p_1)}.$$

Having obtained a formula for two-stage pure linear inference, an extension to n-stage (involving $(n+1)$ events) is desired.

Consider the general case of $(n+1)$ nodes in pure linear form as illustrated by Figure 2–15. In order to compare results with the formula of Figure 2–13, the node names are the same, except for D^* in Figure 2–12. The name D_n is more general than D^*, which refers specifically to a report on the next node in the "chain". The usual notation (for this thesis) will be used, such that D_i will refer to the node/variable name, with states $D_i = d_i$ or $D_i = \overline{d_i}$, until a direct comparison with Martin's (1980) formula is required.

Define p_i, q_i, $i = 1, 2, \ldots, n$, such that

$$p_i = \frac{\Pr(d_i \mid d_{i-1})}{\Pr(d_i \mid \overline{d_{i-1}})} \qquad \text{and} \qquad q_i = \frac{\Pr(\overline{d_i} \mid d_{i-1})}{\Pr(\overline{d_i} \mid \overline{d_{i-1}})},$$

where either $p_i < 1 < q_i$ or $p_i > 1 > q_i$ for each i.

From this definition, and in a similar manner to the derivation of (2.3) and (2.4),

$$\Pr(d_i \mid d_{i-1}) = \frac{p_i(q_i-1)}{(q_i-p_i)}, \qquad \Pr(d_i \mid \overline{d_{i-1}}) = \frac{(q_i-1)}{(q_i-p_i)}, \tag{2.5}$$

$$\Pr(\overline{d_i} \mid d_{i-1}) = \frac{q_i(1-p_i)}{(q_i-p_i)} \quad \text{and} \quad \Pr(\overline{d_i} \mid \overline{d_{i-1}}) = \frac{(1-p_i)}{(q_i-p_i)}. \tag{2.6}$$

Now introduce new notation such that

$$d_i(1) \equiv d_i \qquad \text{and} \qquad d_i(0) \equiv \overline{d_i}, \qquad \text{for } i = 0, 1, 2, \ldots, n, \tag{2.7}$$

so the expressions at (2.6) can be written generally (for $i = 1, 2, \ldots, n$) as

$$\Pr(d_i(j_i) \mid d_{i-1}(j_{i-1})) = p_i^{j_i j_{i-1}} q_i^{(1-j_i)j_{i-1}} (1-p_i)^{1-j_i} (q_i-1)^{j_i} (q_i-p_i)^{-1} \tag{2.8}$$

where $j_i, j_{i-1} = 0, 1$.

We require a general formula for

$$\frac{\Pr(d_n \mid d_0)}{\Pr(d_n \mid \overline{d_0})} = \frac{\Pr(d_n(1) \mid d_0(1))}{\Pr(d_n(1) \mid d_0(0))}$$

in order to compare with the case where $n = 4$ as in Figure 2–13. Note that the method of Martin (1980) does not enable such a likelihood ratio to be expressed compactly as a formula for the general n-stage pure linear structure, although the pattern is clear.

For any $i = 1, 2, \ldots, n$,

$$\begin{aligned}
\Pr(d_i \mid d_0) &= \Pr(d_i(1) \mid d_0(1)) \\
&= \Pr(d_i(1) \mid d_{i-1}(1), d_0(1)) \Pr(d_{i-1}(1) \mid d_0(1)) + \\
&\qquad \Pr(d_i(1) \mid d_{i-1}(0), d_0(1)) \Pr(d_{i-1}(0) \mid d_0(1)) \\
&= \Pr(d_i(1) \mid d_{i-1}(1)) \Pr(d_{i-1}(1) \mid d_0(1)) + \\
&\qquad \Pr(d_i(1) \mid d_{i-1}(0)) \Pr(d_{i-1}(0) \mid d_0(1)) \\
&\qquad\qquad \textit{(by conditional independence)} \\
&= \sum_{j_{i-1}=0}^{1} \Pr(d_i(1) \mid d_{i-1}(j_i)) \Pr(d_{i-1}(j_i) \mid d_0(1)).
\end{aligned} \tag{2.9}$$

By analogy with the two-stage case, and by repeated application of expression (2.9), $\Pr(d_n(1) \mid d_0(1))$ can be calculated:

$$\begin{aligned}
\Pr(d_n(1) \mid d_0(1)) &= \sum_{j_{n-1}=0}^{1} \Pr(d_n(1) \mid d_{n-1}(j_n)) \Pr(d_{n-1}(j_n) \mid d_0(1)) \\
&= \sum_{j_{n-1}=0}^{1} \Bigg\{ \Pr(d_n(1) \mid d_{n-1}(j_n)) \times \qquad \textit{(by (2.9))} \\
&\qquad \sum_{j_{n-2}=0}^{1} \Pr(d_{n-1}(j_{n-1}) \mid d_{n-2}(j_{n-2})) \Pr(d_{n-2}(j_{n-2}) \mid d_0(1)) \Bigg\} \\
&= \sum_{j_{n-1}=0}^{1} \sum_{j_{n-2}=0}^{1} \Bigg\{ \Pr(d_n(1) \mid d_{n-1}(j_n)) \times \\
&\qquad \Pr(d_{n-1}(j_{n-1}) \mid d_{n-2}(j_{n-2})) \Pr(d_{n-2}(j_{n-2}) \mid d_0(1)) \Bigg\} \\
&= \sum_{j_{n-1}=0}^{1} \sum_{j_{n-2}=0}^{1} \Bigg\{ \Pr(d_n(1) \mid d_{n-1}(j_n)) \Pr(d_{n-1}(j_{n-1}) \mid d_{n-2}(j_{n-2})) \times \\
&\qquad \sum_{j_{n-3}=0}^{1} \Pr(d_{n-2}(j_{n-2}) \mid d_{n-3}(j_{n-3})) \Pr(d_{n-3}(j_{n-3}) \mid d_0(1)) \Bigg\}
\end{aligned}$$

and so on until

$$\begin{aligned}
&\Pr(d_n(1) \mid d_0(1)) = \\
&\quad \sum_{j_{n-1}=0}^{1} \sum_{j_{n-2}=0}^{1} \cdots \sum_{j_1=0}^{1} \Bigg\{ \Pr(d_n(1) \mid d_{n-1}(j_n)) \Pr(d_{n-1}(j_{n-1}) \mid d_{n-2}(j_{n-2})) \times \cdots \\
&\qquad\qquad \cdots \times \Pr(d_1(j_1) \mid d_0(1)) \Bigg\} \\
&= \sum_{j_{n-1}=0}^{1} \sum_{j_{n-2}=0}^{1} \cdots \sum_{j_1=0}^{1} \prod_{i=1}^{n} \Pr(d_i(j_i) \mid d_{i-1}(j_{i-1})) \qquad \text{where } j_0 = 1,\ j_n = 1.
\end{aligned}$$

Note that from (2.8) the term $(q_i - p_i)^{-1}$ occurs for each $\Pr(d_i(j_i) \mid d_{i-1}(j_{i-1}))$, and does not depend on j_i.

Finally, define the function Z where

$$\begin{aligned}
Z(i, j_0, j_n) &= \Pr(d_i(j_i) \mid d_{i-1}(j_{i-1})) \times (q_i - p_i) \\
&= p_i^{j_i j_{i-1}} q_i^{(1-j_i) j_{i-1}} (1 - p_i)^{1-j_i} (q_i - 1)^{j_i}
\end{aligned}$$

(by (2.8)) for $i = 1, 2, \ldots, n$; j_0 and j_n are included in the inputs to function Z since they are not summed out, and ultimately cause the difference between, say,

$\Pr(d_n(1)|d_0(1))$ and $\Pr(d_n(1)\,|\,d_0(0))$. Hence we have

$$\Pr(d_n(1)\,|\,d_0(1)) = \sum_{j_{n-1}=0}^{1}\sum_{j_{n-2}=0}^{1}\cdots\sum_{j_1=0}^{1}\prod_{i=1}^{n}\left\{Z(i,1,1)\times(q_i-p_i)^{-1}\right\}.$$

For $\Pr(d_n(1)\,|\,d_0(0))$, the working will be the same as above with $\Pr(d_1(j_1)\,|\,d_0(0))$ replacing $\Pr(d_1(j_1)\,|\,d_0(1))$. Thus

$$\Pr(d_n(1)\,|\,d_0(0)) = \sum_{j_{n-1}=0}^{1}\sum_{j_{n-2}=0}^{1}\cdots\sum_{j_1=0}^{1}\prod_{i=1}^{n}\left\{Z(i,1,0)\times(q_i-p_i)^{-1}\right\},$$

so that the $(q_i-p_i)^{-1}$ terms can be cancelled, giving the required likelihood ratio

$$\frac{\Pr(d_n\,|\,d_0)}{\Pr(d_n\,|\,\overline{d_0})} = \frac{\displaystyle\sum_{j_{n-1}=0}^{1}\sum_{j_{n-2}=0}^{1}\cdots\sum_{j_1=0}^{1}\prod_{i=1}^{n}Z(i,1,1)}{\displaystyle\sum_{j_{n-1}=0}^{1}\sum_{j_{n-2}=0}^{1}\cdots\sum_{j_1=0}^{1}\prod_{i=1}^{n}Z(i,1,0)}. \tag{2.10}$$

When $n=4$, as in Figure 2–13, and using Martin's (1980) notation of capital letters for states plus letting $D^* = D_4 = d_4$, expression (2.10) becomes

$$\frac{\Pr(D^*\,|\,D_0)}{\Pr(D^*\,|\,\overline{D_0})} = \frac{\displaystyle\sum_{j_3=0}^{1}\sum_{j_2=0}^{1}\sum_{j_1=0}^{1}\prod_{i=1}^{4}Z(i,1,1)}{\displaystyle\sum_{j_3=0}^{1}\sum_{j_2=0}^{1}\sum_{j_1=0}^{1}\prod_{i=1}^{4}Z(i,1,0)}. \tag{2.11}$$

Clearly expression (2.11) is a far more compact formula than that in Figure 2–13, and involves only one form of input, i.e. likelihood ratios. Martin's expression needs a mixture of likelihood ratios and conditional probabilities.

Bifurcation

The word *bifurcation* is used here to describe the structure in a network representing the case where one variable is conditional on the values of two other variables. Consider three events taken from the forensic science example—A, B and F, as illustrated in Figure 2–16. The nodes represent:

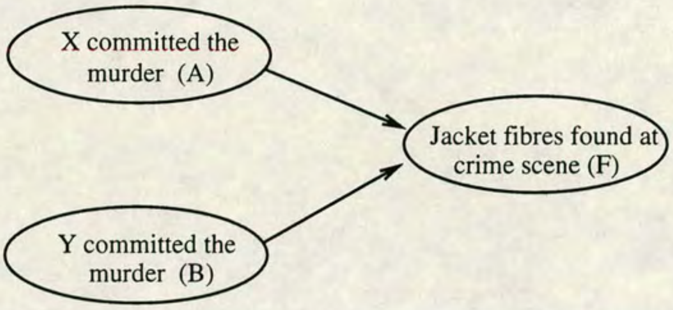

X committed the murder (A)
Jacket fibres found at crime scene (F)
Y committed the murder (B)

Figure 2–16: *Segment of forensic science example—bifurcation.*

- A: X committed the murder;
- B: Y committed the murder; and
- F: Fibres from a jacket similar to the one found in the possession of X are found at the crime scene.

Given this subnetwork, the following likelihood ratios can be defined:

$$p_1 = \frac{\Pr(f \mid a, b)}{\Pr(f \mid \overline{a}, b)}, \qquad q_1 = \frac{\Pr(\overline{f} \mid a, b)}{\Pr(\overline{f} \mid \overline{a}, b)}, \tag{2.12}$$

$$p_2 = \frac{\Pr(f \mid \overline{a}, \overline{b})}{\Pr(f \mid a, \overline{b})} \quad \text{and} \quad q_2 = \frac{\Pr(\overline{f} \mid \overline{a}, \overline{b})}{\Pr(\overline{f} \mid a, \overline{b})}. \tag{2.13}$$

The interpretation of these likelihood ratios is perhaps less clear than those for the pure linear case; consider for example p_1. This can be considered as the answer to the question "given that Y was involved in the murder, how much more likely is it that fibres from X's jacket were found at the scene of the crime if X was involved in the murder than if X was *not* involved?"

Ratios of this kind may be easier quantities to define than conditional probabilities such as those arising from the question "what is the probability of fibres from X's jacket being found at the scene of the crime given that Y was involved in the murder while X was not?"

These definitions allow calculation of the conditional probabilities:

$$\Pr(f \mid a, b) = \frac{p_1(q_1 - 1)}{(q_1 - p_1)}, \qquad \Pr(f \mid a, \overline{b}) = \frac{(q_2 - 1)}{(q_2 - p_2)},$$

$$\Pr(f \mid \overline{a}, b) = \frac{(q_1 - 1)}{(q_1 - p_1)} \quad \text{and} \quad \Pr(f \mid \overline{a}, \overline{b}) = \frac{p_2(q_2 - 1)}{(q_2 - p_2)}.$$

From these, other likelihood ratios follow, such as

$$\frac{\Pr(f \mid a, b)}{\Pr(f \mid a, \overline{b})} = \frac{p_1(q_1 - 1)(q_2 - p_2)}{(q_1 - p_1)(q_2 - 1)}.$$

As an example suppose that an expert has defined the likelihood ratios required above as $p_1 = 12$, $q_1 = 3/10$, $p_2 = 1/7$ and $q_2 = 3$. This gives the conditional probabilities

$$\Pr(f \mid a, b) = 0.718, \qquad \Pr(f \mid a, \overline{b}) = 0.700,$$

$$\Pr(f \mid \overline{a}, b) = 0.060, \qquad \Pr(f \mid \overline{a}, \overline{b}) = 0.100;$$

other likelihood ratios can be calculated directly from the ratios:

$$\frac{\Pr(f \mid a, b)}{\Pr(f \mid a, \overline{b})} = \frac{12 \times (0.3 - 1) \times (3 - \frac{1}{7})}{(0.3 - 12) \times (3 - 1)} = \frac{-24}{-23.4} = 1.026.$$

Extensions to Bifurcation

The results for bifurcation can be extended to the case where an event is conditional on ν other events, as illustrated by Figure 2–17.

Note that from this diagram, it would appear that any pair of variables A_i and A_j are conditionally independent given B. This is true if Figure 2–17 represents a *complete* influence diagram for a model; however, taking it only as part of a larger network, A_i and A_j cannot generally be assumed conditionally independent given B. In fact, there may even be an arc connecting A_i and A_j directly, indicating they are dependent on each other.

Given that all variables are binary, it is clear that there will be 2^ν different conditional probabilities of the form $\Pr(b \mid a_1(j_1), a_2(j_2), \ldots, a_\nu(j_\nu))$, where $j_i =$

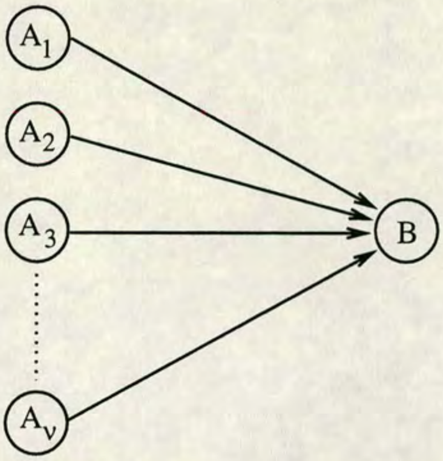


Figure 2–17: *One event conditioned on ν other events.*

0, 1 for $i = 1, 2, \ldots, \nu$, and the $a_i(j_i)$ notation is used for a_i and $\overline{a_i}$ as at (2.7). Hence 2^ν likelihood ratios will need to be defined—$2^{\nu-1}$ pairs of "p"s and "q"s, similar to those at (2.12) and (2.13).

Let $\mathcal{A}$ be the set of all permutations of one from each pair $(a_i(0), a_i(1))$ for $i = 1, 2, \ldots, \nu$. This set will have 2^ν elements, denoted α_k for $k = 1, 2, \ldots, 2^\nu$. Consider two elements of $\mathcal{A}$, α_l and α_m say, that are used to define p_s and q_s, i.e.

$$p_s = \frac{\Pr(b \,|\, \alpha_l)}{\Pr(b \,|\, \alpha_m)} \qquad \text{and} \qquad q_s = \frac{\Pr(\overline{b} \,|\, \alpha_l)}{\Pr(\overline{b} \,|\, \alpha_m)}.$$

In the same way as for bifurcation, it can be shown that

$$\Pr(b \,|\, \alpha_l) = \frac{p_s(q_s - 1)}{(q_s - p_s)} \qquad \text{and} \qquad \Pr(b \,|\, \alpha_m) = \frac{(q_s - 1)}{(q_s - p_s)}.$$

Given further elements of $\mathcal{A}$ such as α_x and α_y with likelihood ratios

$$p_r = \frac{\Pr(b \,|\, \alpha_x)}{\Pr(b \,|\, \alpha_y)} \qquad \text{and} \qquad q_r = \frac{\Pr(\overline{b} \,|\, \alpha_x)}{\Pr(\overline{b} \,|\, \alpha_y)},$$

similar working enables calculation of, for example,

$$\frac{\Pr(b \,|\, \alpha_l)}{\Pr(b \,|\, \alpha_x)} = \frac{p_s(q_s - 1)(q_r - p_r)}{p_r(q_s - p_s)(q_r - 1)}.$$

It is worth at this stage making some observations on the definition of likelihood ratios needed as input here. Note that there must be $2^{\nu-1}$ definitions of p_i

and corresponding q_i. These likelihood ratios need two conditional probabilities each, and these pairs can be ordered in many ways. The object will be to arrange the pairs of conditional probabilities so that the evaluation of the likelihood ratios (often subjective) can be made as clear and as simple as possible. The ratios can be defined so that for each one, only one of the conditioning variables changes between the numerator and the denominator. In other words, take α_l and α_m as above, for example; the state indicators j_i at each $a_i(j_i)$ will be the same for α_l as for α_m for all i except one.

Furthermore, given a structure such as that in Figure 2–17, it can be arranged that the "one different" conditioning variable is the same *for each likelihood ratio*, for any ν. This can be shown by a simple proof by induction:

1. Take the set $\mathcal{A}$ as above, with $\nu = 1$. Thus $\mathcal{A}$ has $2^1 = 2$ elements, $\alpha_1 = a_1(1)$ and $\alpha_2 = a_1(0)$. The likelihood ratio p equals $\Pr(b \mid a_1(1)) / \Pr(b \mid a_1(0))$, and so the required arrangement can be found for $\nu = 1$.

2. Assume a suitable arrangement holds for ν conditioning variables. There exist 2^ν permutations α_k in $\mathcal{A}$, and $2^{\nu-1}$ likelihood ratios $p_s = \Pr(b \mid \alpha_{k_1}) / \Pr(b \mid \alpha_{k_2})$. Now consider the addition of a new conditioning variable, $A_{\nu+1}$. Let $\mathcal{A}^*$ be the set of all permutations of one from each pair $(a_i(0), a_i(1))$ for $i = 1, 2, \ldots, \nu+1$, so that $\mathcal{A}^*$ has $2^{\nu+1}$ elements α^*_κ. For each of the 2^ν required likelihood ratios $p^*_s = \Pr(b \mid \alpha^*_{\kappa_1}) / \Pr(b \mid \alpha^*_{\kappa_2})$, take a separate $\alpha_k \in \mathcal{A}$, and define $\alpha^*_{\kappa_1} = \{\alpha_k, a_{\nu+1}(1)\}$ and $\alpha^*_{\kappa_2} = \{\alpha_k, a_{\nu+1}(0)\}$. Thus each ratio has only *one* conditioning variable differing in state between top and bottom, and the α^*_κ clearly are all the necessary permutations.

3. The required arrangement is thus possible for any $\nu = 1, 2, \ldots$.

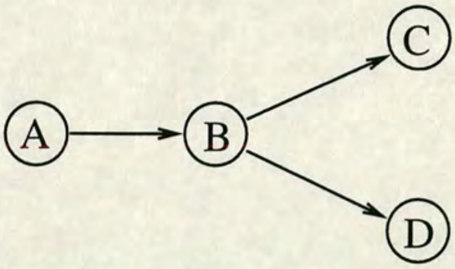


Figure 2–18: *Analyse this structure as two pure linear cases.*

General Structure

The above methods for dealing with likelihood ratios can be combined for more general structures. Some examples are given here.

Consider the structure of the subnetwork in Figure 2–18. This can be treated as two separate pure linear structures; $A \longrightarrow B \longrightarrow C$ and $A \longrightarrow B \longrightarrow D$. Likelihood ratios such as $\dfrac{\Pr(d\,|\,a)}{\Pr(d\,|\,\overline{a})}$ can thus be calculated using the methods described above.

The structure in Figure 2–19, taken as a subnetwork of the forensic science example, gives rise to defined likelihood ratios

$$p_a = \frac{\Pr(a\,|\,r)}{\Pr(a\,|\,\overline{r})}, \quad q_a = \frac{\Pr(\overline{a}\,|\,r)}{\Pr(\overline{a}\,|\,\overline{r})}, \quad p_b = \frac{\Pr(b\,|\,r)}{\Pr(b\,|\,\overline{r})}, \quad q_b = \frac{\Pr(\overline{b}\,|\,r)}{\Pr(\overline{b}\,|\,\overline{r})},$$

$$p_1 = \frac{\Pr(f\,|\,a,b)}{\Pr(f\,|\,a,\overline{b})}, \quad q_1 = \frac{\Pr(\overline{f}\,|\,a,b)}{\Pr(\overline{f}\,|\,a,\overline{b})}, \quad p_2 = \frac{\Pr(f\,|\,\overline{a},b)}{\Pr(f\,|\,\overline{a},\overline{b})} \quad \text{and} \quad q_2 = \frac{\Pr(\overline{f}\,|\,\overline{a},b)}{\Pr(\overline{f}\,|\,\overline{a},\overline{b})}.$$

The idea of analysing more general structures, such as that of Figure 2–19, is to build up from the basic forms, i.e. pure linear, bifurcation and so on. Here, consider two pure linear segments $R \longrightarrow A$ and $R \longrightarrow B$ and the bifurcation between A, B and F. The ratio of interest here is $\frac{\Pr(f\,|\,r)}{\Pr(f\,|\,\overline{r})}$; this can be written in terms of likelihood ratios,—take $\Pr(f\,|\,r)$ for example:

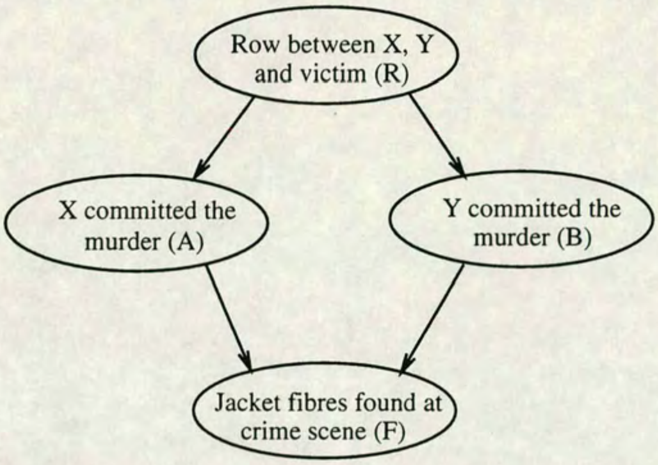

Row between X, Y and victim (R)
X committed the murder (A)
Y committed the murder (B)
Jacket fibres found at crime scene (F)

Figure 2–19: *Subnetwork of forensic science example.*

$$\begin{aligned}\Pr(f\,|\,r) &= \{\Pr(a\,|\,r)\times\Pr(f\,|\,a,r)\}+\{\Pr(\overline{a}\,|\,r)\times\Pr(f\,|\,\overline{a},r)\}\\ &= \{\Pr(a\,|\,r)\times\Pr(b\,|\,r)\times\Pr(f\,|\,a,b)\}+\{\Pr(a\,|\,r)\times\Pr(\overline{b}\,|\,r)\times\Pr(f\,|\,a,\overline{b})\}+\\ &\qquad\{\Pr(\overline{a}\,|\,r)\times\Pr(b\,|\,r)\times\Pr(f\,|\,\overline{a},b)\}+\{\Pr(\overline{a}\,|\,r)\times\Pr(\overline{b}\,|\,r)\times\Pr(f\,|\,\overline{a},\overline{b})\},\end{aligned}$$

where conditional independence ensures $\Pr(f\,|\,a,b,r) = \Pr(f\,|\,a,b)$ and so on. Each term in the last expression above follows from earlier results, so that eventually

$$\begin{aligned}\Pr(f\,|\,r) = \frac{1}{(q_a-p_a)(q_b-p_b)}\Bigg\{&\frac{p_a(q_a-1)(q_1-1)}{(q_1-p_1)}\left[p_bp_1(q_b-1)+q_b(1-p_b)\right]+\\ &\frac{q_a(1-p_a)(q_2-1)}{(q_2-p_2)}\left[p_bp_2(q_b-1)+q_b(1-p_b)\right]\Bigg\}\end{aligned}$$

and a similar expression is gained for $\Pr(f\,|\,\overline{r})$, enabling calculation of the desired likelihood ratio.

Naturally enough there is no simple expression for a likelihood ratio in the general case, but note that even the expression for $\Pr(f\,|\,r)$ above is a summation of simple fractions, and does not display the "expanding out" tendency of that in Figure 2–13. CASPRO (Martin, 1980) will correctly calculate a likelihood ratio, but if an expression is required, writing it in terms of other likelihood ratios seems a more feasible prospect.

There is also a programming benefit. The CASPRO algorithm requires an unknown number of nested loops. While this can be done without *too* much difficulty, the above procedure avoids this problem altogether.

Chapter 3

Structure Learning

3.1 Introduction

Elsewhere in this thesis, it is assumed that a graphical model has been defined fully enough for analysis to proceed. That is, given a model, the other chapters deal with the task of drawing inferences from it.

This chapter deals with problem of recovering a conditional independence graph directly from empirical observations. This is often the job of an expert designing an expert system (a process known as *knowledge acquisition*), and can be somewhat subjective. The automatic recovery of the structure of a model is termed *structure learning*.

Chow and Liu (1968) developed a method, described in section 3.2, of recovering a (skeleton) tree structure from a data set of discrete variables. The section also looks at using kernel methods to estimate probabilities rather than sample frequencies, and presents a comparison.

Section 3.3 studies structure learning with continuous variables, and compares a kernel method with the use of correlation coefficients.

George Rebane extended the Chow and Liu algorithm to *polytrees*, that is trees with arrows on some or all of the edges, in Rebane and Pearl (1987)—see §3.4.

Finally, recovering general network structure (to allow undirected loops) is more difficult, but some ideas are given in §3.5.

3.2 Recovering tree structures

Chow and Liu (1968)[1] showed that a joint distribution can be optimally approximated by a *tree-dependent* distribution obtained from the marginal probabilities of pairs of variables.

Definition 5 *A distribution* $\mathrm{Pr}^t(\mathbf{x})$ *is said to be* tree-dependent *relative to the tree t if it can be written as product of pair-wise conditional probability distributions*

$$\mathrm{Pr}^t(\mathbf{x}) = \prod_{i=1}^{n} \mathrm{Pr}(x_i \mid x_{\mathrm{pa}(i)}),$$

where $X_{\mathrm{pa}(i)}$ *is the parent of* X_i *in some orientation of the tree. The root* X_1*, for which* $\mathrm{Pr}(x_1 \mid x_0) = \mathrm{Pr}(x_1)$*, can be chosen arbitrarily.*

With a large enough sample, the probabilities $\mathrm{Pr}(x_i \mid x_j)$ can be estimated well from the data, so that $\mathrm{Pr}^t(\mathbf{x})$ can be obtained. The question of interest is, given an estimated distribution $\mathrm{Pr}(\mathbf{x})$, what is the tree-dependent distribution $\mathrm{Pr}^t(\mathbf{x})$ that "best" approximates $\mathrm{Pr}(\mathbf{x})$?

[1] Their work was in the area of character recognition, and was concerned with reducing storage space of joint distributions.

Chow and Liu (1968) chose the Kullback-Liebler cross entropy measure (Kullback and Liebler, 1951) as the "distance" between the two distributions, that is

$$D(\mathrm{Pr}, \mathrm{Pr}^t) = \sum_{\mathbf{x}} \mathrm{Pr}(\mathbf{x}) \log \frac{\mathrm{Pr}(\mathbf{x})}{\mathrm{Pr}^t(\mathbf{x})}.$$

This measure is nonnegative and becomes zero when the two distributions are the same.

Chow and Liu (1968) then show that the tree-dependent distribution $\mathrm{Pr}^t(\mathbf{x})$ is an optimal approximation to $\mathrm{Pr}(\mathbf{x})$ if and only if the tree t has maximum weight, where the weight on the branch (X_i, X_j) is defined by the mutual information measure

$$I(X_i, X_j) = \sum_{x_i, x_j} \mathrm{Pr}(x_i, x_j) \log \frac{\mathrm{Pr}(x_i, x_j)}{\mathrm{Pr}(x_i)\,\mathrm{Pr}(x_j)} \geq 0. \tag{3.1}$$

For an optimal Pr^t, the tree t is known as a *maximum weight spanning tree* (MWST).

The MWST for n variables can be found using the following algorithm, by Chow and Liu (1968):

1. From the observed data, estimate the distributions $\mathrm{Pr}(x_i, x_j)$ for all variable pairs.

2. Using the probabilities from step 1, compute all possible branch weights and order them by magnitude.

3. Assign the two branches with largest weights to the tree t.

4. Examine the next largest branch, and add it to the tree unless it forms a loop, in which case discard it and examine the next largest.

5. Repeat step 4 until $n - 1$ branches have been selected; the spanning tree has now been constructed.

It might arise that two branches have equal weights, while only one of them can be chosen (since taking both would create a loop). Ties like this can be broken arbitrarily, and the resulting tree will still be a MWST. It would be advisable however to consider all MWSTs resulting from such ties.

This algorithm uses only second order statistics, which are estimated easily from the observations, and the tree is recovered in $O(n)$ steps.

If the conditional probabilities along the branches of the tree are also estimated from the data, then the distribution and structure of the tree recovered by the MWST algorithm converges with probability 1 to the true underlying distribution, if that distribution is tree-dependent—see Chow and Wagner (1973).

Step 1 of the algorithm requires the estimation of probabilities from the data. A straightforward way to do this is to use the sample frequencies for each pair of variables, a method used by Chow and Liu (1968), Pearl (1988), Gammerman (1990) and Gammerman and Luo (1991). Another possibility is to use kernel methods. The subsection that follows explains how to use a kernel function to estimate the required probabilities, and §3.2.2 compares the results for kernels with those obtained using sample frequencies. The use of bivariate smoothing parameters is proposed in §3.2.3 and examined in §3.2.4.

3.2.1 Estimating Probabilities from Data with a Kernel Function

The process of structure learning involves the calculation of information measures as at (3.1). If the joint probability $\Pr(x_i, x_j)$ equals zero for some combination of values of X_i and X_j, then the contribution to the information measure by that combination will also be zero. If the data set is small, however, it may be that the actual probability of the combination is non-zero, but by chance it has not occurred yet. The use of sample frequencies will set the probability to zero; a method that takes into account the size of the data set might therefore be useful.

Kernel probability distribution estimation is one such method. For a general review of kernel methods see Silverman (1986).

Kernel-type estimators are defined as estimators of the form

$$f_m(x) = \frac{1}{mh_m}\sum_{i=1}^{m} K\left(\frac{x - X_i}{h_m}\right),$$

where $f_m(x)$ is the estimate of the true distribution $f(x)$, $X_1, X_2, \ldots, X_m$ is a random sample of size m from $f(x)$, K is a suitable density function, and h_m is the *smoothing parameter* (or *bandwidth*), with $h_m \to 0$ as $m \to \infty$.

Aitken (1979) defines the following kernel function for a discrete variable X_j with $c_j + 1$ unordered categories, $X_j = 0, 1, \ldots, c_j$:

$$K_j(u_j \mid x_j, \lambda_j) = \begin{cases} \lambda_j & \text{if } u_j = x_j \\ \frac{1-\lambda_j}{c_j} & \text{if } u_j \neq x_j \end{cases}$$

for $j = 1, 2, \ldots, p$, where p is the number of variables, u_j and x_j are particular values of variable X_j, and λ_j is the smoothing parameter for X_j.

In order for K_j to be a density function, it is necessary that

$$1 \geq \lambda_j \geq \frac{1}{c_j + 1} \qquad \forall j.$$

The function for an instantiation $\mathbf{u} = (u_1, u_2, \ldots, u_p)'$ of a vector of possible outcomes $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$, given a vector of smoothing parameters $\boldsymbol{\lambda}$, is

$$\begin{aligned} K(\mathbf{u} \mid \mathbf{x}, \boldsymbol{\lambda}) &= \prod_{j=1}^{p} K_j(u_j \mid x_j, \lambda_j) \\ &= \prod_{j=1}^{p} \left\{ \lambda_j^{1-\Delta(x_j,u_j)} \left(\frac{1-\lambda_j}{c_j}\right)^{\Delta(x_j,u_j)} \right\} \end{aligned}$$

where

$$\Delta(x_j, u_j) = \begin{cases} 0 & \text{if } x_j = u_j \\ 1 & \text{if } x_j \neq u_j \end{cases}.$$

Thus the probability of $\mathbf{u}$ given a data set D is

$$\Pr(\mathbf{u} \mid D) = \frac{1}{m}\sum_{i=1}^{m} K(\mathbf{u} \mid D, \boldsymbol{\lambda}),$$

and the probability of a particular variable X_j taking a specified value x_j can be obtained by summing all $\Pr(\mathbf{u} \mid D)$ for which $u_j = x_j$.

For a binary-valued variable X_j, the kernel function becomes

$$K_j(u_j \mid x_j, \lambda_j) = \begin{cases} \lambda_j & \text{if } u_j = x_j \\ 1 - \lambda_j & \text{if } u_j = \overline{x_j} \end{cases} \tag{3.2}$$

(where x_j is one of 0 or 1 and $\overline{x_j}$ is the opposite) so that in the case where all the variables are binary-valued (see Aitchison and Aitken (1976) for example),

$$K(\mathbf{u} \mid \mathbf{x}, \boldsymbol{\lambda}) = \prod_{j=1}^{p} \left\{ \lambda_j^{1-\Delta(x_j,u_j)} (1 - \lambda_j)^{\Delta(x_j,u_j)} \right\} \tag{3.3}$$

with

$$1 \geq \lambda_j \geq \frac{1}{2} \qquad \forall j.$$

For the rest of this chapter, all discrete variables will be assumed binary-valued.

As stated previously, the motivation for using kernel estimates of probability here is to try and lessen the effect of a combination of values of two variables not occurring in the data set.

It is the case that the kernel estimate of the probability of a particular value of a *single* variable turns out to be non-zero (unless the smoothing parameter equals 1) even when that value does not occur in the data, unlike with sample frequency estimates. However, this non-zero probability actually *cancels out* in the information measure calculation at (3.1), causing $I()$ to be zero (which is intuitively obvious, in any case), as will now be demonstrated.

Consider a data set D, with p variables and observations $x_{i1}, x_{i2}, \ldots, x_{ip}$, $i = 1, 2, \ldots, m$, where w.l.o.g. x_{i1} is always equal to 1 and never 0. Define an instantiation $\mathbf{u}$ which has $u_1 = 0$, but is otherwise arbitrary. Thus from (3.2)

$$K_1(u_1 \mid x_{i1}, \lambda_1) = (1 - \lambda_1) \qquad \forall i.$$

Hence by summing (3.3) over $\mathcal{V}_{-1}$, the state space of $X_2, X_3, \ldots, X_p$,

$$\Pr(X_1 = 0 \mid D, \boldsymbol{\lambda}) \;=\; \sum_{\mathcal{V}_{-1}} \frac{1}{m} \sum_{i=1}^{m} (1 - \lambda_1) \prod_{j=2}^{p} \lambda_j^{1-\Delta(x_{ij},u_j)} (1 - \lambda_j)^{\Delta(x_{ij},u_j)}$$

$$
\begin{aligned}
&= (1-\lambda_1)\sum_{\mathcal{V}_{-1}}\frac{1}{m}\sum_{i=1}^{m}\prod_{j=2}^{p}\lambda_j^{1-\Delta(x_{ij},u_j)}(1-\lambda_j)^{\Delta(x_{ij},u_j)} \\
&= (1-\lambda_1)\sum_{\mathcal{V}_{-1}}\Pr(X_2,X_3,\ldots,X_p\,|\,D,\boldsymbol{\lambda}) \\
&= (1-\lambda_1)
\end{aligned}
$$

since the total probability for all possible values of $X_2, X_3, \ldots, X_p$ must be 1. Note that $\Pr(X_1 = 1\,|\,D,\boldsymbol{\lambda}) = \lambda_1$ by a similar calculation.

Now suppose, again w.l.o.g., that it is desired to calculate the information measure $I(X_1, X_2)$. Then the probabilities $\Pr(X_1, X_2\,|\,D,\boldsymbol{\lambda})$ and $\Pr(X_2\,|\,D,\boldsymbol{\lambda})$ are also required. For $X_1 = 0$, with $\mathcal{V}_{-1,2}$ the state space of $X_3, X_4, \ldots, X_p$,

$$
\begin{aligned}
\Pr(X_1=0,X_2\,|\,D,\boldsymbol{\lambda}) &= \sum_{\mathcal{V}_{-1,2}}\frac{1}{m}\sum_{i=1}^{m}\Bigg\{(1-\lambda_1)\times\lambda_2^{1-\Delta(x_{i2},u_2)}(1-\lambda_2)^{\Delta(x_{i2},u_2)} \\
&\qquad\times\prod_{j=3}^{p}\lambda_j^{1-\Delta(x_{ij},u_j)}(1-\lambda_j)^{\Delta(x_{ij},u_j)}\Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
&= (1-\lambda_1)\sum_{\mathcal{V}_{-1,2}}\frac{1}{m}\Bigg\{\lambda_2\sum_{i_1=1}^{m_1}\prod_{j=3}^{p}\lambda_j^{1-\Delta(x_{i_1 j},u_j)}(1-\lambda_j)^{\Delta(x_{i_1 j},u_j)} \\
&\qquad+(1-\lambda_2)\sum_{i_2=1}^{m_2}\prod_{j=3}^{p}\lambda_j^{1-\Delta(x_{i_2 j},u_j)}(1-\lambda_j)^{\Delta(x_{i_2 j},u_j)}\Bigg\} \\
&= (1-\lambda_1)\frac{1}{m}\Bigg\{\lambda_2\sum_{\mathcal{V}_{-1,2}}\sum_{i_1=1}^{m_1}\prod_{j=3}^{p}\lambda_j^{1-\Delta(x_{i_1 j},u_j)}(1-\lambda_j)^{\Delta(x_{i_1 j},u_j)} \\
&\qquad+(1-\lambda_2)\sum_{\mathcal{V}_{-1,2}}\sum_{i_2=1}^{m_2}\prod_{j=3}^{p}\lambda_j^{1-\Delta(x_{i_2 j},u_j)}(1-\lambda_j)^{\Delta(x_{i_2 j},u_j)}\Bigg\} \\
&= (1-\lambda_1)\frac{1}{m}\Bigg\{\lambda_2\sum_{\mathcal{V}_{-1,2}}\Pr(X_3,X_4,\ldots,X_p\,|\,D,\boldsymbol{\lambda}) \\
&\qquad+(1-\lambda_2)\sum_{\mathcal{V}_{-1,2}}\Pr(X_3,X_4,\ldots,X_p\,|\,D,\boldsymbol{\lambda})\Bigg\} \\
&= (1-\lambda_1)\frac{1}{m}\left\{\lambda_2 m_1+(1-\lambda_2)m_2\right\} \qquad (3.4)
\end{aligned}
$$

where $u_2 = x_{i2}$ on m_1 occasions and $u_2 \neq x_{i2}$ on m_2 occasions, with $m_1 + m_2 = m$. Following similar calculations to the above, it can also be seen that

$$
\Pr(X_1=1,X_2\,|\,D,\boldsymbol{\lambda}) = \lambda_1\frac{1}{m}\{\lambda_2 m_1+(1-\lambda_2)m_2\}, \qquad \text{and}
$$

$$\Pr(X_2 \mid D, \boldsymbol{\lambda}) \;=\; \frac{1}{m}\{\lambda_2 m_1 + (1-\lambda_2)m_2\}.$$

So now these probabilities can be put into (3.1):

$$I(X_1, X_2) = \sum_{X_1, X_2} \Pr(X_1, X_2 \mid D, \boldsymbol{\lambda}) \log \frac{\Pr(X_1, X_2 \mid D, \boldsymbol{\lambda})}{\Pr(X_1 \mid D, \boldsymbol{\lambda}) \Pr(X_2 \mid D, \boldsymbol{\lambda})}$$

$$= \sum_{X_2} \left\{ \Pr(X_1 = 0, X_2 \mid D, \boldsymbol{\lambda}) \log \frac{\Pr(X_1 = 0, X_2 \mid D, \boldsymbol{\lambda})}{\Pr(X_1 = 0 \mid D, \boldsymbol{\lambda}) \Pr(X_2 \mid D, \boldsymbol{\lambda})} \right. \tag{3.5}$$
$$\left. + \Pr(X_1 = 1, X_2 \mid D, \boldsymbol{\lambda}) \log \frac{\Pr(X_1 = 1, X_2 \mid D, \boldsymbol{\lambda})}{\Pr(X_1 = 1 \mid D, \boldsymbol{\lambda}) \Pr(X_2 \mid D, \boldsymbol{\lambda})} \right\}.$$

The first log expression at (3.5) becomes

$$\log \frac{(1-\lambda_1)\frac{1}{m}\{\lambda_2 m_1 + (1-\lambda_2)m_2\}}{(1-\lambda_1) \times \frac{1}{m}\{\lambda_2 m_1 + (1-\lambda_2)m_2\}} = \log 1 = 0,$$

and the second also

$$\log \frac{\lambda_1 \frac{1}{m}\{\lambda_2 m_1 + (1-\lambda_2)m_2\}}{\lambda_1 \times \frac{1}{m}\{\lambda_2 m_1 + (1-\lambda_2)m_2\}} = \log 1 = 0,$$

so that $I(X_1, X_2) = 0$ as required.

It should be clear from the above that the reason the information measure is zero is because the $(1-\lambda_1)$ term can be taken out of all summations, and thus cancels in the log expressions. If, however, two variables (w.l.o.g. X_1 and X_2) both take the values 0 and 1 individually but the combination $X_1 = 0, X_2 = 0$ (for example) never occurs in the data set D, then the information measure $I(X_1, X_2)$ will not necessarily be zero. This is because the $(1-\lambda_1)$ term (nor indeed a $(1-\lambda_2)$ term) cannot be taken out of the summations, and will not therefore cancel in the log expression. A simple numerical counter-example would also show this.

The smoothing parameters are estimated using pseudo-maximum likelihood. Each parameter λ_j, $j = 1, 2, \ldots, p$ is estimated by finding

$$\max_{\lambda_j} \prod_{i=1}^{m} \Pr(x_{ij} \mid D \setminus \{x_{ij}\}, \lambda_j) \tag{3.6}$$

where D is the data set and x_{ij} is an element of the data set, observation i of variable j.

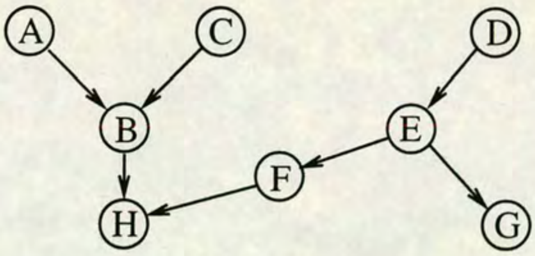


Figure 3–1: *Test structure for comparing sample frequencies and kernels approaches.*

3.2.2 Example

The performance of both the sample frequencies and the kernels approaches to structure learning will now be compared.

The Chow and Liu (1968) MWST algorithm applies for tree-dependent structures. Such structures are however very limited in their applications; only one variable is permitted to be a root. Section 3.4 will consider *polytrees*, trees that can contain more than one root variable (and hence allow a variable to have more than one parent). The test model in this section is a polytree, however; this is to allow the recovery of the MWST to be studied now, and the method of recovering the directionality of the edges to be studied in §3.4 with the same test example. That the MWST algorithm will recover the skeleton of a polytree is confirmed by Theorem 3 of Pearl and Dechter (1989), called Theorem 2 here and seen in §3.4.

The test model has its structure displayed in Figure 3–1, and its probability distributions shown in Table 3–1. The model is defined so that for a small number of simulated observations, it is likely that not all combinations of pairs of variables will occur. For example, $\Pr(\overline{a}, \overline{b}) = 0.9253$, and the remaining combinations of A and B have probabilities less than 0.03. Thus it is hoped that the kernels approach will perform better than the sample frequency approach, since it will not automatically assign the value zero to estimated probabilities of combinations that do not occur in the data.

The resulting skeleton trees (or parts of trees) from the structure learning

$\Pr(a) = 0.05$	$\Pr(d) = 0.25$
$\Pr(c) = 0.40$	$\Pr(g \mid e) = 0.80$
	$\Pr(g \mid \overline{e}) = 0.50$
$\Pr(b \mid a, c) = 0.85$	$\Pr(f \mid e) = 0.02$
$\Pr(b \mid a, \overline{c}) = 0.20$	$\Pr(f \mid \overline{e}) = 0.10$
$\Pr(b \mid \overline{a}, c) = 0.05$	
$\Pr(b \mid \overline{a}, \overline{c}) = 0.01$	$\Pr(h \mid b, f) = 0.04$
	$\Pr(h \mid b, \overline{f}) = 0.08$
$\Pr(e \mid d) = 0.95$	$\Pr(h \mid \overline{b}, f) = 0.25$
$\Pr(e \mid \overline{d}) = 0.85$	$\Pr(h \mid \overline{b}, \overline{f}) = 0.40$

Table 3–1: *Table of distributions for model used to compare sample frequencies and kernels approaches.*

process are shown in Figure 3–2. Clearly the results are very similar; both methods score the same number of correct links at each stage. Note that some of the graphs have less than 7 links—this is because a number of the information measures were zero. Even though the Chow and Liu algorithm suggests that ties should be broken at random, doing so when the variable pairs in question have zero-valued information measures seems inappropriate.

Both methods have only 3 correct links for 50 observations; in the test sample, variables A and B take the value 0 throughout, so it is not surprising that the information measures with those as one of the pair should all be zero.

The only difference of any kind occurs for the 100 observation set. While the two methods give the same correct links, they disagree on the choice of one incorrect link. Sample frequencies chose C–F, whereas kernels chose C–H. It might be argued that C is "closer" to H in the original tree, and hence that the kernels method has performed slightly better, but this is debatable. The combination $(A = a, G = \overline{g})$ did not not occur in the sample of size 100 (while all other pairwise combinations of A and G did), so it was hoped that while sample

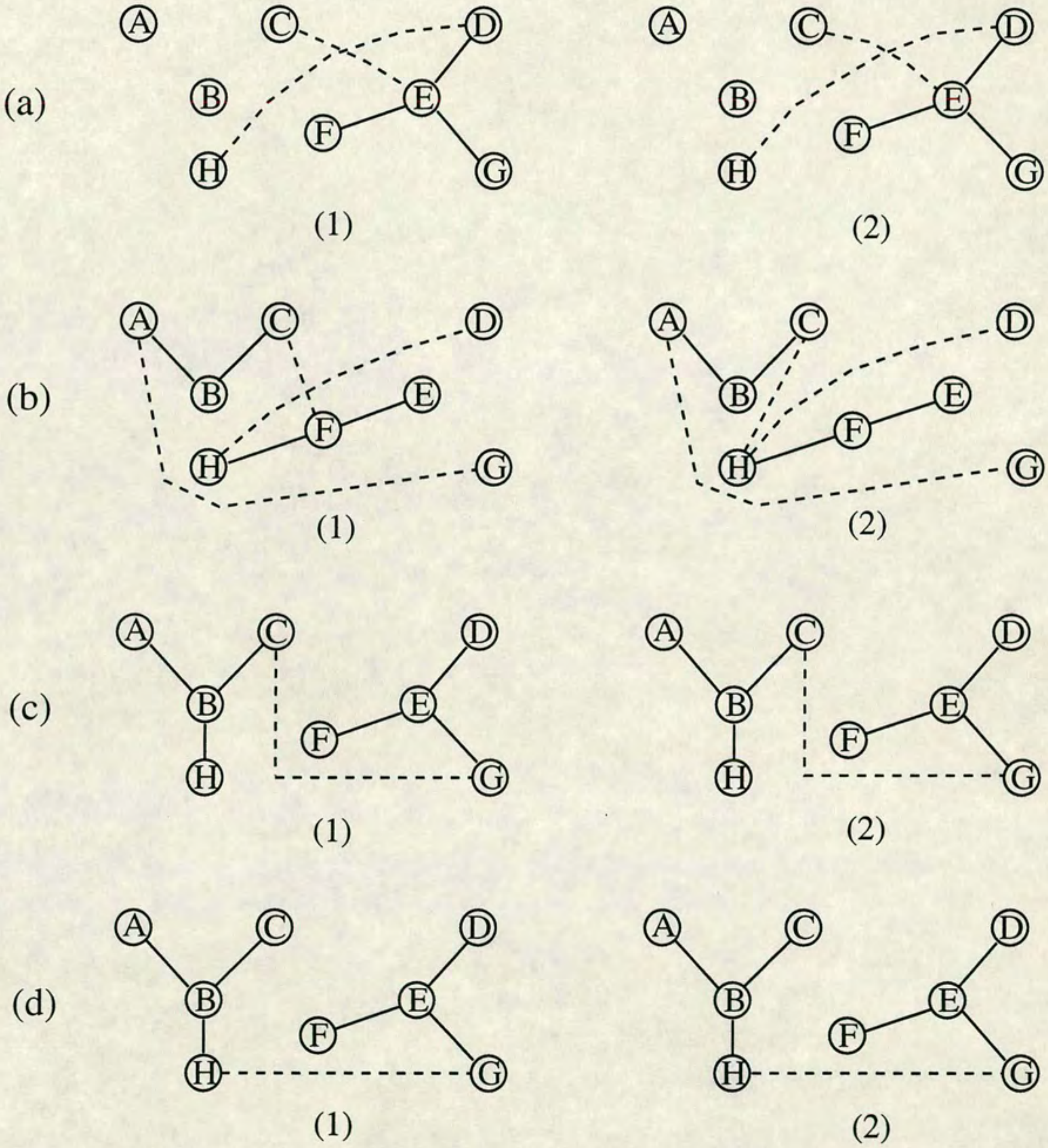


Figure 3–2: *The results of structure learning using (1) sample frequencies and (2) kernels. The numbers of observations were (a) 50, (b) 100, (c) 1000 and (d) 2000. The solid lines represent correct links, while the dotted lines are incorrect links. The graphs at (a) have less than 7 links shown; this is due to information measures of zero.*

frequencies included the erroneous A–G link, the kernels approach might not. Figure 3–2 shows this not to be the case.

Both methods returned exactly the same tree for the larger sample sizes. This is to be expected, since the estimated smoothing parameters tend to 1 as the sample size tends to infinity[2]. If the smoothing parameters are all equal to 1, then the kernels method will be exactly equivalent to the sample frequencies method.

In conclusion, the kernels method does not seem to have performed very much better than the sample frequencies method. The results shown here are very typical of other examples considered; the kernels approach never performed worse in practice, yet it didn't select correct links that the sample frequencies missed. Given the extra time and effort needed for the kernels (including estimating the smoothing parameters), my experiments seem to show that the method using sample frequencies is perfectly adequate.

3.2.3 Bivariate Smoothing Parameters

Since the Chow and Liu algorithm involves pairwise joint probabilities, it might be instructive to consider the use of bivariate smoothing parameters. The single variable marginal probabilities will be estimated as before, but the pairwise probabilities will be estimated using smoothing parameters based upon *both* of the variables in question.

Previously the joint probability of two variables X_1 and X_2 was estimated by[3]

[2] The expression to be maximised at (3.6) tends to zero as n tends to infinity for $\lambda \neq 1$.

[3] Derived using similar calculations to (3.4).

$$\begin{aligned}\Pr(X_1, X_2 \mid D, \lambda_1, \lambda_2) &= \frac{1}{m}\{\lambda_1\lambda_2 m_1 + \lambda_1(1-\lambda_2)m_2 \\ &\qquad + (1-\lambda_1)\lambda_2 m_3 + (1-\lambda_1)(1-\lambda_2)m_4\}\end{aligned}$$

where m_1 is the number of observations in D where both the particular X_1 and X_2 match x_{i1} and x_{i2} respectively; m_2 is the number where X_1 matches and X_2 doesn't; m_3 where X_2 matches and X_1 doesn't; and m_4 where neither match.

The smoothing parameters λ_1 and λ_2 in this pairwise probability calculation can be replaced by λ_{12}; this seems reasonable since a pairwise λ_{jk} can be estimated by

$$\overset{\max}{\lambda_{jk}} \prod_{i=1}^{m} \Pr(x_{ij}, x_{ik} \mid D \setminus \{x_{ij}, x_{ik}\}, \lambda_{jk}),$$

and this parameter is estimated by the observations on the pair of variables for which we are trying to find the joint probabilities.

3.2.4 Example

Applying the Chow and Liu algorithm to the example of section 3.2.2 gives the MWSTs in Figure 3–3.

It can be noted that the trees for 1000 and 2000 observations here are the same as those discovered before bivariate smoothing parameters were introduced. The tree produced for the 50 observation data set however is different from either the sample frequencies or univariate smoothing parameter trees. These methods (compared in section 3.2.2) each gave only (the same) five non-zero information measures, and hence only five links each. However, the use of bivariate smoothing parameters gave all but one of the information measures non-zero values[4]. Hence

[4] Unfortunately, since there is a link A–B, $I(A, B) = 0$; this is the case since $A = \overline{a}$ and $B = \overline{b}$ for the whole data set.

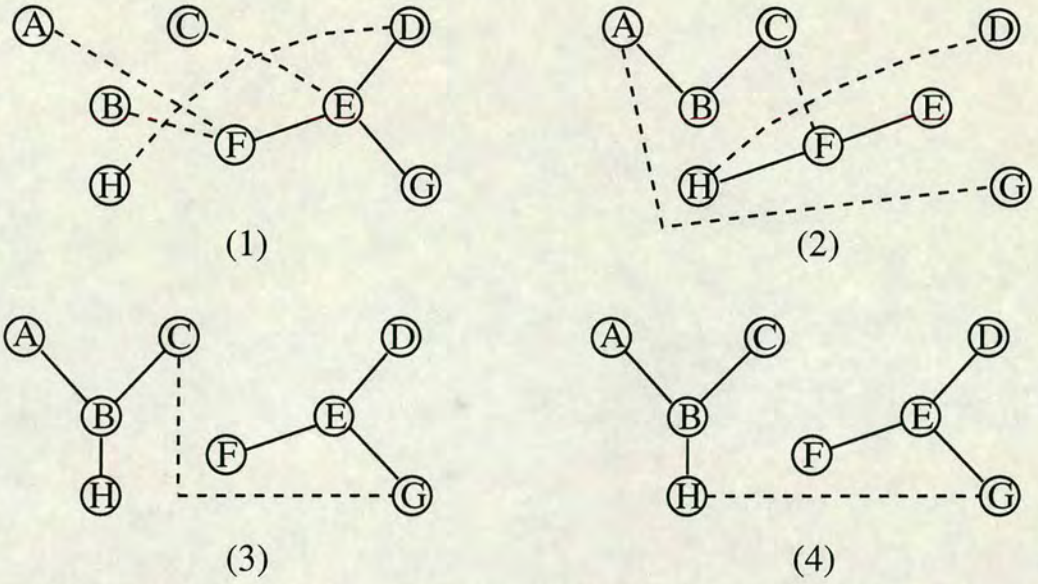


Figure 3–3: The results of structure learning using bivariate smoothing parameters for (1) 50, (2) 100, (3) 1000 and (4) 2000 observations.

the MWST here, at (1) in Figure 3–3, has a full seven links. The two extra links are however incorrect links.

The tree recovered for 100 observations is equivalent to the tree recovered by the sample frequencies method. There is nothing too significant in this; for other data sets, the bivariate smoothing parameter method more closely resembled the other kernel approach.

In §3.2.2 it was noted that the extra effort required for the univariate smoothing parameter kernel approach did not seem to be justified, since there was little or no evidence of an improvement in performance. For the bivariate smoothing parameter case, the same reasoning applies to an extent, especially if the sample size is not small. For a network with p variables, $\frac{1}{2}p(p-1)$ bivariate smoothing parameters must be estimated, along with the p univariate parameters. With large p, and large sample size m, the parameters can take a considerable amount of time to estimate—and even then they are likely to all be very close to 1, and hence the method will be nearly equivalent to the sample frequencies approach.

The only worthwhile point about the use of bivariate smoothing parameters seems to be the fact that it recovers many more non-zero information measures than the other methods. This results in a tree with more links, even though they have tended (with other data sets as well) to be incorrect joins. Thus this apparent benefit of using bivariate smoothing parameters should be viewed somewhat sceptically.

3.3 Continuous Variables

The Chow and Liu algorithm can be altered slightly to apply to a set of continuous variables. The discrete kernel function of the last section can be replaced by a choice of continuous kernels, and this will enable a comparison of the MWST algorithm with a similar routine using not information measures, but sample correlation coefficients.

The information measure for two continuous variables Y_1 and Y_2 is defined as

$$I(Y_1, Y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2) \log \frac{f(y_1, y_2)}{f(y_1)f(y_2)} dy_1 dy_2 \geq 0, \tag{3.7}$$

where the function f is the p.d.f. of Y_1 and Y_2.

The algorithm to find the MWST t with optimal tree-dependent distribution f^t for n continuous variables, analogous to the discrete version, is:

1. Compute all possible branch weights and order them by magnitude.

2. Assign the two branches with the largest weights to the tree t.

3. Examine the next largest branch, and add it to the tree unless it forms a loop, in which case discard it and examine the next largest.

4. Repeat step 3 until $n - 1$ branches have been selected; the spanning tree has now been constructed.

The calculation of information measures for continuous variables is more complicated than for discrete variables, since the summation at (3.1) becomes a double integration, which in practice must be done numerically.

At step 1 of the algorithm, the branch weights must be calculated using an estimate of f, based on a data set D with m observations. Let the conditional kernel $g(y_1, y_2 \,|\, D, \lambda_1, \lambda_2)$ be a kernel estimate of $f(y_1, y_2)$, and $g(y_i \,|\, D, \lambda_i)$ estimates for $f(y_i)$.

The function g is given the forms

$$\begin{aligned} g(y_1, y_2 \,|\, D, \lambda_1, \lambda_2) &= \frac{1}{m}\sum_{k=1}^{m} L(y_1 \,|\, D, \lambda_1) L(y_2 \,|\, D, \lambda_2), \qquad \text{and} \\ g(y_i \,|\, D, \lambda_i) &= \frac{1}{m}\sum_{k=1}^{m} L(y_i \,|\, D, \lambda_i) \qquad \text{for } i = 1, 2. \end{aligned}$$

For ease of reading, conditioning on D, λ_1 and λ_2 will be dropped from function g. There are a number of choices for the kernel function L (Silverman, 1986); consider initially the Normal kernel, so that

$$L(y_i \,|\, D, \lambda_i) = \frac{1}{\lambda_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2\lambda_i^2}\,(y_i - d_{ik})^2\right\} \qquad \text{for } k = 1, 2, \ldots, n, \;\; i = 1, 2,$$

where d_{ik} is the k-th observation of variable Y_i. The calculation of the expression in the log function at (3.7) can be written:

$$\frac{g(y_1, y_2)}{g(y_1)g(y_2)} = \frac{\frac{1}{m}\sum_{k=1}^{m} L(y_1 \,|\, D, \lambda_1) L(y_2 \,|\, D, \lambda_2)}{\left\{\frac{1}{m}\sum_{k=1}^{m} L(y_1 \,|\, D, \lambda_1)\right\}\left\{\frac{1}{m}\sum_{k=1}^{m} L(y_2 \,|\, D, \lambda_2)\right\}}$$

$$= \frac{m\sum_{k=1}^{m}\left\{\frac{1}{\lambda_1\sqrt{2\pi}}\exp\left(\frac{-1}{2\lambda_1^2}(y_1 - d_{1k})^2\right) \times \frac{1}{\lambda_2\sqrt{2\pi}}\exp\left(\frac{-1}{2\lambda_2^2}(y_2 - d_{2k})^2\right)\right\}}{\sum_{k=1}^{m}\left\{\frac{1}{\lambda_1\sqrt{2\pi}}\exp\left(\frac{-1}{2\lambda_1^2}(y_1 - d_{1k})^2\right)\right\}\sum_{k=1}^{m}\left\{\frac{1}{\lambda_2\sqrt{2\pi}}\exp\left(\frac{-1}{2\lambda_2^2}(y_2 - d_{2k})^2\right)\right\}}$$

$$= \frac{m\sum_{k=1}^{m}\exp\left(\frac{-1}{2\lambda_1^2}(y_1 - d_{1k})^2 - \frac{1}{2\lambda_2^2}(y_2 - d_{2k})^2\right)}{\sum_{k=1}^{m}\exp\left(\frac{-1}{2\lambda_1^2}(y_1 - d_{1k})^2\right)\sum_{k=1}^{m}\exp\left(\frac{-1}{2\lambda_2^2}(y_2 - d_{2k})^2\right)}.$$

The denominator above can be rewritten as

$$\sum_{k=1}^{m} \exp\left(\frac{-1}{2\lambda_1^2}(y_1 - d_{1k})^2 - \frac{1}{2\lambda_2^2}(y_2 - d_{2k})^2\right) + \sum_{p=1}^{m}\sum_{\substack{q=1\\p\neq q}}^{m} \exp\left(\frac{-1}{2\lambda_1^2}(y_1 - d_{1p})^2 - \frac{1}{2\lambda_2^2}(y_2 - d_{2q})^2\right).$$

Now let

$$\begin{aligned} a_k(y_1, y_2) &= \exp\left(\frac{-1}{2\lambda_1^2}(y_1 - d_{1k})^2 - \frac{1}{2\lambda_2^2}(y_2 - d_{2k})^2\right),\ k = 1, 2, \ldots, m, \quad \text{and} \\ a_{pq}(y_1, y_2) &= \exp\left(\frac{-1}{2\lambda_1^2}(y_1 - d_{1p})^2 - \frac{1}{2\lambda_2^2}(y_2 - d_{2q})^2\right),\ p, q = 1, 2, \ldots, m;\ p \neq q, \end{aligned}$$

so that

$$\begin{aligned} \frac{g(y_1, y_2)}{g(y_1)g(y_2)} &= \frac{m \sum_{k=1}^{m} a_k(y_1, y_2)}{\sum_{k=1}^{m} a_k(y_1, y_2) + \sum_{p=1}^{m}\sum_{\substack{q=1\\p\neq q}}^{m} a_{pq}(y_1, y_2)} \\ &= m \left(1 + \frac{\sum_{p=1}^{m}\sum_{\substack{q=1\\p\neq q}}^{m} a_{pq}(y_1, y_2)}{\sum_{k=1}^{m} a_k(y_1, y_2)}\right)^{-1}. \end{aligned}$$

Taking logs gives

$$\log \frac{g(y_1, y_2)}{g(y_1)g(y_2)} = \log m - \log\left\{1 + \frac{\sum_{p=1}^{m}\sum_{\substack{q=1\\p\neq q}}^{m} a_{pq}(y_1, y_2)}{\sum_{k=1}^{m} a_k(y_1, y_2)}\right\}.$$

Putting this expression into the definition of the information measure (3.7), we get

$$I(Y_1, Y_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1, y_2)\left[\log m - \log\left\{1 + \frac{\sum_{p=1}^{m}\sum_{\substack{q=1\\p\neq q}}^{m} a_{pq}(y_1, y_2)}{\sum_{k=1}^{m} a_k(y_1, y_2)}\right\}\right] dy_1 dy_2$$

$$= \log m - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1,y_2) - \log\left\{1 + \frac{\sum_{p=1}^{m}\sum_{\substack{q=1\\p\neq q}}^{m} a_{pq}(y_1,y_2)}{\sum_{k=1}^{m} a_k(y_1,y_2)}\right\} dy_1dy_2$$

$$= \log m - \frac{1}{2\pi m\lambda_1\lambda_2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sum_{k=1}^{m} a_k(y_1,y_2)\log\left\{1 + \frac{\sum_{p=1}^{m}\sum_{\substack{q=1\\p\neq q}}^{m} a_{pq}(y_1,y_2)}{\sum_{k=1}^{m} a_k(y_1,y_2)}\right\} dy_1dy_2$$

since g is a density function, and

$$g(y_1,y_2) = \sum_{k=1}^{m} a_k(y_1,y_2) \times (2\pi m\lambda_1\lambda_2)^{-1}.$$

This form of the information measure can aid calculation, but note the double integration required. This is computationally expensive, and has to be done for each combination of 2 from the n variables. In §3.3.1, this approach is compared with a similar method for structure learning which uses correlation coefficients.

An alternative to the Normal kernel is Epanechnikov's quadratic kernel (see Silverman, 1986), defined thus:

$$J(y_i \mid D,\lambda_i) = \frac{3}{\lambda_i 4\sqrt{5}}\left\{1 - \frac{1}{5}\left(\frac{y_i - d_{ik}}{\lambda_i}\right)^2\right\} \qquad \text{for } k = 1,2,\ldots,n, \quad i = 1,2,$$

when $|y_i - d_{ik}| < \sqrt{5}\lambda_i$, and 0 otherwise. The information measure $I(Y_1,Y_2)$ in this case is equal to the expression

$$\log m - \frac{9}{80m\lambda_1\lambda_2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sum_{k=1}^{m} a_k(y_1,y_2)\log\left\{1 + \frac{\sum_{p=1}^{m}\sum_{\substack{q=1\\p\neq q}}^{m} a_{pq}(y_1,y_2)}{\sum_{k=1}^{m} a_k(y_1,y_2)}\right\} dy_1dy_2$$

where

$$\begin{aligned} a_k(y_1,y_2) &= \left(1 - \frac{1}{5\lambda_1^2}(y_1 - d_{1k})^2\right) \times \left(1 - \frac{1}{5\lambda_1^2}(y_2 - d_{2k})^2\right) \qquad \text{and} \\ a_{pq}(y_1,y_2) &= \left(1 - \frac{1}{5\lambda_1^2}(y_1 - d_{1p})^2\right) \times \left(1 - \frac{1}{5\lambda_1^2}(y_2 - d_{2q})^2\right). \end{aligned}$$

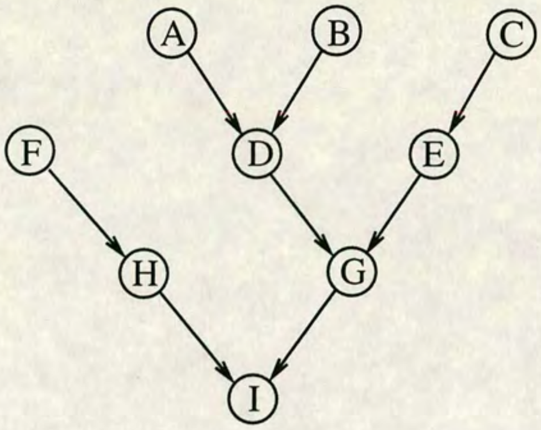


Figure 3–4: *Tree for test of continuous variable structure learning.*

3.3.1 Example

This section applies the information measures algorithm to an example for comparison and a similar method using correlation coefficients. The tree of Figure 3–4 represents nine continuous variables, with conditional Gaussian distributions defined as follows:

$$A \sim N(10,4), \qquad B \sim N(8,9), \qquad C \sim N(15,16),$$

$$D \sim N(7+a-b/2,4), \qquad E \sim N(3+c,4), \qquad F \sim N(0,1),$$

$$G \sim N(3-d+e,4), \qquad H \sim N(7+f,16), \qquad I \sim N(2g-h,4).$$

The trees recovered by the structure learning process are shown in Figure 3–5, for a data set of 100 simulations for each variable. The correlation algorithm is the same as that for information measures, but with product-moment correlation coefficients replacing the weights.

As is seen at first glance, the algorithm using information measures performs very poorly indeed, while the correlation coefficients tree is perfect. Both information measures trees show the same wildly incorrect structure. The information

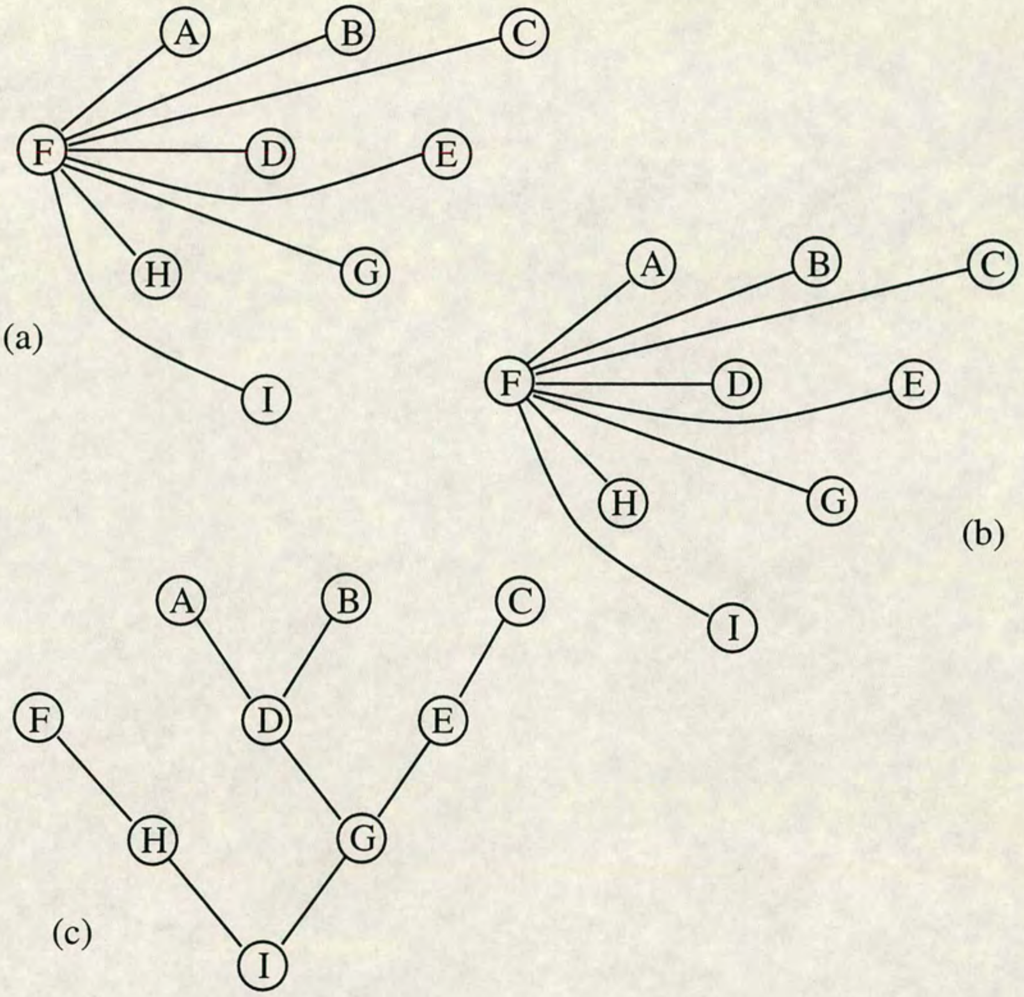


Figure 3–5: *Trees recovered by the structure learning process on continuous data, using: (a) information measures with a Normal kernel function; (b) information measures with a quadratic kernel; and (c) correlation coefficients.*

measures programs also take vastly more time to run than the correlation programs, due to the double integrations of kernel functions involved. Hence I would certainly recommend the correlation approach over information measures.

3.4 Recovering Polytrees

Once the structure of the (skeleton) tree has been ascertained, the next step is to attempt to recover the directions of arrows on the edges, i.e. to form a polytree, by examining the properties of $\Pr(\mathbf{x})$.

If the graphical structure of the true underlying model can be represented by a polytree, then $\Pr(\mathbf{x})$ has the form

$$\Pr(\mathbf{x}) = \prod_{i=1}^{n} \Pr(x_i \,|\, x_{\mathrm{pa}_1(i)}, x_{\mathrm{pa}_2(i)}, \ldots, x_{\mathrm{pa}_r(i)}), \tag{3.8}$$

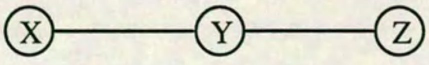
where the $x_{\mathrm{pa}_k(i)}$, for $k = 1, 2, \ldots, r$, are the parents of x_i in the polytree. The variable x_i may, of course, have no parents. Also, the parents of each variable are mutually independent, i.e.

$$\Pr(x_i \,|\, x_{\mathrm{pa}_1(i)}, x_{\mathrm{pa}_2(i)}, \ldots, x_{\mathrm{pa}_r(i)}) = \prod_{k=1}^{r} \Pr(x_{\mathrm{pa}_j(i)}) \qquad \text{for all } i.$$

The following theorem, adapted from Pearl and Dechter (1989), is a consequence of this:

Theorem 2 *If the conditional independencies of a model can be represented by a polytree, then the MWST algorithm of Chow and Liu (1968) unambiguously recovers the skeleton of the polytree.*

Unfortunately, it is not always possible to recover a single unique polytree from $\Pr(\mathbf{x})$. Consider three nodes, X, Y and Z, with tree structure



which give rise to the following three possible combinations of arrows:

1. $X \rightarrow Y \rightarrow Z$, or $X \leftarrow Y \leftarrow Z$,

2. $X \leftarrow Y \rightarrow Z$,

3. $X \rightarrow Y \leftarrow Z$.

Type 1 and type 2 are indistinguishable in terms of independence structure, but type 3 can be identified since X and Z are marginally independent. Hence it is possible to only partially identify the directions of arrows on a skeleton tree. Note that a positive test for independence of X and Z implies that type 3 should supply the directions of the arrows.

The polytree recovery algorithm originated in Rebane and Pearl (1987), and consists of these steps:

1. Generate a MWST using the procedure in §3.2.

2. Search the internal nodes (i.e. nodes with more than one neighbour) of the skeleton, beginning with the outermost layer and working inward, until a multi-parent node (such as Y in type 3 above) is found by a test for independence.

3. When a multi-parent node is found, determine the direction of arrows on its edges by the independence test.

4. For each node having at least one incoming arrow, find the directions of its remaining edges again using the independence test.

5. Repeat steps 2 to 4 until no further arrows can be added to the edges.

6. Any edges that remain undirected will require analysis beyond the scope of the data itself.

7. From $\Pr(\mathbf{x})$ compute the conditional probabilities of equation (3.8).

Kullback (1959) gives a test for independence of two categorical variables with m observations using an information measure. For variables X (with r categories) and Z (with c categories), we can test the two hypotheses

- H_0: Two variables X and Z are pairwise independent, and
- H_1: Two variables X and Z are not pairwise independent,

with (where the s_{ij} are the cell counts, $i = 1, 2, \dots, r$, $j = 1, 2, \dots, c$)

$$2I(H_0, H_1) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} s_{ij} \log\left(\frac{ms_{ij}}{s_{i.}s_{.j}}\right) \tag{3.9}$$

which has the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. Replacing the cell counts at (3.9) with probabilities gives

$$\begin{aligned} 2I(H_0, H_1) &= 2\sum_{i=1}^{r}\sum_{j=1}^{c} mp_{ij} \log\left(\frac{m \times mp_{ij}}{mp_{i.} \times mp_{.j}}\right) \\ &= 2m\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log\left(\frac{p_{ij}}{p_{i.} \times p_{.j}}\right) \\ &= 2mI(X, Z). \end{aligned}$$

Pearl (1988) states that "probabilistic analysis is indeed sensitive only to covariations, so it can never distinguish genuine causal dependencies from spurious correlations". While Chow and Liu's MWST algorithm will find a *best* approximation to $\Pr(\mathbf{x})$, no such guarantee applies for polytrees. For these reasons, automatic structure learning should not be relied upon on its own, but should serve as an aid to expert opinion.

3.4.1 Example

The polytree recovery mechanism is now applied to the test example of section 3.2.2. The χ^2 95% point for 1 degree of freedom is 3.841, and this is used for every test (3.9). Note that I am not considering the implications of performing multiple hypothesis tests for this very exploratory study.

The χ^2 statistics for any relevant pairs of nodes are shown in Table 3–2; where a pair of variables has a statistic greater than 3.841, they are marked with a $*$ for clarity. The resulting attempts at polytrees appear in Figure 3–6. Some edges are shown with two conflicting arrows; this is because different tests between pairs of variables can contradict each other when applying the algorithm.

As can be seen from Figure 3–6, the polytree recovery process seems to require a fairly large number of observations. By the 2000 observation data set, the process has correctly placed arrows on three arcs of the tree: $A \rightarrow B$; $C \rightarrow B$; and $D \rightarrow E$. The remaining edges have conflicting opposite arrows—this may due, in part at least, to the erroneous link H–G included in the MWST.

Again, there seems to be little, if any, difference between the sample frequencies and kernel methods.

Clearly then, while the structure learning process can aid the discovery of the directions of arrows within a tree, the final polytree will probably need to be decided upon by "experts". Pearl (1988) refers to polytrees as *causal polytrees*, and suggests that a consideration of the nature of the causality of the events behind the variables is necessary in order to fully understand and define a polytree and its related distributions.

	Sample Frequencies		Kernels	
Observations	Nodes	χ^2	Nodes	χ^2
50	C D	0.2740	C D	0.1643
	C G	0.7180	C G	0.4401
	C F	0.6617	C F	0.3438
	D F	1.0171	D F	1.3381
	D G	0.2740	D G	0.1643
	F G ∗	5.9809	F G	2.7436
	E H	0.0198	E H	0.0107
100	C E	0.3291	C D	0.4477
	E H	0.0874	C F	1.0550
	C H	1.4743	D F	1.1138
	D F	1.7700	E H	0.0749
	B F	0.1856	B H	0.4257
	A C	0.2207	A C	0.1144
	B G	0.6499	B G	0.4499
1000	D F ∗	5.1142	D F ∗	4.8451
	D G	0.1761	D G	0.1714
	F G	1.4743	F G	1.4156
	C E	0.0042	C E	0.0040
	B G	0.7463	B G	0.7181
	A C ∗	5.2579	A C ∗	4.9008
	A H ∗	6.3594	A H ∗	5.8552
	C H	3.2672	C H	2.9882
2000	D F	1.0828	D F	1.0584
	D G	1.5504	D G	1.5260
	F G ∗	5.2290	F G ∗	5.1302
	E H	0.0113	E H	0.0110
	B G	0.1133	B G	0.1112
	A H ∗	4.9652	A H ∗	4.7839
	C H	2.3826	C H	2.2536
	A C	0.0206	A C	0.0197

Table 3–2: *χ^2 statistics for polytree recovery example.*

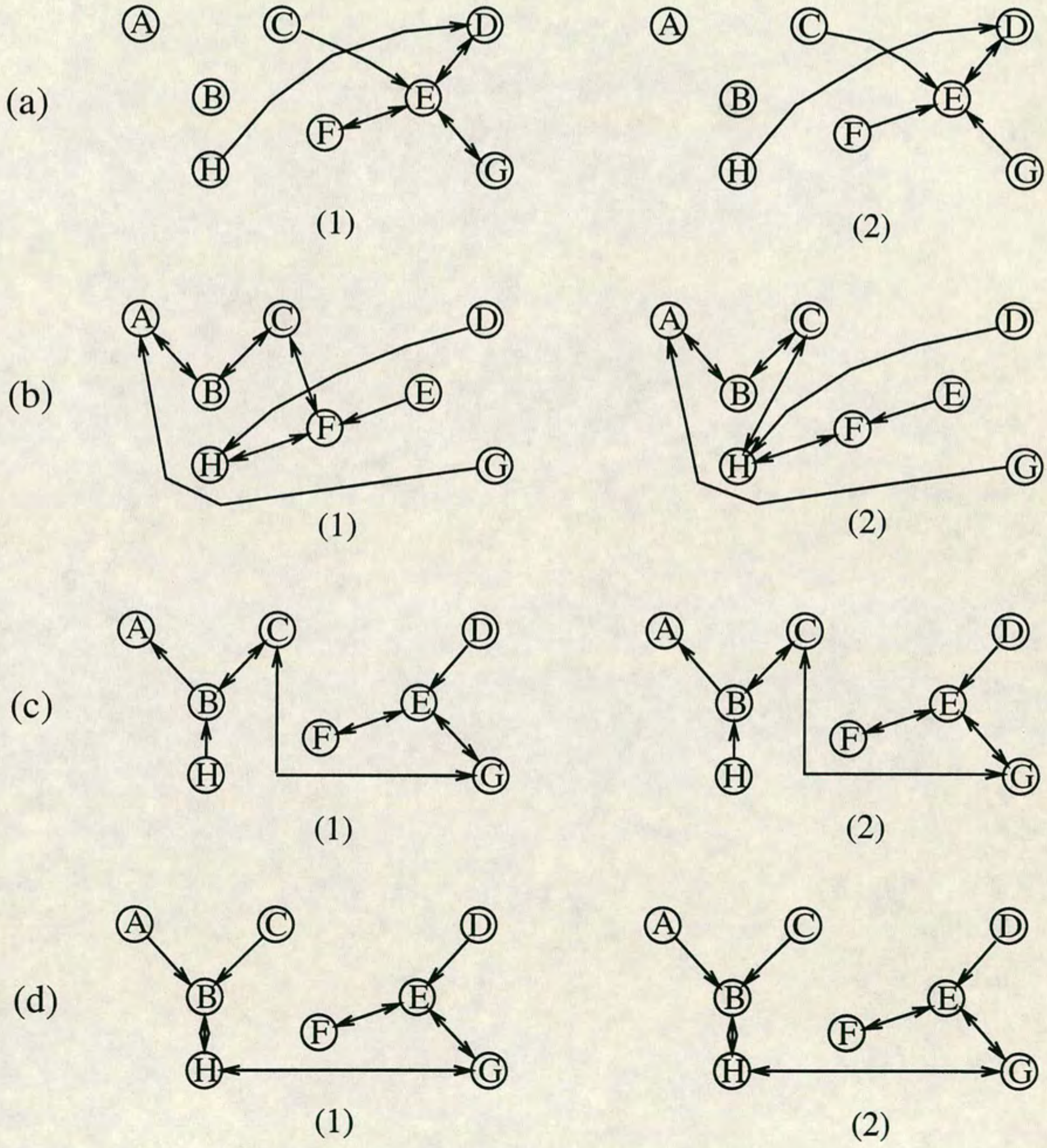


Figure 3–6: *Polytrees recovered by (1) sample frequencies and (2) kernels for (a) 50, (b) 100, (c) 1000, and (d) 2000 observations.*

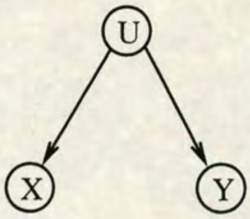


Figure 3–7: *An over-simplistic method could add a bogus link between X and Y.*

3.5 Recovering General Structure

The previous sections of this chapter showed how a tree structure can be drawn from a set of data, given that the data has a tree-dependent distribution.

Recovering a general graph structure from data is more complicated. Note that an apparently straightforward method would be to produce the same list of pairwise information measures as for trees, and use some sort of significance test (such as that at (3.9)) to decide whether or not to include an edge. There would be no checking for undirected loops; just directed loops would be excluded. This approach is too simplistic—consider the dependency structure of Figure 3–7, for example. Under certain circumstances (for instance, if X and Y have very similar conditional distributions given U), the pairwise information measures can produce many spurious edges based on mere correlations, and in this case X and Y might be joined. The tree recovery algorithm will most likely reject the X–Y link, having already joined X–U and Y–U.

Pearl (personal communication) proposes a different method. Essentially the idea is to determine for each pair of variables X and Y a subset $\mathcal{S}_{XY}$ that "shields" X from Y. If no such subset exists, then an edge is added between X and Y. An arrow points from X to Y if there exists a variable Z linked to Y but not X, such that it is not true to say that X is independent of Z given $\mathcal{S}_{XY} \cup Y$—see Pearl (1988), §8.2.3, for motivation.

Pearl recommends identifying for each pair X and Y the set of four or less variables $\mathcal{S}_{XY}$ that gives the *lowest mutual information measure* $I_{\min}(X, Y \mid \mathcal{S}_{XY})$. If $I_{\min}$ is below some threshold, do not connect X and Y, but if not, do connect them.

If the number of variables is not too large however, and the number of observations in the data set is large enough by comparison, then there seems in theory at least no reason not to look beyond four variables for $\mathcal{S}_{XY}$. Clearly however, there may be problems with time of computation for large models.

A possible algorithm for recovering general graph structure is thus:

1. Choose a pair of variables X and Y.

2. Compute $I(X, Y \mid \mathcal{S}_{XY})$ for all possible subsets $\mathcal{S}_{XY}$ of variables other than X or Y, and determine $I_{\min}(X, Y \mid \mathcal{S}_{XY})$.

3. Test $2mI_{\min}$ against the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, where r and c are the number of values that X and Y can take respectively. If $2mI_{\min}$ is above the χ^2 value (at an appropriate level) then draw an edge between X and Y.

4. Repeat steps 1 to 3 for all pairs of variables.

From Pearl (1988), it can be seen that

$$I(X, Y \mid \mathcal{S}_{XY}) = \sum_{x,y,s_{xy}} \Pr(x, y, s_{xy}) \log \frac{\Pr(x, y \mid s_{xy})}{\Pr(x \mid s_{xy}) \Pr(y \mid s_{xy})}.$$

Thus now, unlike the MWST algorithm, we have more than second-order calculations to carry out. Indeed for a data set with a large number of variables, we will definitely need a limit on the size of $\mathcal{S}_{XY}$. The following example has five variables, so that $\mathcal{S}_{XY}$ has a natural size limit of three in any case.

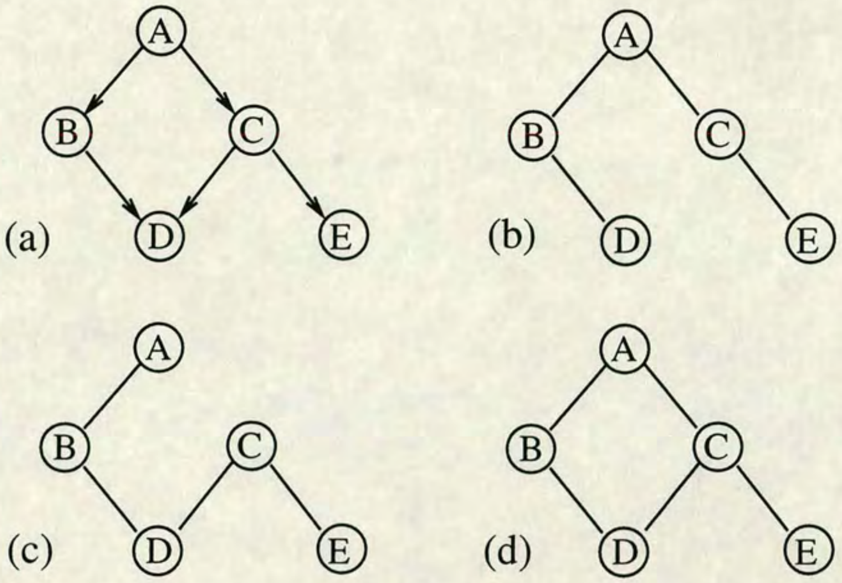


Figure 3–8: *Test example and results for general structure recovery: (a) is the original structure of the example from Pearl (1988); the remainder are the recovered structures for (b) 50, (c) 100, and (d) 1000 observations.*

3.5.1 Example

To test the general structure recovery algorithm, observations were generated for Pearl's metastatic cancer example (Pearl, 1988, and this thesis, section 1.4.2). Figure 3–8 shows the original influence diagram of the example and the recovered graphs from data sets of size 50, 100 and 1000. Table 3–3 shows the relevant χ^2 statistics for each pair of variables for each data set; once again significant results have been marked with an asterisk.

The probabilities needed for the calculations in the above algorithm were calculated using sample frequencies, given the comments in section 3.2.

The 95% point of the χ^2 distribution is 3.841, and so if the information measure is greater than this value for a particular pair of variables, then an edge is added between them. The general structure algorithm seems to work well; the data sets of size 50 and 100 give rise to structures that are only one edge away

Variables		χ^2 statistic for		
		50 obsns	100 obsns	1000 obsns
A	B	* 10.5297	* 6.6572	* 123.2624
A	C	* 7.3858	* 11.7502	* 19.2709
A	D	1.0357	0.5353	2.7400
A	E	0.1186	0.0001	0.1404
B	C	0.5289	0.7463	1.5271
B	D	* 21.9315	* 39.3421	* 397.7299
B	E	1.5464	0.8062	1.0268
C	D	2.2100	* 4.6584	* 34.1915
C	E	* 5.4144	* 5.3281	* 8.2746
D	E	0.6973	2.2889	0.1976

Table 3–3: *χ^2 statistics for the pairs of variables with the general structure recovery example.*

from being perfect, while the graph recovered from the 1000 observation set is entirely correct.

3.6 Conclusions

This chapter has studied the process of recovering structure from data.

In section 3.2, the use of kernel probability estimates in place of sample frequencies within the Chow and Liu algorithm did not prove too successful, even when bivariate smoothing parameters were considered. The kernel methods also consumed a large amount of computing time, relative to the sample frequencies; hence the latter method seems preferable.

Similar comments can be made for the continuous kernels of section 3.3. A far simpler tree recovery program which calculated product-moment correlation coefficients was very much faster, and performed very much better.

Recovering directions for the edges was shown to be a rather more difficult task in section 3.4. Only for the larger sample sizes did the algorithm begin to produce "successful" results.

Finally, the general structure recovery algorithm described in section 3.5 performed very well on the metastatic cancer example. For relatively few observations, the structure recovered was nearly fully correct.

Chapter 4

Stochastic Simulation in Standard Mixed Graphical Models

4.1 Introduction

This purpose of this chapter is to extend the stochastic simulation scheme of Pearl (1988) to models with both discrete and continuous variables.

Pearl's stochastic simulation procedure is an alternative to the exact computation methods of Chapter 1 for models containing only discrete variables. Given a graphical model, with defined directed acyclic graph and conditional probability tables, the method uses random experiments to draw inferences on the marginal probability distributions of discrete variables.

The current chapter extends this discrete scheme to include continuous variables within the model, and these are given conditional Gaussian distributions—as with the example from Lauritzen (1992) described in section 1.5. Such an extended model is termed a *standard model.* Chapter 5 will deal with non-Gaussian distributions and *non-standard* models.

The computational scheme of Lauritzen (1992) enables calculation of marginal probabilities, means and variances of the variables. Normand (1993) commented that only propagating means and variances for mixed models is unsatisfactory. Unfortunately, exact computation of the marginal density functions of continuous variables is generally forbiddingly complex.

The stochastic simulation scheme however allows us to obtain estimates of the marginal densities from simulated values. Unlike the exact methods, strong triangulation of the graph is not necessary. The computations themselves (for the standard model) are straightforward, and consideration of each node in a network involves only its neighbours and its childrens' parents. Kernel methods can then be applied to obtain estimates of the probability density functions.

Section 4.2 describes the stochastic simulation routine of Pearl (1988). The routine in fact uses a Gibbs sampler, but this is not mentioned by Pearl. The corresponding procedure for mixed (standard) models follows in section 4.3.

For some models, analysis by stochastic simulation will require the process to be split into separate runs, called *chains* here. Section 4.4 discusses this matter. Section 4.5 applies the methods of §4.3 and §4.4 to the waste incinerator example from Lauritzen (1992).

Mixed models as in Lauritzen (1992) do not allow discrete variables to have continuous variables as parents in the influence diagram. Finally in this chapter, section 4.6 discusses a possible method of including such cases in the stochastic simulation procedure.

Neither this nor Chapter 5 will consider in detail stopping procedures and convergence for the simulation. These topics are currently the subject of much debate. For example, see: Ritter and Tanner (1992); Smith and Roberts (1993), Besag and Green (1993) and discussion; Gelman and Rubin (1992), Geyer (1992) and comments/rejoinders.

4.2 Discrete Models

Pearl (1988) shows how an influence diagram and associated probabilities representing a graphical model with (only) discrete variables can be used to generate random samples of combinations of events. These samples are governed by the structure of the diagram and by the conditional probabilities given. The marginal probability of each event can then estimated by the percentage of samples in which the event takes each value. The discrete variables discussed from this point will be binary, but extensions to variables with greater than two categories should be obvious.

Henrion (1986) introduced a scheme, called *logic sampling*, where random values are assigned to the variables by passing through the influence diagram in a "top-down" fashion. The procedure first assigns values to variables represented by root nodes—that is, nodes with no parents. A value is generated for a non-root variable once values have been assigned to all its parents. Once new values have been generated for all of the variables, the procedure goes back to the "top" and starts again.

Although this scheme operates in a pleasing, logical way through the network, there is no way to account for variables that have already had their values observed. If the simulated value for an observed variable does not match the observed value, then that simulation run must be discarded. Hence this method can prove very inefficient.

A better approach therefore would be to clamp the appropriate variables to their observed values, and to conduct a simulation on the remaining variables. This is how Pearl's stochastic simulation scheme works.

Assume we have a graph containing m nodes representing m (binary) discrete variables. For each variable S_i, $i = 1, 2, \ldots, m$, we wish to calculate $\Pr(s_i | R_{S_i})$, where R_{S_i} is a realisation of the set of all the variables bar S_i. We can say

$$\Pr(s_i|R_{S_i}) = \frac{\Pr(s_i, R_{S_i})}{\Pr(R_{S_i})} = \frac{\Pr(s_1, s_2, \ldots, s_m)}{\Pr(R_{S_i})}.$$

Since $\Pr(R_{S_i})$ is independent of the value of S_i, we can write

$$\begin{aligned} \Pr(s_i|R_{S_i}) &= \alpha \Pr(s_1, s_2, \ldots, s_m) \\ &= \alpha \prod_{l=1}^{m} \Pr(s_l|U_{S_l}) \\ &= \alpha \Pr(s_i|U_{S_i}) \prod_{j=1}^{\beta} \Pr(c_j|U_{C_j}) \prod_{k=1}^{\gamma} \Pr(g_k|U_{G_k}) \end{aligned}$$

where α is a normalising constant, U_X is the set of variables whose corresponding nodes are parents of a node X, C_j is one of β variables whose corresponding nodes are children of the node for S_i, and G_k is one of γ variables whose corresponding nodes are not children of the S_i node, nor the S_i node itself. Note that (for example) the dependence of C_j on i has been suppressed from the notation.

Clearly $\prod_{k=1}^{\gamma} \Pr(g_k|U_{G_k})$ is also independent of the value of S_i. Finally, as can be seen as Theorem 1 in Pearl (1988),

$$\Pr(s_i|R_{S_i}) = \alpha \Pr(s_i|U_{S_i}) \prod_{j=1}^{\beta} \Pr(c_j|U_{C_j}). \tag{4.1}$$

We can then calculate $\Pr(s_i|R_{S_i})$ by solving for α.

Initially, we need to assign values to unobserved variables. The choice of values here can be arbitrary and will only affect the rate of convergence.

Stochastic simulation then proceeds according to the following algorithm, adapted from Pearl (1987):

1. Compute $\Pr(s_i|R_{S_i})$ for an S_i.

2. Generate a random number r from a Uniform distribution between 0 and 1.

3. Compare r with $\Pr(s_i|R_{S_i})$. If $r \leq \Pr(s_i|R_{S_i})$ then let $S_i = 1$; else let $S_i = 0$.

4. Record the state of S_i.

5. Repeat steps 1–4 for each S_i in the model. Note that the order of the S_i's is not important. This has now completed a *simulation run.*

6. Repeat steps 1–5 for a specified number of simulation runs.

7. Use the recorded values to estimate the marginal probabilities of the variables S_i.

There is an alternative method of calculating the estimates of marginal probabilities; see Gelfand and Smith (1990). Rather than merely averaging the simulated values, we can take the mean of the conditional probabilities $\Pr(s_i|R_{S_i})$. This latter method should give slightly faster convergence.

As stated earlier, this is essentially Gibbs sampling, with $\Pr(s_i|R_{S_i})$ as the Gibbs sampler. For a formal definition of Gibbs sampling, due originally to Geman and Geman (1984), see section 5.2.

4.3 Standard Models

This section defines the *standard* mixed graphical model and describes a stochastic simulation for such a model.

For a standard mixed graphical model: the distribution for a discrete variable (for now assuming its node has parental nodes[1] that are all representing discrete

[1] Such parents will be referred to as *discrete parents*; similarly if (for example) a variable Z is referred to as having "parents" X and Y then this will mean that the node representing Z has the nodes representing X and Y as parents in the influence diagram; X and Y will be termed *parent variables.* A similar comment can be made regarding "children".

variables) is defined by a probability (or a set of probablities for a discrete variate with more than two states) conditional on the values taken by its parent variables. A continuous variable has a different conditional Gaussian distribution for each combination of values of any discrete parent variables, where the variance is fixed for each combination and the mean is a linear function of the values of any continuous parent variables[2].

Again assume we have a graph with m nodes representing the variables $\{S_i\}$, $i = 1, \ldots, m$. Variables can now be either discrete or continuous. When considering each S_i in turn, there are now two possibilities: if S_i is discrete we need $\Pr(s_i|R_{S_i})$, and if S_i is continuous we need the conditional probability density function $f(s_i|R_{S_i})$.

This density can be written

$$f(s_i|R_{S_i}) \quad = \quad \frac{f(s_i, R_{S_i})}{f(R_{S_i})} \quad = \quad \frac{f(s_1, s_2, \ldots, s_m)}{f(R_{S_i})}.$$

Since our directed graph is acyclic, we can use the conditional independence assumptions inherent in the graphical structure to *order* the nodes, and hence the underlying variables, such that

$$f(s_1, s_2, \ldots, s_m) = \prod_{l=1}^{m} f(s_l|U_{S_l}),$$

where U_X is the set of parents of X. This is a consequence of the recursive factorisation identity (Whittaker, 1990).

This enables the joint density to be written as

$$f(s_1, s_2, \ldots, s_m) = f(s_i|U_{S_i}) \prod_{j=1}^{\beta} f(c_j|U_{C_j}) \prod_{k=1}^{\gamma} f(g_k|U_{G_k}),$$

where again C_j is one of β child variables of S_i, and G_k is one of γ variables which are not child variables of S_i, nor the variable S_i itself.

[2] Models which include non-Gaussian distributions and non-linear mean functions are studied in Chapter 5.

Note that $f(R_{S_i})$ will have a similar form, but without $f(s_i|U_{S_i})$. Note also that the random variables $G_1, \ldots, G_\gamma$ are conditionally independent of S_i because of the assumptions in the construction of the graph, so that $f(g_k|U_{G_k})$ terms will be independent of S_i, and can be cancelled, giving

$$f(s_i|R_{S_i}) = \frac{f(s_i|U_{S_i}) \prod_{j=1}^{\beta} f(c_j|U_{C_j})}{\int f(s_i|U_{S_i}) \prod_{j=1}^{\beta} f(c_j|U_{C_j})\, ds_i}. \tag{4.2}$$

Expanding the integral in the expression at (4.2), we get:

$$\begin{aligned} \int f(s_i|U_{S_i}) \prod_{j=1}^{\beta} f(c_j|U_{C_j})\, ds_i \; &= \qquad (4.3) \\ \int_{-\infty}^{\infty} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{s_i-\mu_i}{\sigma_i}\right)^2\right\} &\prod_{j=1}^{\beta} \frac{1}{\sigma_{C_j}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{s_{C_j}-\mu_{C_j}}{\sigma_{C_j}}\right)^2\right\} ds_i, \end{aligned}$$

where σ_i is the standard deviation of S_i, and may depend on values taken by discrete parent variables but is assumed to be independent of values taken by continuous parent variables; and μ_i is the mean function of S_i. The values s_{C_j}, μ_{C_j} and σ_{C_j} are the value, the mean function and the standard deviation respectively of the j-th child variable of S_i $(j = 1, 2, \ldots, \beta)$.

Calculation of this integral is complicated by the fact that S_i is part of each of the mean functions μ_{C_j}, i.e.

$$\mu_{C_j} = u_{C_j} s_i + v_{C_j}, \tag{4.4}$$

where u_{C_j} is the coefficient of S_i in the mean function of the child variable C_j of S_i, and v_{C_j} is the remainder of the mean function, composed of terms not involving S_i.

Putting (4.4) into (4.3) and integrating, we get

$$\begin{aligned}\int f(s_i|U_{S_i}) \prod_{j=1}^{\beta} f(c_j|U_{C_j})\, ds_i &= \\ & \frac{1}{(2\pi)^{\beta/2}} \cdot \frac{1}{\sigma_i} \cdot \prod_{j=1}^{\beta} \frac{1}{\sigma_{C_j}} \cdot \left(\sqrt{\frac{1}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}^2}{\sigma_{C_j}^2}} \right)^{-1} \times \\ \exp\left\{ \frac{1}{2} \left[\left(\frac{1}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}^2}{\sigma_{C_j}^2} \right)^{-1} \cdot \left(\frac{\mu_i}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}(s_{C_j} - v_{C_j})}{\sigma_{C_j}^2} \right)^2 - \left(\frac{\mu_i^2}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{(s_{C_j} - v_{C_j})^2}{\sigma_{C_j}^2} \right) \right] \right\}.\end{aligned}$$

Similarly, the numerator on the right hand side of (4.2) is equal to

$$\begin{aligned}f(s_i|U_{S_i}) \prod_{j=1}^{\beta} f(c_j|U_{C_j}) &= \\ & \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{s_i - \mu_i}{\sigma_i} \right)^2 \right\} \prod_{j=1}^{\beta} \frac{1}{\sigma_{C_j}\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{s_{C_j} - \mu_{C_j}}{\sigma_{C_j}} \right)^2 \right\} \\ &= \frac{1}{(2\pi)^{\frac{\beta+1}{2}}} \cdot \frac{1}{\sigma_i} \cdot \prod_{j=1}^{\beta} \frac{1}{\sigma_{C_j}} \times \\ \exp\left\{ -\frac{1}{2} \left[\sqrt{\frac{1}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}^2}{\sigma_{C_j}^2}}\, s_i - \left(\sqrt{\frac{1}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}^2}{\sigma_{C_j}^2}} \right)^{-1} \cdot \left(\frac{\mu_i}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}(s_{C_j} - v_{C_j})}{\sigma_{C_j}^2} \right) \right]^2 \right\} \times \\ \exp\left\{ \frac{1}{2} \left[\left(\frac{1}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}^2}{\sigma_{C_j}^2} \right)^{-1} \cdot \left(\frac{\mu_i}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}(s_{C_j} - v_{C_j})}{\sigma_{C_j}^2} \right)^2 - \left(\frac{\mu_i^2}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{(s_{C_j} - v_{C_j})^2}{\sigma_{C_j}^2} \right) \right] \right\}.\end{aligned}$$

Putting these results into (4.2), the Gibbs sampler $f(s_i|R_{S_i})$ becomes

$$\begin{aligned}f(s_i|R_{S_i}) &= \frac{1}{\sqrt{2\pi}} \cdot \left(\sqrt{\frac{1}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}^2}{\sigma_{C_j}^2}} \right) \times \\ \exp\left\{ -\frac{1}{2} \left[\sqrt{\frac{1}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}^2}{\sigma_{C_j}^2}}\, s_i - \left(\sqrt{\frac{1}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}^2}{\sigma_{C_j}^2}} \right)^{-1} \cdot \left(\frac{\mu_i}{\sigma_i^2} + \sum_{j=1}^{\beta} \frac{u_{C_j}(s_{C_j} - v_{C_j})}{\sigma_{C_j}^2} \right) \right]^2 \right\}\end{aligned}$$

which is a Gaussian probability density function.

This leads to the following result (which is a particular case of a general result on conditional distributions for a multivariate Normal distribution), which first appeared in Brewer *et al.* (1992):

Result 1 *The distribution of each continuous variable S_i in a standard mixed graphical association model, conditional on the state of all other variables, is Normal with mean*

$$\left(\frac{1}{\sigma_i^2}+\sum_{j=1}^{\beta}\frac{u_{C_j}^2}{\sigma_{C_j}^2}\right)^{-1}\cdot\left(\frac{\mu_i}{\sigma_i^2}+\sum_{j=1}^{\beta}\frac{u_{C_j}(s_{C_j}-v_{C_j})}{\sigma_{C_j}^2}\right),$$

and variance

$$\left(\frac{1}{\sigma_i^2}+\sum_{j=1}^{\beta}\frac{u_{C_j}^2}{\sigma_{C_j}^2}\right)^{-1},$$

where β is the number of child variables of S_i, s_{C_j} is the state of child variable C_j of S_i, μ_i is the mean function of S_i, $\mu_{C_j} = u_{C_j}s_i + v_{C_j}$ is the mean function of child variable C_j of S_i, σ_i^2 is the variance of S_i, and $\sigma_{C_j}^2$ is the variance of child variable C_j of S_i.

Note that for a continuous variable with no discrete parents, the variance in Result 1 will be constant. For a continuous variable with ρ discrete parents, there could be up to 2^ρ different values of the variance, because there will be (up to) 2^ρ different conditional Gaussian distributions (assuming discrete variables are binary). This fact can be used to reduce computation time.

The expression for $\Pr(s_i \,|\, R_{S_i})$ (S_i discrete) will change with the introduction of continuous variables. Adapting (4.1) to allow S_i to have continuous child variables, the equivalent version of (4.2), where S_i has β discrete child variables C_j and δ continuous child variables C_k, is

$$\Pr(s_i|R_{S_i}) = \frac{\Pr(s_i|U_{S_i})\prod_{j=1}^{\beta}\Pr(c_j|U_{C_j})\prod_{k=1}^{\delta}f(c_k|U_{C_k})}{\int\Pr(s_i|U_{S_i})\prod_{j=1}^{\beta}\Pr(c_j|U_{C_j})\prod_{k=1}^{\delta}f(c_k|U_{C_k})}.$$

Since the denominator here is independent of S_i, it can be replaced with a constant to give the following result:

Result 2 *The probability distribution of each discrete variable S_i in a standard mixed graphical association model, conditional on the state of all other variables,* *is*

$$\Pr(s_i|R_{S_i}) = \alpha \Pr(s_i|U_{S_i}) \prod_{j=1}^{\beta} \Pr(c_j|U_{C_j}) \prod_{k=1}^{\delta} f(c_k|U_{C_k}) \tag{4.5}$$

where α is a normalising constant, R_{S_i} is a realisation of the set of all variables except S_i, U_{S_i} is the set of parent variables of S_i; there are β discrete child variables C_j of S_i and δ continuous child variables C_k of S_i.

Thus for a mixed standard model, the stochastic simulation algorithm becomes:

1. Choose a variable S_i; if S_i is discrete, then calculate $\Pr(s_i|R_{S_i})$. If S_i is continuous, calculate the mean and variance of the Gaussian distribution for $S_i \mid R_{S_i}$, as per Result 1.

2. If S_i is discrete, generate a random number r from a Uniform distribution between 0 and 1. Compare r with $\Pr(s_i|R_{S_i})$. If $r \leq \Pr(s_i|R_{S_i})$ then let $S_i = 1$; else let $S_i = 0$.

3. If S_i is continuous, generate a value from the appropriate Gaussian distribution for the continuous S_i.

4. Record the state of S_i.

5. Repeat steps 1–4 for each S_i in the model.

6. Repeat steps 1–5 for a specified number of simulation runs.

7. Use the recorded values to estimate the marginal probabilities of discrete S_i, and the marginal means and variances of continuous S_i. Also, if required, use kernel function estimation to obtain estimates of the probability density functions of the continuous variables from the simulated values.

An example using (a version of) this algorithm will appear in section 4.5.

4.3.1 Observed Variables

The exact computational algorithms described in Chapter 1 deal with observed values by propagating the effect of the observation through the structure of the network. The stochastic simulation procedure however requires a different approach:

Discrete variable observed : If a discrete variable X is observed as having the value $X = x$ (where $x = 0$ or $x = 1$) then $\Pr(x \,|U_X)$ will be set equal to 1, and $\Pr(\overline{x}, |U_X)$ will be set to 0.

Continuous variable observed : If a continuous variable Y is observed as taking the value y, then $f(\,y\,|U_Y)$ becomes merely a function of the values of the *parents* of Y.

Variables X and Y would themselves be omitted in the simulation runs; $\Pr(X \,|U_X)$ and $f(\,y\,|U_Y)$ would be used when generating a value for a parent variable of X and Y respectively. The simulation otherwise proceeds as normal, and the correct results are obtained.

4.4 Chains

It has been well documented in recent Gibbs sampling literature that it can be extremely important, especially when examining a new problem, to perform separate simulation runs from several starting points—see Gelman and Rubin (1992), for example.

This section will illustrate, by means of two very simple examples, that investigative use of multiple simulation runs is relevant to the stochastic simulation procedures, and will provide an adapted version of the algorithm of section 4.3 for multiple runs. These simulations runs are referred to as *chains* here.

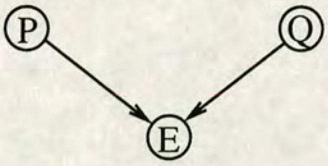


Figure 4–1: *Influence diagram for probabilities of 0 or 1 example.*

$\Pr(p) = 0.2$		$\Pr(q) = 0.4$	
$\Pr(e \mid p, q) = 1$	$\Pr(e \mid p, \overline{q}) = 1$	$\Pr(e \mid \overline{p}, q) = 1$	$\Pr(e \mid \overline{p}, \overline{q}) = 0$

Table 4–1: *Conditional probabilities for probabilities of 0 or 1 example.*

4.4.1 Conditional Probabilities Defined as 0 or 1

In a model where some of the defined conditional probabilities take the values 0 or 1, a single run of the simulation can get "stuck"; that is, certain combinations of values of variables, while possible in themselves, may be impossible to reach from other combinations *via the simulation procedure.* To see this, here is a simple example.

Figure 4–1 shows 3 nodes representing the variables E, P and Q, and the associated probabilities are in Table 4–1. The variables take binary values denoted e and $\overline{e}$, for example. Note that combinations $(q, p, \overline{e})$, $(q, \overline{p}, \overline{e})$, $(\overline{q}, p, \overline{e})$ and $(\overline{q}, \overline{p}, e)$ are logically impossible in the simulation process. For example, if nodes Q and P are in states q and p respectively, then E will be in state e since $\Pr(e|p, q) = 1$.

Suppose $(\overline{q}, \overline{p}, \overline{e})$ is chosen initially. Then from equation (4.1) it is clear that the state of each node will never change during simulation (since $\Pr(e|\overline{p}, \overline{q}) = 0$). Similarly, if one of (q, p, e), $(q, \overline{p}, e)$ and $(\overline{q}, p, e)$ is chosen, the system can move between them but it will never reach $(\overline{q}, \overline{p}, \overline{e})$. Thus the system is in two isolated *chains* of combinations of states; if the nodes have states in a combination

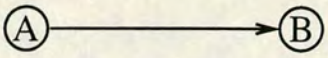


Figure 4–2: *Simple two-node graph to show need for chains in a mixed model.*

belonging to one chain, then they will never be in a combination belonging to the other chain.

A single run of the simulation therefore may not produce the correct marginal probabilities.

4.4.2 Values of Conditional Densities Close to 0

To see that a similar problem can occur with mixed graphical models, consider the following example, with influence diagram in Figure 4–2. The conditional distributions are defined by $\Pr(a) = 0.8$; $B \mid a \sim N(-1, 0.0064)$; and $B \mid \overline{a} \sim N(0.5, 0.0025)$. To conduct stochastic simulation on this artificial network, suppose initially we set $A = a$ and $B = -1.1$. To generate a value for node A, we need the conditional probabilities given by

$$\begin{aligned} \Pr(a \mid b) &= \alpha \Pr(a) f(b \mid a) \\ &\text{and} \\ \Pr(\neg a \mid b) &= \alpha \Pr(a) f(b \mid \neg a) \end{aligned}$$

The problem is that

$$f(b \mid \neg a) = \phi\left(\frac{-1.1 - 0.5}{\sqrt{0.0025}}\right) = \phi\left(\frac{-1.6}{0.05}\right) = \phi(-32),$$

which is effectively zero[3]. Since $f(b \mid a) = 0.054$, node A will *definitely* take state a in the first simulation run. Obtaining an (effectively) non-zero value for $f(b \mid \neg a)$ is an extremely unlikely event. When generating a value for B, the same problem

[3] In fact, $\phi(-32) = e^{-512}/\sqrt{2\pi}$, which is effectively zero on most computers.

exists; the marginal distribution for B is a mixture of two Gaussian distributions, and these are disconnected, in that the simulation cannot move from one mode to the other. The simulation will remain stuck at $A = a$.

4.4.3 Multiple Chains

Multiple chains can be used to solve the problem of a locked system. The variables must be split into two sets: D, the set of discrete variables; and C, the set of continuous variables. A chain H_i is defined for each member of the state space of D. The stochastic simulation procedure is then applied to the variables in C for each chain, with the results weighted by the probability of each chain. In order to obtain a true set of observations of the model, we can simply run different numbers of simulations for each chain, in proportion to $\Pr(H_i)$.

An algorithm for the multiple chains is as follows:

1. Identify the chains H_i by considering the state space of D.

2. Choose a particular chain.

3. Apply the stochastic simulation for mixed models procedure to the continuous variables, C.

4. Record the means and variances for the chain.

5. Repeat steps 2–4 until all the chains have been processed.

6. Calculate the marginal means and variances, using the $\Pr(H_i)$ if necessary.

Note that this method requires calculation of the marginal probabilities of the discrete variables, but this is usually a simple matter.

4.4.4 Observed Variables with Multiple Chains

If a model requires use of the multiple chain algorithm, then dealing with observed variables poses a slight problem. The probabilities of the chains will change, and there is no way to calculate these correctly by the stochastic simulation method.

In this case, the exact propagation procedure of Lauritzen (1992) must be used in order to obtain updated information on the variables in D, from which the new $\Pr(H_i)$ can be calculated. The algorithm of §4.4.3 can then be applied to C.

Even though an exact computational procedure must be used here, the simulation process still has the benefit of providing observations from which estimates of the marginal densities can be obtained.

4.5 Example

This section will apply stochastic simulation to the example from Lauritzen (1992). Since this example has been presented already in Chapter 1 of this thesis, only a brief summary of the model is given here, for ease of reading.

The influence diagram is shown in Figure 4–3, and the variables are defined thus:

Burning regime (Discrete) $\Pr(B = b) = 0.85$.

Filter state (Discrete) $\Pr(F = f) = 0.95$.

Type of waste (Discrete) $\Pr(W = w) = 2/7$.

Filter efficiency (Continuous)

$$\mathcal{L}(E \mid f, \overline{w}) \quad = \quad N(-3.2, 0.00002)$$

$$\begin{aligned}\mathcal{L}(E\,|\,\overline{f},\overline{w}) &= N(-0.5, 0.0001)\\ \mathcal{L}(E\,|\,f,w) &= N(-3.9, 0.00002)\\ \mathcal{L}(E\,|\,\overline{f},w) &= N(-0.4, 0.0001).\end{aligned}$$

Emission of dust (Continuous)

$$\begin{aligned}\mathcal{L}(D\,|\,b,w,e) &= N(6.5+e, 0.03)\\ \mathcal{L}(D\,|\,b,\overline{w},e) &= N(6.0+e, 0.04)\\ \mathcal{L}(D\,|\,\overline{b},w,e) &= N(7.5+e, 0.1)\\ \mathcal{L}(D\,|\,\overline{b},\overline{w},e) &= N(7.0+e, 0.1).\end{aligned}$$

Concentration of CO$_2$ (Continuous)

$$\mathcal{L}(C\,|\,b) = N(-2, 0.1) \qquad \text{and} \qquad \mathcal{L}(C\,|\,\overline{b}) = N(-1, 0.3).$$

Penetrability of light (Continuous) $\mathcal{L}(L\,|\,d) = N(3 - d/2, 0.25)$.

Metal in waste (Continuous)

$$\mathcal{L}(M_i\,|\,w) = N(0.5, 0.01) \qquad \text{and} \qquad \mathcal{L}(M_i\,|\,\overline{w}) = N(-0.5, 0.005).$$

Emission of metal (Continuous) $\mathcal{L}(M_o\,|\,d, m_i) = N(d + m_i, 0.002)$.

Analysis of this example by stochastic simulation will fail unless the multiple chains algorithm is used. To see this, consider a subset of the model—variables F, W and E. Assume initial values for simulation $F = f$, $W = \overline{w}$ and $E = -3.2$. Then

$$\Pr(W = \overline{w}) = (2/7) \times \phi\left(\frac{-3.2 - (-3.9)}{0.00002}\right) = (2/7) \times \phi(-35000),$$

and hence effectively zero. It should be clear then that W will never be set to $\overline{w}$ from the given starting points; similarly F will never reach $\overline{f}$. Hence the system is locked, and multiple chains are needed.

This example contains three binary variables, and hence the state space of D has $2^3 = 8$ elements. A separate chain of the simulation will be run for each.

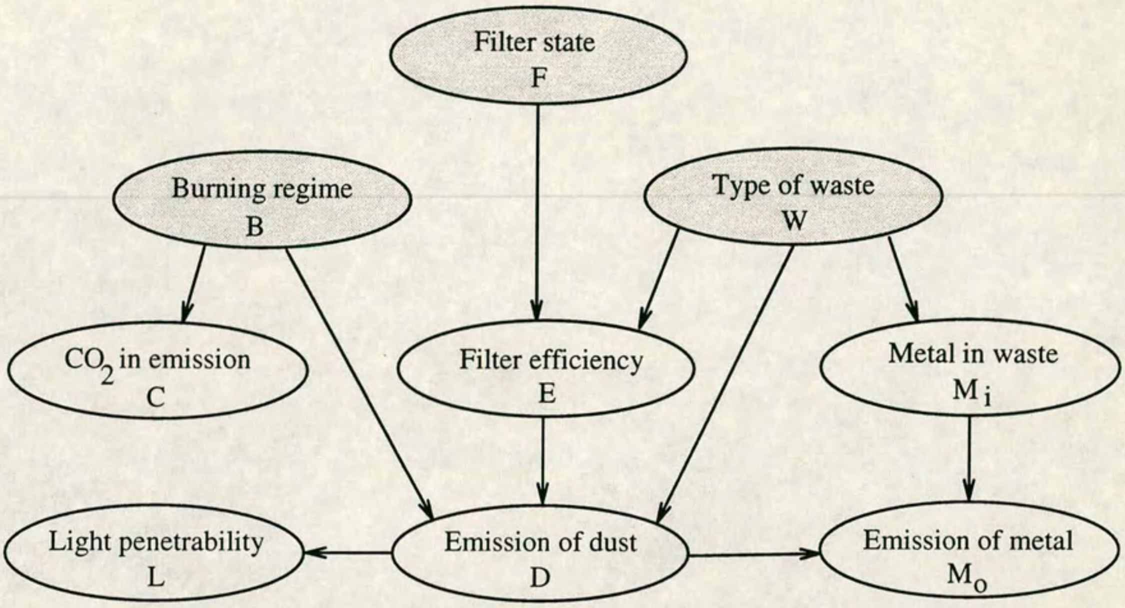


Figure 4–3: *Lauritzen's waste incinerator example.*

Chains	States of Variables	$\Pr(H_t)$
H_1	$b\ f\ w$	646/2800
H_2	$b\ f\ \overline{w}$	1615/2800
H_3	$b\ \overline{f}\ w$	34/2800
H_4	$b\ \overline{f}\ \overline{w}$	85/2800
H_5	$\overline{b}\ f\ w$	114/2800
H_6	$\overline{b}\ f\ \overline{w}$	285/2800
H_7	$\overline{b}\ \overline{f}\ w$	6/2800
H_8	$\overline{b}\ \overline{f}\ \overline{w}$	15/2800

Table 4–2: *Probabilities of chains for the Lauritzen (1992) example.*

	Simulation Runs								Correct	
Node	2800		5600		11200		22400		Values	
	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
E	-3.254	0.503	-3.254	0.502	-3.254	0.502	-3.254	0.503	-3.25	0.50
D	3.042	0.629	3.079	0.588	3.043	0.599	3.040	0.596	3.04	0.59
C	-1.850	0.261	-1.850	0.249	-1.850	0.254	-1.850	0.258	-1.85	0.26
L	1.479	0.424	1.460	0.387	1.479	0.399	1.480	0.404	1.48	0.40
M_i	-0.212	0.214	-0.210	0.213	-0.211	0.211	-0.214	0.211	-0.21	0.21
M_o	2.830	0.789	2.870	0.716	2.832	0.740	2.826	0.745	2.83	0.74

Table 4–3: *Results of simulation for the Lauritzen (1992) example.*

The elements of D and the probabilities of the respective chains H_i are shown in Table 4–2. Note the common denominator of 2800 with the $\Pr(H_i)$; this means that if the overall number of simulation runs is a multiple of 2800, then the chains can be run with numbers in proportion to the $\Pr(H_i)$, and hence the simulated values will represent a true set of observations from the model.

The results of the simulation are shown in Table 4–3. The 22400 simulation takes around a minute with a C program using NAG C subroutines on a Sequent Symmetry Unix mainframe. As can be seen from the table, good results are obtained even for such a short time.

Figures 4–4 and 4–5 show plots of the estimated marginal density functions for the six continuous variables. Kernel methods, using Gaussian kernel function (since the data was generated using Gaussian distributions in the first place), were used to obtain these estimates.

The structure of some of the marginal densities is easy to see; for example, the densities of E and M_i are mixtures of four and two Gaussian densities respectively. By inspection of the model as defined in addition to the kernel plots, it can be seen that C and D also have Gaussian mixture densities, while L and M_o have marginal density functions which are convolutions of Gaussian densities.

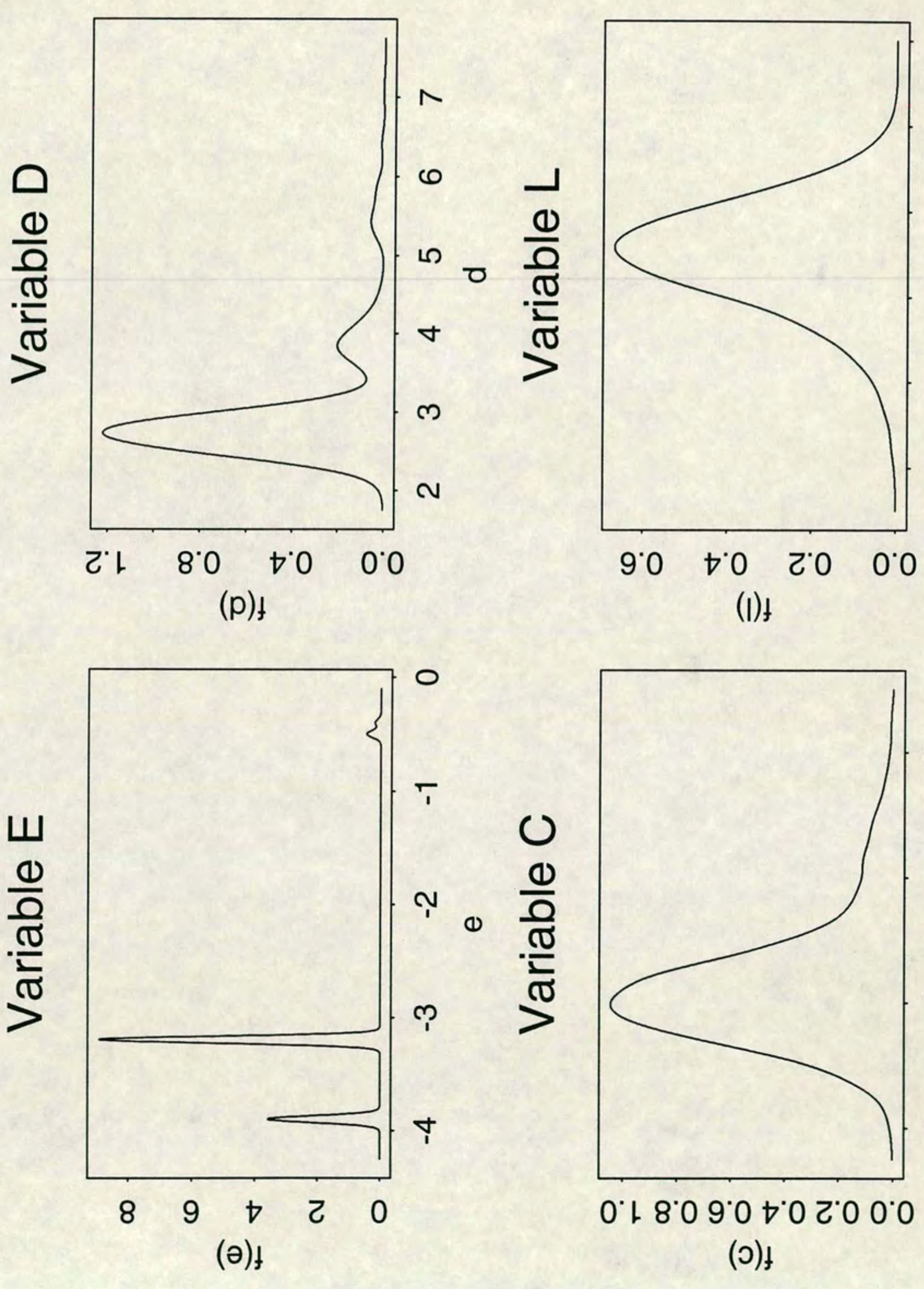


Figure 4–4: *Kernel density estimates for variables E, D, C and L for simulated values obtained by stochastic simulation.*

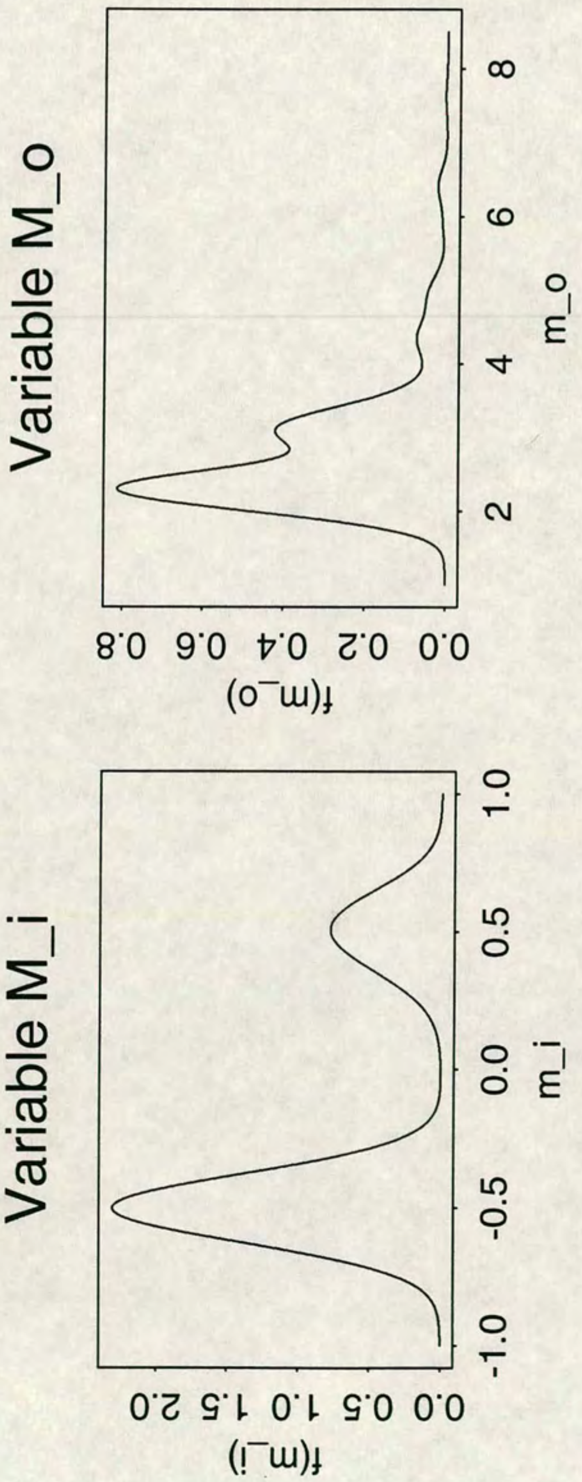


Figure 4–5: *Kernel density estimates for variables M_i and M_o for simulated values obtained by stochastic simulation.*

Chains	States of Variables	$\Pr(H_t)$
H_1	$f\, w$	49/200
H_2	$f\, \overline{w}$	147/200
H_3	$\overline{f}\, w$	1/200
H_4	$\overline{f}\, \overline{w}$	3/200

Table 4–4: *Updated probabilities of chains for the Lauritzen (1992) example with observed variables.*

4.5.1 Observed Variables

The above analysis could have been done by logic sampling; the use of the Gibbs sampler is necessary when variables have their values observed. For the Lauritzen (1992) example, assume now that two variables have been observed; the burning regime is stable ($B = b$) and the log light penetrability is $L = 1.1$.

As stated in section 4.4.4, the exact propagation routine of Lauritzen (1992) must be used to obtain the updated marginal probabilities of the remaining discrete variables, F and W in this case; it turns out that $\Pr(f) = 0.98$ and $\Pr(w) = 0.25$. There are now only 4 chains; they and their respective probabilities are shown in Table 4–4.

The common denominator of 200 for the $\Pr(H_i)$ means that the number of simulation runs should be a multiple of 200 in order to obtain a genuine sample from the model as a whole. However since $\Pr(H_3) = 1/200$, for example, and also given that 200 is a rather low number in any case for simulation such as this, performing well over 200 would seem sensible.

The results of the simulation, along with the updated marginals, appear in Table 4–5. The corresponding kernel density estimates for the five remaining continuous variables are shown in Figures 4–6 and 4–7.

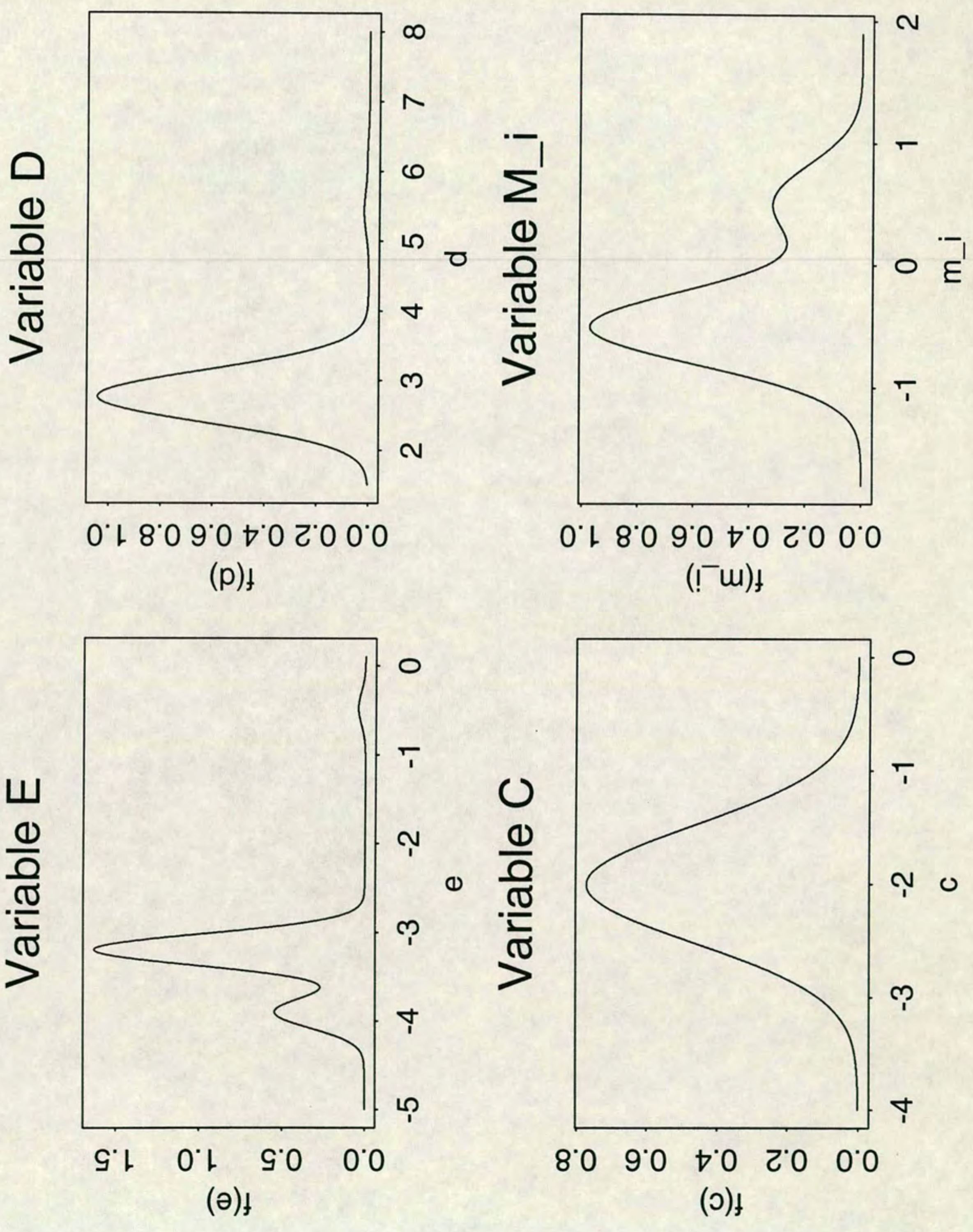


Figure 4–6: *Kernel density estimates for variables E, D, C and M_i after observing values for B and L.*

	Simulation Runs				Correct	
Nodes	20000		40000		Values	
	μ	σ^2	μ	σ^2	μ	σ^2
E	-3.322	0.238	-3.317	0.239	-3.32	0.24
D	2.846	0.181	2.845	0.194	2.84	0.17
C	-2.002	0.100	-2.000	0.100	-2.00	0.10
M_i	-0.251	0.191	-0.251	0.194	-0.25	0.19
M_o	2.595	0.298	2.592	0.289	2.58	0.28

Table 4–5: *Results of simulation on the Lauritzen (1992) example after observing values for B and L.*

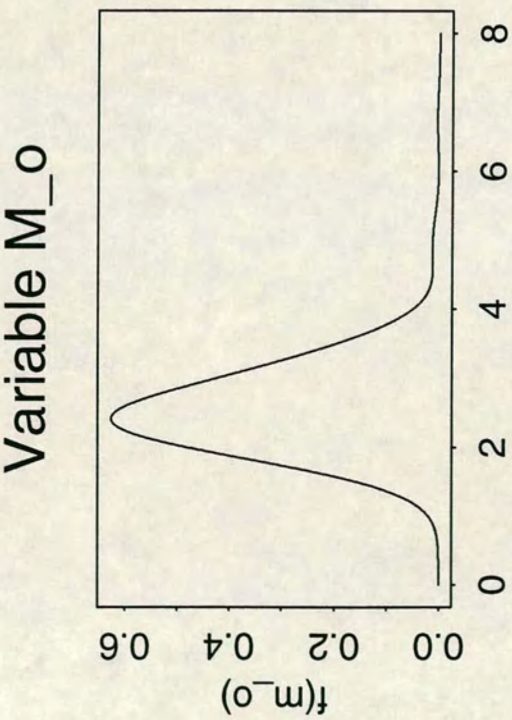


Figure 4–7: *Kernel density estimates for variable M_o after observing values for B and L.*

4.6 Discrete Nodes with Continuous Parents

Thus far in this thesis, discrete variables have not been allowed to have continuous parents in the influence diagram. Lauritzen (1992) does suggest a way of dealing with such cases, but this is just an approximation. A method of incorporating these into the stochastic simulation is now suggested here. The proposed approach is, admittedly, not very realistic in terms of modelling, and may be very difficult to apply in practice; it is however a very simple way of dealing with the current problem.

Consider a discrete node T with one or more continuous parents (and any number of discrete parents). We define a Gaussian distribution for J conditional on the values of its parents as usual. For each value of J generated, we then calculate $\Pr(T|R_T)$ from

$$\Pr(T=1|R_T) = \frac{\exp j}{1+\exp j} \tag{4.6}$$

and generate a value for variable T using this probability. This would enable us to estimate the marginal probability $\Pr(T)$ in the usual way.

The determination of $\Pr(T)$ by exact methods requires the determination of the mean of $\Pr(T|R_T)$ which is

$$E\{\Pr(T|R_T)\} = \int_{-\infty}^{\infty} \left(\frac{\exp j}{1+\exp j}\right) g(j)\, dj,$$

where $g(j)$ is the marginal density function of J. If we know $g(j)$ exactly, then we could estimate this integral using numerical integration. However, it is not always feasible to calculate $g(j)$ exactly; so this scheme with the dummy continuous variable is not generally applicable to exact methods.

While the inclusion of an extra step in the simulation process may seem odd, it does allow the type of situation in question here to be included within the current framework in a very straightforward manner.

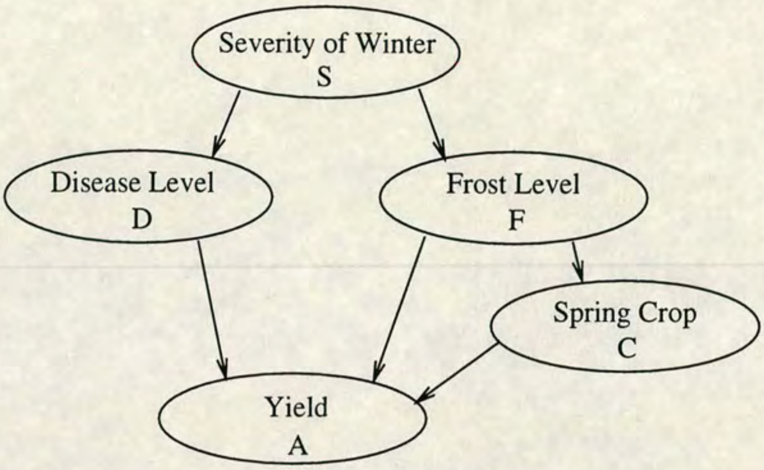


Figure 4–8: *Influence diagram for crop forecast example.*

4.6.1 Example

The following example is adapted from Aitken and Gammerman (1990) and concerns prediction of the yield of a certain crop. This example is artificial, and so not too much meaning should be read into the choice of mean functions.

> The adjustment to the forecast of the yield of an autumn sown crop depends on disease levels and frost damage, which both in turn depend on the severity of the winter. It is possible to sow again with a spring crop, dependent on the amount of frost damage. This second crop would also affect the yield forecast.

This situation is illustrated by the directed acyclic graph in Figure 4–8.

We wish to be able to answer such questions as: suppose the winter was very severe; by how much should the predicted yield be adjusted?

The model is defined by an agricultural expert thus:

Severity of the winter (Continuous) Denoted by S. We let $S \sim N(5, 4)$, representing the average daily temperature in °C.

	Simulation Runs (000's)						Correct	
Node	50		100		200		Values	
	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
S	4.946	3.981	5.040	4.082	5.011	4.021	5.0	4
D	14.925	12.735	15.075	13.357	14.995	13.127	15.0	13
F	3.153	13.280	2.906	13.363	2.949	13.200	3.0	13
J	6.317	57.235	5.795	57.796	5.893	56.750	6.0	56
A	-11.246	39.011	-11.682	40.959	-11.548	39.153	-11.5	39

Table 4–6: Simulation for crop forecasting example

Disease level (Continuous) Denoted by D. We let $D \mid s \sim N(10+s, 9)$, representing the level of parasites.

Frost level (Continuous) Denoted by F. We let $F \mid s \sim N(8-s, 9)$, representing the number of days with frost.

Spring crop (Discrete) There are two states—sow or not sow. Let θ be the probability of sowing the spring crop, i.e. $\Pr(C=c)$. Define

$$J \mid f \sim N(2f, 4),$$

and hence C using the procedure at equation (4.6) above with C replacing T.

Yield (Continuous) Denoted by A. We let $A \mid d, f, j \sim N(0.5-d-f+j, 1)$.

Table 4–6 shows the results of simulation on the model. The numbers of simulation runs are large because the variances are large, and hence convergence is slow. Even so, the C program took only three minutes for the 200,000 simulations. The estimates for θ were 0.791, 0.766 and 0.776 for the 50,000, 100,000 and 200,000 simulation runs respectively, compared with a value obtained by numerical integration of 0.759 ; note that in this example, the marginal density function assigned to J is Gaussian.

4.7 Conclusions

This chapter has described the use of a stochastic simulation algorithm for inference on marginal distributions in mixed graphical models. It has been shown that the simulation, via a Gibbs sampler, can be performed in a relatively straightforward manner.

The use of a simulation method enables estimation of the marginal density distributions by kernel methods; an estimate of this kind can be important, since (from Lauritzen, 1992) "[i]n general both the density itself and the problem of its computation can be forbiddingly complex."

Lauritzen (1992) says of his approximate method for dealing with discrete variables that have continuous parents that "...the error of approximation is negligible compared to the general uncertainty involved in the model building itself." This claim certainly relates to the simulation algorithms presented here, although admittedly the word "negligible" may be a bit strong.

Chapter 5

Stochastic Simulation in Non-Standard Mixed Graphical Models

5.1 Introduction

This chapter presents a study of stochastic simulation on *non-standard models*—that is, graphical models which are not limited to conditional Gaussian distributions and linear mean functions.

The simulation scheme for non-standard models is not as straightforward as the scheme for standard models; in Chapter 4 it was shown that generating a new value for a continuous variable in a standard model meant simply generating a value from a particular Gaussian distribution.

For non-standard models, the probability density function for generating a new value will most likely *not* be Gaussian. As will be demonstrated, it will probably not be a common density function at all, and may only be (practically) known to a factor of an unknown constant; because of this it can be necessary to use a further Markov chain Monte Carlo (MCMC) technique to obtain a new value.

Note that this MCMC step occurs *within* the Gibbs sampling process that drives the simulation. Such a combination of MCMC methods [1] was termed by Smith and Roberts (1993) a *hybrid strategy.*

The choice of MCMC technique is discussed in section 5.2, and also Gibbs sampling, used in this and the previous chapter, is formally defined. Section 5.3 presents some examples. The first two are very simple and just show how the methodology can be applied. The third is a serious example based on forecasting energy demand.

5.2 Hybrid Strategies

This section discusses the procedures required for stochastic simulation on non-standard graphical models. It begins by formally defining Gibbs sampling, and then describes the additional MCMC methods needed in this chapter, also justifying their use.

5.2.1 Gibbs Sampling

A systematic form of the Gibbs sampling algorithm, adapted from Smith and Roberts (1993), follows:

Let $\pi(\mathbf{x}) = \pi(x_1, x_2, \ldots, x_k)$ be an unknown joint density, and for $i = 1, 2, \ldots, k$ let $\pi(x_i \,|\, x_{-i})$ denote defined full conditional densities for each x_i, where $x_{-i} \equiv \{x_j : \forall j \neq i\}$.

[1] Since Gibbs sampling is a MCMC method.

Choose arbitrary starting values $\mathbf{x}^0 = (x_1^0, x_2^0, \ldots, x_k^0)$. Then generate successive values from the full conditional distributions as so:

$$\begin{array}{lll} x_1^1 & \text{from} & \pi(x_1 \,|\, x_{-1}^0) \\ x_2^1 & \text{from} & \pi(x_2 \,|\, x_1^1, x_3^0, x_4^0, \ldots, x_k^0) \\ x_3^1 & \text{from} & \pi(x_3 \,|\, x_1^1, x_2^1, x_4^0, \ldots, x_k^0) \\ & \vdots & \\ x_k^1 & \text{from} & \pi(x_k \,|\, x_{-1}^1). \end{array}$$

This has completed a transition from $\mathbf{x}^0$ to $\mathbf{x}^1$. Repeating this cycle produces a sequence $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^t, \ldots$ which is a realisation of a Markov chain.

The key feature of this algorithm is that inferences can be made on the joint density $\pi(\mathbf{x})$ by sampling only from the conditional densities $\pi(x_i \,|\, x_{-i})$.

For the mixed graphical models studied here, the conditionals $\pi(x_i \,|\, x_{-i})$ used to draw new values are the Gibbs samplers $\Pr(s_i \,|\, R_{S_i})$ and $f(s_i \,|\, R_{S_i})$; as was shown in Chapter 4, these can be calculated from the given information on a model—the distributions of variables conditional on the values of parent variables. Clearly then, the stochastic simulation procedure is an application of Gibbs sampling.

5.2.2 Non-Standard Graphical Models

With standard graphical models, generating new values from the Gaussian conditional densities $f(s_i \,|\, R_{S_i})$ is straightforward; with a non-standard model, sampling new values is not as simple in general.

Consider a non-standard model, where the conditional density functions defined for each continuous variable are denoted

$$f_{S_i}(s_i \,|\, U_{S_i}),$$

and the mean functions are specified as

$$g_{S_i}(U_{S_i}),$$

where U_{S_i} is the set of parents of S_i, and the functions f_{S_i} and g_{S_i} are specific to variable S_i; a model can contain a mixture of different types of conditional distribution and mean function.

An equivalent version of Result 2 for non-standard models is:

Result 3 *The probability distribution of each discrete variable S_i in a non-standard mixed graphical association model, conditional on the state of all other variables, is*

$$\Pr(s_i|R_{S_i}) = \alpha \Pr(s_i|U_{S_i}) \prod_{j=1}^{\beta} \Pr(c_j|U_{C_j}) \prod_{k=1}^{\delta} f_{C_k}(c_k|U_{C_k})$$

where α is a normalising constant, R_{S_i} is a realisation of the set of all nodes except S_i, U_{S_i} is the set of parents of S_i; there are β discrete children C_j of S_i and δ continuous children C_k of S_i; and the conditional density defined for a child C_k is f_{C_k}.

There is no equivalent result of Result 1 for non-standard models. The expression at (4.2) becomes

$$f(s_i|R_{S_i}) = \frac{f_{S_i}(s_i|U_{S_i}) \prod_{j=1}^{\beta} f_{C_j}(c_j|U_{C_j})}{\int f_{S_i}(s_i|U_{S_i}) \prod_{j=1}^{\beta} f_{C_j}(c_j|U_{C_j})\, ds_i}.$$

With a model in which the f_{S_i} are all conditional Gaussian distributions, calculation of the integral proves simple; there is however no such general expression for other definitions. The integral can be evaluated numerically, but this is computationally very expensive, and it does not necessarily mean that the resulting $f(s_i|R_{S_i})$ will be easy to sample from. Thus it will be necessary to regard the integral as an unknown constant. This leads to

$$f(s_i|R_{S_i}) = \alpha f_{S_i}(s_i|U_{S_i}) \prod_{j=1}^{\beta} f_{C_j}(c_j|U_{C_j}) \tag{5.1}$$

where the integral has been replaced by α^{-1}.

It is worth noting at this point that when analysing a non-standard model, Result 1 *does* apply to a continuous variable if: (a) the variable has a conditional Gaussian distribution and linear mean function; and (b) the variable has no children with non-Gaussian conditional distributions or non-linear mean functions. Also, if a continuous variable has no children then a value can be generated for it directly from its conditional distribution (if possible) since (5.1) merely becomes

$$f(s_i|R_{S_i}) = f_{S_i}(s_i|U_{S_i}).$$

There are various candidates for sampling from $f(s_i|R_{S_i})$. Using ordinary rejection sampling (Ripley, 1987) is generally not feasible, as this method requires the determination of the maximum value of the ratio of two density functions at each Gibbs step and this is forbiddingly expensive. An alternative form of rejection sampling, Adaptive Rejection Metropolis Sampling (ARMS) (Gilks *et al.*, 1992), which removes the necessity for the calculation of the maximum value by creating its own "envelope", was tried on some test non-standard models; while initially the method seemed promising, once a variable was introduced with continuous parents *and* children, ARMS failed, unfortunately in contradiction to the comments in Brewer and Aitken (1993).

Two methods did however prove successful; the Metropolis-Hastings (M-H) algorithm and the auxiliary variables method. Sections 5.2.3 and 5.2.4 will now define these procedures.

5.2.3 Metropolis-Hastings

At each step of the simulation, the Metropolis-Hastings algorithm is used to obtain *one* sample from the Gibbs sampler. The following description has been adapted from Smith and Roberts (1993).

Suppose we wish to generate a sample from a density $\pi(x)$. The M-H algorithm works by constructing a Markov chain with equilibrium distribution $\pi(x)$ and transition probability $p(x, x')$ of moving from the current state x to the proposed new state x', which is drawn from a generating distribution $q(x, x')$—for the moment an arbitrary transition function. With probability $a(x, x')$, x' is accepted, or else the next step in the Markov chain is set to the current value x. The transition probabilities are given by:

$$p(x, x') = \begin{cases} q(x, x')a(x, x') & \text{if } x' \neq x \\ 1 - \sum_{x''} q(x, x'')a(x, x'') & \text{if } x' = x \end{cases}$$

and

$$a(x, x') = \begin{cases} \min\left\{\frac{\pi(x')q(x', x)}{\pi(x)q(x, x')}, 1\right\} & \text{if } \pi(x)q(x, x') > 0 \\ 1 & \text{if } \pi(x)q(x, x') = 0. \end{cases}$$

The choice of generating distribution q is not necessarily straightforward. Smith and Roberts (1993) note that "[i]nsight from general importance sampling methodology suggests that Student t- or split Student t-forms with small degrees of freedom are likely to be good candidates."

Translating this description to the terminology of mixed graphical models, consider the generation of a new value for S_i. The target distribution $\pi(x)$ is the Gibbs sampler $f(s_i|R_{S_i})$. A proposed new value s_i' is drawn from q, and then $a(s_i, s_i')$ is compared with a random number generator to see if S_i is set to s_i' or the current value s_i. After deciding on a value for S_i, the next variable in the simulation run is considered.

5.2.4 Auxiliary Variables

The following description of the auxiliary variables method has been adapted from Besag and Green (1993).

The method involves augmenting a variable x with one or more additional variables $\mathbf{u} = \{u_j,\ j = 1, 2, \ldots, k\ ; k \geq 1\}$, which, given x, are conditionally independent. The joint distribution of x and the $\{u_j\}$ is defined by taking $\pi(x)$ as the marginal for x and specifying $\pi(u_j|x)$. A Markov chain is then constructed: first the u_j are drawn from $\pi(u_j|x)$; then x' is generated given the u_j and the current state x. Besag and Green (1993) show that choosing a suitable transition function (e.g. $\pi(x'|\mathbf{u})$) preserves $\pi(x)$ as the equilibrium distribution.

Besag and Green (1993) go on to consider the case when $\pi(x)$ can be written in the form

$$\pi(x) = \alpha\,\pi_0(x)\prod_j b_j(x), \tag{5.2}$$

where it is straightforward to sample from $\pi_0(x)$. Introducing an auxiliary variable u_j for each term $b_j(x)$ and defining $\pi(u_j|x)$ to be the uniform distribution on $[0, b_j(x)]$ gives

$$\begin{aligned} \pi(x,\mathbf{u}) &= \pi(x)\prod_j \pi(u_j|x) \qquad \textit{(using conditional independence)} \\ &= \alpha\,\pi_0(x)\prod_j b_j(x)\{\mathrm{I}\,[0 \le u_j \le b_j(x)]b_j(x)^{-1}\} \qquad (5.3) \\ &= \alpha\,\pi_0(x)\,\mathrm{I}\,[\cap_j\{0 \le u_j \le b_j(x)\}], \end{aligned}$$

where I[] is the indicator function. Sampling from $\pi(x,\mathbf{u})$ now just involves sampling from $\pi_0(x)$ and imposing the constraints $\{b_j(x) \ge u_j\}$ by rejection. Note that the constant α can be incorporated into one of the product terms: it cancels out with $b_j(x)b_j(x)^{-1}$ at (5.3). Also consider sampling $u_j^* = \alpha u_j$ from $[0, \alpha b_j(x)]$; this is the same as sampling u_j from $[0, b_j(x)]$, and α also cancels in the rejection test $\alpha b_j(x) \ge u_j^* = \alpha u_j$. Thus we can write

$$\pi(x,\mathbf{u}) = \pi_0(x)\,\mathrm{I}\,[\cap_j\{0 \le u_j \le b_j(x)\}].$$

Now consider a non-standard graphical model. Note that (5.2) is in the same form as (5.1) with $\pi(x) = f(s_i|R_{S_i})$, $\pi_0(x) = f_{S_i}(s_i|U_{S_i})$ and $b_j(x) = f_{C_j}(c_j|U_{C_j})$. To sample from $f(s_i|R_{S_i})$ therefore, first generate the β auxiliary variables u_j from $U[0, f_{C_j}(c_j|U_{C_j})]$ (using the current value s_i to calculate $f_{C_j}(c_j|U_{C_j})$), and then get s_i' from the density $f_{S_i}(s_i|U_{S_i})$. The rejection tests are performed, and

s_i' is only accepted if $f_{C_j}(c_j|U_{C_j}) \geq u_j$ (using s_i') for each $j = 1, \ldots, \beta$. If it is rejected, the procedure is repeated until a proposed s_i' is accepted, and then the next variable in the simulation run is considered.

5.3 Examples

The examples in this section consider two types of "non-standard variable":

- A variable that has a conditional Gaussian distribution, but with a *quadratic* mean function; that is, a variable S_i with ϵ continuous parents $P_1, P_2, \ldots, P_\epsilon$ has a mean function g_{S_i} of the form

$$g_{S_i}(p_1, p_2, \ldots, p_\epsilon) = \phi + \sum_{t=1}^{\epsilon} \psi_t p_t + \theta_t p_t^2.$$

- A variable S_i that has a conditional Gamma distribution $\mathrm{Ga}(\alpha, \beta)$, which has mean $\alpha\beta$ and variance $\alpha\beta^2$, i.e.

$$f_{S_i}(s_i|U_{S_i}) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}.$$

  A (linear) mean function μ_{S_i} and variance σ^2 are defined for this variable; the parameters α and β can then be calculated as

$$\alpha = \frac{\mu_{S_i}^2}{\sigma^2} \qquad \text{and} \qquad \beta = \frac{\sigma^2}{\mu_{S_i}}.$$

  Note that during the simulation, when generating a value for S_i, new values of α and β must be calculated for each iteration, since the state of parent variables, and hence the value of the mean function, may have changed.

The first two examples that now follow have very simple graphical structures, and are very small. This is so that the exact marginal means and variances can be calculated, and compared with the results obtained from simulation. The third is a more serious application of the methodology described in this chapter.

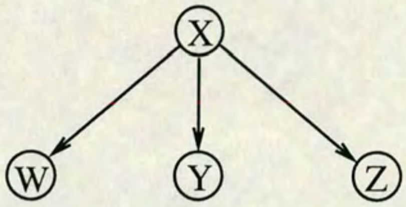


Figure 5–1: *Structure for quadratic relationship and conditional Gamma examples.*

5.3.1 Quadratic Relationship Example

This example has the influence diagram shown in Figure 5–1, and four variables are defined as follows, some having quadratic mean functions:

$$X \sim N(4,1);$$

$$W \,|\, x \sim N(1+x+x^2,1); \quad Y \,|\, x \sim N(3-x-x^2,0.5); \quad Z \,|\, x \sim N(7+x+2x^2,1);$$

This example has been analysed using both the Metropolis-Hastings and the auxiliary variables approaches. Note that generating values for variables W, Y and Z merely involves sampling from their defined conditional Gaussian distributions with the current value of variable X put into the relevant mean function.

For the M-H algorithm, the Student t-distribution with 5 degrees of freedom was used as the generating distribution q, as this appeared to give good results. The t-distribution was centred on the current value of the variable in question; the M-H method generally produces a rather "slow moving" chain, so this is a not unreasonable tactic.

For the auxiliary variables, the function $\pi_0(x)$ of equation (5.2) in this case is simply a Gaussian probability density function.

The simulation results are shown, along with the correct marginal means and variances, in Table 5–1. The numbers of simulation runs needed here were

	Simulation Runs (000's)						Correct	
Node	100		1000		2000		Values	
(M-H)	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
X	3.82	0.49	4.33	0.87	3.95	0.90	4	1.0
W	19.95	39.53	24.95	77.83	21.44	71.76	22	84.0
Y	14.29	23.66	18.29	48.37	15.54	43.27	16	51.5
Z	41.06	138.18	50.57	276.66	43.93	253.32	45	298.0
(Aux.)	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
X	3.79	0.85	3.96	0.96	4.00	0.98	4	1.0
W	20.12	70.10	21.56	78.23	22.00	83.37	22	84.0
Y	14.54	41.49	15.65	47.55	16.00	51.25	16	51.5
Z	41.46	246.72	44.16	276.88	45.00	295.85	45	298.0

Table 5–1: *Marginal means and variances for quadratic relationship model using both M-H and auxiliary variables approaches.*

very large; the quadratic relationships "magnify" the variances. Whilst for a given number of simulation runs the auxiliary variable method seems to perform slightly better, the rejection tests mean that the M-H version runs faster.

Using a C program on a Sequent Computer Systems model S81 under a Unix operating system, and using NAG C random number generators, the 100,000 runs for M-H took around 4 CPU minutes, and for auxiliary variables about 5 CPU minutes. There was very little problem with implementation in this case; for example, with the M-H, although q was chosen to be the t_5-distribution, other numbers of degrees of freedom seemed to work equally well.

5.3.2 Conditional Gamma Example

This example has variables given a conditional Gamma distribution. The structure of the influence diagram is the same as for the previous example, and thus is also illustrated by Figure 5–1. The (conditional) means and variances for the variables are defined as follows:

X: mean 2, variance 2;

W: mean $(2+x)$, variance $(2+x)$;

Y: mean x, variance x;

Z: mean $(1+x)$, variance $(1+x)$.

Note that in this case, since the means and variances are equal for each variable, the value for β with each Gamma distribution will be 1. The corresponding distributions for the variables are

$$X \sim \mathrm{Ga}(2,1);$$

$$W \,|\, x \sim \mathrm{Ga}(2+x,1); \qquad Y \,|\, x \sim \mathrm{Ga}(x,1); \qquad Z \,|\, x \sim \mathrm{Ga}(1+x,1).$$

Note that when performing stochastic simulation on this model, values for W, Y and Z are generated by sampling directly from the relevant Gamma density. As in section 5.3.1, one of the M-H or auxiliary variables methods must be used for X.

For the auxiliary variables method, here the function $\pi_0(x)$ of equation (5.2) is a Gamma probability density function; routines exist for sampling from Gamma distributions (see Ripley, 1987, for example), so this should not prove a problem.

The choice of generating distribution q for the Metropolis-Hastings algorithm was more difficult. The t-distribution (centred on the current value as before) with various numbers of degrees of freedom was tried, along with a Gaussian distribution with a selection of variances. None of these proved successful; even for very large numbers of simulation runs, the marginal means and variances were incorrect. In fact, it seemed as though the simulation was converging to marginal means and variances different from the correct values, obtained by direct calculation (possible with this very simple example).

	Simulation Runs (000's)						Correct	
Node	10		20		50		Values	
(M-H)	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
X	1.93	1.73	1.97	2.09	1.98	1.94	2	2
W	3.90	5.58	3.97	6.09	3.98	5.90	4	6
Y	1.92	3.67	1.98	4.19	1.99	3.91	2	4
Z	2.92	4.73	2.98	5.04	2.99	4.98	3	5
(Aux.)	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
X	1.91	1.68	1.97	1.95	2.00	1.97	2	2
W	3.87	5.55	3.95	5.82	3.99	5.96	4	6
Y	1.92	3.58	1.96	3.90	2.00	3.95	2	4
Z	2.92	4.62	2.96	4.90	3.00	4.97	3	5

Table 5–2: *Marginal means and variances for conditional Gamma model using both M-H and auxiliary variables approaches.*

This problem was solved by taking q to be a Gamma distribution, with α and β equal to the current value of the mean and variance (if necessary, calculated as functions of continuous parent variables) respectively; in this case, $\alpha = 2$ and $\beta = 1$ for variable X.

The resulting marginal means and variances obtained are shown in Table 5–2. For this example, it can be seen that in terms of number of simulation runs, M-H and auxiliary variables perform quite similarly, but the auxiliary variables approach still has the edge on this basis. The respective times for 50,000 runs are also close: 2 CPU minutes for M-H, and a shade over 2 CPU minutes for auxiliary variables.

It should be noted at this stage that while the implementation of the auxiliary variables approach proved very straightforward for this example, that for the Metropolis-Hastings was not so; it took many hours of programming until a successful q was found.

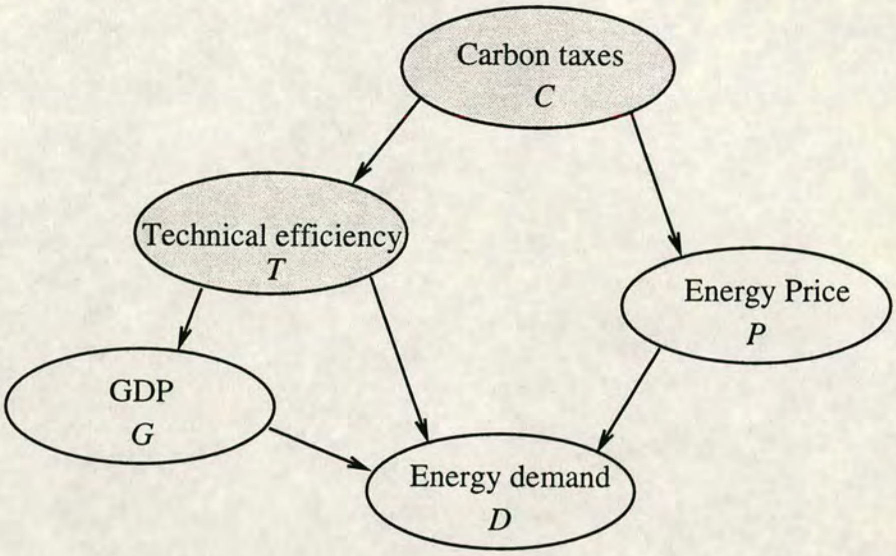


Figure 5–2: *Influence diagram for energy forecasting example.*

5.3.3 Energy Demand Example

This example incorporates both of the variable types studied in sections 5.3.1 and 5.3.2, and concerns the forecasting of energy demand. The influence diagram is illustrated in Figure 5–2.

> A forecaster of national energy requirements sees energy demand being driven by changes in gross domestic production (GDP), because energy is a necessary input for production, and by price. The forecaster also believes that events which trigger large price increases also stimulate extra research on energy-efficient technology leading to reduction in GDP elasticity. The discrete variable "technical efficiency" has two possible levels—efficiency does or does not improve. The other discrete variable (considered important by the forecaster), carbon taxes, also has two levels—the taxes may or may not be introduced and their introduction will affect prices and may also directly stimulate improvements in efficiency.

The variables within the system are defined by the energy expert thus:

Carbon tax (Discrete) Denoted by C, taking values $c =$ *taxes introduced* and $\overline{c} =$ *taxes not introduced*, with $\Pr(C = c) = 0.10$.

Technical efficiency (Discrete) Denoted by T, taking values $t =$ *some change in technical efficiency* and $\overline{t} =$ *no change in technical efficiency*, with $\Pr(t \,|\, c) = 0.20$ and $\Pr(t \,|\, \overline{c}) = 0.01$.

Energy price (Continuous) Denoted by P, with distributions:

$$P \,|\, C = c \;\sim N(5.836, 0.2) \qquad\qquad P \,|\, C = \overline{c} \;\sim N(5.558, 0.2).$$

Gross Domestic Product (Continuous) Denoted by G, and given a Gamma distribution. $G \,|\, T = t$ has mean 2.334 and variance 0.04, giving $\mathrm{Ga}(136.2, 0.01714)$; $G \,|\, T = \overline{t}$ has mean 2.266 and variance 0.04, giving $\mathrm{Ga}(128.4, 0.01765)$.

Energy demand (Continuous) Denoted by D, with distributions:

$$D \,|\, T = t, p, g \;\sim N(1.091 - 0.115p + 0.40g + 0.017g^2, 0.05)$$

$$D \,|\, T = \overline{t}, p, g \;\sim N(1.091 - 0.115p + 0.45g + 0.015g^2, 0.05).$$

The simulation procedure for this example will be studied in detail. The Gibbs sampler for the two discrete variables are (taking R_X to mean the set of all the variables except X)

$$\Pr(C \,|\, R_C) = a_1 \Pr(C) \Pr(t \,|\, C) f_P(p \,|\, C)$$

and

$$\Pr(T \,|\, R_T) = a_2 \Pr(T \,|\, c) f_G(g \,|\, T) f_D(d \,|\, T, p, g).$$

where are a_1 and a_2 are the appropriate constants. Note that since f_G is a Gamma density, calculating $f_G(g \,|\, T)$ for $T = t$ and $T = \overline{t}$ requires the evaluation of a Gamma integral, which is clearly computationally expensive. However, since for this particular example there are only two possible values of the Gamma parameter α, it will only be necessary to calculate the integral twice. The NAG C subroutine library contains a function for evaluating Gamma integrals.

The generation of new values for variable D is straightforward, since it has no children. The relevant mean is calculated for the current values of T, P and G and then the new D is sampled from the Gaussian distribution.

Variables P and G are of interest to us. They have "non-standard" children, and will need the hybrid strategies. The Gibbs samplers for these variables are

$$f(P \mid R_P) = a_3 f_P(P \mid c) f_D(d \mid t, P, g)$$

and

$$f(G \mid R_G) = a_4 f_G(G \mid t) f_D(d \mid t, p, G).$$

Thus for the auxiliary variables method, the function $\pi_0(x)$ of equation (5.2) is a Gaussian density for P and a Gamma density for G. Values are sampled from these distributions, and constraints from variable D (via f_D) are imposed by rejection as described in section 5.2.4.

The generating functions q for the Metropolis-Hastings are: the Student t-distribution with 5 degrees of freedom for P; and either $\text{Ga}(136.2, 0.01714)$ or $\text{Ga}(128.4, 0.01765)$ for G, when $T = t$ or $T = \bar{t}$ respectively. These are chosen as a consequence of the exploratory studies described in §5.3.1 and §5.3.2.

With a non-standard model, it may not be possible to calculate the marginal means and variances exactly[2]. However, it is possible to obtain good estimates (for models with as yet no variables observed) using logic sampling (Henrion, 1986), as described in section 4.2, with very many simulation runs. Table 5–3

[2] After all, this is one of the main reasons for using stochastic simulation in the first place.

			Simulation Runs (000's)					Logic
Method	Node	Marginal	1	5	10	20	50	Sampling
Auxiliary	C	Prob.	0.102	0.103	0.101	0.100	0.100	0.100
variables	T	Prob.	0.033	0.030	0.030	0.029	0.029	0.029
	P	μ	5.589	5.596	5.587	5.587	5.584	5.586
		σ^2	0.210	0.207	0.204	0.211	0.208	0.212
	G	μ	2.282	2.270	2.265	2.267	2.267	2.268
		σ^2	0.038	0.040	0.040	0.040	0.039	0.040
	D	μ	1.556	1.544	1.538	1.545	1.543	1.544
		σ^2	0.071	0.062	0.065	0.064	0.063	0.064
Metropolis	C	Prob.	0.094	0.101	0.101	0.099	0.100	0.100
-Hastings	T	Prob.	0.029	0.029	0.028	0.029	0.029	0.029
	P	μ	5.566	5.579	5.587	5.588	5.586	5.586
		σ^2	0.212	0.216	0.211	0.217	0.213	0.212
	G	μ	2.278	2.265	2.268	2.267	2.266	2.268
		σ^2	0.038	0.039	0.039	0.039	0.041	0.040
	D	μ	1.550	1.538	1.543	1.540	1.542	1.544
		σ^2	0.063	0.062	0.064	0.063	0.064	0.064

Table 5–3: *Simulation for energy demand example using both auxiliary variables and Metropolis-Hastings.*

shows the results of simulation for the M-H and auxiliary variables approaches, and compares these with the results obtained with logic sampling.

Here we see that in terms of number of simulation runs, the two approaches have a similar performance level. Table 5–4 displays the time taken for the C programs performing the simulation (again using NAG C subroutines) in CPU seconds. The times taken are clearly very similar as well.

Another useful comparison is to consider the percentages of proposed new values that are actually accepted by the two methods. The comparison is not straightforward, since the M-H keeps the old value if the rejection test fails, while the auxiliary variables method suggests new values until one is accepted.

	Simulation Runs (000's)				
Method	1	5	10	20	50
Auxiliary variables	5.1	24.6	52.3	115.6	205.1
Metropolis-Hastings	4.4	24.1	44.9	104.6	203.5

Table 5–4: *Times to run for both methods in CPU seconds (each entry an average of 5 runs).*

For 50,000 simulation runs, the auxiliary variables procedure accepted 54% of proposed values for P and 35% for G, while the M-H accepted 50% for P and 80% for G. While this suggests that the M-H is doing a lot better, the relative simplicity of the rejection tests with auxiliary variables cancels out the time that might be saved here.

Figures 5–3 to 5–5 display plots of the kernel density estimates for the continuous variables P, G and D. Figure 5–3 shows the estimates using simulated observations from logic sampling; Figure 5–4 shows the estimates for Metropolis-Hastings; and Figure 5–5 shows those for the auxiliary variables approach. The kernel density estimation does not clearly reveal the fact that for G and P, the marginal densities are just mixtures of two Gamma and two Gaussian densities respectively. This is because the two distributions in each case are very close together; making the window-width smaller simply increases the noise level in the plots, and hence any genuine "bimodal" characteristics become indecipherable. Also note that there seems to be very little (if any) difference in the plots between the three methods.

There seems little to choose between the two methods with these experiments. It is certainly the case that once the procedures have been set up appropriately, Metropolis-Hastings and auxiliary variables perform equally well. However, the M-H took far longer to set up; the choice of generating distribution q was the sticking point.

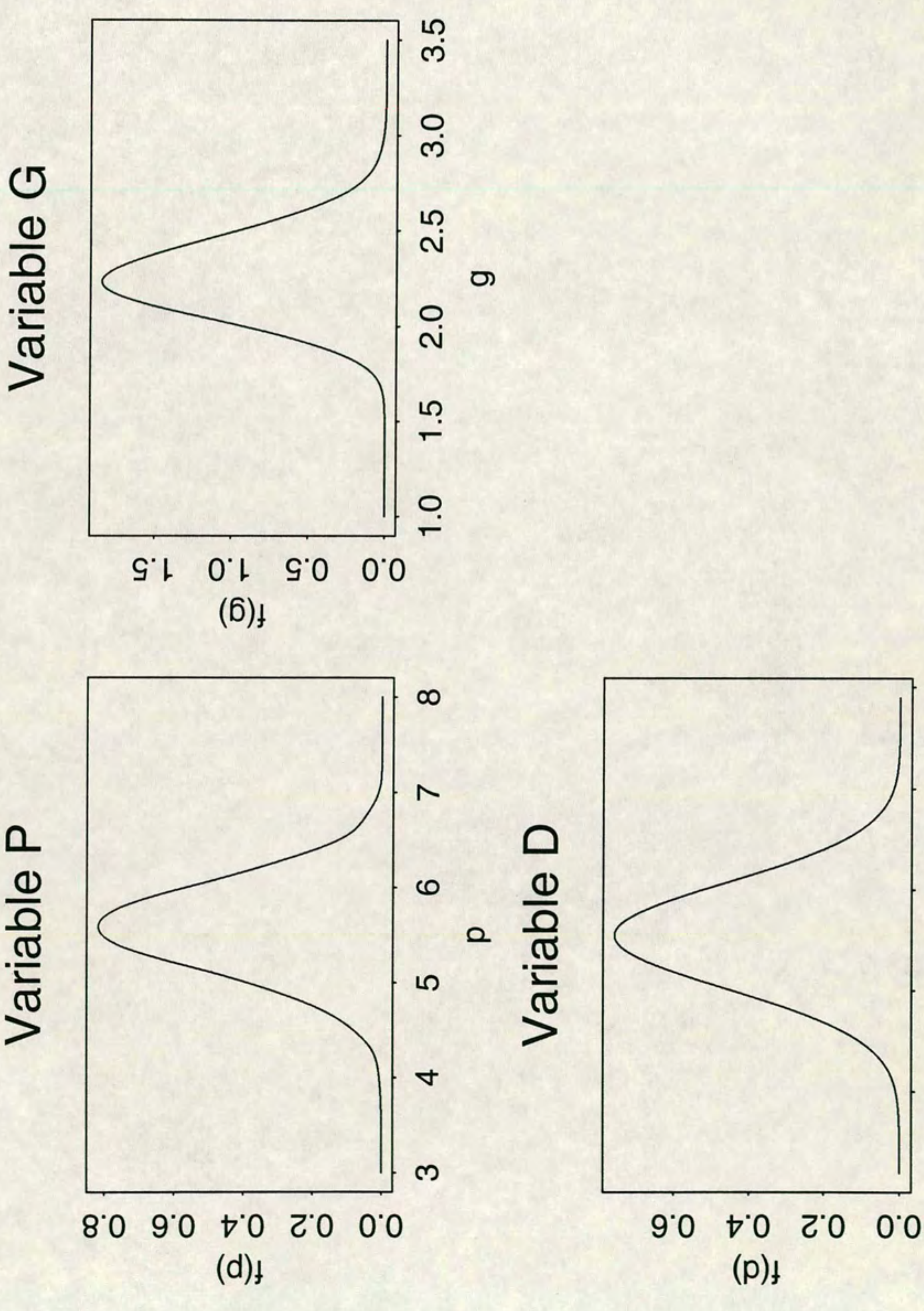


Figure 5–3: *Kernel density estimates for variables P, G and D from simulated values obtained by logic sampling.*

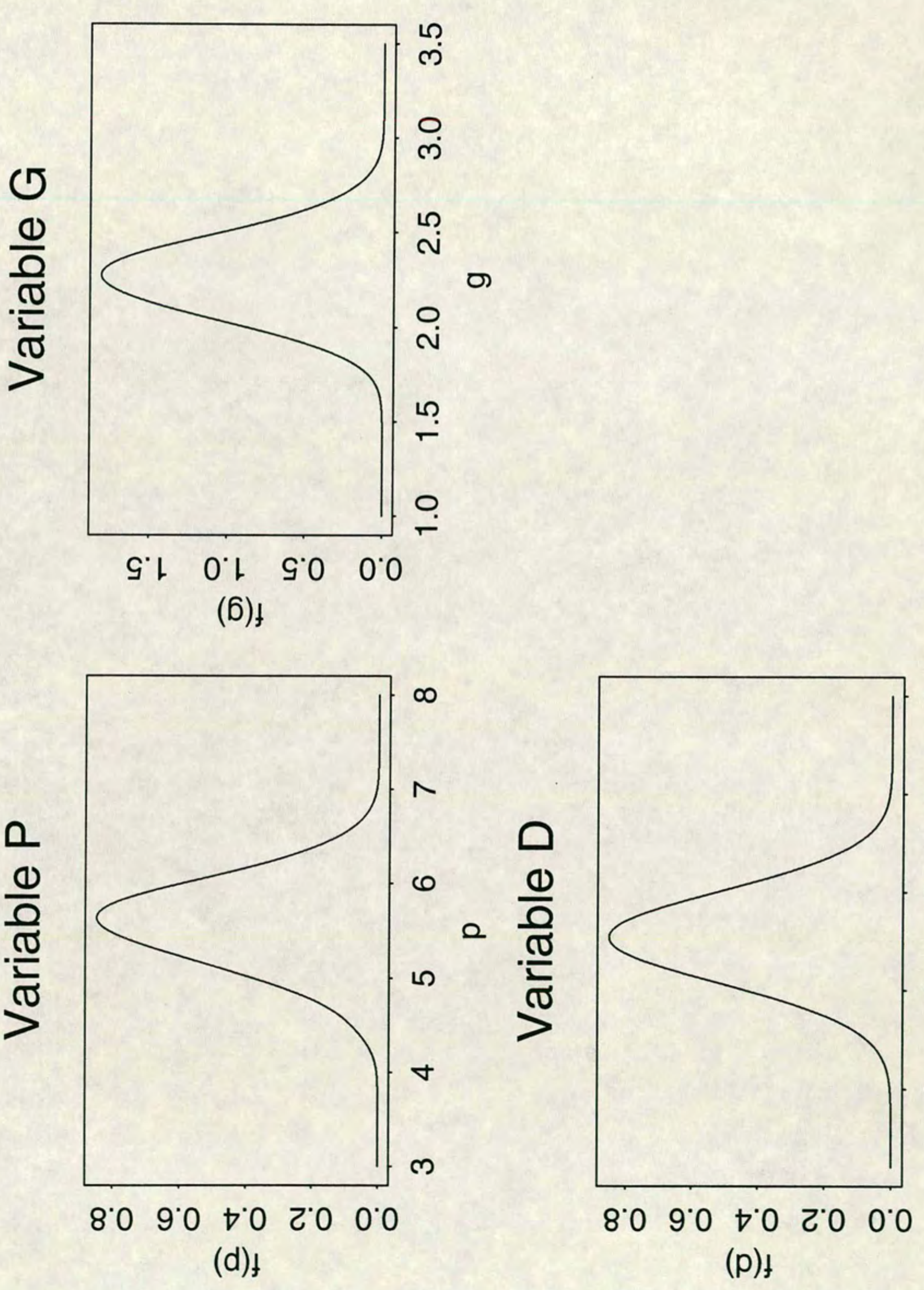


Figure 5–4: *Kernel density estimates for variables P, G and D from simulated values obtained by the Metropolis-Hastings method.*

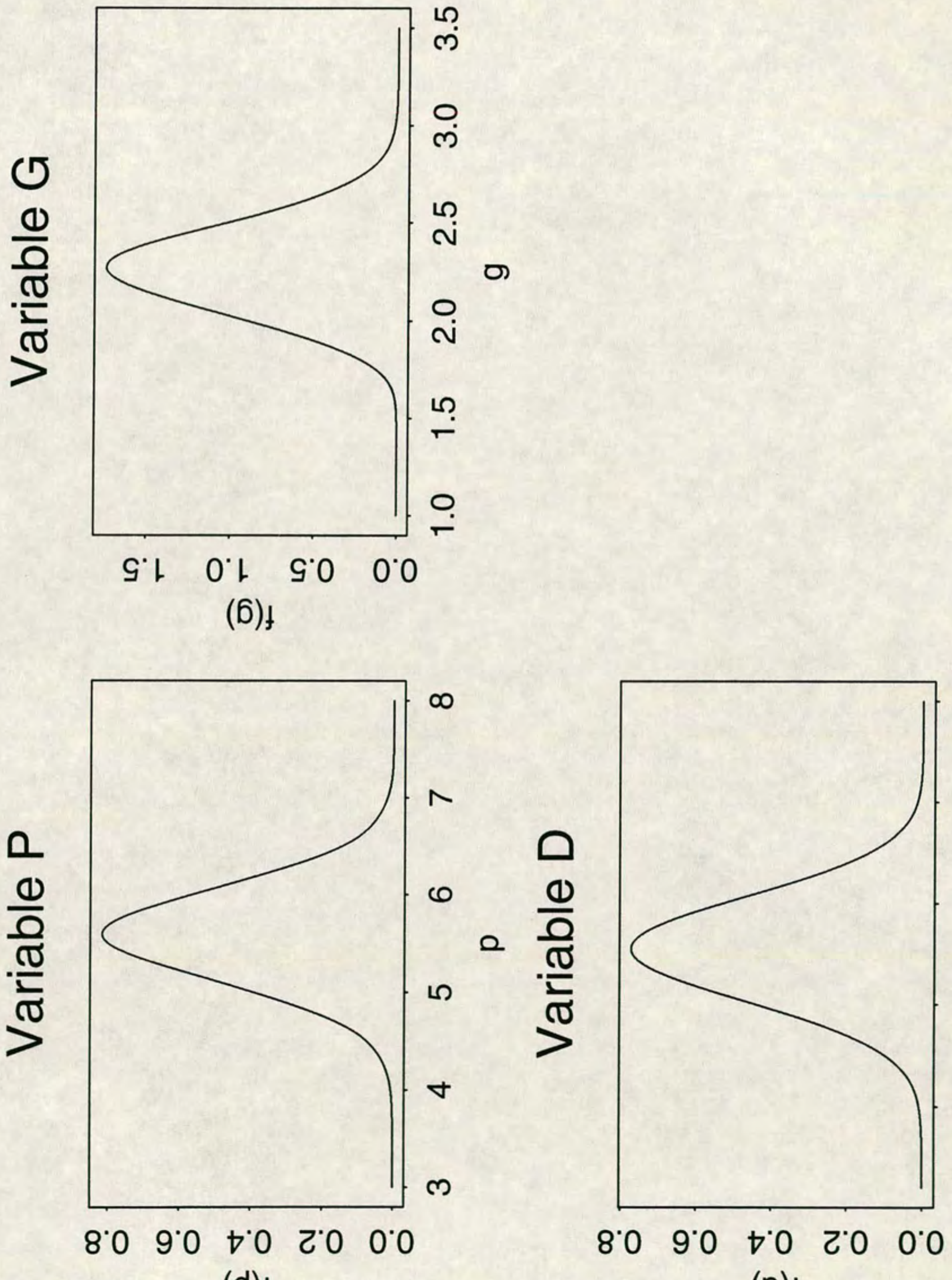


Figure 5–5: *Kernel density estimates for variables P, G and D from simulated values obtained by the auxiliary variables method.*

Finding a suitable q for M-H is not obvious in general, whereas with the auxiliary variables the "generating distribution" $\pi_0(x)$ falls directly out of the relevant formula. For this reason, and given comparable performances by both methods, the auxiliary variables approach is preferred over the Metropolis-Hastings for analysing mixed graphical association models.

5.4 Conclusions

This chapter has extended the methods of stochastic simulation for mixed graphical association models. Use of the Metropolis-Hastings algorithm and an auxiliary variables method for the Gibbs sampling has enabled analysis of models that contain non-Gaussian conditional distributions and non-linear mean functions.

The Metropolis-Hastings and auxiliary variables methods showed comparable levels of performance once implemented. However, it is in the implementation itself that a real difference between the two became clear. While the generating distribution required for auxiliary variables appears as a natural consequence of equation (5.2), that for M-H requires a significant degree of research.

For this reason, the auxiliary variables approach is recommended as a reliable method for analysing non-standard models.

Chapter 6

Conclusions

This thesis has studied aspects of graphical models. Chapter 1 described the notation and properties of such models, and summarised recent developments in their analysis.

Chapter 2 presented a review of the use of graphical models in legal reasoning (and related) literature. It was seen that the use of "graphs" in this context can be traced as far back as 1913; some suggestions were made for converting "degrees of belief" into probabilities to be used in a graphical model.

The use of likelihood ratios is important in legal reasoning, as they may be easier for juries to understand than probabilities. It was shown in section 2.5 that likelihood ratios can be used as input to simple propagation procedures, with certain computational and formulaic advantages over similar work using probabilities as input.

The problem of recovering graphical structure from a data set was the subject of Chapter 3. Kernel methods were applied to the tree structure learning process, but these had very limited (if any) advantages over the far simpler sample frequencies (for discrete variables) and correlation coefficient (for continuous variables) approaches. The kernel methods were, on the whole, rejected due to the much greater time needed for the programs to run. This was especially true for the continuous case, where the kernel methods, needing a number of numerical double integrations of kernel functions, performed very badly indeed.

The recovery of polytrees for discrete models was also not greatly improved by a kernel method. A new algorithm for recovering more general graph structure (to allow undirected loops) was analysed, and proved reasonably successful.

The main work of this thesis was covered in Chapters 4 and 5. In Chapter 4 a stochastic simulation procedure for estimating marginals of variables in graphical models using a Gibbs sampler was described, and extended to cover continuous variables with conditional Gaussian distributions. This procedure was shown to be successful, and to be efficient computationally. The simulated observations allow the application of a non-parametric density estimation method to estimate the marginal probability density functions, something quite difficult to do under the exact computation framework of Chapter 1. A procedure for allowing discrete variables to have continuous parents within the graphical structure in such models was also discussed. A certain problem associated with Gibbs sampling, namely it's occasional inability to move from one mode to another, was shown to be surmountable with a slight adaptation of the stochastic simulation algorithm.

Chapter 5 extended the stochastic simulation scheme further to include variables with non-Gaussian conditional distributions and non-linear mean functions. This scheme needed the inclusion of a further Markov chain Monte Carlo routine to generate a value from the Gibbs sampler; my preference for an approach based on auxiliary variables was noted.

It has been shown therefore that stochastic simulation is a valid and viable method of analysing graphical models, and that further, it allows extensions to the set of models that can be defined.

Bibliography

Aitchison J and Aitken CGG (1976) Multivariate binary discrimination by the kernel method. *Biometrika* **63** No. 3, 413–420.

Aitken CGG (1979) The kernel method of density estimation with applications in discrimination, selection of features, and conditional and marginal distributions. *Ph.D. thesis, University of Glasgow, Glasgow.*

Aitken CGG and Gammerman AJ (1989) Probabilistic reasoning in evidential assessment. *Journal of the Forensic Science Society* **29** 303–316.

Aitken CGG and Gammerman AJ (1990) An Illustrative Example of the Use of Probabilistic Reasoning in Agricultural Forecasting. *Technical Appendix to progress report: DOSES Project B6, Likely Phase 2*, 3–11.

Besag J and Green PJ (1993) Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society* B **55** 25–38.

Birch MW (1963) Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society* **25** 220–223.

Birch MW (1964) The detection of partial association I: the 2×2 case. *Journal of the Royal Statistical Society* **26** 313–324.

Brewer MJ and Aitken CGG (1993) Contribution to discussion of meeting on the Gibbs sampler and other Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society* B **55** No. 1, 69–70.

Brewer MJ, Aitken CGG, Luo Z and Gammerman AJ (1992) Stochastic Simulation in Mixed Graphical Association Models. *In proceedings of COMPSTAT 1992, Eds. Dodge Y and Whittaker J*, 257–262.

Chow CK and Liu CN (1968) Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* **IT-14** 462–467.

Chow CK and Wagner TJ (1973) Consistency of an estimate of tree-dependent probability distributions. *IEEE Transactions on Information Theory* **IT-19** 369–371.

Cooper GF (1984) NESTOR: A computer-based medical diagnostic aid that integrates causal and probabilistic knowledge. *Ph.D. diss., Department of Computer Science, Stanford University, USA.*

Edwards W (1962) Dynamic decision theory and probabilistic information processing. *Human Factors* **4** 59–73.

Edwards W (1991) Influence diagrams, Bayesian imperialism, and the Collins case: an appeal to reason. *Cardozo Law Review* **13** 1025–1074.

Edwards W, Phillips LD, Hays WL and Goodman BC (1968) Probabilistic information processing systems: design and evaluation. *IEEE Transactions on Systems Science and Cybernetics* **4** 248–265.

Edwards W, Schum DA and Winkler RL (1990) Murder and (of?) the likelihood principle: a trialogue. *Journal of Behavioral Decision Making* **3** No. 2, 75–89.

Fairley W and Mosteller F (1974) A conversation about Collins. *University of Chicago Law Review* **41** 242.

Finkelstein M and Fairley W (1970) A Bayesian approach to identification evidence. *Harvard Law Review* **83** 1021.

Gammerman AJ (1990) Computational models of diagnostic reasoning. *UNICOM Seminars, London* 12–15.

Gammerman AJ and Luo Z (1991) Constructing causal trees from a medical database. *Technical Report TR91002, Department of Computer Science, Heriot-Watt University, Edinburgh.*

Gelfand AE and Smith AFM (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85** No. 410, 398–409.

Gelman A and Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statistical Science* **7** No. 4, 457–472.

Geman S and Geman D (1984) Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

Geyer CJ (1992) Practical Markov chain Monte Carlo. *Statistical Science* **7** No. 4, 473–482.

Gilks WR, Best NG and Tan KKC (1992) Adaptive rejection sampling from non log-concave densities. *Technical report, Medical Research Council Biostatistics Unit, Cambridge.*

Goodman LA (1970) The multivariate analysis of qualitative data. *Journal of the American Statistical Association* **65** 226–256.

Haberman SJ (1974) The analysis of frequency data. *University Press, Chicago, USA.*

Henrion M (1986) Propagating uncertainty by logic sampling in Bayes' networks. *Technical report, Department of Engineering and Public Policy, Carnegie-Mellon University, USA.*

Jensen FV, Lauritzen SL and Olesen KG (1990) Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly* **5** 269–282.

Kiiveri H and Speed TP (1982) Structural analysis of multivariate data: a review. *In* Leinhardt S *(Ed.) Sociological Methodology. Jossey Bass, San Francisco.*

Kiiveri H, Speed TP and Carlin JB (1984) Recursive causal models. *Journal of the Australian Mathematics Society* A **36** 30–52.

Koehler (1991) Probabilities in the courtroom: an evaluation of the objections and policies. *In* Kagehiro D and Laufer W (eds.) *Handbook of Psychology and Law.*

Kullback S (1959) Information theory and statistics. *New York, Wiley.*

Kullback S and Liebler RA (1951) Information and sufficiency. *Annals of Mathematical Statistics* **22** 79–86.

Lauritzen SL (1992) Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association* **87** No. 420, 1098–1108.

Lauritzen SL and Spiegelhalter DJ (1988) Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society* B **50** No. 2, 157–224.

Lauritzen SL and Wermuth N (1984) Mixed interaction models. *Research Report. R-84-8. Institute of Electronic Systems, University of Aalborg, Denmark.*

Lauritzen SL and Wermuth N (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics* **17** 31–57.

Leimer HG (1989) Triangulated graphs with marked vertices. *In* Graph theory in memory of GA Dirac *(LD Andersen* et al.*, eds.), Annals of Discrete Mathematics* **41** 311–324.

Luo Z (1992) A probabilistic reasoning and learning system based on Bayesian Belief Networks. *Ph.D. thesis, Department of Computer Science, University of Heriot Watt, Edinburgh.*

Martin AW (1980) A general algorithm for determining likelihood ratios in cascaded inference. *Research Report. #80-03. Department of Psychology, Rice University, Houston, Texas 77001, USA.*

Morgan E (1961) Basic problems of evidence. pp 185–186.

Niles HE (1922) Correlation, causation and Wright theory of "Path Coefficients". *Genetics* **7** 258–273.

Normand S-L (1993) Comment to Spiegelhalter DJ, Dawid AP and Lauritzen SL (1993) "Bayesian analysis in expert systems". *Statistical Science* **8** No. 3, 263–265.

Pearl J (1987) Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence* **32** 245–257.

Pearl J (1988) Probabilistic reasoning in intelligent systems. *San Mateo, California, Morgan Kaufmann.*

Pearl J and Dechter R (1989) Learning structure from data: a survey. *Technical Report R-132, Computer Science Department, University of California, Los Angeles, California, USA.*

Rebane G and Pearl J (1987) The recovery of causal poly-trees from statistical data. *Proceedings of 3rd Workshop on Uncertainty in AI, Seattle*, 222–228.

Ripley BD (1987) Stochastic simulation. *Wiley, Chichester, England.*

Ritter C and Tanner MA (1992) Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy Gibbs Sampler. *Journal of the American Statistical Association* **87** No. 419, 861–868.

Silverman BW (1986) Density estimation for statistics and data analysis. *Chapman and Hall, London.*

Schum DA (1987) Evidence and inference for the intelligence analyst. *Lanham, Maryland, University Press of America.*

Schum DA (1989a) Inference networks and their many subtle properties. *Information and Decision Technologies.*

Schum DA (1989b) Knowledge, probability, and credibility. *Journal of Behavioral Decision Making* **2** No. 1, 39–62.

Smith AFM and Roberts GO (1993) Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society* B **55** 3–24.

Tillers P and Schum DA (1988) Charting new territory in judicial proof: beyond Wigmore. *Cardozo Law Review* **9** No. 3, 907–966.

Tribe L (1971) Trial by mathematics: precision and ritual in the legal process. *Harvard Law Review* **84** 1329.

Twining W (1984) Taking facts seriously. *Journal of Legal Education* **34** 22-42.

Wermuth N and Lauritzen SL (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society* B **52**, No. 1, 21–50.

Whittaker J (1990) Graphical models in applied multivariate statistics. *Chichester, UK, Wiley.*

Wigmore JH (1913) The problem of proof. *Illinois Law Review* **VIII** No. 2, 77–103.

Wigmore JH (1937) The science of judicial proof (3rd ed.). *Little, Brown and Co., Boston.*

Wright S (1921) Correlation and causation. *Journal of Agricultural Research* **20** 557–585.