# A Tutorial on Learning Bayesian Networks

David Heckerman
heckerma@microsoft.com

March 1995 (Revised July 1995)

Technical Report
MSR-TR-95-06

Microsoft Research
Advanced Technology Division
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

**Abstract**

We examine a graphical representation of uncertain knowledge called a Bayesian network. The representation is easy to construct and interpret, yet has formal probabilistic semantics making it suitable for statistical manipulation. We show how we can use the representation to learn new knowledge by combining domain knowledge with statistical data.

# 1 Introduction

Many techniques for learning rely heavily on data. In contrast, the knowledge encoded in expert systems usually comes solely from an expert. In this paper, we examine a knowledge representation, called a *Bayesian network,* that lets us have the best of both worlds. Namely, the representation allows us to learn new knowledge by combining expert domain knowledge and statistical data.

A Bayesian network is a graphical representation of uncertain knowledge that most people find easy to construct and interpret. In addition, the representation has formal probabilistic semantics, making it suitable for statistical manipulation (Howard, 1981; Pearl, 1988). Over the last decade, the Bayesian network has become a popular representation for encoding uncertain expert knowledge in expert systems (Heckerman et al., 1995a). More recently, researchers have developed methods for learning Bayesian networks from a combination of expert knowledge and data. The techniques that have been developed are new and still evolving, but they have been shown to be remarkably effective in some domains (Cooper and Herskovits 1992; Aliferis and Cooper 1994; Heckerman et al. 1995b).

Using Bayesian networks, the learning process goes as follows. First, we encode the existing knowledge of an expert or set of experts in a Bayesian network, as is done when building a probabilistic expert system. Then, we use a database to update this knowledge, creating one or more new Bayesian networks. The result includes a refinement of the original expert knowledge and sometimes the identification of new distinctions and relationships. The approach is robust to errors in the knowledge of the expert. Even when expert knowledge is unreliable and incomplete, we can often use it to improve the learning process.

Learning using Bayesian networks is similar to that using neural networks. The process employing Bayesian networks, however, has two important advantages. One, we can easily encode expert knowledge in a Bayesian network and use this knowledge to increase the efficiency and accuracy of learning. Two, the nodes and arcs in learned Bayesian networks often correspond to recognizable distinctions and causal relationships. Consequently, we can more easily interpret and understand the knowledge encoded in the representation.

This paper is a brief tutorial on Bayesian networks and methods for learning them from data. In Sections 2 and 3 we discuss the Bayesian philosophy and the representation. In Sections 4 through 7, we describe methods for learning the probabilities and structure of a Bayesian network. In Sections 8 and 9, we discuss methods for identifying new distinctions about the world and integrating these distinctions into a Bayesian network. We restrict our discussion to Bayesian and quasi-Bayesian methods for learning. An interesting and often effective non-Bayesian approach is given by Pearl and Verma (1991) and Spirtes et al. (1993). Also, we limit our discussion to problem domains where variables take on discrete states. More general techniques are given in Buntine (1994) and Heckerman et al. (1995b).

## 2 The Bayesian Philosophy

Before we discuss Bayesian networks and how to learn them from data, it will help to review the Bayesian interpretation of probability. A primary element of the language of probability (Bayesian or otherwise) is the event. By event, we mean a state of some part of our world in some time interval in the past, present, or future. A classic example of an event is that a particular flip of a coin will come up heads. A perhaps more interesting event is that gold will close at more than $400 per ounce on January 1, 2001.

Given an event $e$, the prevalent conception of its probability is that it is a measure of the frequency with which $e$ occurs, when we repeat many times an experiment with possible outcomes $e$ and $\bar{e}$ (not $e$). A different notion is that the probability of $e$ represents the *degree of belief* held by a person that the event $e$ will occur in a single experiment. If a person assigns a probability of 1 to $e$, then he believes with certainty that $e$ will occur. If he assigns a probability of 0 to $e$, then he believes with certainty that $e$ will not happen. If he assigns a probability between 0 and 1 to $e$, then he is to some degree unsure about whether or not $e$ will occur.

The interpretation of a probability as a frequency in a series of repeat experiments is traditionally referred to as the *objective* or *frequentist* interpretation. In contrast, the interpretation of a probability as a degree of belief is called the *subjective* or *Bayesian* interpretation, in honor of the Reverend Thomas Bayes, a scientist from the mid-1700s who helped to pioneer the theory of probabilistic inference (Bayes 1958; Hacking, 1975). As we shall see in Section 4, the frequentist interpretation is a special case of the Bayesian interpretation.

In the Bayesian interpretation, a probability or belief will always depend on the state of knowledge of the person who provides that probability. For example, if we were to give someone a coin, he would likely assign a probability of 1/2 to the event that the coin would

show heads on the next toss. If, however, we convinced that person that the coin was weighted in favor of heads, he would assign a higher probability to the event. Thus, we write the probability of $e$ as $p(e|\xi)$, which is read as the probability of $e$ *given* $\xi$. The symbol $\xi$ represents the state of knowledge of the person who provides the probability.

Also, in this interpretation, a person can assess a probability based on information that he *assumes* to be true. For example, our coin tosser can assess the probability that the coin would show heads on the eleventh toss, under the assumption that the same coin comes up heads on each of the first ten tosses. We write $p(e_2|e_1,\xi)$ to denote the probability of event $e_2$ *given* that event $e_1$ is true and given background knowledge $\xi$.

Many researchers have written down different sets of properties that should be satisfied by degrees of belief (e.g., Cox 1946; Good 1950; Savage 1954; DeFinetti 1970) From each of the lists of properties, these researchers have derived the same rules—the rules of probability. Two basic rules, from which other rules may be derived, are the *sum rule*, which says that for any event $e$ and its complement $\bar{e}$,

$$p(e|\xi) + p(\bar{e}|\xi) = 1$$

and the *product rule*, which says that for any two events $e_1$ and $e_2$,

$$p(e_1, e_2|\xi) = p(e_2|e_1,\xi)\ p(e_1|\xi)$$

where $p(e_1, e_2|\xi)$ denotes the probability that $e_1$ and $e_2$ are true given $\xi$.

Other commonly used rules are often expressed in terms of variables rather than events. A *variable* takes on values from a collection of mutually exclusive and collectively exhaustive states, where each state corresponds to some event. A variable may be *discrete*, having a finite or countable number of states, or it may be *continuous*. For example, a two-state or *binary* variable can be used to represent the possible outcomes of a coin flip; whereas a continuous variable can be used to represent the weight of the coin. In this paper, we use lower-case letters (usually near the end of the alphabet) to represent single variables and upper-case letters to represent sets of variables. We write $x = k$ to denote that variable $x$ is in state $k$. When we observe the state for every variable in set $X$, we call this set of observations a state of $X$, and write $X = k$. Sometimes, we leave the state of a variable or set of variables implicit. The *probability distribution over a set of variables* $X$, denoted $p(X|\xi)$, is the set of probabilities $p(X = k|\xi)$ for all states of $X$.[1]

[1] When $X$ contains only continuous variables, $p(X|\xi)$ is sometimes called a probability density. When $X$ contains both discrete and continuous variables, $p(X|\xi)$ is sometimes called a generalized probability density. We do not make these distinctions in this paper.

One common rule of probability is *Bayes' theorem*:

$$p(X|Y,\xi) = \frac{p(Y|X,\xi)}{p(Y|\xi)}\, p(X|\xi) \quad \text{for } p(Y|\xi) > 0$$

Here, $p(X|\xi)$ is the probability distribution of $X$ before we know $Y$, and $p(X|Y,\xi)$ is the probability distribution of $X$ after we know $Y$. These distributions are sometimes called the *priors* and *posteriors* of $X$, respectively. In many cases, only the relative posterior of $X$ is of interest. In this case, Bayes' theorem is written

$$p(X|Y,\xi) = c\; p(Y|X,\xi)\; p(X|\xi)$$

where $c$ is a normalization constant. Another rule is the *chain rule*:

$$p(x_1,\ldots,x_n|\xi) = \prod_{i=1}^{n} p(x_i|x_1,\ldots,x_{i-1},\xi)$$

We shall see that this rule is important in the definition of Bayesian networks. Also, we have the *generalized sum rule*:

$$\sum_Y p(X,Y|\xi) = p(X|\xi)$$

where $\sum_Y$ is a generalized sum that includes integrals when some or all of the variables in $Y$ are continuous. Finally, we have the *expansion rule*:

$$p(X|\xi) = \sum_Y p(X|Y,\xi)\; p(Y|\xi)$$

The Bayesian philosophy extends to decision making under uncertainty in a discipline known as *decision theory*. In general, a decision has three components: what a decision maker can do (his *alternatives*), what he knows (his *beliefs*), and what he wants (his *preferences*). In decision theory, we use a *decision variable* to represent a set of mutually exclusive and exhaustive alternatives, Bayesian probabilities to represent a decision maker's beliefs, and *utilities* to represent a decision maker's preferences. Decision theory has essentially one rule: *maximize expected utility* (MEU). This rules says that, given a set of mutually exclusive and exhaustive alternatives, a decision maker should (1) assign a utility to every possible outcome of every possible alternative, (2) assign (Bayesian) probabilities to every possible outcome given every possible alternative, and (3) choose the alternative that maximizes his expected utility. Several researchers have shown that this rule follows from sets of compelling axioms (e.g., von Neumann and Morgenstern 1947; Savage, 1954).

In practice, decision making under uncertainty can be a difficult task. In fact, researchers have shown that people often violate the MEU rule (Tversky and Kahneman 1974). The

deviations are so significant and predictable that some researchers have come to reject the rule (Kahneman et al., 1982). Many other researchers, however, argue that the axioms used to derive the MEU rule are too compelling to reject, and argue further that people's deviations from the rule make the use of decision-theoretic concepts and representations all the more important (Howard 1990). If there's any doubt, this author has the latter view.

## 3 Bayesian Networks

A *problem domain* (or *universe*) is just a set of variables. A Bayesian network is a model of the (usually uncertain) relationships among variables in a domain. More precisely, given a domain of variables $U = \{x_1, \ldots, x_n\}$, the *joint probability distribution for* $U$ is a probability distribution over all the states of $U$. A Bayesian network for $U$ represents a joint probability distribution for $U$. The representation consists of a set of local conditional probability distributions, combined with a set of assertions of conditional independence that allow one to construct the global joint distribution from the local distributions.

To illustrate the representation, let us consider the domain of troubleshooting a car that won't start. The first step in constructing a Bayesian network is to decide what variables and states to model. One possible choice of variables for this domain is *Battery* ($b$) with states good and bad, *Fuel* ($f$) with states not empty and empty, *Gauge* ($g$) with states not empty and empty, *Turn Over* ($t$) with states yes and no, and *Start* ($s$) with states yes and no. Of course, we could include many more variables (as we would in a real-world example). Also, we could model the states of one or more of these variables at a finer level of detail. For example, we could let *Gauge* be a continuous variable with states ranging from 0% to 100%.

The second step in constructing a Bayesian network is to construct a directed acyclic graph that encodes assertions of conditional independence. We call this graph the *Bayesian-network structure.* Given a domain $U = \{x_1, \ldots, x_n\}$, we can write the joint probability distribution of $U$ using the chain rule of probability as follows:

$$p(x_1, \ldots, x_n|\xi) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}, \xi). \tag{1}$$

Now, for every $x_i$, there will be some subset $\Pi_i \subseteq \{x_1, \ldots, x_n\}$ such that $x_i$ and $\{x_1, \ldots, x_n\}$ are conditionally independent given $\Pi_i$. That is,

$$p(x_i|x_1, \ldots, x_{i-1}, \xi) = p(x_i|\Pi_i, \xi) \tag{2}$$

These conditional independencies define the Bayesian-network structure. The nodes in the

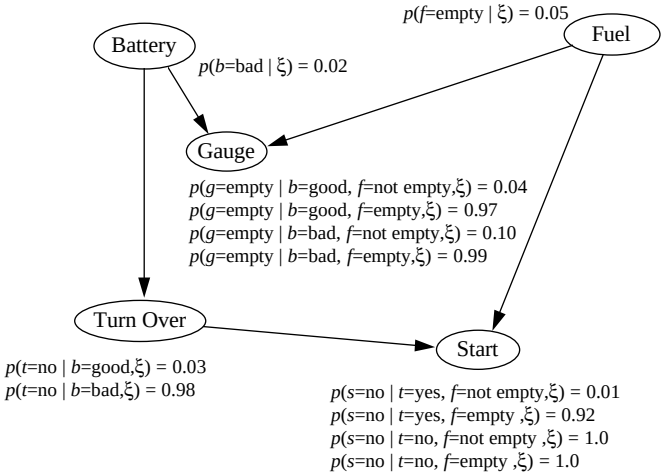


Figure 1: A Bayesian-network for troubleshooting a car that won't start. Arcs are drawn from cause to effect. The local probability distribution(s) associated with a node are shown adjacent to the node.

structure correspond to variables in the domain. The parents of $x_i$ correspond to the set $\Pi_i$.

In our example, using the ordering $b$, $f$, $g$, $t$, and $s$, we have the conditional independencies

$$p(f|b,\xi) = p(f|\xi) \qquad p(t|b,f,g,\xi) = p(t|b,\xi) \qquad p(s|b,f,g,t,\xi) = p(s|f,t,\xi) \tag{3}$$

Consequently, we obtain the structure shown in Figure 1.

The final step in constructing a Bayesian network is to assess the local distributions $p(x_i|\Pi_i,\xi)$—one distribution for every state of $\Pi_i$. These distributions for our automobile example are shown in Figure 1. Combining Equations 1 and 2, we see that a Bayesian network for $U$ always encodes the joint probability distribution for $U$.

A drawback of Bayesian networks as defined is that network structure depends on variable order. If the order is chosen carelessly, the resulting network structure may fail to reveal many conditional independencies in the domain. As an exercise, the reader should construct a Bayesian network for the automobile troubleshooter domain using the ordering $(s,t,g,f,b)$. Fortunately, in practice, we can often readily assert causal relationships among variables in a domain, and can use these assertions to construct a Bayesian-network structure without preordering the variables. Namely, to construct a Bayesian network for a given set of variables, we draw arcs from cause variables to their immediate effects. In almost all

cases, doing so results in a Bayesian network whose conditional-independence implications are accurate. For example, the network in Figure 1 was constructed using the assertions that *Gauge* is the direct causal effect of *Battery* and *Fuel*, *Turn Over* is the direct causal effect of *Battery*, and *Start* is the direct causal effect of *Turn Over* and *Fuel*.

Because a Bayesian network for any domain determines a joint probability distribution for that domain, we can—in principle—use a Bayesian network to compute any probability of interest. For example, suppose we want to compute the probability distribution of *Fuel* given that the car doesn't start. From the rules of probability we have

$$p(f|s = \text{no}, \xi) = \frac{p(f, s = \text{no}|\xi)}{p(s = \text{no}|\xi)} = \frac{\sum_{b,g,t} p(b, f, g, t, s = \text{no}|\xi)}{\sum_{b,f,g,t} p(b, f, g, t, s = \text{no}|\xi)} \tag{4}$$

In a real-world problem with $n$ variables, this approach is not feasible, because it entails summing over $2^n$ or more terms. Fortunately, we can exploit the conditional independencies encoded in a Bayesian network to make this computation more efficient. In this case, given the conditional independencies in Equation 3, Equation 4 becomes

$$p(f|s = \text{no}, \xi) = \frac{p(f|\xi) \sum_b p(b|\xi) \sum_t p(s = \text{no}|t, f, \xi) p(t|b, \xi) \sum_g p(g|b, f, \xi)}{\sum_f p(f|\xi) \sum_b p(b|\xi) \sum_t p(s = \text{no}|t, f, \xi) p(t|b, \xi) \sum_g p(g|b, f, \xi)} \tag{5}$$

That is, conditional independence produces a decomposition of the joint probability distribution that can be used in conjunction with the distributive law to reduce the dimensionality of the computations.

The general problem of computing probabilities of interest from a (possibly implicit) joint probability distribution is called *probabilistic inference*. All exact algorithms for probabilistic inference in Bayesian networks exploit conditional independence roughly as we have described, although with different twists. For example, Howard and Matheson (1981), Olmsted (1983) , and Shachter (1988) developed an algorithm that reverses arcs in the network structure until the answer to the given probabilistic query can be read directly from the graph. In this algorithm, each arc reversal corresponds to an application of Bayes' theorem. Pearl (1986) developed a message-passing scheme that updates the probability distributions for each node in a Bayesian network in response to observations of one or more variables. Lauritzen and Spiegelhalter (1988) created an algorithm that first transforms the Bayesian network into a tree where each node in the tree corresponds to a subset of variables in the domain. The algorithm then exploits several mathematical properties of this tree to perform probabilistic inference. Most recently, D'Ambrosio (1991) developed an inference algorithm that simplifies sums and products symbolically, as in the transformation from Equation 4 to Equation 5.

Although we can exploit assertions of conditional independence in a Bayesian network for probabilistic inference, exact inference in an arbitrary Bayesian network is NP-hard (Cooper,

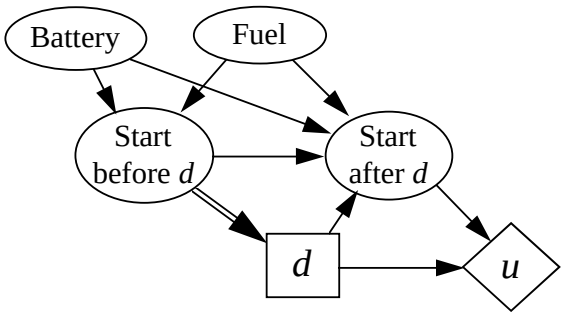


Figure 2: A influence diagram (structure) for a decision of whether to change the battery, get fuel, or do nothing. The square node $d$ represents our alternatives. The diamond node $u$ represents our utilities of all the possible outcomes. The double-line arc from *Start* to $d$ represents the assertion that we know whether or not the car starts when we make decision $d$.

1990). Even approximate inference (for example, using Monte-Carlo methods) is NP-hard (Dagum and Luby, 1994). For many applications, however, the networks are small enough (or can be simplified sufficiently) so that these complexity results are not fatal. For those applications where the usual inference methods are impractical, researchers are developing techniques that are custom tailored to particular network topologies (Heckerman 1989; Suermondt and Cooper, 1991), or particular inference queries (Ramamurthi and Agogino 1988; Shachter et al. 1990; Jensen and Andersen 1990) .

The *influence diagram* is an extension of the Bayesian-network representation to decision problems. Like the Bayesian network, an influence diagram contains nodes representing uncertain variables and arcs representing probabilistic dependence. In an influence diagram, these constructs are called *chance nodes* and *relevance arcs*, respectively. In addition, influence diagrams may contain *decision nodes*, which represent decision variables, and at most one *utility node,* which represents a decision maker's preferences. Also, influence diagrams may contain *information arcs,* which indicate what is known at the time a decision is made.

For example, in our troubleshooting domain, suppose we have the options to replace the battery, get fuel for the car, or do nothing. An influence diagram for this decision is shown in Figure 2. The square node $d$ is a decision node and represents our alternatives. The diamond node $u$ is the utility node. The arcs pointing to the chance (oval) nodes are relevance arcs. The arc from *Start* to $d$ is an information arc. Its presence asserts that, at the time we make the decision, we know whether or not the car starts. In general,

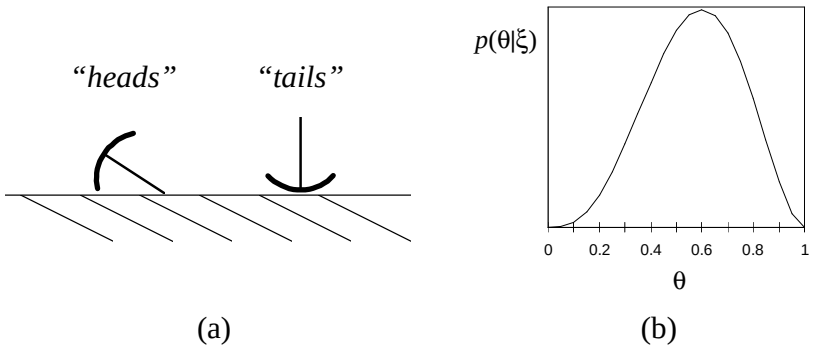


Figure 3: (a) The outcomes of a thumbtack flip. (b) A probability distribution for $\theta$, the physical probability of heads.

information arcs point only to decision nodes, whereas relevance arcs point only to chance nodes.

## 4 Learning Probabilities: The One-Variable Case

Because Bayesian networks have a probabilistic interpretation, we can use traditional techniques from Bayesian statistics to learn these models from data. We discuss these techniques in the remainder of the paper. Several of the techniques that we need can be discussed in the context of learning the probability distribution of a single variable. In this section, we examine this case.

Consider a common thumbtack—one with a round, flat head that can be found in most supermarkets. If we throw the thumbtack up in the air and let in land on a hard, flat surface, it will come to rest either on its point (*heads*) or on its head (*tails*), as shown in Figure 3a.[2] Suppose we give the thumbtack to someone, who then flips it many times, and measures the fraction of times the thumbtack comes up heads. A frequentist would say this long-run fraction is a probability, and would observe flips of the thumbtack to estimate this probability. In contrast, from the Bayesian perspective, we recognize the possible values of this fraction as a variable—call it $\theta$—whose true value is uncertain. We can express our uncertainty about $\theta$ with a probability distribution $p(\theta|\xi)$, and update this distribution as we observe flips of the thumbtack.

We note that, although $\theta$ does not represent a degree of belief, collections of long-run fractions like $\theta$ satisfy the rules of probability. In this paper, we shall refer to $\theta$ as a *physical*

[2]This example is taken from Howard (1970).

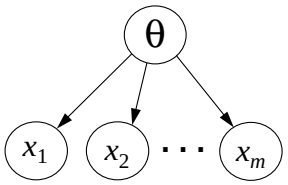


Figure 4: A Bayesian-network structure depicting the conditional independencies associated with a binomial sample.

*probability* (after Good, 1959) to distinguish it from a degree of belief.[3] Figure 3 shows one possible probability distribution for $\theta$.

Now suppose we observe $D = \{x_1, \ldots, x_m\}$, the outcomes of $m$ flips of the thumbtack. We sometimes refer to this set of observations as a *database*. If we knew the value of $\theta$, then our probability for heads on any flip would be equal to $\theta$, no matter how many outcomes we observe. That is,

$$p(x_l = heads | x_1, \ldots, x_{l-1}, D, \xi) = \theta \tag{6}$$

where $x_l$ is the outcome of the $l$th flip of the thumbtack. Similarly, we have

$$p(x_l = tails | x_1, \ldots, x_{l-1}, D, \xi) = 1 - \theta \tag{7}$$

In particular, the outcomes are mutually independent, given $\theta$. We can represent this conditional independence assertion using a Bayesian-network structure, as shown in Figure 4.

In reality, we are uncertain about the value of $\theta$. In this case, we can use the expansion rule to determine our probability that the next toss of the thumbtack will come up heads:

$$p(x = heads | \xi) = \int p(x = heads | \theta, \xi)\ p(\theta | \xi)\ d\theta = \int \theta\ p(\theta | \xi)\ d\theta \equiv E(\theta | \xi)$$

where $E(\theta|\xi)$ denotes the expectation of $\theta$ given $\xi$. That is, our probability for heads on the next toss is just the expectation of $\theta$. Furthermore, suppose we flip the thumbtack once and observe heads. Using Bayes' theorem, the posterior probability distribution for $\theta$ becomes

$$p(\theta | x = heads, \xi) = c\ p(x = heads | \theta, \xi)\ p(\theta | \xi) = c\ \theta\ p(\theta | \xi)$$

where $c$ is some normalization constant. That is, we obtain the posterior distribution for $\theta$ by multiplying its prior distribution by the function $f(\theta) = \theta$ and renormalizing. This

[3] The variable $\theta$ is also referred to as a frequency, objective probability, and true probability.

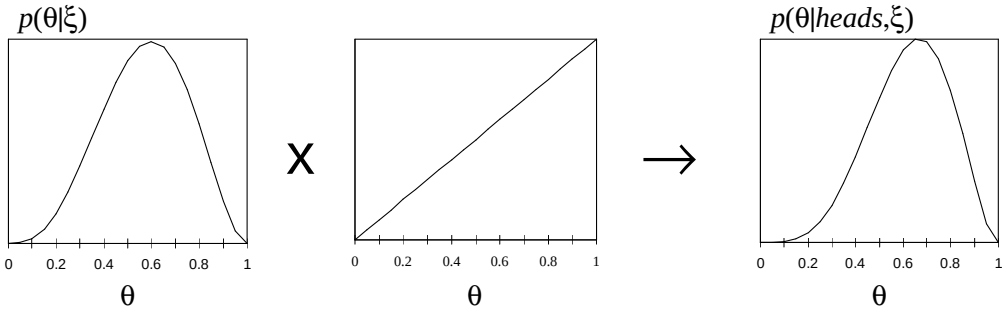


Figure 5: A graphical depiction of the use of Bayes' theorem to compute the posterior probability distribution of the physical probability $\theta$.

procedure is depicted graphically in Figure 5. As expected, the posterior is shifted to the right and is slightly narrower. Similarly, if we observe a single tails, we obtain

$$p(\theta|x = tails, \xi) = c\ (1 - \theta)\ p(\theta|\xi)$$

In general, if we observe $h$ heads and $t$ tails in the database $D$, then we have

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

That is, once we have assessed a prior distribution for $\theta$, we can determine its posterior distribution given any possible database. Note that the order in which we observe the outcomes is irrelevant to the posterior—all that is relevant is the number of heads and the number of tails in the database. We say that $h$ and $t$ are a *sufficient statistic* for the database.

In this simple example, our outcome variable has only two states (*heads* and *tails*). Now, imagine we have a discrete outcome variable $x$ with $r \geq 2$ states. For example, this variable could represent the outcome of a roll of a loaded die ($r = 6$). As in the thumbtack example, we can define the physical probabilities of each outcome, which we denote $\Theta_x = \{\theta_{x=1}, \ldots, \theta_{x=r}\}$. We assume that each state is possible so that each $\theta_{x=k} > 0$. In addition, we have $\sum_{k=1}^{r} \theta_{x=k} = 1$. Also, if we know these physical probabilities, then the outcome of each "toss" of $x$ will be conditionally independent of the other tosses, and

$$p(x_l = k|x_1, \ldots, x_{l-1}, \Theta_x, \xi) = \theta_{x=k} \tag{8}$$

Any database of outcomes $\{x_1, \ldots, x_m\}$ that satisfies these conditions is called an $(r-1)$-*dimensional multinomial sample with physical probabilities* $\Theta_x$ (Good, 1965). When $r = 2$, as in the thumbtack example, the sequence is said to be a *binomial sample*. The concept

of a multinomial sample (and its generalization, the random sample) will be central to the remaining discussions in this paper.

Analogous to the thumbtack example, we have

$$p(x = k|\xi) = \int \theta_{x=k}\ p(\Theta_x|\xi)\ d\Theta_x \equiv E(\theta_{x=k}|\xi) \tag{9}$$

where $p(x = k|\xi)$ is our probability that $x = k$ in the next case. Note that, because $\sum_{k=1}^{r} \theta_{x=k} = 1$, the distribution for $\Theta_x$ is technically a distribution over the variables $\Theta_x \setminus \{\theta_{x=k}\}$ for some $k$ (the symbol $\setminus$ denotes set difference). Also, given any database $D$ of outcomes, we have

$$p(\Theta_x|D, \xi) = c \cdot \prod_{k=1}^{r} \theta_{x=k}^{N_k}\ p(\Theta_x|\xi) \tag{10}$$

where $N_k$ is the number of times $x = k$ in $D$, and $c$ is a normalization constant. Note that the counts $N_1, \ldots, N_r$ are a sufficient statistic for the multinomial sample.

Given a multinomial sample, a user is free to assess any probability distribution for $\Theta_x$. In practice, however, one often uses the Dirichlet distribution because it has several convenient properties. The variables $\Theta_x$ are said to have a *Dirichlet distribution with exponents* $N'_1, \ldots, N'_r$ when the probability distribution of $\Theta_x$ is given by

$$p(\Theta_x|\xi) = \frac{\Gamma(\sum_{k=1}^{r} N'_k)}{\prod_{k=1}^{r} \Gamma(N'_k)} \prod_{k=1}^{r} \theta_{x=k}^{N'_k - 1}, \quad N'_k > 0 \tag{11}$$

where $\Gamma(\cdot)$ is the *Gamma* function, which satisfies $\Gamma(x + 1) = x\Gamma(x)$ and $\Gamma(1) = 1$. When the variables $\Theta_x$ have a Dirichlet distribution, we also say that $p(\Theta_x|\xi)$ *is Dirichlet.* The exponents $N'_k$ must be greater than 0 to guarantee that the distribution can be normalized. Note that the exponents $N'_k$ are a function of the user's state of information $\xi$. When $r = 2$, the Dirichlet distribution is also known as a *beta distribution.* The probability distribution on the left-hand-side of Figure 5 is a beta distribution with exponents $N'_{heads} = 3$ and $N'_{tails} = 2$. The probability distribution on the right-hand-side of the figure is a beta distribution with exponents $N'_{heads} = 4$ and $N'_{tails} = 2$.

From Equation 10, we see that if the prior distribution of $\Theta_x$ is Dirichlet, then the posterior distribution of $\Theta_x$ given database $D = \{x_1, \ldots, x_m\}$ is also Dirichlet:

$$p(\Theta_x|D, \xi) = c \prod_{k=1}^{r} \theta_{x=k}^{N'_k + N_k - 1} \tag{12}$$

We say that the Dirichlet distribution is closed under multinomial sampling, or that the Dirichlet distribution is a *conjugate family of distributions* for multinomial sampling. Also, when $\Theta_x$ has a Dirichlet distribution, the expectation of $\theta_{x=k}$—equal to the probability that

$x = k$ in the first observation—has a simple expression:

$$E(\theta_{x=k}|\xi) = p(x = k|\xi) = \frac{N'_k}{N'} \tag{13}$$

where $N' = \sum_{k=1}^{r} N'_k$. As we shall see, these properties make the Dirichlet a useful prior for learning.

A survey of methods for assessing a beta distribution is given by Winkler (1967). These methods include the direct assessment of the probability distribution using questions regarding relative areas, assessment of the cumulative distribution function using fractiles, assessing the posterior means of the distribution given hypothetical evidence, and assessment in the form of an equivalent sample size. These methods can be generalized with varying difficulty to the non-binary case.

The equivalent-sample-size method generalizes particularly well. The method is based on Equation 13, which says that we can assess a Dirichlet distribution by assessing the probability distribution $p(x|\xi)$ for the next observation, and $N'$. In so doing, we may rewrite Equation 11 as

$$p(\Theta_x|\xi) = c \cdot \prod_{k=1}^{r} \theta_{x=k}^{N' p(x=k|\xi)-1} \tag{14}$$

where $c$ is a normalization constant. Assessing $p(x|\xi)$ is straightforward. Furthermore, the following two observations suggest a simple method for assessing $N'$.

The variance of a distribution for $\Theta_x$ is an indication of how much the mean of $\Theta_x$ is expected to change, given new observations. The higher the variance, the greater the expected change. It is sometimes said that the variance is a measure of a user's *confidence* in the mean for $\Theta_x$. The variance of the Dirichlet distribution is given by

$$Var(\theta_{x=k}|\xi) = \frac{p(x = k|\xi)(1 - p(x = k|\xi))}{N' + 1} \tag{15}$$

Thus, $N'$ is a reflection of the user's confidence.

In addition, suppose we were initially completely ignorant about a domain—that is, our distribution $p(\Theta_x|\xi)$ was given by Equation 11 with each exponent $N'_k = 0$.[4] Suppose we then saw $N'$ cases with sufficient statistics $N'_1, \ldots, N'_r$. Then, by Equation 12, our prior would be the Dirichlet distribution given by Equation 11.

Thus, we can assess $N'$ as an *equivalent sample size*: the number of observations we would have had to have seen starting from complete ignorance in order to have the same confidence in $\Theta_x$ that we actually have. For example, we would obtain the probability

[4] This prior distribution cannot be normalized, and is sometimes called an *improper prior.* To be more precise, we should say that each exponent is equal to some number close to zero.

distribution for $\theta$ in Figure 3 if we assessed $p(heads|\xi)$ to be 3/5 and the equivalent sample size to be five.

So far, we have only considered a variable with discrete outcomes. In general, we can imagine a physical probability distribution over a variable (discrete or continuous) from which database cases are drawn at random. This physical probability distribution typically can be characterized by a finite set of *parameters*. If the outcome variable is discrete, then the physical probability distribution has a parameter corresponding to each physical probability in the distribution (and, herein, we sometimes refer to these physical probabilities as parameters). If the outcome variable is continuous, the physical probability distribution may be (e.g.) a normal distribution. In this case, the parameters would be the mean and variance of the distribution. A database of cases drawn from a physical probability distribution is often called a *random sample.*

Given such a physical probability distribution with unknown parameters, we can update our beliefs about these parameters given a random sample from this distribution using techniques similar to those we have discussed. For random samples from many named distributions—including normal, Gamma, and uniform distributions—there exist corresponding conjugate priors that offer convenient properties for learning probabilities similar to those properties of the Dirichlet. These priors are sometimes referred to collectively as the *exponential family.* The reader interested in learning about these distributions should read DeGroot (1970, Chapter 9).

## 5 Learning Probabilities: Known Structure

The notion of a random sample generalizes to domains containing more than one variable as well. Given a domain $U = \{x_1, \ldots, x_n\}$, we can imagine a multivariate physical probability distribution for $U$. If $U$ contains only discrete variables, this distribution is just a finite collection of discrete physical probabilities. If $U$ contains only continuous variables, this distribution could be (e.g.) a multivariate-normal distribution characterized by a mean vector and covariance matrix. Given a random sample from a physical probability distribution, we can update our priors about the parameters of the distribution. This updating is especially simple when conjugate priors for the parameters are available (see DeGroot 1970).

Now, however, let us consider the following wrinkle. Suppose we know that this multivariate physical probability distribution can be encoded in some particular Bayesian-network structure $B_S$. We may have gotten this information—for example—from our causal knowledge about the domain. In this section, we consider the task of learning the parameters of $B_S$. We discuss only the special case where all the variables in $U$ are discrete and

where the random sample (i.e., database) $D = \{C_1, \ldots, C_m\}$ contains no missing data—that is, each case $C_l$ consists of the observation of *all* the variables in $U$ (we say that $D$ is *complete*). In Section 8, we consider the more difficult problem where $D$ contains missing data. Buntine (1994) and Heckerman and Geiger (1994) discuss the case where $U$ may contain continuous variables.

When a database $D$ is a random sample from a multivariate physical probability distribution that can be encoded in $B_S$, we simply say that $D$ is a random sample from $B_S$. As an example, consider the domain $U$ consisting of two binary variables $x$ and $y$. Let $\theta_{xy}, \theta_{x\bar{y}}, \theta_{\bar{x}y}$, and $\theta_{\bar{x}\bar{y}}$ denote the parameters (i.e., physical probabilities) for the joint space of $U$, where $\theta_{x\bar{y}}$ is the physical probability of the event where $x$ is true and $y$ is false, and so on. (Note that, in using the overbar, we are departing from our standard notation.) Then, saying that $D$ is a random sample from the network structure containing no arc between $x$ and $y$, is the assertion that the parameters of the joint space satisfy the independence constraints $\theta_{xy} = \theta_x\theta_y, \theta_{x\bar{y}} = \theta_x\theta_{\bar{y}}$, and so on, where—for example—$\theta_x = \theta_{xy} + \theta_{x\bar{y}}$ is the physical probability associated with the event where $x$ is true. It is not difficult to show that this assertion is equivalent to the assertion that the database $D$ can be decomposed into two multinomial samples: the observations of $x$ are a multinomial sample with parameter $\theta_x$, and the observations of $y$ are a multinomial sample with parameter $\theta_y$.

As another example, suppose we assert that a database for our two variable domain is a random sample from the network structure $x \rightarrow y$. Here, there are no constraints on the parameters of the joint space. Furthermore, this assertion implies that the database is made up of at most three binomial samples: (1) the observations of $x$ are a binomial sample with parameter $\theta_x$, (2) the observations of $y$ in those cases (if any) where $x$ is true are a binomial sample with parameter $\theta_{y|x}$, and (3) the observations of $y$ in those cases (if any) where $x$ is false are a binomial sample with parameter $\theta_{y|\bar{x}}$. Consequently, the occurrences of $x$ in $D$ are conditionally independent given $\theta_x$, and $y$ in case $C$ are conditionally independent of the other occurrences of $y$ in $D$ given $\theta_{y|x}$, $\theta_{y|\bar{x}}$, and $x$ in case $C$. We can graphically represent the conditional-independence assertions associated with these random samples using a Bayesian-network structure as shown in Figure 6a.

Given the collection of random samples shown in Figure 6a, it is tempting to apply our one-variable techniques to learn each parameter separately. Unfortunately, this approach is not correct when the parameters are dependent as shown in the figure. For example, as we see occurrences of $x$ and update our beliefs about $\theta_x$, our beliefs about $\theta_{y|x}$ and $\theta_{y|\bar{x}}$ will also change. Suppose, however, that all of the parameters are independent, as shown in Figure 6b. Then, provided the database is complete, we can update each parameter separately.

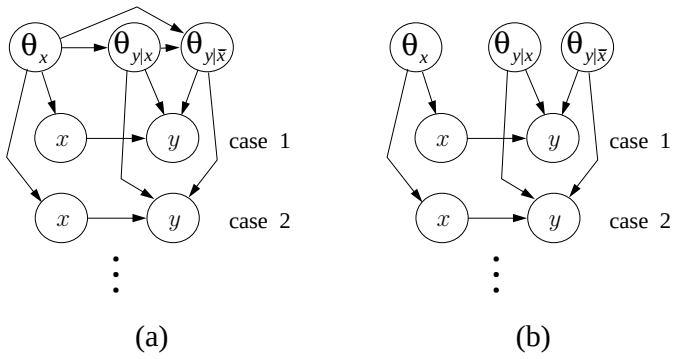


Figure 6: (a) A Bayesian-network structure for a two-binary-variable domain $\{x, y\}$ showing conditional independencies associated with the assertion that the database is a random sample from the structure $x \rightarrow y$. (b) Another Bayesian-network structure showing the added assumption of parameter independence.

In the remainder of this section, we shall assume that all parameters are independent. We call this assumption—introduced by Spiegelhalter and Lauritzen (1990)—*parameter independence*. In Section 8, we discuss methods for handling dependent parameters.

To complete the discussion, we need some notation. Let $B_S^h$ denote the assertion (or hypothesis) that a database $D$ is a random sample from a Bayesian network structure $B_S$. Given this network structure, let $r_i$ be the number of states of variable $x_i$; and let $q_i = \prod_{x_l \in \Pi_i} r_l$ be the number of states of $\Pi_i$. Let $\theta_{ijk}$ denote the physical probability of $x_i = k$ given $\Pi_i = j$. In addition, let

$$\Theta_{ij} \equiv \cup_{k=1}^{r_i} \{\theta_{ijk}\}$$

$$\Theta_{B_S} \equiv \cup_{i=1}^{n} \cup_{j=1}^{q_i} \Theta_{ij}$$

Note that the parameters $\Theta_{B_S}$ in conjunction with $B_S$ determine all the physical probabilities of the joint space.

Let us assume that each variable set $\Theta_{ij}$ has a Dirichlet distribution:

$$p(\Theta_{ij}|B_S^h, \xi) = c \cdot \prod_k \theta_{ijk}^{N'_{ijk}-1} \tag{16}$$

where $c$ is a normalization constant. Then, if $N_{ijk}$ is the number of cases in database $D$ in which $x_i = k$ and $\Pi_i = j$, we obtain

$$p(\Theta_{ij}|D, B_S^h, \xi) = c \cdot \prod_k \theta_{ijk}^{N'_{ijk}+N_{ijk}-1} \tag{17}$$

where $c$ is some other normalization constant. Furthermore, applying Equation 13 to each multinomial sample, we can compute the probability that each $x_i = k$ and $\Pi_i = j$ in $C_{m+1}$, the next case to be seen after the database $D$:

$$p(C_{m+1}|D, B_S^h, \xi) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{N'_{ijk} + N_{ijk}}{N'_{ij} + N_{ij}} \tag{18}$$

where $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$ and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

## 6 Learning Structure

In the previous section, we considered the situation where we are uncertain about the physical probabilities, but certain about the network structure that encodes these probabilities. Now, suppose we are not only uncertain about the probabilities, but also uncertain about the structure that encodes them. As with any set of events, we can express this uncertainty by assigning a prior probability $p(B_S^h|\xi)$ to each possible hypothesis $B_S^h$. Furthermore, we can update these probabilities as we see cases. In so doing, we learn about the structure of the domain.

As in the previous section, let $B_S^h$ denote the (now uncertain) hypothesis that the database $D$ is a random sample from the Bayesian network structure $B_S$. From Bayes' theorem, we have

$$p(B_S^h|D, \xi) = c\ p(B_S^h|\xi)\ p(D|B_S^h, \xi) \tag{19}$$

where $c$ is a normalization constant. Also, from the product rule, we have

$$p(D|B_S^h, \xi) = \prod_{l=1}^{m} p(C_l|C_1, \ldots, C_{l-1}, B_S^h, \xi) \tag{20}$$

We can evaluate each term on the right-hand-side of this equation using Equation 18, under the assumption that the database $D$ is complete. Thus, we obtain the posterior probability of $B_S^h$ given $D$:

$$\begin{aligned} p(B_S^h|D, \xi) &= c \cdot p(B_S^h|\xi) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left\{ \left[ \frac{N'_{ij1}}{N'_{ij}} \cdot \frac{N'_{ij1}+1}{N'_{ij}+1} \cdots \frac{N'_{ij1}+N_{ij1}-1}{N'_{ij}+N_{ij1}-1} \right] \cdot \right. \\ &\quad \left[ \frac{N'_{ij2}}{N'_{ij}+N_{ij1}} \cdot \frac{N'_{ij2}+1}{N'_{ij}+N_{ij1}+1} \cdots \frac{N'_{ij2}+N_{ij2}-1}{N'_{ij}+N_{ij1}+N_{ij2}-1} \right] \cdots \\ &\quad \left. \left[ \frac{N'_{ijr_i}}{N'_{ij}+\sum_{k=1}^{r_i-1} N_{ijk}} \cdot \frac{N'_{ijr_i}+1}{N'_{ij}+\sum_{k=1}^{r_i-1} N_{ijk}+1} \cdots \frac{N'_{ijr_i}+N_{ijr_i}-1}{N_{ij}+N'_{ij}-1} \right] \right\} \\ &= c \cdot p(B_S^h|\xi) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})} \end{aligned} \tag{21}$$

Using these posterior probabilities and Equation 18, we may compute the probability distribution for the next case to be observed after we have seen a database. From the expansion rule, we obtain

$$p(C_{m+1}|D,\xi) = \sum_{B_S^h} p(C_{m+1}|D,B_S^h,\xi)\ p(B_S^h|D,\xi) \tag{22}$$

There are three important points to be made about this approach. One, it can happen that two Bayesian-network structures represent exactly the same sets of probability distributions. We say that the two structures are *equivalent* (Verma and Pearl, 1990). For example, for the three variable domain $\{x, y, z\}$, each of the network structures $x \rightarrow y \rightarrow z$, $x \leftarrow y \rightarrow z$, and $x \leftarrow y \leftarrow z$ represents the distributions where $x$ and $z$ are conditionally independent of $y$. Consequently, these network structures are equivalent. As another example, a *complete network structure* is one that has no missing edges—that is, it encodes no assertions of conditional independence. A domain containing $n$ variables has $n!$ complete network structures: one network structure for each possible ordering of the variables. All complete network structures for a given domain represent the same joint probability distributions—namely, all possible distributions—and are therefore equivalent.

In general, two network structures are equivalent if and only if they have the same structure ignoring arc directions and the same v-structures (Verma and Pearl, 1990). A *v-structure* is an ordered tuple $(x, y, z)$ such that there is an arc from $x$ to $y$ and from $z$ to $y$, but no arc between $x$ and $y$. Using this characterization of network-structure equivalence, Chickering (1995) has created an efficient algorithm for identifying all Bayesian-network structures that are equivalent to a given network structure.

Given that $B_S^h$ is the assertion that the physical probabilities for the joint space of $U$ can be encoded in the network structure $B_S$, it follows that the hypotheses associated with two equivalent network structures must be identical. Consequently, two equivalent network structures must have the same (prior and posterior) probability. For example, in the two variable domain $\{x, y\}$, the network structures $x \rightarrow y$ and $y \rightarrow x$ are equivalent, and will have the same probability. In general, this property is called *hypothesis equivalence.* In light of this property, we should associate each hypothesis with an equivalence class of structures rather than a single network structure, and our methods for learning network structure should actually be interpreted as methods for learning equivalence classes of network structures (although, for the sake of brevity, we often blur this distinction).[5]

[5]Hypothesis equivalence holds provided we interpret Bayesian-network structures simply as representations of conditional independence. Nonetheless, stronger definitions of Bayesian networks exist where arcs have a causal interpretation (e.g., Pearl and Verma, 1991). Heckerman et al. (1995b) argue that, although it is unreasonable to assume hypothesis equivalence when working with causal Bayesian networks, it is often

The second important point about this approach is that, in writing Equation 22, we have assumed that the hypothesis equivalence classes are mutually exclusive. In reality, these hypotheses are not mutually exclusive. For example, in our two-variable domain, both network structures $x \rightarrow y$ and the empty network structure can encode parameters satisfying the equality $\theta_y = \theta_{y|x}$. Therefore, the hypotheses associated with these non-equivalent network structures overlap. Nonetheless, in this approach, we assume that the priors on parameters for any given network structure have bounded densities, and hence the overlap of hypotheses will be of measure zero.

Finally, in writing Equation 22, we have limited ourselves to hypotheses corresponding to assertions that the physical probability distribution of the joint space comes from one particular network structure. We can relax this assumption, assuming that the physical probability distribution can be encoded in a *set* of network structures. In this paper, however, we do not pursue this generalization.

In principle, the approach we have discussed in this section is essentially all there is to learning network structure. In practice, when the user believes that only a few alternative network structures are possible, he can directly assess the priors for the possible network structures and their parameters, and subsequently use Equations 21 and 22 or their generalizations for continuous variables and missing data. For example, Buntine (1994) has designed a software system whereby a user specifies his priors for a set of possible models using Bayesian networks in a manner similar to that shown in Figure 6. The system then compiles this specification into a computer program that learns from a database.

Nonetheless, the number of network structures for a domain containing $n$ variables is more than exponential in $n$. Consequently, when the user cannot exclude almost all of these network structures, there are several issues that must be considered. In particular, computational constraints can prevent us from summing over all the hypotheses in Equation 22. Can we approximate $p(C_{m+1}|D,\xi)$ accurately by retaining only a small fraction of these hypotheses in the sumΓ If so, which hypotheses should we includeΓ In addition, how can we efficiently assign prior probabilities to the many network structures and their parametersΓ In the subsections that follow, we consider each of these issues.

### 6.1 Scoring Metrics

The most important issue is whether we can approximate $p(C_{m+1}|D,\xi)$ well using just a small number of network-structure hypotheses. This question is difficult to answer in theory. Nonetheless, several researchers have shown experimentally that even a single "good"

reasonable to adopt a weaker assumption of *likelihood equivalence,* which says that the observations in a database can not help to discriminate two equivalent network structures.

network structure often provides an excellent approximation (Cooper and Herskovits 1992; Aliferis and Cooper 1994; Heckerman et al., 1995b). We given an example in Section 7. These results are somewhat surprising, and are largely responsible for the great deal of recent interest in learning Bayesian networks.

Given this observation, another important consideration is how to identify "good" network structures. The approach that has been adopted by many researchers is to use a *scoring metric* in combination with a search algorithm. The scoring metric takes prior knowledge, a database, and a set of network structures, and computes the goodness of fit of those structures to the prior knowledge and data. The search algorithm identifies network structures to be scored. In this section, we discuss scoring metrics. In Section 6.4, we discuss search algorithms.

An obvious scoring metric for a single network-structure (equivalence class) is the relative posterior probability of that structure given the database. For example, we can compute $p(D, B_S^h|\xi) = p(B_S^h|\xi)\ p(D|B_S^h, \xi)$ or compute a *Bayes factor*: $p(B_S^h|D, \xi)/p(B_{S0}^h|D, \xi)$ where $B_{S0}^h$ is some reference network structure such as the empty network structure. When we use Equation 21 to compute this relative posterior probability, the scoring metric is sometimes called the *Bayesian Dirichlet* (BD) metric. A network structure with the highest posterior probability is often called a *maximum a posteriori* (MAP) structure. To score a set of distinct network structures $S$ we can use $\sum_{B_S \in S} p(D, B_S^h|\xi)$. Note that practitioners typically compute logarithms of the probabilities to avoid numerical underflow.

Madigan and Raferty (1994) suggest an alternative scoring metric that uses relative posterior probability in conjunction with heuristics based on the principle of Occam's Razor.

Other scoring metrics approximate the posterior-probability metric. In Section 8, we discuss algorithms that can find a local maximum in the probability $p(D|B_S^h, \Theta_{B_S}, \xi)$ as a function of $\Theta_{B_S}$ (the physical probabilities associated with network structure $B_S$). We cannot use such a local maximum as a score for $B_S$, because it will always favor the most complex network structures, which place no constraints on the parameters $\Theta_{B_S}$. Nonetheless, we can use a local maximum of $p(D|B_S^h, \Theta_{B_S}, \xi)$ as a score for $B_S$ if we also penalize structures based on their complexity. Akaike (1974) suggests the scoring metric

$$\log p(D|M, \widehat{\Theta}_{B_S}, \xi) + \mathrm{Dim}(M)$$

where $M$ is a model, $\widehat{\Theta}_{B_S}$ denotes the values of $\Theta_{B_S}$ that maximize the probability, and $\mathrm{Dim}(M)$ is the number of logically independent parameters in $M$. This scoring metric is sometimes called the A information criterion (AIC). For a Bayesian network, the penalty is given by

$$\mathrm{Dim}(B_S) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} q_i(r_i - 1)$$

Schwarz (1978) suggests a similar scoring metric with a penalty term given by $(1/2)\text{Dim}(M)\log(m)$, where $m$ is the number of cases in the database. This metric is sometimes called the Bayesian information criterion (BIC).

Another metric that approximates the posterior-probability metric is minimum description length (MDL) (Rissanen 1987). The MDL of a network structure is the sum of the number of bits required to encode the model (which increases with increasing model complexity) and the number of bits required to encode the database given the model (which decreases with increasing model complexity) relative to a particular coding scheme. We note that, in the limit, as the number of cases in the database approach infinity, the BD metric with uniform priors on structures, BIC, and MDL give the same relative scores (Kass and Raferty, 1993). Unfortunately, in practice, this asymptotic equivalence is rarely achieved.

Finally, we can score structures based on a decision model for the domain. In such cases, we can score a network structure by extending that structure to an influence diagram, which includes a decision model for the domain.

For example, suppose we wish to compute a score for the Bayesian network for medical diagnosis shown in Figure 7a, given a complete database of disease and symptoms. To do so, we extend this network to an influence diagram, such as the one shown in Figure 7b. The decision model represented by this influence diagram includes a single treatment decision and the assertion that the patient's utility depends only on the disease and the treatment decision. We compute a score for the Bayesian network by processing the cases sequentially. Namely, for each case $C_l$ in the database, we use Equation 18 to predict the probability distribution over diseases in that case, given the symptoms in that case and the previous cases $C_1, \ldots, C_{l-1}$. Then, we use the influence diagram to determine the optimal treatment. Next, from the utilities encoded in the influence diagram, we determine the utility of the outcome where the patient is given the chosen treatment, but has the disease recorded in the database for case $C_l$. We compute the score for the Bayesian network structure by summing these utilities over all cases. This metric can be generalized to score multiple network structures. The advantage of this decision-theoretic approach is that it optimizes what we indeed want to optimize: expected utility. The disadvantage is that it requires that we construct an influence diagram (i.e., decision model) for the domain.

### 6.2 Priors on Structures

The posterior-probability and decision-theoretic metrics require that we assign priors to all possible network structures. In this section, we present an efficient method for doing so described by Heckerman et al. (1995b).

The approach requires that the user constructs a *prior-network structure* for the domain.

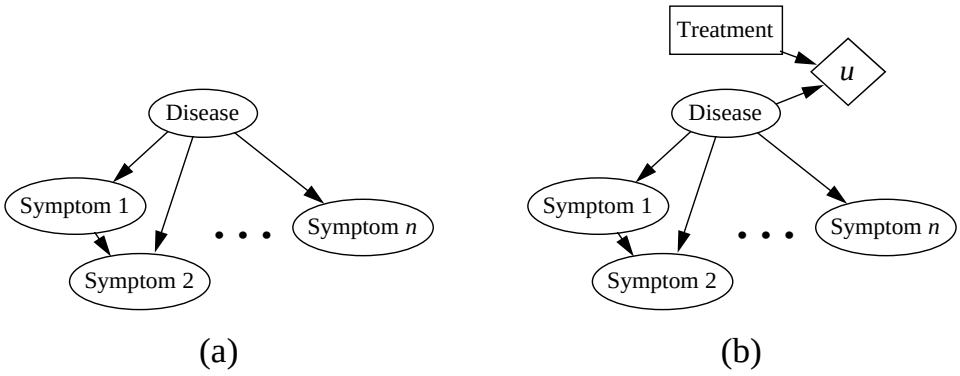


Figure 7: (a) A Bayesian network for medical diagnosis. (b) A corresponding influence diagram for medical treatment.

The method assumes that this structure is a user's "best guess" of the network structure that encodes the physical probabilities.

Given a prior network structure $P$, we compute the prior probability of $B_S$ as follows. For every variable $x_i$ in $U$, let $\delta_i$ denote the number of nodes in the symmetric difference of $\Pi_i(B_S)$ and $\Pi_i(P)$—that is, $(\Pi_i(B_S) \cup \Pi_i(P)) \setminus (\Pi_i(B_S) \cap \Pi_i(P))$. Then, $B_S$ and the prior network differ by $\delta = \sum_{i=1}^{n} \delta_i$ arcs. We compute the prior probability by penalizing $B_S$ by a constant factor $0 < \kappa \leq 1$ for each such arc. That is, we set

$$p(B_S^h|\xi) = c\ \kappa^{\delta} \tag{23}$$

where $c$ is a normalization constant, which we can ignore. Note that this approach assigns equal priors to equivalent network structures only when the prior network structure is empty (see Heckerman et al. [1995b] for a discussion of this point).

This formula is simple, as it requires only the assessment of a prior network structure and a single constant $\kappa$. Nonetheless, if the user is willing, he can provide more detailed knowledge by assessing different penalties for different nodes $x_i$ and for different parent configurations of each node (Buntine, 1991). Another variant of this approach is to allow the user to categorically assert that some arcs in the prior network must be present. We can again use Equation 23, except that we set to zero the priors of network structures that do not conform to these constraints.

### 6.3 Priors on Network Parameters

The posterior-probability and decision-theoretic metrics also require that we assign priors to network parameters for all possible network structures. We can also use this information

to compute the AIC and BIC metrics more efficiently (see Section 8.3). Several authors have discussed similar practical approaches for assigning these priors when many structures are possible (Cooper and Herskovits, 1991, 1992; Buntine, 1991; Spiegelhalter et al., 1993; Heckerman et al., 1995b). In this section, we describe the approach of Heckerman et al.

Their approach is based on a result from Geiger and Heckerman (1995). Namely, if all allowed values of the physical probabilities are possible, then parameter independence and hypothesis equivalence[6] imply that the physical probabilities for complete network structures must have Dirichlet distributions as specified in Equation 16 with the constraint

$$N'_{ijk} = N' \, p(x_i = k, \Pi_i = j | B^h_{S_C}, \xi) \tag{24}$$

where $N'$ is the user's equivalent sample size for the domain, $B^h_{S_C}$ is the hypothesis corresponding to any complete network structure, and $p(x_i = k, \Pi_i = j | B^h_{S_C}, \xi)$ is the user's probability that $x_i = k$ and $\Pi_i = j$ in the first case to be seen in the database.

Under these conditions, the priors on parameters for all complete network structures may be determined by (1) constructing a prior network for the first case to be seen (from which the probabilities in Equation 24 may be computed) and (2) assessing the equivalent sample size (i.e., confidence) in that prior network. In Section 7, we give an example of a prior network.

To determine priors for parameters of incomplete network structures, Heckerman et al. (1995b) use the assumption of *parameter modularity*, which says that given two network structures $B_{S1}$ and $B_{S2}$, if $x_i$ has the same parents in $B_{S1}$ and $B_{S2}$, then

$$p(\Theta_{ij} | B^h_{S1}, \xi) = p(\Theta_{ij} | B^h_{S2}, \xi)$$

for $j = 1, \ldots, q_i$. They call this property parameter modularity, because it says that the distributions for parameters $\Theta_{ij}$ depend only on the structure of the network that is local to variable $x_i$—namely, $\Theta_{ij}$ only depends on $x_i$ and its parents. For example, consider the network structure $x \rightarrow y$ and the empty structure for our two-variable domain with corresponding hypotheses $B^h_{x \rightarrow y}$ and $B^h_{xy}$. In both structures, $x$ has the same set of parents (the empty set). Consequently, by parameter modularity, $p(\theta_x | B^h_{x \rightarrow y}, \xi) = p(\theta_x | B^h_{xy}, \xi)$.

Given the assumptions of parameter modularity and independence, it is a simple matter to construct priors for the parameters of an arbitrary network structure given the priors on complete network structures. In particular, given parameter independence, we construct the priors for the parameters of each node separately. Furthermore, if node $x_i$ has parents $\Pi_i$ in the given network structure, we identify a complete network structure where $x_i$ has these parents, and use parameter modularity to determine the priors for this node. The

[6] Actually, Geiger and Heckerman (1995) proved this result using only likelihood equivalence.

result is a special case of the BD metric, called the BDe metric, that assigns equal scores to equivalent network structures. In Section 7, we illustrate the use of this metric.

### 6.4 Search Methods

In this section, we examine search methods for identifying network structures with high scores. Essentially all such search methods make use of a property of the scoring metrics that we call decomposability. Given a network structure for domain $U$, we say that a measure on that structure is *decomposable* if it can be written as a product of measures, each of which is a function only of one node and its parents. For example, from Equation 21, we see that the probability $p(D|B_S^h, \xi)$ given by the BD metric is decomposable. Consequently, if the prior probabilities of network structures are decomposable (as they are in Equation 23), then so is the BD metric. Thus, we can write

$$p(D, B_S^h|\xi) = \prod_{i=1}^{n} s(x_i|\Pi_i) \tag{25}$$

where $s(x_i|\Pi_i)$ is only a function of $x_i$ and its parents. Most Bayesian and non-Bayesian metrics are decomposable. Given a decomposable metric, 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.

First, let us consider the special case of finding the network structure with the highest score among all structures in which every node has at most one parent. For each arc $x_j \rightarrow x_i$ (including cases where $x_j$ is null), we associate a weight $w(x_i, x_j) \equiv \log s(x_i|x_j) - \log s(x_i|\emptyset)$. From Equation 25, we have

$$\begin{aligned} \log p(D, B_S^h) &= \sum_{i=1}^{n} \log s(x_i|\pi_i) \\ &= \sum_{i=1}^{n} w(x_i, \pi_i) + \sum_{i=1}^{n} \log s(x_i|\emptyset) \end{aligned} \tag{26}$$

where $\pi_i$ is the (possibly) null parent of $x_i$. The last term in Equation 26 is the same for all network structures. Thus, among the network structures in which each node has at most one parent, ranking network structures by sum of weights $\sum_{i=1}^{n} w(x_i, \pi_i)$ or by score has the same result.

Finding the network structure with the highest weight is a special case of a well-known problem of finding *maximum branchings* described—for example—in Evans and Minieka (1991). The problem is defined as follows. A *tree-like network* is a connected directed acyclic graph in which no two edges are directed into the same node. The root of a tree-like network is a unique node that has no edges directed into it. A *branching* is a directed forest

that consists of disjoint tree-like networks. A *spanning branching* is any branching that includes all nodes in the graph. A *maximum branching* is any spanning branching which maximizes the sum of arc weights (in our case, $\sum_{i=1}^{n} w(x_i, \pi_i)$). An efficient polynomial algorithm for finding a maximum branching was first described by Edmonds (1967), later explored by Karp (1971), and made more efficient by Tarjan (1977) and Gabow et al. (1984).

These algorithms can be used to find the branching with the highest score regardless of the metric we use, as long as one can associate a weight with every edge. Therefore, this algorithm is appropriate for any decomposable metric. When using metrics that assign equal scores to equivalent network structures, however, we have

$$s(x_i|x_j)s(x_j|\emptyset) = s(x_j|x_i)s(x_i|\emptyset)$$

Thus, for any two edges $x_i \rightarrow x_j$ and $x_i \leftarrow x_j$, the weights $w(x_i, x_j)$ and $w(x_j, x_i)$ are equal. Consequently, the directionality of the arcs plays no role for such metrics, and the problem reduces to finding the undirected forest for which $\sum w(x_i, x_j)$ is a maximum. This search can be done using a maximum spanning tree algorithm.

Now, let us consider the case where we find the best network from the set of all networks in which each node has no more than $k$ parents. Unfortunately, the problem for $k > 1$ is NP-hard (Chickering et al. 1995). Therefore, it is appropriate to use heuristic search algorithms.

Most of the commonly discussed search methods for learning Bayesian networks make successive arc changes to the network, and employ the property of decomposability to evaluate the merit of each change. The possible changes that can be made are easy to identify. For any pair of variables, if there is an arc connecting them, then this arc can either be reversed or removed. If there is no arc connecting them, then an arc can be added in either direction. All changes are subject to the constraint that the resulting network contains no directed cycles. We use $E$ to denote the set of eligible changes to a graph, and $\Delta(e)$ to denote the change in log score of the network resulting from the modification $e \in E$. Given a decomposable metric, if an arc to $x_i$ is added or deleted, only $s(x_i|\Pi_i)$ need be evaluated to determine $\Delta(e)$. If an arc between $x_i$ and $x_j$ is reversed, then only $s(x_i|\Pi_i)$ and $s(x_j|\Pi_j)$ need be evaluated.

One simple heuristic search algorithm is *local search* (e.g., Johnson, 1985). First, we choose a graph. Then, we evaluate $\Delta(e)$ for all $e \in E$, and make the change $e$ for which $\Delta(e)$ is a maximum, provided it is positive. We terminate search when there is no $e$ with a positive value for $\Delta(e)$. Using decomposable metrics, we can avoid recomputing all terms $\Delta(e)$ after every change. In particular, if neither $x_i$, $x_j$, nor their parents are changed, then $\Delta(e)$ remains unchanged for all changes $e$ involving these nodes as long as the resulting

network is acyclic. Candidates for the initial graph include the empty graph, a random graph, a graph determined by one of the polynomial algorithms described previously in this section, and the prior network.

A potential problem with any local-search method is getting stuck at a local maximum. Methods for escaping local maxima include iterated hill-climbing and simulated annealing. In *iterated hill-climbing*, we apply local search until we hit a local maximum. Then, we randomly perturb the current network structure, and repeat the process for some manageable number of iterations.

In one variant of *simulated annealing* described by Metropolis et al. (1953), we initialize the system at some temperature $T_0$. Then, we pick some eligible change $e$ at random, and evaluate the expression $p = \exp(\Delta(e)/T_0)$. If $p > 1$, then we make the change $e$; otherwise, we make the change with probability $p$. We repeat this selection and evaluation process $\alpha$ times or until we make $\beta$ changes. If we make no changes in $\alpha$ repetitions, then we stop searching. Otherwise, we lower the temperature by multiplying the current temperature $T_0$ by a decay factor $0 < \gamma < 1$, and continue the search process. We stop searching if we have lowered the temperature more than $\delta$ times. Thus, this algorithm is controlled by five parameters: $T_0, \alpha, \beta, \gamma$ and $\delta$. To initialize this algorithm, we can start with the empty graph, and make $T_0$ large enough so that almost every eligible change is made, thus creating a random graph. Alternatively, we may start with a lower temperature, and use one of the initialization methods described for local search.

Other methods for escaping local maxima include best-first search (Korf, 1993) and Gibbs' sampling (see Section 8.2).

# 7 A Real-World Example

Figure 8 illustrates an application of these techniques to the real-world domain of ICU ventilator management, taken from Heckerman et al. (1995b). Figure 8a is a hand-constructed Bayesian network for this domain, called the Alarm network (Beinlich et al., 1989) (the probabilities are not shown). Figure 8c is a database of 10,000 cases that is sampled from the Alarm network. Figure 8b is a hypothetical prior network for the domain. Heckerman et al. (1995b) constructed this network by adding, deleting, and reversing arcs in the Alarm network and by adding noise to the probabilities of the Alarm network.

Figure 8d shows the most likely network structure found by local search initialized with the prior network structure using the BDe metric, an equivalent sample size $N' = 64$, and priors on network structures determined by Equation 23 with $\kappa = 1/(N' + 1)$. Comparing the three network structures, we see that the learned network structure is much closer to that

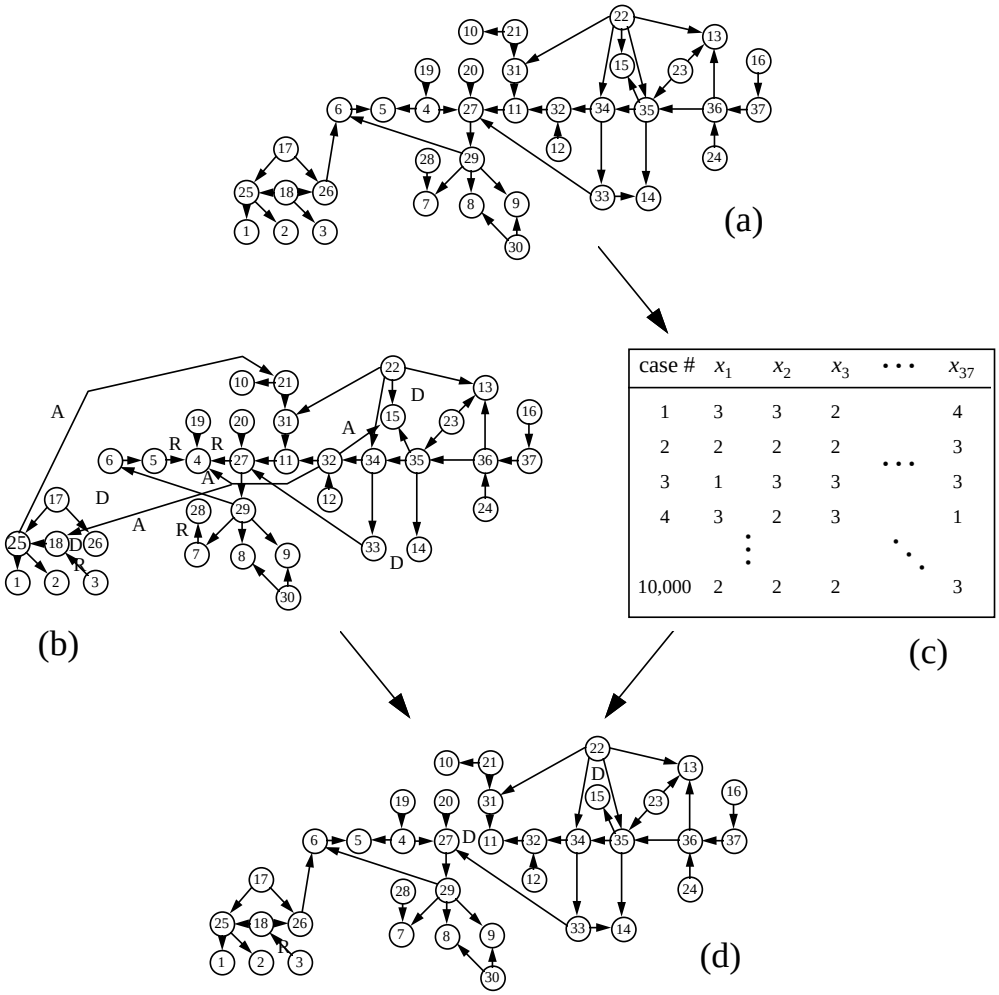


Figure 8: (a) The Alarm network structure. (b) A prior network encoding a user's beliefs about the Alarm domain. (c) A 10,000-case database generated from the Alarm network. (d) The network learned from the prior network and a 10,000 case database generated from the Alarm network. Arcs that are added, deleted, or reversed with respect to those in the Alarm network are indicated with A, D, and R, respectively. (Taken from Heckerman et al., 1995b.)

of the Alarm network than that of the prior network. Furthermore, the joint distribution encoded by the learned network is much closer to that of the Alarm network than that of the prior network. In particular, whereas the cross entropy of the joint distributions of the prior network with respect to that of the Alarm network is 5.6, the cross entropy of the joint distribution of the learned network with respect to that of the Alarm network is 0.03.[7] The learning algorithm has effectively used the database to "correct" the prior knowledge of the user.

# 8 Missing Data

In real databases, observations of one or more variables in one or more cases are typically missing. In this section, we consider extensions to previous methods that can handle missing data. We caution the reader that the methods we discuss assume that whether or not an observation is missing is independent of the actual states of the variables. For example, these methods are not appropriate for a medical database where data about drug response is missing in those patients who became too sick to take the drug. Methods for addressing dependencies in omissions have been explored by (e.g.) Rubin (1978), Robins (1986), and Pearl (1995).

## 8.1 Fill-In Methods

First, let us consider the simple situation where we observe a single incomplete case $C$ in domain $U$. Let $Y$ denote the variables not observed in the case. Under the assumption of parameter independence, we can compute the posterior distribution of $\Theta_{ij}$ as follows:

$$\begin{aligned}
p(\Theta_{ij}|C,\xi) &= \sum_U p(U|C,\xi)\; p(\Theta_{ij}|Y,C,\xi) \\
&= \sum_{x_i,\Pi_i} \left[ \sum_{U\setminus(\{x_i\}\cup\Pi_i)} p(U|C,\xi) \right] p(\Theta_{ij}|Y,C,\xi) \\
&= \sum_{x_i,\Pi_i} p(x_i,\Pi_i|C)\; p(\Theta_{ij}|Y,C,\xi) \\
&= (1 - p(\Pi_i = j|C,\xi))\; \{p(\Theta_{ij}|\xi)\} + \\
&\quad \sum_{k=1}^{r_i} p(x_i = k, \Pi_i = j|C,\xi)\; \{p(\Theta_{ij}|x_i = k, \Pi_i = j,\xi)\} \qquad (27)
\end{aligned}$$

Each term in curly brackets in Equation 27 is a Dirichlet distribution. Thus, unless both $x_i$ and all the variables in $\Pi_i$ are observed in case $C$, the posterior distribution of $\Theta_{ij}$ will

[7] By way of comparison, the cross entropy of an empty network whose probabilities are determined from the marginals of the Alarm network with respect to that of the Alarm network is 13.6.

be a linear combination of Dirichlet distributions. Such distributions are sometimes called *Dirichlet mixtures*; and the probabilities $(1 - p(\Pi_i = j|C, \xi))$ and $p(x_i = k, \Pi_i = j|C, \xi), k = 1, \ldots, r_i$ are called *mixing coefficients*.

If we observe two cases, then the situation becomes more complex, because the computation of the mixing coefficients involves finding the means of Dirichlet mixtures. In general, as shown (e.g.) in Cooper and Herskovits (1992), the computational complexity of the exact computation of $p(D, B_S^h|\xi)$ can be exponential in the number of missing variable entries in the database.

Thus, in practice, we require an approximation. One approach is to approximate each correct posterior distribution $\Theta_{ij}$ with a single Dirichlet distribution, and continue to use Equation 20 along with the formula for the mean of a Dirichlet distribution. Several such approximations have been described in the literature. For example, Titterington (1976) describes a method called *fractional updating*, wherein for each $\Theta_{ij}$, we pretend that we have observed a fractional number of observations corresponding to that parameter set. In particular, he suggests the approximation

$$p(\Theta_{ij}|C, \xi) \approx c \prod_{k=1}^{r_i} \theta_{ijk}^{p(x_i=k, \Pi_i=j|C,\xi)} p(\Theta_{ij}|\xi) \tag{28}$$

One drawback of this method is that it falsely increases the equivalent sample sizes of the Dirichlet distributions associated with each $\Theta_{ij}$, because it replaces each missing datum with a fractional sample. Cowell et al. (1995) suggest an approach that does not have this problem. Namely, they approximate $\Theta_{ij}$ by a single Dirichlet whose means and average variance $\sum_{k=1}^{r_i} Var(\theta_{ijk})/r_i$ are the same as those for the correct Dirichlet mixture. We note that Titterington's approach—unlike Cowell et al.'s method—produces a scoring metric that assigns equal scores to equivalent network structures.

These approximations process the data in the database sequentially, and make use of the assumption of parameter independence and properties of the Dirichlet distribution. Other methods—including Gibbs sampling, the EM algorithm, and gradient descent—process all the data at once, and can handle continuous domain variables and dependent parameters.

## 8.2 Gibbs Sampling

Gibbs sampling, described—for example—by Geman and Geman (1984), is a special case of Markov chain Monte Carlo methods for approximate inference (Hastings, 1970). Given variables $X = \{x_1, \ldots, x_n\}$ with some joint distribution $p(X|\xi)$, we can use a Gibbs sampler to approximate the expectation of any function $f(X)$ as follows. First, we choose an initial state of each of the variables in $X$ somehow. Next, we pick some variable $x_i$, unassign its

current state, and compute its probability distribution given the assignments to the other $n-1$ variables. Then, we sample a state for $x_i$ based on this probability distribution, and compute $f(X)$. Finally, we iterate the previous two steps, keeping track of the average value of $f(X)$. In the limit as the number of samples approach infinity, this average is equal to $E(f(X)|\xi)$ (the expectation of $f(X)$ with respect to the distribution $p(X|\xi)$) provided two conditions are met. First, the Gibbs sampler must be *irreducible*: The probability distribution $p(X)$ must be such that we can eventually sample any possible state of $X$ given any possible initial state of $X$. For example, if $p(X)$ contains no zero probabilities, then the Gibbs sampler will be irreducible. Second, each $x_i$ must be chosen infinitely often. In practice, an algorithm for deterministically rotating through the variables is typically used. Good introductions to Gibbs sampling—including methods for initialization and a discussion of convergence—are given by York (1992) and Neal (1993).

Now, suppose we have a database $D = \{C_1, \ldots, C_m\}$ with missing data, and we want to approximate $p(\Theta_{B_S}|B_S^h, D, \xi)$. One variant of Gibbs sampling for performing this approximation goes as follows. First, we initialize the parameters $\Theta_{B_S}$ somehow. Second, for each case $C_l$ in $D$ containing missing data, we fill in the missing data using the assigned values of $\Theta_{B_S}$. For example, suppose variables $x_3$ and $x_7$ are unobserved in case $C_1$. We fill in $x_3$ by sampling from the distribution $p(x_3|C_1, \Theta_{B_S}, \xi)$, and then fill in $x_7$ by sampling from the distribution $p(x_7|x_3, C_1, \Theta_{B_S}, \xi)$. This step can be done using any standard Bayesian network inference algorithm. Third, we reassign the parameters $\Theta_{B_S}$ according to the posterior distribution $p(\Theta_{B_S}|D', \xi)$, where $D'$ is the completed database. Finally, we iterate the previous two steps, and use the sampled values of $\Theta_{B_S}$ as an approximation for $p(\Theta_{B_S}|B_S^h, D, \xi)$. Buntine (1994) discusses this approach in more detail.

Gibbs sampling can also be used in place of searching over network structures. Namely, we can modify the Gibbs sampler described in the previous paragraph so that it can transition from one network structure to another. For example, after each sampling pass through the parameters and database, the sampler can evaluate the $p(B_S^h, D|\xi)$ for every structure that is "close" to the current network structure (e.g., within one arc addition, deletion, or reversal) and sample a new network structure according to this distribution (see Madigan and Raferty, 1994).

### 8.3 EM Algorithm

The expectation–maximization (EM) algorithm is an approximation algorithm that can find a local maximum of a probability $p(\cdot|\Theta, \xi)$ as a function of parameters $\Theta$ (Dempster et al., 1977). Given a database $D = \{C_1, \ldots, C_m\}$ with missing data, we can approximate $p(D|B_S^h, \xi)$ as the local maximum for $p(D|B_S^h, \Theta_{B_S}, \xi)$ found by the EM algorithm. Like the

Gibbs sampler, the EM algorithm can handle models with missing data, continuous domain variables, and dependent parameters. Although the EM algorithm tends to provide a less accurate approximation, it typically converges more quickly than a Gibbs sampler.

The EM algorithm can be viewed as a deterministic version of the Gibbs sampler. Like the Gibbs sampler, we begin the approximation of $p(D|B_S^h, \xi)$ by assigning values to $\Theta_{B_S}$ somehow. Next, rather than sample a complete database $D'$, we compute the *expected sufficient statistics* for the missing entries in the database. In particular, we compute

$$E(N_{ijk}|\Theta_{B_S}, \xi) = \sum_{l=1}^{m} p(x_i = k, \Pi_i = j|C_l, \Theta_{B_S}, \xi) \tag{29}$$

When $x_i$ and all the variables in $\Pi_i$ are observed in case $C_l$, the term for this case requires a trivial computation: it is either zero or one. Otherwise, we can use any Bayesian network inference algorithm to evaluate the term. This computation is called the *expectation step* of the EM algorithm.

Next, rather than sample new values for $\Theta_{B_S}$, we use the expected sufficient statistics as if they were actual sufficient statistics from a database $D'$, and set the new values of $\Theta_{B_S}$ to be the modes of the posterior distribution $p(\Theta_{B_S}|D', B_S^h, \xi)$. For example, if the parameters $\Theta_{B_S}$ have a Dirichlet distribution, then we have

$$\text{Mode}(\theta_{ijk}|\xi) = \frac{N'_{ijk} + E(N_{ijk}|\Theta_{B_S}, \xi) - 1}{N'_{ij} + E(N_{ij}|\Theta_{B_S}, \xi) - r_i}$$

This mode exists provided $N'_{ijk} + E(N_{ijk}|\Theta_{B_S}, \xi) > 1, k = 1, \ldots, r_i$. (If not, we can use the expectation of $\theta_{ijk}$ in its place.) This assignment is called the *maximization step* of the EM algorithm.

Dempster (1977) showed that, under certain regularity conditions, iteration of the expectation and maximization steps will converge to a local maximum of the probability $p(D|B_S^h, \Theta_{B_S}, \xi)$.

## 8.4 General Optimization Methods

To use Gibbs sampling or the EM algorithm in practice, we need to compute the distributions $p(\Theta_{B_S}|D', B_S^h, \xi)$ efficiently. These computations are efficient provided the parameters $\Theta_{B_S}$ have a Dirichlet distribution or some other distribution from the exponential family. When the distributions do not have this form, we can use general optimization methods to maximize $p(D|B_S^h, \Theta_{B_S}, \xi)$ (Gill et al. 1981; Press et al. 1992).

Many of these approaches—for example, gradient descent, conjugate gradient, and quasi-Newton methods—exploit derivatives of the function to be maximized to speed up convergence. In some situations, these derivatives can be computed efficiently in closed form.

Buntine (1994) discusses such methods in detail. Here, we mention the case where all variables in $U$ are discrete. In this situation, Russell et al. (1994) have shown that

$$\frac{\partial \log p(D|B_S^h, \Theta_{B_S}, \xi)}{\partial \theta_{ijk}} = \frac{E(N_{ijk}|\Theta_{B_S}, \xi)}{\theta_{ijk}}$$

where $E(N_{ijk}|\Theta_{B_S}, \xi)$ is the expected sufficient statistic given by Equation 29. As noted in the previous section, this term may be computed by any standard Bayesian network inference algorithm.

## 9 Learning New Variables

In a database with missing data, a particular variable may be observed in some cases, or it may never be observed. In the latter situation, we say that the variable is *hidden*.

Any of the methods described in the previous section can be used to learn Bayesian networks containing identified hidden variables. The network structure may be fixed and only the physical probabilities uncertain, or both the network structure and parameters may be uncertain. One example of learning the probabilities of a fixed structure with hidden variables is the AutoClass algorithm of Cheeseman and Stutz (1995), which performs *unsupervised classification*. The model underlying the algorithm is a Bayesian network with a single hidden variable whose states correspond to the unknown classes. The number of states of the hidden variable is uncertain and has a prior distribution. Also, this hidden variable renders sets of observable variables conditionally independent. The algorithm searches over variations of this model (including the number of states of the hidden variable), using a version of the EM algorithm to approximate the posterior probability of each model variation.

In addition, we can use methods for learning with missing data to identify (under uncertainty) the existence of new variables. Namely, we hypothesize a mutually exclusive and exhaustive set of Bayesian-network structures, some containing hidden variables and some not. We assign priors to each structure and its parameters, and then update these priors with data using one of the described algorithms for handling missing data.

## 10 Pointers to the Literature

Like all tutorials, this tutorial is incomplete. For those readers interested in learning more about graphical models and methods for learning them, we offer the following additional references. A more detailed guide to the literature can be found in Buntine (1995).

Charniak (1991) provides an easy-to-read introduction to the Bayesian-network representation. Spiegelhalter et al. (1993) and Heckerman et al. (1995b) give simple discussions of methods for learning Bayesian networks for domains containing only discrete variables. Buntine (1994) and Heckerman and Geiger (1994) provide more detailed discussions. Experimental comparisons of different learning approaches can be found in Cooper and Herskovits (1992), Aliferis and Cooper (1994), Lauritzen et al. (1994), Cowell et al. (1995), and Heckerman et al. (1995b).

In addition to directed models, researchers have also explored graphs containing undirected edges as a knowledge representation. These representations are discussed (e.g.) in Lauritzen (1982), Verma and Pearl (1990), and Frydenberg (1990). Bayesian methods for learning such models from data are described by Dawid and Lauritzen (1993) and Buntine (1994).

Finally, several software systems for learning graphical models have been implemented. Thomas, Spiegelhalter, and Gilks (1992) have created a system that takes a learning problem specified as a Bayesian network and compiles this problem into a Gibbs-sampler computer program. Badsberg (1992) and Højsgaard et al. (1994) have built systems that can learn directed, undirected, and mixed graphical models using a variety of scoring metrics.

## Acknowledgments

I thank David Chickering, Eric Horvitz, Chris Meek, Koos Rommelse, and Padhraic Smyth for their comments on earlier versions of this manuscript.

## References

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

[Aliferis and Cooper, 1994] Aliferis, C. and Cooper, G. (1994). An evaluation of an algorithm for inductive learning of Bayesian belief networks using simulated data sets. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA, pages 8–14. Morgan Kaufmann.

[Badsberg, 1992] Badsberg, J. (1992). Model search in contingency tables by CoCo. In Dodge, Y. and Wittaker, J., editors, *Computational Statistics*, pages 251–256. Physica Verlag, Heidelberg.

[Bayes, 1958] Bayes, T. (1958). An essay towards solving a problem in the doctrine of chances. *Biometrika*, 46:293–298. Reprint of original work of 1763.

[Beinlich et al., 1989] Beinlich, I., Suermondt, H., Chavez, R., and Cooper, G. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medicine,* London, pages 247–256. Springer Verlag, Berlin.

[Buntine, 1991] Buntine, W. (1991). Theory refinement on Bayesian networks. In *Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence,* Los Angeles, CA, pages 52–60. Morgan Kaufmann.

[Buntine, 1994] Buntine, W. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225.

[Buntine, 1995] Buntine, W. (1995). A guide to the literature on learning graphical models. Technical Report IC-95-05, NASA Ames Research Center.

[Charniak, 1991] Charniak, E. (1991). Bayesian networks without tears. *AI Magazine*, 12:50–63.

[Cheeseman and Stutz, 1995] Cheeseman, P. and Stutz, J. (1995). Bayesian classification (AutoClass): Theory and results. In Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, *Advances in Knowledge Discovery and Data Mining*, page ΓΓ AAAI Press, Menlo Park, CA.

[Chickering, 1995] Chickering, D. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence,* Montreal, QU. Morgan Kaufmann.

[Chickering et al., 1995] Chickering, D., Geiger, D., and Heckerman, D. (1995). Learning Bayesian networks: Search methods and experimental results. In *Proceedings of Fifth Conference on Artificial Intelligence and Statistics,* Ft. Lauderdale, FL, pages 112–128. Society for Artificial Intelligence in Statistics.

[Cooper, 1990] Cooper, G. (1990). Computational complexity of probabilistic inference using Bayesian belief networks (Research note). *Artificial Intelligence*, 42:393–405.

[Cooper and Herskovits, 1992] Cooper, G. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

[Cooper and Herskovits, 1991] Cooper, G. and Herskovits, E. (January, 1991). A Bayesian method for the induction of probabilistic networks from data. Technical Report SMI-91-1, Section on Medical Informatics, Stanford University.

[Cowell et al., 1995] Cowell, R., Dawid, A., and Sebastiani, P. (1995). A comparison of sequential learning methods for incomplete data. Technical Report 135, Department of Statistical Science, University College London.

[Cox, 1946] Cox, R. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13.

[Dagum and Luby, 1994] Dagum, P. and Luby, M. (1994). On the approximation of probabilistic inference. Technical Report May, Section on Medical Informatics, Stanford University School of Medicine, Stanford, CA.

[D'Ambrosio, 1991] D'Ambrosio, B. (1991). Local expression languages for probabilistic dependence. In *Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence,* Los Angeles, CA, pages 95–102. Morgan Kaufmann.

[de Finetti, 1970] de Finetti, B. (1970). *Theory of Probability.* Wiley and Sons, New York.

[DeGroot, 1970] DeGroot, M. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38.

[Edmonds, 1967] Edmonds, J. (1967). Optimum branching. *J. Res. NBS*, 71B:233–240.

[Evans and Minieka, 1991] Evans, J. and Minieka, E. (1991). *Optimization algorithms for networks and graphs.* Marcel Dekker Inc., New York.

[Frydenberg, 1990] Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, 17:333–353.

[Gabow et al., 1984] Gabow, H., Galil, Z., and Spencer, T. (1984). Efficient implementation of graph algorithms using contraction. In *Proceedings of FOCS.*

[Geiger and Heckerman, 1995] Geiger, D. and Heckerman, D. (Revised February, 1995). A characterization of the Dirichlet distribution applicable to learning Bayesian networks. Technical Report MSR-TR-94-16, Microsoft, Redmond, WA.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–742.

[Gill et al., 1981] Gill, P., Murray, W., and Wright, M. (1981). *Practical Optimization*. Academic Press, New York.

[Good, 1950] Good, I. (1950). *Probability and the Weighing of Evidence*. Hafners, New York.

[Good, 1959] Good, I. (1959). Kinds of probability. *Science*, 129:443–447.

[Good, 1965] Good, I. (1965). *The Estimation of Probabilities*. MIT Press, Cambridge, MA.

[Hacking, 1975] Hacking, I. (1975). *The Emergence of Probability*. Cambridge University Press, New York.

[Hastings, 1970] Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

[Heckerman, 1989] Heckerman, D. (1989). A tractable algorithm for diagnosing multiple diseases. In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, Windsor, ON, pages 174–181. Association for Uncertainty in Artificial Intelligence, Mountain View, CA. Also in Henrion, M., Shachter, R., Kanal, L., and Lemmer, J., editors, *Uncertainty in Artificial Intelligence 5*, pages 163–171. North-Holland, New York, 1990.

[Heckerman and Geiger, 1994] Heckerman, D. and Geiger, D. (December, 1994). Learning Bayesian networks. Technical Report MSR-TR-95-02, Microsoft, Redmond, WA.

[Heckerman et al., 1995b] Heckerman, D., Mamdani, A., and Wellman, M. (1995a). Real-world applications of Bayesian networks. *Communications of the ACM*, 38.

[Heckerman et al., 1995a] Heckerman, D., Geiger, D., and Chickering, D. (1995b). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, to appear.

[Højsgaard et al., 1994] Højsgaard, S., Skjøth, F., and Thiesson, B. (1994). User's guide to BIOFROST. Technical report, Department of Mathematics and Computer Science, Aalborg, Denmark.

[Howard, 1970] Howard, R. (1970). Decision analysis: Perspectives on inference, decision, and experimentation. *Proceedings of the IEEE*, 58:632–643.

[Howard, 1990] Howard, R. (1990). From influence to relevance to knowledge. In Oliver, R. and Smith, J., editors, *Influence Diagrams, Belief Nets, and Decision Analysis*, chapter 1. Wiley and Sons, New York.

[Howard and Matheson, 1981] Howard, R. and Matheson, J. (1981). Influence diagrams. In Howard, R. and Matheson, J., editors, *Readings on the Principles and Applications of Decision Analysis*, volume II, pages 721–762. Strategic Decisions Group, Menlo Park, CA.

[Jensen and Andersen, 1990] Jensen, F. and Andersen, S. (1990). Approximations in Bayesian belief universes for knowledge based systems. Technical report, Institute of Electronic Systems, Aalborg University, Aalborg, Denmark.

[Johnson, 1985] Johnson (1985). How fast is local searchΓ In *FOCS*, pages 39–42.

[Kahneman et al., 1982] Kahneman, D., Slovic, P., and Tversky, A., editors (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York.

[Karp, 1971] Karp, R. (1971). A simple derivation of Edmond's algorithm for optimal branchings. *Networks*, 1:265–272.

[Kass and Raftery, 1993] Kass, R. and Raftery, A. (1993). Bayes factors and model uncertainty. Technical Report 571, Department of Statistics, Carnegie Mellon University, PA.

[Korf, 1993] Korf, R. (1993). Linear-space best-first search. *Artificial Intelligence*, 62:41–78.

[Lauritzen, 1982] Lauritzen, S. (1982). *Lectures on Contingency Tables*. University of Aalborg Press, Aalborg, Denmark.

[Lauritzen and Spiegelhalter, 1988] Lauritzen, S. and Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistical Society B*, 50:157–224.

[Lauritzen et al., 1994] Lauritzen, S., Thiesson, B., and Spiegelhalter, D. (1994). Diagnostic systems created by model selection methods: A case study. In Cheeseman, P. and Oldford, R., editors, *AI and Statistics IV*, volume Lecture Notes in Statistics, 89, pages 143–152. Springer-Verlag, New York.

[Madigan and Raftery, 1994] Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89:1535–1546.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

[Neal, 1993] Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

[Olmsted, 1983] Olmsted, S. (1983). *On representing and solving decision problems*. PhD thesis, Department of Engineering-Economic Systems, Stanford University.

[Pearl, 1986] Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29:241–288.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

[Pearl, 1995] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, to appear.

[Pearl and Verma, 1991] Pearl, J. and Verma, T. (1991). A theory of inferred causation. In Allen, J., Fikes, R., and Sandewall, E., editors, *Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452. Morgan Kaufmann, New York.

[Press et al., 1992] Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C*. Cambridge University Press, New York.

[Ramamurthi and Agogino, 1988] Ramamurthi, K. and Agogino, A. (1988). Real time expert system for fault tolerant supervisory control. In Tipnis, V. and Patton, E., editors, *Computers in Engineering*, pages 333–339. American Society of Mechanical Engineers, Corte Madera, CA.

[Rissanen, 1987] Rissanen, J. (1987). Stochastic complexity (with discussion). *Journal of the Royal Statistical Society, Series B*, 49:223–239 and 253–265.

[Robins, 1986] Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure results. *Mathematical Modeling*, 7:1393–1512.

[Rubin, 1978] Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58.

[Russell et al., 1994] Russell, S., Binder, J., and Koller, D. (1994). Adaptive probabilistic networks. Technical Report CSD-94-824, University of California, Berkeley.

[Savage, 1954] Savage, L. (1954). *The Foundations of Statistics*. Dover, New York.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

[Shachter, 1988] Shachter, R. (1988). Probabilistic inference and influence diagrams. *Operations Research*, 36:589–604.

[Shachter et al., 1990] Shachter, R., Andersen, S., and Poh, K. (1990). Directed reduction algorithms and decomposable graphs. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence,* Boston, MA, pages 237–244. Association for Uncertainty in Artificial Intelligence, Mountain View, CA.

[Spiegelhalter et al., 1993] Spiegelhalter, D., Dawid, A., Lauritzen, S., and Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8:219–282.

[Spiegelhalter and Lauritzen, 1990] Spiegelhalter, D. and Lauritzen, S. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605.

[Spirtes et al., 1993] Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.

[Suermondt and Cooper, 1991] Suermondt, H. and Cooper, G. (1991). A combination of exact algorithms for inference on Bayesian belief networks. *International Journal of Approximate Reasoning*, 5:521–542.

[Tarjan, 1977] Tarjan, R. (1977). Finding optimal branchings. *Networks*, 7:25–35.

[Thomas et al., 1992] Thomas, A., Spiegelhalter, D., and Gilks, W. (1992). Bugs: A program to perform Bayesian inference using Gibbs sampling. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 4*, pages 837–842. Oxford University Press.

[Titterington, 1976] Titterington, D. (1976). Updating a diagnostic system using unconfirmed cases. *Applied Statistics*, 25:238–247.

[Tversky and Kahneman, 1974] Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131.

[Verma and Pearl, 1990] Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence,* Boston, MA, pages 220–227. Morgan Kaufmann.

[von Neumann and Morgenstern, 1947] von Neumann, J. and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.

[Winkler, 1967] Winkler, R. (1967). The assessment of prior distributions in Bayesian analysis. *American Statistical Association Journal*, 62:776–800.

[York, 1992] York, J. (1992). *Bayesian methods for the analysis of misclassified or incomplete multivariate discrete data*. PhD thesis, Department of Statistics, University of Washington, Seattle.