

---

---

# Learning Probabilistic Networks

---

**Paul J Krause**

PHILIPS RESEARCH LABORATORIES  
CROSSOAK LANE  
REDHILL, SURREY RH1 5HA  
UNITED KINGDOM

## Abstract

A probabilistic network is a graphical model that encodes probabilistic relationships between variables of interest. Such a model records qualitative influences between variables in addition to the numerical parameters of the probability distribution. As such it provides an ideal form for combining prior knowledge, which might be limited solely to experience of the influences between some of the variables of interest, and data. In this paper, we first show how data can be used to revise initial estimates of the parameters of a model. We then progress to showing how the structure of the model can be revised as data is obtained. Techniques for learning with incomplete data are also covered. In order to make the paper as self contained as possible, we start with an introduction to probability theory and probabilistic graphical models. The paper concludes with a short discussion on how these techniques can be applied to the problem of learning causal relationships between variables in a domain of interest.

## 1. Motivation

Two well cited tutorials on learning probabilistic networks are extant. The first by Buntine [11] is primarily an annotated bibliography. The second by Heckerman (available in its most extended form as [37]) is quite technical. There does seem to be a need for a more accessible introductory paper. This paper aims to provide a broad survey of the field in a limited space, and as such has had to try and strike a balance between technical content and accessibility. The primary goal has been to produce a paper which will provide a reader with an introduction to the approaches that are being taken and the main issues arising: the references cited will need to be consulted in order to obtain full technical details.

A personal acknowledgement is also needed. This has been brought to the beginning of the paper in order to be clear about the lineage of this tutorial from the outset. This paper initially started as a joint effort between myself and David Heckerman. Because of this I have drawn on David's tutorials [36] and [37] at several points. In addition David has provided extensive comments on three early versions. However, at several points in this paper I put forward my own perspective of subjective probabilities as "expert judgements" of a true, physical, probability, rather than as a degree of belief. Whilst David encouraged me to air this viewpoint, it is not one to which he personally subscribes. In addition, as things turned out I essentially ended up writing the entire paper; although the topics of several sections are the same as David's tutorials, the content, wording and presentation style are very different. For these reasons, we agreed that I should be sole author. I mention this at some length as merely to "thank David Heckerman for

helpful comments on earlier drafts” would be a poor acknowledgement of his help and support; I simply could not have produced a paper that was worthy of submission without this help. Good discussions of subjective probabilities as degrees of belief can be found in, for example, [7], [36] and [37], and again I emphasise that this paper is necessarily but a first step on the road to understanding this subject. Look upon it as an exercise in scientific journalism.

## **2. Aims of the paper**

During the last two decades there has been a steadily expanding interest in the use of rigorous probabilistic inference as a technique for the development of expert systems. One track of research has addressed the development of efficient inference techniques for use in complex large-scale knowledge models. Of course, the use of such an inference technique presupposes the availability of a valid knowledge model. A second track of research has been targeted at the development of techniques to enable a combination of expert knowledge and data to be used to develop and further refine the desired knowledge models. This second track has rapidly matured during the first half of this decade. As a result, we feel the time is ripe to provide a tutorial overview of the current state of the art in the second track. This paper aims to provide such an overview in a form which will be accessible to those without detailed technical knowledge of the area.

In order to make the paper as self contained as possible, we start with a succinct revision of basic probability theory. The primary agenda behind this revision is to emphasise the role of probability as a measure. Sometimes, as we shall see, the probability of an event can be measured directly. But at other times, the data for such a measurement may not be available and we will need to rely initially on probabilities elicited from an expert, or experts. These numerical parameters are not, however, the sole components of a knowledge model. There will always be some structure to the influences between variables in a knowledge model. This structure is of key importance and will be the subject of sections 4 and 5.

The main question which we wish to answer in this paper is: what information can we use to derive the probabilities and structure for a knowledge model? If we use expert judgement to elicit the structure and the associated probabilities, can we update the probabilities in the light of experience? Can we update the structure? Indeed, could we even learn the structure and the probabilities directly from statistical data? Sections 6, 7 and 8 provide positive answers to the main and these ancillary questions. We will see that the influences between variables in a knowledge model can often be usefully considered in terms of “causal” influence when eliciting the structure of a model. If we are able to learn structure, the question that naturally arises is, can we learn causality? This final question will be discussed in section 9.

## **3. Introduction to Probability**

### **3.1 Probability as a measure**

A basic familiarity with probability theory is assumed for the purposes of this paper. However, for completeness we give the following definitions.

When using probability, one may talk in terms of: the probability that a cancer patient will respond to a certain form of chemotherapy; the probability that a projectile might hit a region of space; the probability of observing a string of three identical outcomes in six dice throws. We shall use the general term *sample point* to refer to the “things” we are talking about; an

abstraction of a cancer patient, a geometric point, a chance outcome.

A *sample space*, or *universe*, is the set of all possible sample points in a situation of interest. It is usual to use  $\Omega$  to designate a specific sample space. The sample points in a sample space must be mutually exclusive and collectively exhaustive.

A *probability measure*,  $p(\cdot)$ , is a function on *subsets* of a sample space  $\Omega$ . These subsets are called *events*. We can refer to the values of  $p(A)$ ,  $p(A \cup B)$ ,  $p(\Omega)$  as the probabilities of the respective events (for  $A, B \subseteq \Omega$ ). But the function  $p(\cdot)$  is a *measure* with the following properties:

**Definition:**

A *probability measure* on a sample space  $\Omega$  is a function mapping subsets of  $\Omega$  to the interval  $[0, 1]$  such that:

1. For each  $A \subseteq \Omega$ ,  $p(A) \geq 0$ .
2.  $p(\Omega) = 1$
3. For any countably infinite collection of *disjoint* subsets of  $\Omega$ ,  $A_k$ ,  $k = 1, \dots$ ,

$$p\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} p(A_k)$$

In general, we will need to check that the sets (events) themselves satisfy certain properties to ensure that they are *measurable*. More details on this can be obtained, for example, from [14], [95].

Sometimes we can estimate the probabilities we want by counting (the ratio of the number of cancer patients cured to the total number treated, for example) or some other form of direct measurement. In this sense, we are saying that probability is an attribute or a property of the real world. The term *physical probability* is often used to denote this interpretation of probability [32].

### 3.2 The reliability of measurements

Of course, any act of measurement has an element of imprecision associated with it. So, we would expect the probabilities obtained by measurement also to be imprecise; strictly, any physical probability should be represented by a distribution of possible values. In general, the more information we have, the tighter will be the distribution. Sometimes, however, we will have no direct physical measurements by which to estimate a probability. An example of such a case might be when one is asked to toss a coin one has never seen before and judge the probability that it will land heads up. If one believes the coin to be fair, an estimate of 1/2 for this physical probability would seem reasonable. For more complex situations, the value elicited for such a probability may vary from subject to subject; perhaps dependent on the level or relevance of expertise of the subjects.

Sometimes, a probability elicited in this way is taken as a measure of an expert's *belief* that a certain situation will arise. This can lead to extensive discussions as to whether experts and others do, or should, use the laws of probability to update their beliefs as new evidence becomes available. To avoid such discussions, one might just take the elicitation of such probabilities as an act of expert judgement. Whichever view one takes, probability theory offers a

gold standard by which the probability estimates may be revised in the light of experience. This is discussed clearly and at a fundamental level in Bernardo and Smith, [7], Chapter 4. It is also a theme expanded on through much of this paper.

We should be clear at this stage that our interest is in using whatever sources of information are available to us to produce a model of the real world. We are not interested in modelling the expert. Nevertheless, expert judgement can be a useful starting point for the ultimate derivation of physical probabilities.

### 3.3 Bayes' theorem

We have so far concentrated largely on the static aspects of probability theory. But probability is a dynamic theory; it provides a mechanism for coherently revising the probabilities of events as evidence becomes available. Conditional probability and Bayes' theorem play a central role in this. Again for completeness, we include a brief discussion of these here.

We write  $p(A | B)$  to represent the probability of event A (an hypothesis) conditional on the occurrence of some event B (evidence). If we are counting sample points, we are interested here in the fraction of events B for which A is also true; we are switching our attention from the universe  $\Omega$  to the universe B. From this it should be clear that (with the comma denoting conjunction of events):

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

Although often written in the form  $p(A, B) = P(A | B)p(B)$  and referred to as the “product rule”, this is in fact the simplest form of Bayes' Theorem. It is important to realise that this form of the rule is not, as often stated, a definition. Rather, it is a theorem derivable from simpler assumptions.

Bayes' Theorem can easily be rewritten in a form which tells us how to obtain a posterior probability in a hypothesis A after observation of some evidence B, given the *prior* probability in A and the likelihood of observing B were A to be the case:

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

This simple formula has immense practical importance on a domain such as diagnosis. It is often easier to elicit the probability, for example, of observing a symptom given a disease than that of a disease given a symptom. Yet, operationally it is usually the latter which is required.

In its general form, Bayes' Theorem is stated as follows.

#### **Proposition**

Suppose  $\bigcup_n A_n = \Omega$  is a partition of a sample space into disjoint sets. Then:

$$p(A_n | B) = \frac{p(B | A_n)p(A_n)}{\sum_n p(B | A_n)p(A_n)}$$

It is important to appreciate that Bayes' Theorem is as applicable at the 'meta-level' as it is at the domain level. It can be used to handle the case where the hypothesis is a proposition in the knowledge domain (a specific disease, perhaps) and the evidence is observation of some condition (perhaps a symptom). However, it can also handle the case where a hypothesis is that a parameter in a knowledge model has a certain value (or distribution of values) or that the model has a certain structure, and the evidence is some incoming case data.

### 3.4 Probability distributions on sets of variables

The points in a sample space may be very concrete. If we are considering an epidemiological study, they may be people; in the case of quality control on an assembly line, they may be specific electronic components. As such, they may possess certain qualities about which we are interested and which may be observed or measured in some way. So, for example, an electronic logic gate may be functional or non-functional, or it may have a certain weight (a distinction which may be of less interest from a purely functional point of view). We will refer to such a distinction, about which we may be uncertain, as a *variable*. A variable has a set of *states* corresponding to a mutually exclusive and exhaustive set of events. It may be *discrete*, in which case it has a finite or countable number of states, or it may be *continuous*. For example, we may use a discrete (binary) variable to represent the possible functioning or otherwise of a logic gate selected from a production line, and a continuous variable to represent its weight.

Strictly, a variable<sup>1</sup> (on  $\Omega$ ) is a function,  $X$  say, from sample points  $\omega \in \Omega$  to a domain representing the qualities or distinctions of interest. An element of randomness in  $X(\omega)$  is induced by the selection "at random" of the sample point  $\omega$ ; a specific logic gate, or a specific throw of a die. Once the sample point has been chosen, the outcome  $X(\omega)$  is fixed and can be measured, or otherwise determined. However, it is often the case that the elements of the underlying sample space can be implicitly understood, in which case an unadorned capital letter is used to represent the variable. Following this custom, in the remainder of this paper we will use upper-case letters to represent single variables (e.g.  $X$ ,  $Y$  and  $Z$ ). Lower case letters will be used for states, with, for example,  $X = x$  denoting that variable  $X$  is in state  $x$ .

Of course, some qualities of a sample point may be easier to determine than others. For example, although we can readily determine whether a logic gate is functional or not, determining the underlying cause of a non-functional gate is not normally possible without some invasive inspection of the device. But here experience might help us. Suppose we had carefully analysed a hundred non-functional devices and found sixty-five with a faulty bond between the encapsulated silicon chip and one of the connecting pins, and forty-five with a malfunctioning chip. Then given a new observation of a non-functioning gate, we can use these statistics to predict the chances of that gate having specific states for these not-so-easily observable qualities.

Real world problems are typically more complex than this. To move a little closer to a real example, figure 1 lists a set of variables which will have specific states for some person of interest (after [53]). The states of some of these variables, such as  $a$  (visit Asia) or  $s$  (Smoking), may be easy to determine. Our goal is to predict the most likely states of those variables that are harder to observe directly, such as  $l$  (Lung cancer) or  $b$  (Bronchitis), using the observed

---

1. The term "random variable" is often used. Here we reserve the term "random variable" for the situation where there are repeated observations, which is not strictly the case for the variables used in probabilistic networks as discussed in this paper.

A: visit to Asia  
 B: Bronchitis  
 D: Dyspnoae  
 L: Lung cancer  
 S: Smoking  
 T: Tuberculosis  
 X: positive X-ray

Figure 1: An hypothetical set of variables associated with a sample space of humans (after [53]).

states. We can do this if we are able to elicit a probability distribution  $p(A, B, D, L, S, T, X)$  over all the variables of interest. Yet even if each variable has only two states we will need to elicit  $(2^7-1)$  distinct values in order to define the probability distribution completely. This would require a massive data collection exercise if we were to hope to use physical probabilities, or alternatively make unreasonable demands on the domain experts if we were to think in terms of eliciting probabilities from them. Yet even this is a “toy” problem in relation to some of the real applications that are being built.

The problem is that in defining a *joint probability distribution*, such as  $p(A, B, D, L, S, T, X)$ , we need to assign probabilities to all possible events. We can, however, make the knowledge elicitation problem much more tractable if we exploit the structure that is very often implicit in the domain knowledge. The next two sections will expand on this.

#### 4. Graph theory

Very many problem domains can be structured through using a graphical representation. Essentially, one identifies the concepts or items of information which are *relevant* to the problem at hand (nodes in a graph), and then makes explicit the *influences* between concepts. This section introduces some of the terminology associated with the use of graphs.

At its most abstract, a graph  $\mathcal{G}$  is simply a collection of vertices  $V$  and edges  $E$  between vertices:  $\mathcal{G} = (V, E)$ . We can associate a graph  $\mathcal{G}$  with a set of variables  $U = \{X_1, X_2, \dots, X_n\}$  by establishing a one to one relationship between the nodes in the graph and the variables in  $U$ . One might, for example, label the nodes from 1 to  $n$ , with nodes being associated with the appropriately subscripted variable in  $U$ . An edge  $e(i,j)$  might be *directed* from node  $i$  to node  $j$ . In this case, the edge  $e(j,i)$  cannot simultaneously belong to  $E$  and we say that node  $i$  is a *parent* of its *child* node  $j$ . If both  $e(i,j)$  and  $e(j,i)$  belong to  $E$ , we say the edge is undirected<sup>1</sup>.

For example, the graph shown in figure 2 is associated with the set of variables introduced in

---

1. There are graphs in which there can be directed arcs from  $i$  to  $j$  and from  $j$  to  $i$ . However, we do not consider them in this paper.

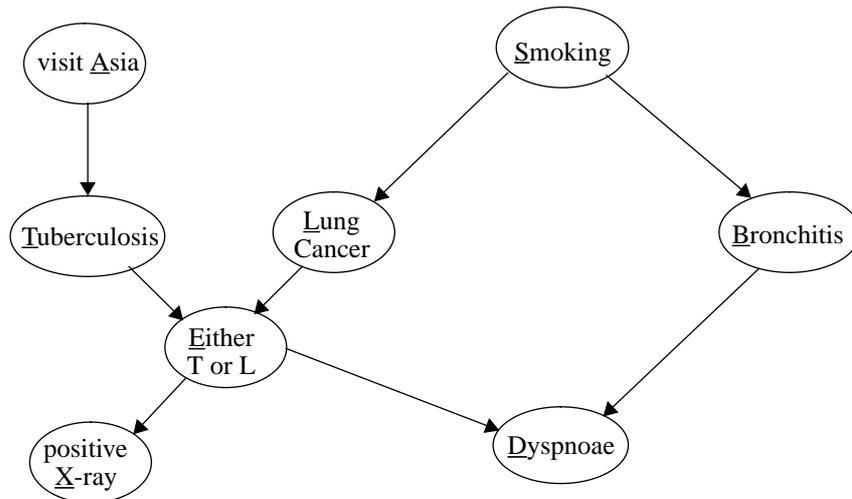


Figure 2: A graph associated with the variable set of figure 1. An extra node (Either T or L) has been introduced in order to simplify the graph. The example is due to Lauritzen and Spiegelhalter [53].

figure 1. Here, one additional node, E(ither T or L), has been introduced in order to simplify the graph. Note that in this graph all the edges are directed, and there are no nodes from where it is possible to follow a sequence of directed edges and return to the starting point (no *directed cycles*).

A graph which contains only directed edges is known as a *directed graph*. Those graphs which contain no directed cycles have been particularly studied in the context of probabilistic expert systems. These are referred to as *directed acyclic graphs* (DAGs).

A graph which contains only undirected edges is known as an *undirected graph*. These have not received so much attention in the expert systems literature, but are of importance in statistical modelling. Graphs with a mixture of directed and undirected edges are referred to as *mixed graphs* (e.g. [72]).

As mentioned in the opening to this section, the important point about a graphical representation of a set of variables is that the edges can be used to indicate *relevance* or *influences* between variables. Absence of an edge between two variables, on the other hand, provides some form of independence statement; nothing about the state of one variable can be inferred by the state of the other

There is a direct relationship between the independence relationships that can be expressed graphically and the independence relationships that can be defined in terms of probability distributions.

## 5. Independence

### 5.1 Forms of independence

The notions of *independence* and *conditional independence* are a fundamental component of

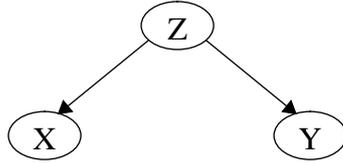


Figure 3: X is conditionally independent of Y given Z.

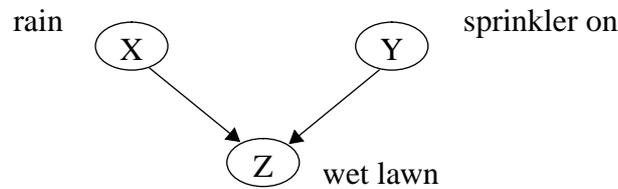


Figure 4: X and Y are conditionally dependent given Z.

probability theory. It is this combination of qualitative information with the quantitative information of the numerical parameters that makes probability theory so expressive. Detailed studies of conditional independence properties can be found in Dawid [21] and Pearl [67]. For completeness, we include definitions of the basic notions here.

We shall use the notation introduced by Dawid. Let X and Y be variables. Then  $X \perp\!\!\!\perp Y$  denotes X and Y are independent. The corresponding probabilistic expression of this is:

$$p(x, y) = p(x)p(y).$$

Now introduce a further variable Z. Then  $X \perp\!\!\!\perp Y \mid Z$  denotes that X is conditionally independent of Y given Z. One expression of this in terms of probability distributions is:

$$p(x, y \mid z) = p(x \mid z)p(y \mid z).$$

We can draw a directed acyclic graph that directly encodes this assertion of conditional independence. This is shown in figure 3. A significant feature of the structure in figure 3 is that we can now decompose the joint probability distribution for the variables X, Y and Z into the product of terms involving at most two variables:

$$p(x, y, z) = p(x, y \mid z)p(z) = p(x \mid z)p(y \mid z)p(z).$$

As a concrete example, think of the variable Z as representing a disease such as measles. The variables X and Y represent distinct symptoms; perhaps “red spots” and “Koplik’s spots”<sup>1</sup> respectively. Then if we observe the disease (measles) as present, the probability of either symptom being present is determined. Actual confirmation of one symptom being present will not alter the probability of occurrence of the other.

A different scenario is illustrated in figure 4. Here X and Y are marginally independent, but

---

1. Small white spots found inside the mouth.



Figure 5: X and Y are conditionally independent given Z.

conditionally *dependent* given Z. This is best illustrated with another simple (often used) example. Both “rain” (X) and “sprinkler on” (Y) may cause the lawn to become wet. Before any observation of the lawn is made, the probability of rain and the probability of the sprinkler being on are independent. However, once the lawn is observed to be wet, confirmation of it raining may influence the probability of the sprinkler being on (they are, by and large, alternative causes). This is an example of “explaining away” (p. 447 [80]); the presence of one cause making an alternative less likely. See for [67] further discussion of this.

Figure 4 provides the graphical representation of this situation. The probability distribution can again be factorised. This time as:

$$p(x, y, z) = p(z | x, y)p(x)p(y)$$

Note that this is again making use of the (marginal) independence of X and Y ( $p(x, y) = p(x)p(y)$ ).

A third and final example completes the cases of interest. The disease X (“Kawasaki disease”) is known to cause the pathological process Z (“myocardial ischemia”). This, in turn, has an associated symptom Y (“chest pain”). Here, if through some additional test, myocardial ischemia is diagnosed as present, observation of chest pain will have no further influence on the probability of Kawasaki disease being the underlying cause of the expression of the pathological process. That is, X and Y are again conditionally independent given Z. In this case, the probability distribution factorises as:

$$p(x, y, z) = p(y | z)p(z | x)p(x)$$

A pattern is emerging. We can now provide a generalisation of what is happening.

## 5.2 Factorised distributions

We have just seen a series of examples of cases where a directed acyclic graph has admitted a simple factorisation of the joint probability distribution. In these examples, each factor is a conditional probability involving just parent and child nodes. This is a general property of directed acyclic graphs [53].

### Proposition

Let  $U = \{X_1, X_2, \dots, X_n\}$  have an associated graph  $\mathcal{G} = (V, E)$ , where  $\mathcal{G}$  is a DAG. Then the joint probability distribution  $p(U)$  admits a direct factorisation:

$$p(u) = \prod_{i=1}^n p(x_i | pa(x_i)) \tag{eqn. 1.}$$

Here  $pa(x_i)$  denotes a value assignment for the parents of  $x_i$ .

A slightly larger example will help to bring out the importance of this property. In figure 2 we took the variables listed in figure 1 and generated a DAG which represents the influences between these variables (one additional variable has been introduced to simplify the graph). A factorisation of the probability distribution can now be written down directly using equation 1:

$$p(a, b, d, e, l, s, t, x) = p(a)p(s)p(t|a)p(l|s)p(b|s)p(e|l, t)p(d|e, b)p(x|e)$$

The net result is that the probability distribution for a large set of variables may be represented by a product of conditional probability relationships between small clusters of semantically related propositions. Now, instead of needing to elicit a joint probability distribution over a set of complex events, the problem is broken down into the assessment of these conditional probabilities as *parameters* of the graphical representation.

The population of these relationships with numerical parameters is now much more tractable. It should now be clear that the elicitation of this qualitative graphical structure is of fundamental importance in easing the problem of eliciting the probability distribution for a knowledge model.

### 5.3 d-separation and Markov properties

It is important to establish the precise relationships between the topology of a graph and the independence properties of the associated probability distribution. Pearl's d-separation [67] provides a straightforward technique for reading off the independence relationships from a graph. The use of d-separation will be illustrated in this section, and then some pointers given to the literature on related studies of independence properties. Finally, the notion of Markov equivalence will be introduced. This last is of major importance in connection with learning the structure of probabilistic networks.

d-separation (d for *directional*) can be stated as follows (e.g. [70]). In this definition, a *path* is any contiguous sequence of edges ignoring directions.

#### **Definition**

For any three disjoint subsets  $\mathbf{X}$ ,  $\mathbf{Y}$ ,  $\mathbf{Z}$  of nodes in a DAG  $\mathcal{G}$ ,  $\mathbf{Z}$  is said to d-separate  $\mathbf{X}$  from  $\mathbf{Y}$  if there is no path from a node in  $\mathbf{X}$  to a node in  $\mathbf{Y}$  along which the following conditions hold:

- every node with converging arrows is in  $\mathbf{Z}$  or has a descendent in  $\mathbf{Z}$ ;
- every other node is outside  $\mathbf{Z}$ .

This is best illustrated by means of an example. Figure 6 illustrates a simple DAG which is a composition of the earlier structures. Let  $\mathbf{X} = \{2\}$  and  $\mathbf{Y} = \{3\}$ . First of all, we will consider the case where  $\mathbf{Z} = \{1\}$ . d-separation is a little hard to apply when it is first encountered because it is defined in terms of a negative condition ("there is no path such that ..."). This simply means that in establishing whether d-separation holds we must consider all possible paths between the two sets of variables. In this case there are two:  $2 \leftarrow 1 \rightarrow 3$ , and  $2 \rightarrow 4 \leftarrow 3$ . In the first case, the node in  $\mathbf{Z}$  has *no* converging arrows; in the second the node with converging arrows and its descendent are *outside*  $\mathbf{Z}$ . That is, neither path satisfies the two conditions (they are said to be *blocked* by  $\mathbf{Z}$ ) and  $\mathbf{X}$  and  $\mathbf{Y}$  are d-separated by  $\mathbf{Z} = \{1\}$ . However, consider now  $\mathbf{Z}' = \{1, 5\}$ . Here a descendant of node 4, node 5, is in  $\mathbf{Z}'$ . This potentially opens up a pathway between  $\mathbf{X}$  and  $\mathbf{Y}$ ; if we learn the value of 5, its causes, 2 and 3, will be rendered dependent. In this case

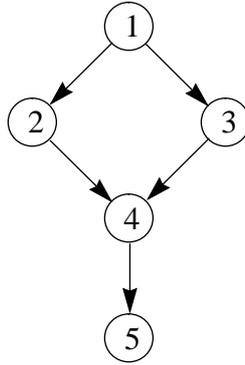


Figure 6: A simple DAG composing the structures of figures 3, 4 and 5.

path  $2 \rightarrow 4 \leftarrow 3$  is said to be active, and  $\mathbf{X}$  and  $\mathbf{Y}$  are not d-separated by  $\mathbf{Z}' = \{1,5\}$ .

Some further worked examples of d-separation can be found in [45], and we recommend that the reader examine additional examples to gain further understanding of d-separation. But the general principal is that it provides a mechanism for reading off the dependencies and independencies implicit in a DAG. Furthermore, we have the following result due to Geiger and Pearl [30]:

### Theorem

For every DAG  $\mathcal{G}$ , there exists a probability distribution  $p$  such that for each triple  $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$  of disjoint subsets of variables the following holds:

$$\mathbf{Z} \text{ d-separates } \mathbf{X} \text{ and } \mathbf{Y} \text{ iff } \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$$

This shows that no rule for detecting independence in a DAG can improve on d-separation in terms of completeness.

Note that the equivalence cannot be taken in the other direction; there are some probability distributions with independence properties that cannot be represented using a DAG. See [87] for some examples of this, and [27], [104] for more general discussions on modelling probability distributions using graphical representations.

We should mention that the literature contains a number of alternative criteria for assessing independence. One overview of these can be found in Lauritzen et al. [54], which contains a detailed study of the independence properties of factorisable distributions. They show that for a directed acyclic graph  $\mathcal{G}$  and probability measure  $p$  (subject to certain constraints), the following properties are equivalent:

- $p$  admits a recursive factorisation according to  $\mathcal{G}$ ;
- $p$  obeys the global directed Markov property, relative to  $\mathcal{G}$ ;
- $p$  obeys the local directed Markov property, relative to  $\mathcal{G}$ .

The *directed local Markov Property* is easily expressed in words as: any variable is conditionally independent of its nondescendants, given its parents. In the case of the global directed

Markov property, two sets of variables  $A$  and  $B$  are conditionally independent given a third,  $S$ , if  $S$  separates  $A$  and  $B$  in graph  $\mathcal{G}$ . Here, *separates* denotes that all paths between  $A$  and  $B$  pass through  $S$ . However, we do not wish to dwell on these criteria here. The important point is to emphasise that Lauritzen et al. also demonstrate an equivalence between the global directed Markov property and Pearl's d-separation. They then extend the generality of their results by dropping reference to all features of probability other than conditional independence. All these (equivalent) criteria can be thought of as generalising to an *algebra of relevance*, not just one of probability. Further discussion of this in the specific context of d-separation can be found in, e.g. [30], [70], [71].

#### 5.4 The implications for building implementable probabilistic models

The lessons from this section can be summarised quite succinctly. Firstly, graphs may be used to represent qualitative influences in a domain. Secondly, the conditional independence statements implied by the graph can be used to factorise the associated probability distribution. This factorisation can then be exploited to (a) ease the problem of eliciting the global probability distribution, and (b) allow the development of computationally efficient algorithms for updating probabilities on receipt of evidence.

We are not too concerned with the development of computationally efficient algorithms in this paper. Of more interest is the method for the construction of probabilistic models implied by the above. This factorises into two stages:

- *Qualitative stage*: Consider the general relationships between the variables of interest in terms of the relevance of one variable to another in specified circumstances.
- *Quantitative stage*: Numerical specification of the parameters of the model.

#### **Definition:**

In order to fully specify the model, one needs a graph  $\mathcal{G}$ , and the set of parameters  $\Theta$  which make up the factorised probability distribution. We shall refer to  $\mathcal{M} = (\mathcal{G}, \Theta)$  as a *graphical model*.

This certainly helps, but it is not a complete solution. In general, there will still be a large number of parameters to elicit. What is more, given that the assignment of values to these parameters will be based on varying amounts of information and experience, they will be of varying precision and reliability. It should also be emphasised that there is nothing sacred about the underlying graphical representation; there will almost always be an element of subjectivity in its specification.

A number of questions immediately arise. If there is some imprecision in the values of the parameters, can we update them in the light of experience? Equally, can we update the structure in the light of experience? Indeed, to remove any element of subjectivity, could we even learn the structure and the parameters directly from statistical data? In the remainder of the paper, we will first look at some aids to the knowledge elicitation problem, and then progress to reviewing the work that has been carried out to date to address these last three questions.

## 5.5 Some pointers to further reading on decomposable models and factorisable distributions

The primary focus of this paper is on exploiting the relationships between graphical representations and factorisable distributions for knowledge representation and elicitation. However, these relationships have also been exploited to develop computationally efficient techniques for revising probability distributions in the light of evidence. Basically the revision of the global probability distribution is decomposed into a sequence of local computations by exploiting the relevance properties implied by the model. Three general approaches are the arc reversal/node reduction technique of Shachter [83], the message passing algorithm due to Pearl [66], and the “clique tree” approach of Lauritzen and Spiegelhalter [53]. Shachter [84] demonstrated that the arc reversal/node reduction technique, Pearl’s algorithm and the Lauritzen & Spiegelhalter algorithm are essentially identical for any DAG.

The Lauritzen and Spiegelhalter algorithm was further developed by Jensen and others to form the basis of the HUGIN expert system shell. Jensen et al. [47] extends earlier work restricted to singly connected trees to cover multiply connected trees (in which there may be more than one path between any two nodes) by introducing a secondary structure called a junction tree. Jensen et al. [46] provides an accessible account of the algorithm used as the basis of the HUGIN shell.

In a slightly different vein, Smyth et al. [87] reviews the applicability and utility of techniques for graphical modelling to hidden Markov models (HMMs). Its particular merit in the context of this paper is that it contains an up to date and self-contained review of the basic principles of probabilistic independence networks. Short discussions of parameter estimation and model selection techniques are also included.

Lauritzen [52] provides an up to date text on graphical models.

## **6. The problem of knowledge elicitation**

The central topic of this paper is the learning of parameters, and ultimately the structure, of a knowledge model from data. However, there are a number of techniques which have been developed which enable at least an initial model to be constructed using the experience of domain experts. We will review these briefly in this section.

The *similarity graph* technique of Heckerman [34] exploits the clinical technique of *differential diagnosis* to elicit directed graphical structures relating diseases which are similar. Using an iterative technique, a graph is built up for each differential diagnosis and the resulting graphs superimposed to provide a complete model of the specified domain.

Nilsson [64] suggested that if the number of parameter assessments is insufficient to specify a joint distribution uniquely, then maximum entropy arguments may be used to complete the distribution.

There are a number of techniques that can be used to assist in the elicitation of probabilities; Good [32] and Winkler [103] are well cited references in this context. However, these can be extremely time consuming for a large network, and rather intimidating for the expert(s) from whom the parameters are being elicited. Druzdzel and van der Gaag [25] offer a “non-invasive” technique for eliciting the network parameters which will accommodate whatever probabilistic information the expert is willing and comfortable to provide. This information may range from something as precise as a value, or range of values, for a parameter, through to much vaguer qualitative influences between pairs of variables (a specific combination of values for two vari-

ables may be seen as making some state more likely, for example). This information is then applied as constraints on the *distribution hyperspace* of all possible joint distributions over a set of variables. Second-order probability distributions are then obtained over the parameters (probabilities) to be assessed (provided the constraints are consistent). In general, the elicitation of the second-order probability distribution on the parameters may be an iterative process. The expert(s) may need to be “confronted” with any inconsistencies in the constraints in order to refine the elicitation. This seems to be a very interesting approach, but at the time of writing has yet to be tried out on a substantial problem [Druzdzal, *pers. comm.*].

This is to emphasise that eliciting probabilities directly from experts is still an active area of research. Once elicited, one is not necessarily sure about their reliability. Learning techniques can be used to critique probabilities used as parameters in a network, and revise them if necessary.

## 7. Learning probabilities from data

### 7.1 Known structure with all variables observable

The most straightforward learning situation to consider is that where the structure is (believed to be) known and data is obtainable on all variables in the network. In this section, we will first demonstrate the basic techniques in the context of learning the probability distribution of a single variable before moving onto a network of variables.

We will consider the flipping of a common drawing pin (thumb tack). The pin is thrown into the air and allowed to land again on a hard flat surface. This event has two possible outcomes; a “heads” where the drawing pin lands with its point touching the surface, and a “tails” where the pin lands on its back (see Figure 7). Suppose someone carries out a long series of such events and measures the fraction of flips where the outcome is heads. From a frequentistic perspective, the long-run fraction of head outcomes is a probability. The measurements carried out by our observer provide an *estimate* of this probability.

Let us define a variable  $\Theta$  whose true value  $\theta$  corresponds to the long run fraction of head events (we will use greek characters when the variables refer to parameters of a network). We can express our uncertainty about  $\theta$  given the current state,  $\xi$ , of evidence concerning  $\theta$  by defining a probability density function  $p(\theta | \xi)$  on its possible values<sup>1</sup>. This can then be updated as further flips of the drawing pin are observed. Figure 7 shows one possible probability density function for  $\theta$ .

Suppose we now flip the drawing pin again and observe another heads. Using Bayes’ theorem, the posterior probability becomes

$$p(\theta | X = \text{heads}, \xi) = c p(X = \text{heads} | \theta, \xi) p(\theta | \xi)$$

where  $c$  is some normalisation coefficient. Note that if we *know* the value of  $\theta$ , then the probability of a heads on any flip would be equal to  $\theta$ , no matter how many outcomes we observe. That is

$$p(X = \text{heads} | \theta, \xi) = \theta$$

Hence

$$p(\theta | X = \text{heads}, \xi) = c \theta p(\theta | \xi)$$

---

1. A probability density function represents a probability distribution over a continuous set of outcomes.

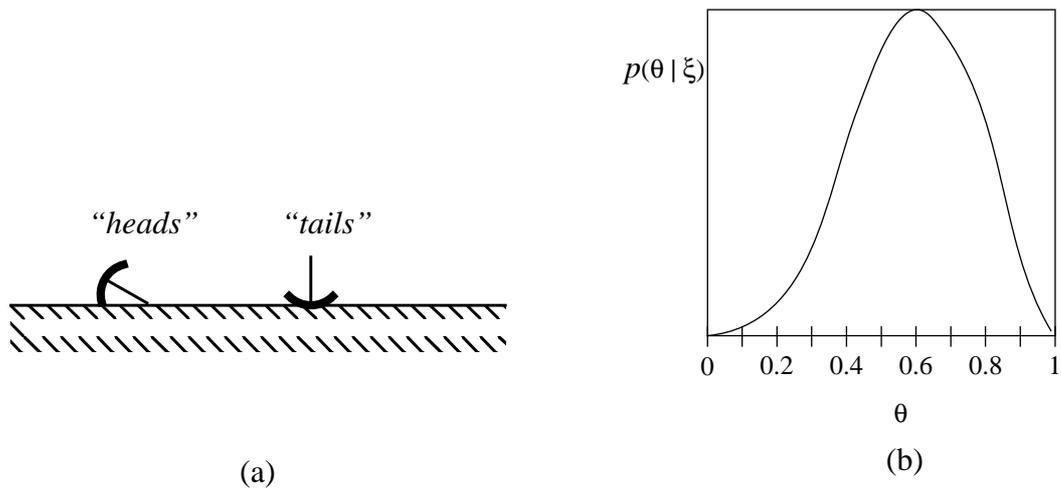


Figure 7: (a) The two possible outcomes from flips of a drawing pin (“thumb tack”). (b) A probability density function for the parameter value  $\theta$ , the long run fraction of heads associated with a drawing pin. Redrawn after [37].

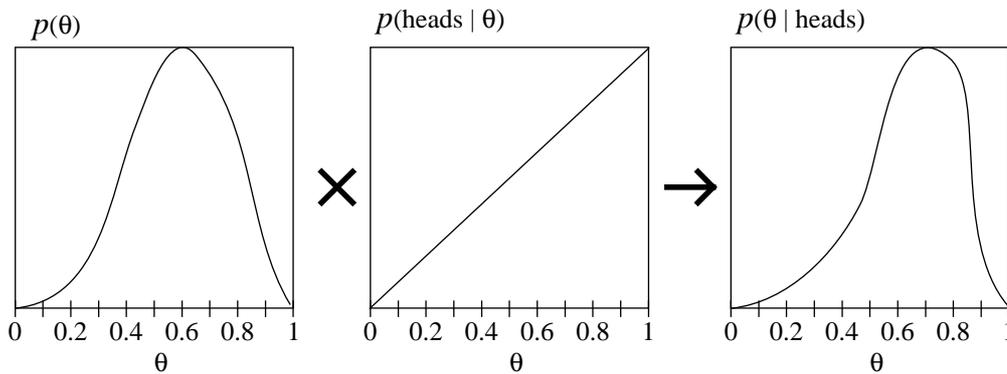


Figure 8: Revision of the probability density function  $p(\theta | \xi)$  given observation of a “heads” (redrawn after [37]).

That is, we obtain the posterior probability density for  $\theta$  by multiplying its prior density by the function  $f(\theta) = \theta$  and renormalising. This procedure is depicted graphically in Figure 8. As we might expect, the resulting probability density function is slightly tighter and shifted slightly to the right.

It is quite straightforward to see that in general<sup>1</sup>, if we observe  $h$  heads and  $t$  tails, then

$$p(\theta | h \text{ heads}, t \text{ tails}, \xi) = c \theta^h (1 - \theta)^t p(\theta | \xi)$$

That is, once we have assessed a prior density function for  $\theta$ , all that is relevant in a random

1. *Provided that*; (1) observations are independent given  $\Theta = \theta$ , and (2) the probability  $p(X = \text{heads} | \theta)$  is constant.

sample of events is the number of heads and the number of tails.  $h$  and  $t$  are said to be *sufficient statistics* for this model; they provide a summarisation of the data that is sufficient to compute the posterior from the prior.

So far, we have only considered the case where the outcome variable has two states. The results generalise quite straightforwardly to cover situations where the outcome variable has  $r > 2$  states. In this case, we denote the probabilities of the outcomes  $\Theta_x = \{ \theta_{x=1}, \dots, \theta_{x=r} \}$ , where each state is assumed possible so that  $\theta_{x=k} > 0$  for all  $1 \leq k \leq r$ . In addition, we have

$\sum_{k=1}^r \theta_{x=k} = 1$ . If we know these probabilities, then the outcome  $x_i$  of any event will be inde-

pendent of the outcomes of all other events (Figure 9). Any database  $D = \{x_1, \dots, x_m\}$  of outcomes that satisfies these conditions is called a random sample from an  $(r-1)$ -dimensional multinomial distribution with parameters  $\Theta_x$  [32]. In the specific case where  $r = 2$  the sequence is often referred to as a *binomial sample*.

Referring back to our earlier example, we have seen how the probability density function for  $\theta$  may be updated given a binomial sample and some prior density function for  $\theta$ . There are no formal constraints on the choice of the prior density function  $p(\theta | \xi)$ . However, in practice there are certain advantages to using what is known as the *beta distribution* in the two state case. A variable  $\Theta$  is said to have a beta distribution with hyperparameters<sup>1</sup>  $\alpha, \beta$  when its probability density function is given by

$$p(\theta | \alpha, \beta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \alpha, \beta > 0 \quad \text{eqn. 2.}$$

This is the density function that was used in Figure 7(b), with  $\alpha = 3$  and  $\beta = 2$ . The first property to notice is that on observing a heads in our drawing pin example, the prior distribution is revised to become

$$p(\theta | \text{heads}, \alpha, \beta) = c\theta^{(\alpha+1)-1}(1-\theta)^{\beta-1}$$

In general, after  $h$  heads and  $t$  tails, we will have:

$$p(\theta | h \text{ heads}, t \text{ tails}, \alpha, \beta) = c\theta^{h+\alpha-1}(1-\theta)^{t+\beta-1} \quad \text{eqn. 3.}$$

Clearly, the result after sampling remains a beta distribution.

The second property to notice is that the expectation with respect to this distribution has a simple form:

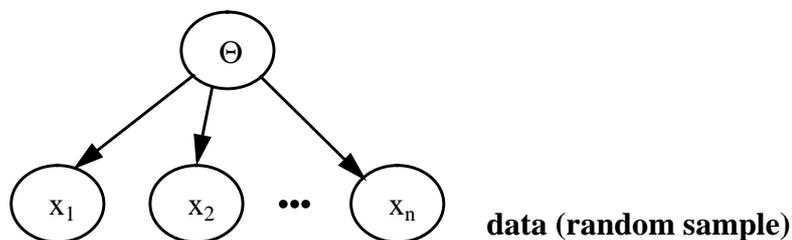


Figure 9:

---

1. The term hyperparameter is used to distinguish  $\alpha$  and  $\beta$  from the parameter  $\theta$ .

$$E_{p(\theta | \alpha, \beta)}(\theta) = \int \theta p(\theta | \alpha, \beta) d\theta = \frac{\alpha}{\alpha + \beta} \quad \text{eqn. 4.}$$

Note that for clarity, E is subscripted with the distribution used for expectation. We may take this as the probability of observing a heads on the next flip, in the drawing pin example. In particular, with the prior distribution of Figure 7,  $E_{p(\theta | \alpha, \beta)}(\theta) = 3/5$ . After observation of one heads, the revised density function on the right hand side of Figure 8 has expectation 4/6.

Note that we may think of the denominator of Equation 4 as an *equivalent sample size*. That is, the prior has been assessed “as if”  $\alpha + \beta$  observations had been made.

The generalisation of the beta distribution to  $r > 2$  states is known as the *Dirichlet distribution* (with hyperparameters  $\alpha_1, \dots, \alpha_r$ ). The above two properties generalise fully. In particular, we say that the set of Dirichlet distributions obtained as a Dirichlet prior is updated is a *conjugate family of distributions* for sampling from a multinomial distribution.

Further details of the Dirichlet distribution can be found in [102]. For simplicity we will refer here to a Dirichlet distribution with hyperparameters  $\alpha_1, \dots, \alpha_r$  as

$$\mathcal{D}[\alpha_1, \dots, \alpha_r]$$

We now have the basic machinery needed to learn the probabilities for a network in the situation where a network structure  $S$  is assumed known and where all the variables are observable, *provided* we make one additional assumption. Let  $S^h$  denote the hypothesis that the joint probability can be factored according to  $S$ , and  $\mathcal{V}$  be the set of case variables in the network  $S$ . Spiegelhalter and Lauritzen [88] introduced the idea of considering the conditional probabilities of a network such as  $S$  as being generated by a set of parameters  $\{\theta_v | V \in \mathcal{V}\}$ . The parameters  $\theta_v$  are uncertain, with some prior distribution  $p(\theta_v | S^h)$ . Then, given a random data sample  $D = \{x_1, \dots, x_N\}$  we wish to update the overall parameterisation  $\theta$  (whose components are  $\theta_v$ ) given  $S^h$ . The additional simplifying assumption that is made in this situation is that of *parameter independence*. That is, the parameters  $\theta_v$  are assumed a priori independent variables. In this case,  $p(\theta | S^h) = \prod_v p(\theta_v | S^h)$  and the joint distribution for the variables  $\mathcal{V}$  and parameters  $\theta$ , given  $S^h$  can be written as:

$$p(\mathcal{V}, \theta | S^h) = \prod_v p(V | pa(V), \theta_v, S^h) p(\theta_v | S^h) \quad \text{eqn. 5.}$$

From equation 5 it is clear that for any node  $V \in \mathcal{V}$ ,  $\theta_v$  can be considered to be just another parent of  $V$  in an extended network. Figure 10 gives an example of such a parameterised network for the often used “visit to Asia” example.

In addition to parameter independence, we assume:

- No missing data;
- Conjugate priors.

If we can obtain data on each variable  $V \in \mathcal{V}$ , then assuming that the parameters are independent, we can update each parameter  $\theta_v$  independently using just the same techniques as we used in the single variable case. In particular, if each parameter set has a Dirichlet prior distribution, then a posteriori, the distribution remains Dirichlet.

Consider a node  $V$  with  $K$  states. For a specific parent configuration  $pa(V)^*$  the variable will be parameterised as:

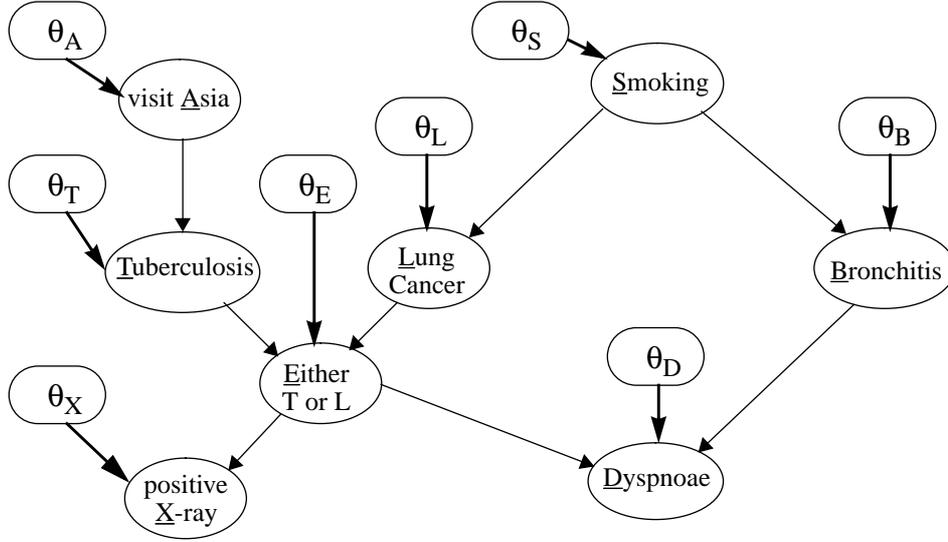


Figure 10: Parameterisation of the “Visit to Asia” example to account for uncertainty in the values of the conditional probabilities for variables in the network (after [88]).

$$p(V \mid \text{pa}(V)^*, \theta_{V|\text{pa}(V)^*}, S^h) = \theta_{V|\text{pa}(V)^*, S^h} \quad \text{eqn. 6.}$$

We assume a Dirichlet prior distribution  $\mathcal{D}[\alpha_1, \dots, \alpha_K]$  is given for  $\theta_{V|\text{pa}(V)^*, S^h}$ . In order to *predict* the outcome of the next case, we can use as before the expectations for the parameters as the conditional probability of each variable. In particular for the node  $V$ , for each state  $k \in K$ :

$$p(V_k \mid \text{pa}(V)^*, S^h) = E_{\mathcal{D}}[\theta_{V|\text{pa}(V)^*, S^h}]_k = \alpha_k / \alpha \quad \alpha = \sum_{j=1}^K \alpha_j \quad \text{eqn. 7.}$$

The expectation is taken over the distribution  $\mathcal{D} = \mathcal{D}[\alpha_1, \dots, \alpha_K]$ . After the occurrence of a case, if we then observe  $V$  to have been in the  $j$ th state and  $\text{pa}(V)$  to have taken the configuration  $\text{pa}(V)^\ddagger$ , we update the parameter as:

$$\theta_{V|\text{pa}(V)^\ddagger} \mid V_j, \text{pa}(V)^\ddagger, S^h \sim \mathcal{D}[\alpha_1, \dots, \alpha_j + 1, \dots, \alpha_K] \quad \text{eqn. 8.}$$

Note that the lesson of this section can be quite simply summarised. That is, the case of learning parameters from complete data with known structure can be reduced to maintaining sufficient statistics for certain distributions (the Dirichlet being one example).

This outlines the updating scheme for a single observation. Details of the generalisation to a database of  $N$  observations can be found in [36]. Although this is quite straightforward, it is often the case that certain variables in a network may be physically impossible or too expensive to observe with certainty. Techniques to handle such cases are considered in the next section.

## 7.2 Known structure with incomplete data

In the case of medical diagnosis, one may well elicit a model in which some of the variables

cannot be unintrusively observed. During routine use, the values of these variables could be too expensive (in terms of cost or of risk to the patient) to observe on a regular basis. Analogous situations can arise in most other domains, and so there is a need to extend the work of the previous section to cover the situation where a subset of the variables in the model are unobservable. For simplicity, we will restrict the discussion in this section to the case where the availability or otherwise of an observation is independent of the actual states of the variables. (Methods for addressing dependencies on omissions have been studied in, for example, [69], [75] and [78]. We will come back to this in the next section).

The two variable case illustrates the general situation quite simply. We have the variables  $X_1$  and  $X_2$  with states  $\{x_1, \bar{x}_1\}$  and  $\{x_2, \bar{x}_2\}$  respectively.  $X_2$  is observed to be in state  $x_2$ , whilst the state of  $X_1$  is unknown. Hence there are two possible completions of the database,  $\{x_1, x_2\}$  and  $\{\bar{x}_1, x_2\}$ ; one for each possible variable instantiation of the unknown  $X_1$ . Then the posterior distribution for  $\theta_{2|1}$  is obtained by summing over the possible states of  $X_1$  as follows:

$$p(\theta_{2|1} | x_2) = p(\theta_{2|1} | x_1, x_2)p(x_1 | x_2) + p(\theta_{2|1} | \bar{x}_1, x_2)p(\bar{x}_1 | x_2)$$

In this case, the terms  $p(\theta_{2|1} | x_1, x_2)$  and  $p(\theta_{2|1} | \bar{x}_1, x_2)$  would be Beta distributions, and the resulting posterior distribution is a “mixed distribution” with *mixing coefficients*  $p(x_1 | x_2)$  and  $p(\bar{x}_1 | x_2)$ .

In general (see, e.g. [36], [88]), the posterior distribution for each of the parameters  $\theta_v$  will be a linear combination of Dirichlet distributions, or a *Dirichlet mixture*, with mixing coefficients that will need to be computed from the prior distributions. What one is doing is computing the joint posterior distribution given each possible completion of the database, and then mixing these over all possible completions. Note that this means that the posterior over parameters are no longer independent. In addition, as the process continues and further cases are observed, one faces a combinatorial explosion (e.g. [19], [89]); in general, the complexity of the exact computation is exponential in the number of missing variable entries in the data base, [18]. Thus, in practice some approximation is required.

A number of such approximations have been described in the literature. One can broadly divide them into two classifications; deterministic and stochastic. In an early example, referred to as quasi-Bayes [86] or fractional updating [97], for each unobserved variable it was assumed that a fractional number of instances of each state of that variable had been observed. This early work has been criticised because it falsely increases the equivalent sample sizes [6], and hence implicit precision, of the Dirichlet distributions. Consequently, this approach has subsequently been refined so that the approximating distribution attempts to match the moments of the correct mixture [19], [89], [98].

These deterministic approximations process the data in the database sequentially, and also make use of the assumption of parameter independence and properties of the Dirichlet distribution. In contrast, the various stochastic methods process all the data at once, and can handle continuous domain variables and dependent parameters. (However, it should be mentioned that the comparison reported in [19] indicates that the deterministic methods can perform well in comparison with the stochastic methods).

A widely studied stochastic method is Gibbs sampling. This is a special case of the general Markov chain Monte-Carlo methods for approximate inference (e.g. [33], [62]). Gibbs sampling can be used to approximate any function of an initial joint distribution  $p(X)$  provided certain conditions are met. Firstly, the Gibbs sampler must be *irreducible*. That is, the distribution

$p(X)$  must be such that we can sample any possible state of  $X$  given any possible initial state of  $X$ . This condition is satisfied if, for example, the full joint distribution has no zeros (i.e. every variable instantiation is possible). Secondly, strictly each instantiation must be chosen infinitely often. In practice, an algorithm for deterministically rotating through the variables is used to meet this requirement. If these conditions hold, then the average value of the sampled function approaches the expectation with respect to  $p(X)$  with probability 1 as the number of samples tends to infinity [15]. The problem, of course, arises as to how many samples to take in a given situation, and how to estimate the error in the resulting posterior distribution (which is what we are interested in here). Heuristic strategies exist for answering these questions ([62], [73]), although there is no easy general solution.

For more information, an outline of the Gibbs sampling algorithm can be found in [10], whilst [58] and [62] contain good discussions of Gibbs sampling, including methods for initialisation and a discussion of convergence.

An alternative approximation algorithm is the expectation-maximization (EM) algorithm [24]. This can be viewed as a deterministic version of Gibbs sampling, and can be used to search for the *maximum a posteriori* (MAP) estimate of model parameters. The EM algorithm iterates through two steps; the *expectation step* and the *maximisation step*. In the first step, as a complete database is unavailable, the *expected sufficient statistics* for the missing entries of the database  $D$  are computed. The computational effort needed to perform this computation can be intense. However, any Bayesian network inference algorithm may be used for this evaluation (e.g. [51]). In the second step, the expected sufficient statistics are taken as though they were the actual sufficient statistics of a database  $D'$ , and the mean or mode of the parameters  $\theta$  are calculated. This computation is such that the probability of observation of  $D'$  given the network structure and parameters is maximised; hence the term maximisation step.

The EM algorithm is fast, but it has the disadvantage of not providing a distribution over the parameters  $\theta$ . In addition, it was reported in [51] that when a substantial amount of data was missing, the likelihood function has a number of local maxima leading to poor results (this is also a problem with gradient ascent methods). Modifications have been suggested to overcome this, and some authors suggest that a switch also be made to alternative algorithms when near a solution in order to overcome the slow convergence of EM when near local maxima ([10], [61]).

One of the algorithms suggested as an alternative to switch to in the above is *gradient descent*. A demonstration of the pure use of gradient descent is reported in [81]. This paper raises rather an interesting point. Firstly, let us introduce the notion of a *hidden* variable. In the case of a database with missing variables, a particular variable may be observed in some cases, or it may never be observed. The denotation *hidden variable* refers to the latter situation. Now, from the preceding discussion it is clear that learning parameters for a network with hidden variables is more complex than the situation where all the variables are observable *in the same structure*. In contrast, the work reported in [81] indicates that learning parameters for a network containing one or more hidden variables, provided they are cognitively meaningful, can be more statistically efficient than learning parameters for an alternative network in which all the parameters are observable. Note that here, “efficiency” refers to the need for relatively small amounts of data and *not* to speed of computation. Figure 11 indicates the two alternative networks considered. Network structure (a) showed a far more rapid reduction in mean square error per output value as a function of number of training cases than did network structure (b). The lesson to be

learnt is that a soundly structured model is likely to be easier to train as well as having an inherently higher diagnostic accuracy. This demonstrates one of the advantages of probabilistic networks over neural networks; the exploitation of known structural information can dramatically reduce the amount of learning data needed. Basically, the number of parameters needed is reduced and hence learning becomes more statistically efficient.

### 7.3 Systematically missing data

The methods of the preceding section are based on the assumption that whether or not a datum is missing is independent of the state of the world. It can happen that this assumption is invalid, and that certain values of variables are missing systematically throughout a database. This can lead to bias in the parameter estimates [89]. This is a problem which has been of particular interest in the study of causal inference in statistics, and we will outline a widely used solution to this problem here.

Suppose, for example, we are interested in studying the effectiveness of a new treatment on a population suffering from a specific disease or syndrome. (One could think in terms of the trial of a new drug, but there are a number of other forms a treatment could take). Here, the effect of the treatment on an individual is defined as the comparison (usually, the difference) between the value of the outcome if the individual is treated and the value of the outcome if the same individual had not been treated. To assess the influence of the treatment, a randomised trial would typically be carried out. Individuals from the population will be randomly assigned to the treatment, and a comparison made with the remaining individuals who do not undergo the treatment. The difficulty is that if an individual is assigned to a treatment, the value of the outcome had that same individual not been treated will not be available. And vice-versa. This difficulty can be met by introducing the notion of potential outcomes; one tries to “imagine” observing outcomes on an individual in circumstances other than those to which the individual was actually exposed.

Rubin has provided much foundational work on this approach [77], [78]. The basic idea is to build a prior model about when the data will be missing, and then update this model using real-world data. Rubin’s work generalises to more complex situations where the assignment mechanism is more complex than a simple randomised trial [79], and is now widely used in statistics and epidemiology (see [48] for a general discussion of the “Rubin Causal Model”, RCM). Further analysis of the RCM can be found in [4], together with illustration of the use of RCM to

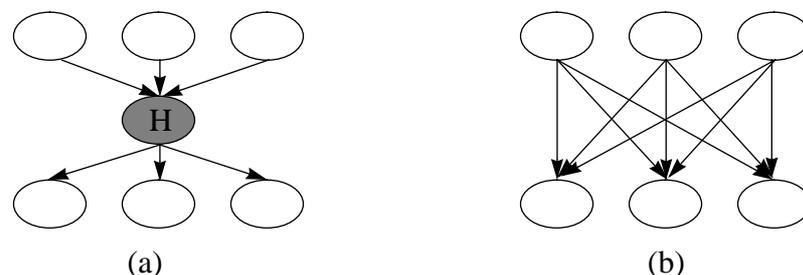


Figure 11: In these diagrams, all nodes except H are three-valued. H is two-valued. (a) represents a probabilistic network with a hidden variable, H. It requires 45 parameters. (b) is the corresponding fully observable network. This requires 168 parameters (after [81]).

estimate the effect of Veteran status in the Vietnam era on mortality.

#### 7.4 Training versus adaptation

At this stage it is worth making clear the distinction between the production of an initial model, and the revision of parameters in an already extant model. There are domains where sufficient data is available to learn the parameters for a network using the database. We may refer to this activity as *training*. However, the more usual case in the field of expert systems is that a combination of expert knowledge and statistical data must be used to elicit the required parameters. Here, one will wish to use incoming case data to revise the parameters of the network; the parameters may be initially imprecisely specified, or even incorrect. Olesen et al. [65] distinguish this from training, preferring the term *adaptation*.

As an example, certain epidemiological data might be available which could be used to obtain some of the probabilities used as parameters in a medical expert system. This data could also be used to provide reliabilities on these probabilities. Expert judgement could be used to provide the remaining parameters, but it is much harder to be sure of how much one can trust these judgements. The learning techniques described in this paper enable *all* these parameters to be adapted as data passes through the system, so that the model more closely represents the true state of the world.

The process of adaptation brings with it some additional requirements. Given that some of the initial parameters may be unreliable, one may wish to rapidly discount a prior distribution on a parameter if incoming case data is in wide disagreement with it. Figure 12 shows an example taken from Spiegelhalter et al. [90] where an expert's assessment was in significant disagreement with the observed data. The parameter of interest is taken from a system for diagnosing lung disease or congenital heart disease in babies. The expert's opinion was that the proportion of cases of lung disease that exhibited grunting would be between 80 and 90%; the mean value being 85%. After sixteen cases with lung disease had been admitted, only four had in fact exhibited grunting. Figure 12 demonstrates that after these sixteen cases, the predictive probability that had the expert's judgement as starting point is still slowly being revised downward, whilst the reference prior is beginning to stabilise at what appears to be the "true" value. Spiegelhalter and Lauritzen [88] and Spiegelhalter and Cowell [89] describe significance tests which have been developed to provide a measure of the discrepancy between the expert's assessment and the observed data. This provides a formal basis for the rapid rejection of the expert's prior assessment in favour of a reference value which would provide better predictions.

A modification of the expert system shell HUGIN uses a technique of *fading* to discount "things learnt a long time ago" to make the system more prone to adapt. In aHUGIN (for *adaptive* HUGIN), the equivalent sample size is discounted by a fading factor  $\theta$  each time a new case is taken into account [65].  $\theta$  is a real number less than, although typically close to, one. For a Dirichlet density, decreasing the equivalent sample size corresponds to flattening the density. This then (rapidly) reduces its influence as case data passes through the system.

#### 7.5 Some further reading on learning parameters

Spiegelhalter & Lauritzen [88] describes some initial work on extending their work reported in Lauritzen and Spiegelhalter [53] to allow imprecision in the probabilities in a network to be accommodated and revised as a database of cases accumulates. The paper discusses the use of

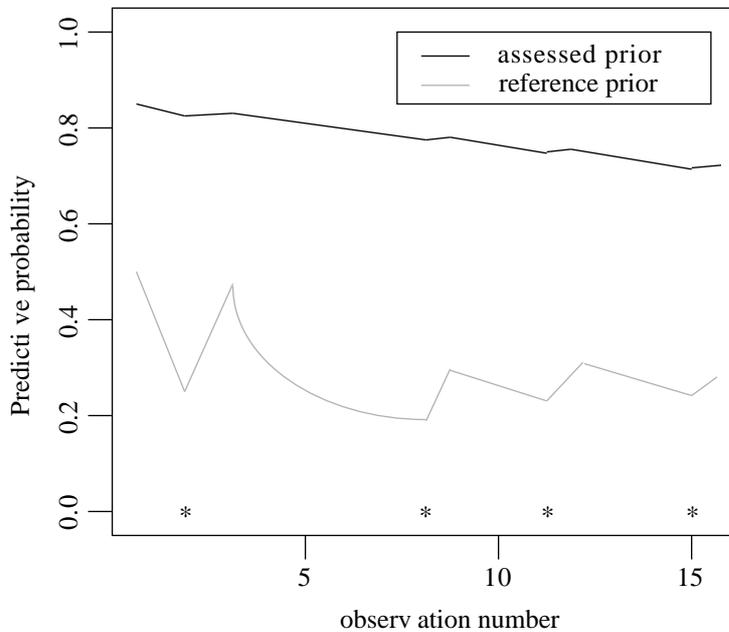


Figure 12: Revisions of predictive probabilities that the next case of lung disease be reported to a general practitioner. Starting point for the top curve is assessed prior, that for the bottom curve is the reference prior. Asterisks mark positive observations. (After Spiegelhalter et al. [90]).

discrete models, models based on Dirichlet distributions and models of the logistic regression type. Spiegelhalter et al. [91] continues the development of the Lauritzen & Spiegelhalter algorithm with a discussion of the use of data to revise the conditional probabilities in a given structure. Details are given on the use of Dirichlet priors for learning about parameters. Data is assumed not to be available on all the nodes in the network. This paper also includes some discussion of the use of data to compare models.

Dawid and Lauritzen [22] introduced the notion of a *hyper Markov law*. This is a probability distribution over a set of probability measures on a multivariate space. Their philosophy is similar to [88], in that the role of the expert is seen as a provider of a prior distribution expressing their uncertainty about the numerical parameters of a graphical model. These parameters may then be revised and eventually superseded, using case data. A major distinct contribution of [22] is to explore the details of this Bayesian approach to learning parameters in the context of *undirected* rather than directed graphs.

## 8. Learning structure from data

The previous section focused on the situation where we were uncertain about the physical probabilities in a network, but certain about the network structure. However, in general one must also allow some uncertainty in the structure of the probabilistic model. In this section we will look at techniques for learning structure as well as parameters. We will first make some general comments, and then look at some of the specific techniques that have been developed.

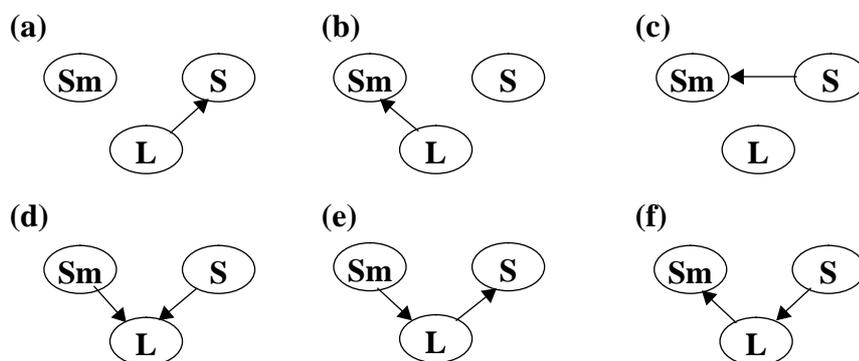


Figure 13: Some possible network structures for describing the influences between **Smoking**, **Sex** and **Lung Cancer**.

A naive approach might be to enumerate all possible network structures and then select that (or those) which maximise some suitable criterion. A Bayesian criterion, for example, will estimate the posterior probability of each network structure given the observed database. Other criteria, such as minimum message length (e.g. [100], and see below), have also been explored. Unfortunately, as the number of nodes in the network increases, it rapidly becomes infeasible to enumerate exhaustively all the possible structures. Even for a number of nodes as small as 10, the number of possible structures is approximately  $4.2 \times 10^{18}$  (the number of probabilistic-network structures for a given number of nodes  $n$  can be determined using a function published by Robinson [76]). So, in general, *unless* one has prior grounds for eliminating significant classes of possible structures, some more statistically efficient search strategy will be required. A simple example can be used to illustrate this, and one other important point. Suppose we are interested in studying the relationships between **Smoking**, **Sex** and **Lung Cancer**. After studying a small number of cases, we might have a database expressed as the relational table, Table 1:

**Table 1: An Example Database of Cases.**

case	Sm	S	L
1	T	M	T
2	F	M	F
3	T	F	F
4	F	F	F
5	T	F	T

There are 25 different networks which might be used to describe this problem as a directed graph. Some of these are shown in Figure 12. Given that a meaning has been given to the variable names, many of the possible structures will not seem intuitively plausible; one would not consider a directed influence from **Smoking** to **Sex** as being at all realistic, for example. How-

ever, if we ignore the semantics of the nodes at this stage, the additional point we wish to make is that several of the possible networks will be equivalent in the sense that they represent equivalent independence statements. Consider, for example, networks (d), (e) and (f) in figure 12. The probability distributions decompose respectively as:

$$\begin{aligned} p_d(\text{Sm}, \text{S}, \text{L}) &= p_d(\text{L} \mid \text{Sm}, \text{S})p_d(\text{Sm})p_d(\text{S}); \\ p_e(\text{Sm}, \text{S}, \text{L}) &= p_e(\text{S} \mid \text{L})p_e(\text{L} \mid \text{Sm})p_e(\text{Sm}); \\ p_f(\text{Sm}, \text{S}, \text{L}) &= p_f(\text{Sm} \mid \text{L})p_f(\text{L} \mid \text{S})p_f(\text{S}) \end{aligned}$$

Apply Bayes rule repeatedly to this last decomposition, and rearrange:

$$p(\text{Sm} \mid \text{L})p(\text{L} \mid \text{S})p(\text{S}) = p(\text{Sm} \mid \text{L})p(\text{L})p(\text{S} \mid \text{L}) = p(\text{S} \mid \text{L})p(\text{L} \mid \text{Sm})p(\text{Sm})$$

It is then clear that networks (e) and (f) have equivalent functional decompositions, and hence express equivalent independence properties.

This notion of *equivalence* of network structures [99] is important and should be kept in mind during the remaining discussion<sup>1</sup>. This leads to an important property which is expected to hold of Bayesian network structures. Let  $B_S^h$  represent the hypothesis that the physical probabilities for some specified joint space can be encoded in the network structure  $B_S$ . Then the hypotheses associated with two equivalent network structures must be identical; the structures consequently having the same prior and posterior probabilities. This property is referred to as *hypothesis equivalence*. The implication is that we should associate each hypothesis  $B_S^h$  with an equivalence class of hypotheses, rather than a single hypothesis. It should be born in mind in most of the following that the learning methods strictly pertain to learning equivalence classes of structures.

Nevertheless, in this example we *can* use expert knowledge to categorically eliminate a large number of the proposed network structures from consideration. All those with arrows from **Lung Cancer** to **Sex**, or from **Smoking** to **Sex** should certainly go. It might also be reasonably safe to remove all those with arrows from **Lung Cancer** to **Smoking**, but this is perhaps beginning to impose a judgement rather than a “technological” constraint. Now, given the remaining possible network structures, let  $B_S^h$  represent the hypothesis that the database D of Table 1, is a random sample from the (equivalence class of a) network structure  $B_S$ . Then, we simply select that hypothesis whose posterior probability given the data is a maximum. If  $\xi$  represents the current background knowledge, as before, and  $c$  is a normalisation constant, then this can be calculated from the data in table 1 using

$$p(B_S^h \mid D, \xi) = c p(B_S^h \mid \xi) p(D \mid B_S^h, \xi) \quad \text{eqn. 9.}$$

Full details for the computation of this expression can be found in [36].

At a high level view, that is all there is to learning structure in probabilistic networks. In practice, if the user believes that only a few network structures are possible, they can directly assess the priors for the possible network structures and their parameters and then compute the posterior probabilities as described. However, if one is trying to learn a model of a situation about which there is little prior knowledge, then some extra mechanism must be in place to control the explosion of possible structures.

---

1. More on testing for and characterising equivalent network structures can be found in [12] and [99].

Note that, in contrast to the previous section, the remainder of this section only covers the situation where complete data is available.

### 8.1 Search and score approaches to learning

A key computational problem is that a rigorous Bayesian approach to learning structure involves averaging over all possible models [37]. Since the number of possible structure hypotheses is more than exponential in the number of variables, this is clearly intractable in general. In some cases, model averaging using Monte-Carlo methods can be statistically efficient and yield good predictions [59]. However, the problem can be simplified even further if positive answers can be provided to the following questions. Can a sufficiently accurate approximation be provided by including only a small fraction of the hypotheses in the sum? If so, which hypotheses?

These are hard questions to address theoretically. However, progress in learning probabilistic networks was significantly advanced when several workers showed experimentally that even a single “good” network structure can offer a sufficiently accurate approximation ([2], [18], [40]). This is one approach which statisticians have been using for decades in the context of other types of models, and is referred to as *model selection*. An alternative is *selective model averaging*, in which a manageable number of “good” models is selected from all possible models. The latter is more complex, because it is advantageous to identify network structures that are significantly different (so at least representative of the whole distribution of models). Partly for this reason, and partly because model selection is by far the most widely discussed in the literature, we will only discuss model selection. The reader is referred to [57] for a discussion of selective model averaging.

So the question then is, how to identify “good” hypotheses? A widely used general strategy is to augment the scoring criterion with a search algorithm. As before, prior knowledge, a database and a set of network structures are taken, and the goodness of fit of those structures to the prior knowledge and the data is computed according to some criterion. The search algorithm is used to identify those structures to be scored.

Given the earlier discussions, an obvious choice of a scoring criterion for a network structure (equivalence class) is the relative posterior probability of that structure given the database. One might, for example, compute either  $p(D, B_S^h | \xi) = p(B_S^h | \xi) p(D | B_S^h, \xi)$ , or the relative posterior probability  $p(B_S^h | D, \xi) / p(B_{S_0}^h | D, \xi)$ , where  $B_{S_0}^h$  is some reference network structure. If this is computed using the assumption of Dirichlet distributions, this is sometimes referred to as the *Bayesian Dirichlet* (BD) criterion.

A number of alternative criteria have been proposed. A scoring metric that uses relative posterior probability in conjunction with heuristics based on the principle of Occam’s Razor was proposed by Madigan and Raftery [57]. Others include the A information criterion (AIC) [1], the Bayesian information criterion (BIC) [82] and minimum description length (MDL, or sometimes minimum message length) [74]. These last two differ only by a minus sign and asymptotically approximate the BD criterion. That is, in the limit as the number of cases in the database approaches infinity, BIC and MDL give the same scores as the BD metric with uniform priors on structures [49]. This is rarely achieved in practice due to limits on the sizes of datasets that are available. However, BIC is easy to use and does not require the evaluation of prior distributions. Consequently, it can be a practical criterion to use in the appropriate cir-

cumstances [50].

A potential difficulty with the posterior-probability criteria is the need to assign prior probabilities to each possible network structure. However, a statistically efficient (in terms of data demands) method for doing this has been described by Heckerman et al. [40]. This requires only the assessment of a *prior network structure* for the domain, the user’s “best guess”, and a single constant. However, if a user is willing, they can provide more detailed knowledge by assessing different penalties for different nodes, and for different parent configurations for each node [9]. Alternatively, they might categorically assert that some arcs in the prior network must be present. Similar practical approaches to the necessary assignment of priors to parameters for all possible network structures have been discussed in [9], [17], [18], [40], [91].

## 8.2 Search strategies

Having selected a scoring criterion, such as BD, and assessed the relevant priors, one then needs to use some search strategy to seek out a preferred network structure. In the case of a network structure with the highest posterior probability, this is often called the maximum a posteriori (MAP) structure (again, strictly an equivalence class of structures). All such search methods make use of the decomposability property of the scoring criteria. Given a network structure for a domain, we say that a measure on that structure is *separable* if it can be written as a product of measures, each of which is a function of only one node and its parents. Most metrics for complete databases are separable.

The BD criterion is one example of a separable criterion [36]. Thus, we can write:

$$p(D, B_S^h | \xi) = \prod_{i=1}^n s(X_i | \Pi_i) \quad \text{eqn. 10.}$$

where  $s(X_i | \Pi_i)$  is only a function of  $X_i$  and its parents. So, given such a separable criterion, we can compare the score for two network structures that differ by the addition or deletion of arcs pointing to  $X_i$  by computing only the term  $s(X_i | \Pi_i)$  for both structures. If an arc between, for example, nodes  $X_i$  and  $X_j$  is reversed then the terms  $s(X_i | \Pi_i)$  and  $s(X_j | \Pi_j)$  need to be calculated. A general technique for search strategies is then to make successive arc changes to the network, and employ the property of decomposability to evaluate the merit of each change.

Cooper & Herskovits [18] describes the use of a greedy search algorithm for identifying the most probable structure given some test data. Aliferis & Cooper [2] evaluates the accuracy of K2, a specific instantiation of a greedy search algorithm, using simulated data. The use of simulated data from well-specified (gold-standard) models allows the accuracy with which a specific technique learns the model structure to be measured directly. The alternative is to measure the accuracy indirectly by assessing the predictive accuracy of the resulting model [42]. Using their direct measurements, Aliferis & Cooper report that the mean percentage of correctly found arcs was 91.6%, whilst the mean ratio of superfluous arcs was 4.7%. Further empirical studies leading to refined search algorithms can be found in [13] and [92].

## 8.3 Assumptions made to enable the computation of criteria

In the general case, given a database of cases  $D$ , we wish to learn the structure  $S$  (a directed acyclic graph), and the parameters  $\theta$  (conditional probability distributions) of a network, where each node in the structure corresponds to an element from a given domain of variables. In order to compute the BD criterion in closed form a number of hypotheses need to be made [31]:

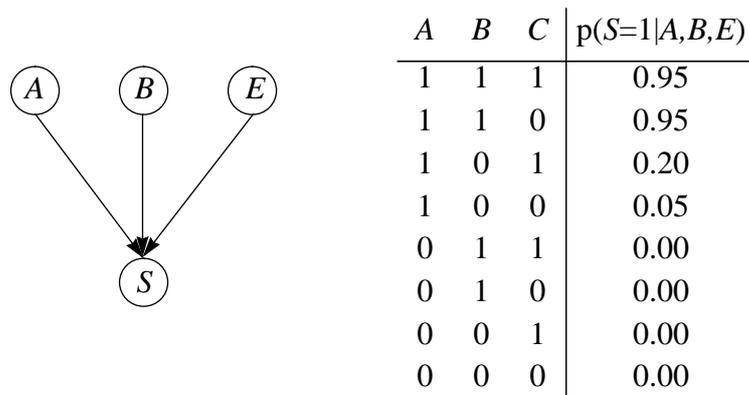


Figure 14: A simple network structure and its associated conditional probability table.

- the database  $D$  is a multinomial sample from some network  $(S, \theta)$ ;
- for each network, the parameters associated with one node are independent of the parameters associated with other nodes (global independence), and the parameters associated within a node given one instance of its parents are independent of the parameters of that node given other instances of its parent nodes;
- if a node has the same parents in two distinct networks then the distribution of the parameters associated with this node are identical in both networks (parameter modularity);
- each case is complete;
- the distribution of the parameters associated with each node is Dirichlet.

As indicated earlier, the last two assumptions are made so that the distributions of the parameters stay within the same conjugate family of distributions, with sampling. Interestingly, the first three assumptions, together with one additional assumption do in fact imply the fifth assumption. This was demonstrated by Geiger and Heckerman [31]. The additional assumption is that data can not help to discriminate two network structures that encode the same sets of probability distributions (likelihood equivalence).

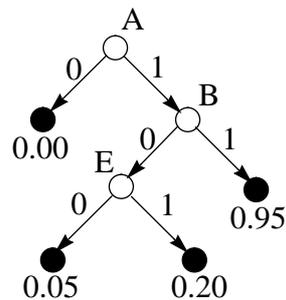
As has been indicated, a wide variety of learning algorithms are extant. Buntine [10] provides a framework from which a wide variety of data analysis and learning algorithms can be constructed. Here graphical models are used at a meta-level to represent the task of learning object level models. Buntine hopes that this work will form the basis of a computational theory of Bayesian learning.

#### 8.4 Refinement of Local Model Structure

Although the global structure of probabilistic networks can be exploited to reduce the number of parameters that need to be elicited or learned, the broad search strategies of the preceding section still assume that the number of parameters is *locally* exponential in the number of parents of a node. This can be illustrated using the simple example of figure 14 (one might think of the nodes  $A$ ,  $B$ ,  $E$ , and  $S$  as corresponding to the events “alarm armed”, “burglary”, “earthquake” and “load alarm sound”, respectively). The naive representation of the conditional

A	B	C	$p(S=1 A,B,E)$
1	1	1	0.95
1	1	0	0.95
1	0	1	0.20
1	0	0	0.05
*			0.00

(a)



(b)

Figure 15: Two alternative representations of a local CPT structure. The default table in (a) requires 5 parameters, whilst the decision tree of (b) requires 4.

probability table (CPT) on the right hand side of the figure requires  $2^3=8$  parameters. Yet, by inspecting the table, we can see that there is scope for a representation which uses fewer parameters. Firstly, we have a default condition that if the alarm is not armed ( $A=0$ ) there will be no alarm sound ( $S=0$ ) whatever the state of the other two parents. In addition, if the alarm is armed and there is a burglary ( $B=1$ ), then the load alarm sound has the same probability of occurrence whether an earthquake is happening or not.

Figure 15 shows two alternative representations of this local structure which use fewer parameters than the original structure; a default table requiring 5 parameters and a decision tree with 4. In the default table, a single default probability is provided for all those parameters that are not explicitly listed. In the case of the decision tree, each leaf describes a probability for node S. The arcs are labelled with the possible states of each parent node, and one travels down the tree taking the appropriate branches to find the probability of S given the selected states of the parent nodes. The default table captures only the default condition, whilst the decision table also captures the overriding impact of a burglary whenever the alarm is armed; hence the difference in the parameters required.

There is an important benefit to be gained by modifying the learning algorithms so that they can recognise such a local structure where appropriate. Since fewer parameters are needed, the estimation of these parameters is more efficient for a given amount of data. Furthermore, most scoring criteria aim for a balance between complexity of the preferred network and the accuracy with which it represents the frequencies in the database from which it is learnt. This means that in the example just given, a simpler network might be preferred using, say, just two of the three possible parents as this will reduce the number of parameters needed in the naive tabular representation of the CPT. Thus the search algorithm may show a bias against the “true” more complex structure.

Friedman and Goldszmidt [28] have proposed two modifications to the minimum description length (MDL) criterion for learning probabilistic networks. These allow for a compact encoding of default tables and decision trees, respectively. Both criteria are separable (see Section 8.2) and so may be used in conjunction with a search strategy employing local additions, removals and reversals of edges to learn a structure in a similar way to those already discussed.

Friedman and Goldszmidt then carried out a series of experiments to evaluate the effectiveness of these new criteria. These experiments supported the hypotheses that the structures learnt using the new criteria reduce error by preferring more complex structures in those situations where the associated CPT can be described using fewer than exponential parameters. In addition, the estimation of these parameters was more robust as they are based on relatively larger samples. Analogous changes to the BD metric are also proposed, although not explored.

The compact representation of local CPT structure is a very important topic. Further discussions of the fundamental knowledge representation issues can be found in [9], [38], [67] and [94].

## 9. Can we learn causation, not just correlation, from data?

This final section is somewhat more controversial than the work described in the preceding sections. We wish to discuss the possibility of learning causation from data. The controversy arises partly because the term “causation” is open to interpretation; it has been the subject of much philosophical debate over the last two hundred years. Nevertheless, statistics are regularly used to demonstrate causality in some sense. For example, controlled experiments are used to see whether a new drug “causes” side effects or not. The micro-molecular mechanisms underlying the causality may not be understood, but the “molar level” statistical correlations may be sufficiently convincing for the drug to be withdrawn. A more concrete example was the inference of a causal relationship between tuberculosis and certain sources of drinking water in London at the turn of this century. This led to improvements in sanitation which had a far more significant impact on the health of the population than most discoveries in health care. So, causal models implied by data can be extremely valuable, and any advances in the development of techniques for deriving them is to be welcomed.

With regard to the aforementioned controversy, the author of this paper has in the past been somewhat confused by the use of the word “causality” in connection with graphical models. This is perhaps due to an overemphasis on what may be termed *micromediation*: the specification of causal connections at fine levels of detail and using fine timescales. It is helpful to remember instead that many of the causal links in a graphical model (both learnt and provided) are expressed at the *molar* level: in terms of large and often complex objects (populations; distributions). Cook and Campbell [16] provide a set of statements which captures well most of the key features of the molar view of causality:

- Causal assertions are meaningful at the molar level even when the ultimate micromediation is not known.
- Because they are contingent on many other conditions and causal laws, molar causal laws are fallible and hence probabilistic.
- The effects in molar laws can be the result of multiple causes; one might (casually) speak of *the* cause having been learnt, but that does not necessarily imply the *only* cause.
- While it is easiest for molar causal laws to be detected in closed systems with controlled conditions, field research involves mostly open systems.
- Effects follow causes in time, even though they may be instantaneous at the level of micromediation.
- Some causal laws can be reversed, with cause and effect interchangeable.
- The paradigmatic assertion in causal relationships is that the manipulation of a cause will result in the manipulation of an effect.

The last statement is of particular significance. Causal information of this kind is needed in order to make predictions in the face of intervention. Conversely, as in the case of good experimental science, causal information can be learned with interventional data. There are cases in the social sciences and in medicine where interventional data can be obtained through, for example, randomised trials. However, there are also many cases where this is impractical, or even immoral. The alternative then is to try to learn causal information from data alone.

The notion of conditional independence provides a link between learning graphical models from data and learning causality. For many domains, there is a direct connection between lack of cause and conditional independence. This connection is sometimes called the *causal Markov assumption* [93]:

- A domain with causal relationships given by graph  $\mathcal{G}$  exhibits the conditional independence assumptions determined by d-separation applied to  $\mathcal{G}$ .

For example, measles “causes” Koplik’s spots, measles “causes” red spots, but the probabilities of the two symptoms are independent once measles is confirmed to be present. Several authors have reported that the causal Markov assumption is appropriate for many domains (e.g. [35], [67], [93]). Learning causality is essentially just the inversion of this; given certain conditional independence assertions learned from data, determine those graphs which could have produced those assertions. By the causal Markov assumption, those graphs must include the true causal graph.

Caution is needed here. It is possible that two or more alternative network structures may record the same conditional independence assertions. Consider, for example, the set of variables  $\mathbf{X} = \{X, Y, Z\}$  and the three possible network structures  $X \rightarrow Y \rightarrow Z$ ,  $X \leftarrow Y \rightarrow Z$ , and  $X \leftarrow Y \leftarrow Z$ . All three structures represent only the independence assertion that  $X$  and  $Z$  are conditionally independent given  $Y$ . In this sense, they are equivalent structures. This was illustrated in Section 8. Verma and Pearl define two Bayesian-network structures for a set of variables  $\mathbf{X}$  as *independence equivalent* if they represent the same set of independence assertions [99]. Verma and Pearl also provide a simple condition for assessing whether two network structures are independence equivalent. Strictly, we need the stronger notion of *distribution equivalence*; two network structures for  $\mathbf{X}$  are distribution equivalent if they encode the same set of probability distributions for  $\mathbf{X}$ . However, for the purposes of this discussion, the important point is to remember that one is strictly learning equivalence classes of structures rather than individual structures. Further discussion of the above two forms of equivalence can be found in [37].

In [60] Meek defines a *complete causal explanation* of a dependency model  $\mathcal{M}$  (a set of conditional independence statements) as a directed acyclic graph  $\mathcal{G}$  where the set of conditional independence facts entailed by  $\mathcal{G}$  is exactly the set of facts in  $\mathcal{M}$ . It follows that if  $\mathcal{G}$  is a complete causal explanation of a dependency model  $\mathcal{M}$ , and  $\mathcal{G}'$  is equivalent to  $\mathcal{G}$ , then  $\mathcal{G}'$  must also be a complete causal explanation of  $\mathcal{M}$ . Although one cannot discriminate between equivalent causal explanations, one can usefully ask, what are the causal relationships that are common to every causal explanation of the set of independence facts in a sample? Meek provides algorithms for answering this question in the case where the causal explanations are consistent with some background knowledge consisting of a set of directed edges that are required, and a set of directed edges that are forbidden.

Because of its import in epidemiology, economics, the social sciences and other “soft” sciences, learning causal networks has a long history. The seminal references for this topic are the book by Spirtes et al. [93], and the paper by Verma & Pearl [99]. We shall refer to this general approach under *independence search-based* techniques for learning causality (ISC), to discriminate them from the Bayesian approaches to learning structure of the previous section. The techniques appear very different. However, as the data set approaches infinity there is an asymptotic correspondence. So, it would be foolish to regard the two approaches as competitors; learning causality is a deep problem which warrants viewing from different perspectives.

A possible concern with the use of collected data to learn causality is that a fundamental premise of scientific methodology is that some form of intervention or experimentation is necessary in order to discover causality. For example, in the case of smoking and lung cancer of section 8, one might conclude that smoking had a causal relationship with lung cancer if by intervening to stop smoking, one reduced the incidence of lung cancer. However, Spirtes et al. [93] and Pearl and Verma [71] have shown that, *under certain assumptions*, passive observation is sufficient to determine all or a part of (equivalence classes of) the causal structure of a system under study<sup>1</sup>. As mentioned earlier in this section, this is the goal one is aiming for.

Wermuth and Lauritzen [101] discuss the use of graphical chain models in the formulation of *substantive research hypotheses*. Their interest was from the social sciences, to produce statistical models “capturing characteristics, behaviour, abilities, attitudes of people or historical and environmental conditions”. In [101] they describe simple criteria for identifying equivalent statistical models from graphs. This is an important point to take on board when discussing the learning of causality, as it is not possible to differentiate between alternative substantive research structures if they correspond to equivalent graphical models.

The Bayesian learning techniques discussed in this paper can be viewed as providing a “soft” version of the independence search-based techniques, using some expert judgement to initiate the search. The required technical material, using a Bayesian approach, has essentially all been covered in the preceding sections. So the main motivation behind this section was to air the issues rather than describe further technology. For further reading, a small scale, but informative, example of learning causality using real-world data can be found in [37]. This data has also been studied using non-Bayesian techniques by Wittaker [104] and by Spirtes *et al.* [93]. The papers on the Rubin Causal Model referred to in section 7.3 are also relevant, although this work does not make significant use of graphical models as yet.

A number of problems still remain to be solved. For example, the techniques work best if all variables in the domain of interest are observable. Difficulties arise if hidden variables are permitted. By way of illustration, the graph  $X \leftarrow a \rightarrow Z \leftarrow b \rightarrow Y$  implies the same set of conditional independence statements on the variable  $X, Y, Z$  as the graph  $X \rightarrow Z \leftarrow Y$ . However, the former does not present  $X$  as a cause of  $Z$ , whilst the latter does. The independence search-based techniques can be used to identify when an association must be attributed to a hidden common cause, [71], [93]. The Bayesian techniques, on the other hand, still have difficulties when hidden variables are involved.

---

1. These are the Causal Markov Assumption, and a further assumption called *faithfulness* [92]. In the Bayesian learning case, the latter follows from the assumption that the parameters have a probability density function [37].

## 10. Conclusions

This paper has described an area of research in expert systems which is rapidly maturing, and which is beginning to find significant real-world application. Application of the techniques described in this paper have been described by Madigan and Raftery [57], Lauritzen et al. [55], Singh and Provan [85], and Friedman and Goldszmidt [28]. In addition, a number of research groups have developed software systems specifically for learning graphical models. The work described by Spirtes et al. [93] has been further developed into a program called TETRAD II for learning about cause and effect. This is reported in Scheines et al. Badsberg [5] and Højsgaard et al. [43] have built systems which can learn mixed graphical models (chain graphs) using a variety of criteria for model selection. A widely used benchmark in the learning literature is a system called BUGS, which was created by Thomas, Spiegelhalter and Gilks [96]. BUGS takes a learning problem specified as a Bayesian network and compiles this problem into a Gibbs-sampler computer program. Finally, a software package developed by Radford Neal is available on the internet (<http://www.cs.toronto.edu/~radford/fbm.software.html>). Neal's package supports Bayesian regression and classification models [63]. A web page listing currently available commercial and research software for learning belief networks can be found at <http://bayes.stat.washington.edu/almond/belfit.html>.

This all goes to show that learning probabilistic networks from data is a healthy and buoyant research area with great potential for application. We hope that this paper will go some way to raising awareness of the possibilities for exploiting this work.

## 11. References

- [1] Akaike H. 1974. A New Look at Statistical Model Identification. *IEEE Trans. Automatic Control*, **19**, 716-723.
- [2] Aliferis C.F. & Cooper G.F. 1994 An evaluation of an algorithm for inductive learning of Bayesian belief networks using simulated data sets. In: *Uncertainty in Artificial Intelligence - Proceedings of the 10th Conference*, 8-14.
- [3] Allen J., Fikes. R. & Sandewall E. 1991 *KR-91, Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*. Cambridge, MA: Morgan Kauffman.
- [4] Angrist J.D., Imbens G.W. & Rubin D.B. 1996 Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, **91**, 444-455.
- [5] Badsberg J. 1992. Model search in contingency tables in CoCo. In: Dodge Y. & Wittaker J. (eds), *Computational Statistics*, Heidelberg: Physica Verlag, 251-256.
- [6] Bernardo J.M. and Giron F.J. 1988. A Bayesian analysis of simple mixture problems. In *Bayesian Statistics 3* (Bernardo J.M., DeGroot M.H., Lindley D.V. and Smith A.F.M., eds), Oxford University Press, 67-78.
- [7] Bernardo J.M. & Smith A.E.M. 1994 *Bayesian Theory*. Chichester: John Wiley.
- [8] Besnard P. & Hanks S. 1995. *Uncertainty in Artificial Intelligence: Proceedings of the Eleventh Conference*, San Fransisco: Morgan Kaufmann.
- [9] Buntine W. 1991. Theory Refinement in Bayesian Networks. In *Proc. Seventh Conference on Uncertainty in Artificial Intelligence*, San Francisco: Morgan Kaufmann, 52-60.
- [10] Buntine W. 1994 Operations for Learning with Graphical Models. *Journal of Artificial Intelligence*, **2**, 159-225.

- [11] Buntine W. 1996 A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, **8**, 195-210.
- [12] Chickering D. 1995 A Transformational Characterisation of Equivalent Bayesian Network Structures. In Besnard & Hanks (eds), 87-98.
- [13] Chickering D. 1996 Learning equivalence classes of Bayesian-network structures. In Horvitz E. & Jensen F. (eds.).
- [14] Chung K.L. 1979 *Elementary Probability Theory with Stochastic Processes*. New York: Springer-Verlag.
- [15] Çinlar E. 1975. *Introduction to Stochastic Processes*. Prentice Hall.
- [16] Cook T.D. & Campbell D.T. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago: Rand McNally College Publishing Company.
- [17] Cooper G.F. & Herskovits E. 1992 *Bayesian Method for the Induction of Probabilistic Networks from Data*. Technical Report SMI-91-1, Section on Medical Informatics, Stanford University.
- [18] Cooper G.F. & Herskovits E. 1992 Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, **9**, 309-347.
- [19] Cowell R.G., Dawid A.P. and Sebastiani P. 1996. *A comparison of sequential learning methods for incomplete data*. In *Bayesian Statistics 5*, pp. 533-542, Oxford: Clarendon Press.
- [20] Cox D. 1993. Causality and Graphical Models. *Bull. Int. Stat. Inst.*, Proc. 49th Session 1, 365-372.
- [21] Dawid A.P. 1979 Conditional Independence in Statistical Theory. *J.R.. Statist. Soc. B*, **49**, 1-31.
- [22] Dawid A.P. & Lauritzen S.L. 1993 Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *The Annals of Statistics*, **21**, 1272-1317.
- [23] DeGroot M.H. 1970 *Optimal Statistical Decisions*. McGraw-Hill: New York.
- [24] Dempster A., Laird N. & Rubin D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, **39**, 1-38.
- [25] Druzdzel M.J. & van der Gaag L.C. 1995 Elicitation of Probabilities for Belief Networks: Combining Qualitative and Quantitative Information. In: Besnard & Hanks (eds), 141-148.
- [26] Dubois D., Wellman M.P., D'Ambrosio B.D. & Smets P. 1992 *Uncertainty in Artificial Intelligence: Proceedings of the Eighth Conference*, San Mateo, CA: Morgan Kaufmann.
- [27] Edwards D. 1995 *Introduction to Graphical Modelling*. New York: Springer-Verlag.
- [28] Friedman N. & Goldszmidt M. 1996a. Learning Bayesian Networks with Local Structure. In *UAI '96*.
- [29] Friedman N. & Goldszmidt M. 1996b. Building classifiers using Bayesian networks. In *Proceedings of AAAI-96*, Menlo Park, CA: AAAI Press, 1277-1284.
- [30] Geiger D. & Pearl J. 1988. On the logic of causal models. *Proc. 4th Workshop on Uncertainty in AI*, St Paul, Minn., 136-147.
- [31] Geiger D. & Heckerman D. 1995 A Characterisation of the Dirichlet Distribution with Application to Learning Bayesian Networks. In: Besnard & Hanks (eds), 196-207.
- [32] Good I.J. 1965 *The Estimation of Probabilities*. Cambridge, MA: MIT Press.

- [33] Hastings W. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97-109.
- [34] Heckerman D. 1990 Probabilistic similarity networks. *Networks*, **20**, 607-636.
- [35] Heckerman D. Mamdani A. & Wellman M. 1995 Real-world applications of Bayesian networks, *Comm. ACM*, **38**.
- [36] Heckerman D. 1996a Bayesian Networks for Knowledge Discovery. In: Fayyad U.M., Piatetsky-Shapiro G., Smyth P and Uthurusamy R. (eds), *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press, 273-305.
- [37] Heckerman D. 1996b. *A Tutorial on Learning with Bayesian Networks*. Technical Report MSR-TR-95-06, Microsoft Corporation, Redmond, USA.
- [38] Heckerman D. & Breese J.S. 1994. A new look at causal independence. In Lopez de Mantaras & Poole (eds), 286-292.
- [39] Heckerman D. & Shachter R. 1995 Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research*, **3**, 405-430.
- [40] Heckerman D., Geiger D. & Chickering D. 1995. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, **20**, 197-243.
- [41] Henrion M., Shachter R.D., Kanal L. & Lemmer J.F. (eds.) 1990 *Uncertainty in Artificial Intelligence 5*. Amsterdam: Elsevier Science Publishers B.V. (North Holland).
- [42] Herskovits E. 1991 *Computer-Based Probabilistic-Network Construction*. PhD thesis, Stanford University.
- [43] Højsgaard S., Skjøth F. and Thiesson B. 1994. *User's guide to BIOFROST*. Technical report, Department of Mathematics and Computer Science, Aalborg, Denmark.
- [44] Horvitz E. & Jensen F. (eds.) 1996. *Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Conference*, San Fransisco: Morgan Kaufmann.
- [45] Jensen F.V. 1996 *An introduction to Bayesian Networks*. London: UCL Press Ltd.
- [46] Jensen F.V., Lauritzen S.L. & Olesen K.G. 1989 *Bayesian Updating in Recursive Graphical Models by Local Computations*. Technical Report R 89-15, Department of Mathematics and Computer Science, University of Aalborg, Denmark.
- [47] Jensen F.V., Olesen K.G. & Andersen S.K. 1990 An Algebra of Bayesian Belief Universes for Knowledge-Based Systems. *Networks*, **20**, 637-659.
- [48] Holland P.W. 1986 Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**, 945-960.
- [49] Kass R. & Raftery A. 1993 *Bayes Factors and Model Uncertainty*. Technical Report 571, Dept. of Statistics, Carnegie Mellon University.
- [50] Kass R. & Raftery A. 1995, Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795.
- [51] Lauritzen S.L. 1995. The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, **19**, 191-210.
- [52] Lauritzen S.L. 1996. *Graphical Models*. Oxford: Clarendon Press.
- [53] Lauritzen S.L. & Spiegelhalter D.J. 1988. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. R. Stat. Soc. Ser. B*, **50**, 157-224.
- [54] Lauritzen S.L., Dawid A.P., Larsen B.N. & Leimer H-G. 1990 Independence Properties

- of Directed Markov Fields. *Networks*, **20**, 491-505.
- [55] Lauritzen S.L., Thiesson S. & Spiegelhalter D. 1994. Diagnostic systems created by model selection methods: a case study. In: Cheeseman P. & Oldford R (eds) *AI and Statistics IV*, New York: Springer-Verlag Lecture Notes in Statistics vol 89, 143-152.
- [56] Lopez de Mantaras R. & Poole D. (eds) 1994. *Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference*, San Fransisco: Morgan Kaufmann.
- [57] Madigan D. & Raftery A. 1994. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *J. Am. Statist. Association*, **89**, 1535-1546.
- [58] Madigan D. & York J. 1995. Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215-232.
- [59] Madigan D., Raftery A., Volinsky C. & Hoeting J. 1996. Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models*, Portland OR.
- [60] Meek C. 1995 Causal inference and causal explanation with background knowledge. In: Besnard & Hanks (eds), 403-410.
- [61] Meilijson I. 1989. A fast improvement to the EM algorithm on its own terms. *J. Roy. Statist. Soc. B.*, **51**(1), 127-138.
- [62] Neal R. *Probabilistic inference using Markov chain Monte Carlo methods*. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto.
- [63] Neal R. 1996 *Bayesian Learning for Neural Networks*. New York: Springer-Verlag.
- [64] Nilsson N. 1986 Probabilistic logic. *Artificial Intelligence*, **28**, 71-87.
- [65] Oleson K.G., Lauritzen, S.L. & Jensen F.V. 1992 aHugin: A System Creating Adaptive Causal Probabilistic Networks. In: Dubois et al. (eds), 223-229.
- [66] Pearl J. 1986 A constraint-propagation approach to probabilistic reasoning. In: Kanal L.N. and Lemmer J.F. (eds), *Uncertainty in Artificial Intelligence*, Amsterdam, North-Holland, 3718-382.
- [67] Pearl J. 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, Ca: Morgan Kauffman.
- [68] Pearl J. 1995 On the Testability of Causal Models with Latent and Instrumental Variables. In: Besnard & Hanks (eds), 435-443.
- [69] Pearl J. 1995 Causal diagrams for empirical research. *Biometrika*, **82**, 669-710.
- [70] Pearl J., Geiger D. & Verma T. 1990. The Logic of Influence Diagrams. In: Oliver R.M & Smith J.Q. (eds) 1990, *Influence Diagrams, Belief Nets and Decision Analysis*, Chichester: John Wiley.
- [71] Pearl J. & Verma T.S. 1991 A theory of inferred causation. In: Allen J., Fikes R. & Sandewall E. (eds), 441-452.
- [72] Richardson T. 1997. Extensions of undirected and acyclic, directed graphical models. In: *Proc, 6th Conference on Artificial Intelligence in Statistics*, Ft Lauderdale, 407-419.
- [73] Ripley B. 1987. *Stochastic Simulation*. Chichester: John Wiley & Sons.
- [74] Rissanen J. 1987. Stochastic Complexity (with discussion). *J. Roy. Statist. Soc. B*, **49**, 223-239.
- [75] Robins J. 1986 A new approach to causal inference in mortality studies with sustained exposure results. *Mathematical Modelling*, **7**, 1393-1512.

- [76] Robinson R.W. 1977 Counting unlabelled acyclic digraphs. In: Dold A. & Eckmann B. *Lecture notes in Mathematics 622: Combinatorial Mathematics V*. Berlin: Springer-Verlag.
- [77] Rubin D.B. 1974 Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**, 688-701.
- [78] Rubin D.B. 1978 Bayesian inference for causal effects: The role of randomisation. *Annals of Statistics*, **6**, 34-58.
- [79] Rubin D.B. 1991 Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. *Biometrics*, **47**, 1213-1234.
- [80] Russell S. and Norvig P. 1995. *Artificial Intelligence: A Modern Approach*. New Jersey: Prentice-Hall.
- [81] Russell S., Binder J., Koller D. & Kanazawa K. 1995 Local learning in probabilistic networks with hidden variables. In: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, San Fransisco: Morgan Kaufmann, 1146-1152.
- [82] Schwarz G. 1978. Estimating the Dimensions of a Model. *Annals of Statistics*, **6**, 461-464.
- [83] Shachter R.D. 1986. Evaluating Influence Diagrams. *Operations Research*, **34**, 871-882.
- [84] Shachter R.D. 1990 Evidence absorption and propagation through evidence reversals. In: *Uncertainty and Artificial Intelligence 5*. Henrion et al. (eds), 75-190.
- [85] Singh M. and Provan G. 1995. *Efficient learning of selective Bayesian network classifiers*. Technical report MS-CS-95-36, Computer and Information Science Department, University of Pennsylvania, Philadelphia, PA.
- [86] Smith A.F.M. and Makov U.E. 1978 A quasi-Bayes sequential procedure for mixtures. *J. R. Statist. Soc. Ser. B*, **40**, 106-111.
- [87] Smyth P., Heckerman D. & Jordan M.J. 1996 *Probabilistic Independence Networks for Hidden Markov Probability Models*. Microsoft Research Technical Report MSR-TR-96-03.
- [88] Spiegelhalter D.J. & Lauritzen S.L. 1990 Sequential Updating of Conditional Probabilities on Directed Graphical Structures. *Networks*, **20**, 579-605.
- [89] Spiegelhalter D.J. & Cowell R. 1992 Learning in Probabilistic Expert Systems. In *Bayesian Statistics 4*, (Bernardo J.M., Berger J.O., Dawid A.P. and Smith A.F. eds), Oxford University Press, 447-465.
- [90] Spiegelhalter D.J., Harris N., Bull K. & Franklin R. 1991 *Empirical evaluation of prior beliefs about frequencies: methodology and a case study in congenital heart disease*. BAIES Report BR-24, MRC Biostatistics unit, Cambridge, England.
- [91] Spiegelhalter D.J., Dawid A.P., Lauritzen S.L. & Cowell R.G. 1993 Bayesian Analysis in Expert Systems (with discussion). *Statistical Science*, **8**, 219-283.
- [92] Spirtes P. & Meek C. 1995 Learning Bayesian networks with discrete variables from data. In: Proc. First International Conference on Knowledge Discovery and Data Mining, Montreal QU, Morgan Kaufmann.
- [93] Spirtes P., Glymour, C. & Scheines R. 1993 *Causation, Prediction and Search*. New York: Springer-Verlag.
- [94] Srinivas S. 1993. A generalisation of the noisy-or model. In *Uncertainty in Artificial*

- Intelligence: Proceedings of the Ninth Conference*, San Francisco: Morgan Kaufmann, 208-215.
- [95] Taylor S.J. 1966 *Introduction to Measure and Integration*. Cambridge University Press.
- [96] Thomas A., Spiegelhalter D.J. & Gilks W.R. 1992. BUGS: A program to perform bayesian inference using gibbs sampling. *Bayesian Statistics 4*, (Bernardo J.M., Berger J.O., Dawid A.P. and Smith A.F. eds), Oxford University Press, 837-842.
- [97] Titterington D.M. 1976 Updating a diagnostic system using unconfirmed cases. *Applied Statistics*, **25**, 238-247.
- [98] Titterington D.M., Smith A.F.M. and Makov U.E. 1985. *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley.
- [99] Verma T. & Pearl J. 1990. Equivalence and Synthesis of Causal Models. In *Proc. Sixth Conference on Uncertainty in Artificial Intelligence*, San Francisco: Morgan Kaufmann, 220-227.
- [100] Wallace C.S. & Korb K. 1996. Learning a Linear Causal Model by MML. *Proc. UNICOM Seminar on Intelligent Data Management*, Chelsea Village, London, UK.
- [101] Wermuth N. & Lauritzen S.L. 1990 On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models. *J. R. Stat. Soc. B*, **52**, 21-50.
- [102] Wilks S.S. 1963. *Mathematical Statistics*. New York: John Wiley.
- [103] Winkler R. 1967 The assessment of prior distributions in Bayesian analysis. *American Statistical Association Journal*. **62**, 776-800.
- [104] Wittaker J. 1990 *Graphical Models in Applied Multivariate Statistics*. Chichester: John Wiley.