ELSEVIER

Artificial Intelligence 88 (1996) 143-161

Artificial Intelligence

Propagating imprecise probabilities in Bayesian networks *

Gernot D. Kleiter *

Institut für Psychologie, Universität Salzburg, Hellbrunnerstr. 34, 5020 Salzburg, Austria

Received March 1995; revised February 1996

Abstract

Often experts are incapable of providing "exact" probabilities; likewise, samples on which the probabilities in networks are based must often be small and preliminary. In such cases the probabilities in the networks are imprecise. The imprecision can be handled by second order probability distributions. It is convenient to use beta or Dirichlet distributions to express the uncertainty about probabilities. The problem of how to propagate point probabilities in a Bayesian network now is transformed into the problem of how to propagate Dirichlet distributions in Bayesian networks.

It is shown that the propagation of Dirichlet distributions in Bayesian networks with incomplete data results in a system of probability mixtures of beta-binomial and Dirichlet distributions. Approximate first order probabilities and their second order probability density functions are obtained by stochastic simulation. A number of properties of the propagation of imprecise probabilities are discussed by the use of examples. An important property is that the imprecision of inferences increases rapidly as new premises are added to an argument. The imprecision can be used as a pruning criterion in a network to keep the number of variables involved in an inferential argument small. Thus, imprecision may be used as an Ockam's razor in Bayesian networks.

1. Introduction

Bayesian belief networks represent and process probabilistic knowledge. Their representational components belong to one of two domains, a qualitative or a quantitative

* Thanks are due to the Fonds zur Förderung der wissenschaftlichen Forschung, Vienna, for the financial support. Thanks are also due to the hospitality of the Department of Psychology, Bowling Green State University, Ohio, especially to Michael E. Doherty.

* E-mail: gernot.kleiter@sbg.ac.at.

PII S0004-3702(96)00021-5

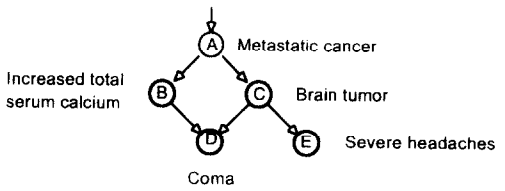

Fig. 1. Bayesian network: Cooper's medical diagnosis example.

Table 1
Weight tables associated with Cooper's example in Fig. 1; the numbers were choosen so that two conditions are fullfilled: (i) the ratios of the weights preserve the probabilities of the original version of Cooper's example and (ii) the total sum of all elementary weights is 120

$\langle A \rangle$		$\langle B, A \rangle$	$\neg a$	a	$\langle C, A \rangle$	$\neg a$	a	$\langle E, C \rangle$	$\neg c$	c
$\neg a$	96	$\neg b$	77	5	$\neg c$	91	19	$\neg e$	44	2
a	24	b	19	19	c	5	5	e	66	8

$\langle D, B, C \rangle$	$\neg b$	$\neg b$	b	b
	$\neg c$	c	$\neg c$	c
$\neg d$	72	1	7	1
d	4	4	27	4

one. The basic qualitative relationships of (conditional) dependence and independence between variables are expressed in the visual language of graph theory. The quantitative specifications of the involved (conditional) probability distributions are organized in tables and attached to the nodes of the graph. The tables are not "visible" in the graphical representation. Consider the example shown in Fig. 1 and Table 1 [7,34,37]. The network in Fig. 1 represents the dependencies in a graphical model. The nodes A–E represent clinical *absent/present* variables like diseases, test results, or symptoms. Table 1 contains the associated quantitative specifications. Assume we have investigated 120 patients suspected of suffering from a specific metastatic form of cancer. It turns out that 24 actually have developed the metastatic form and 96 have not. Of those having the metastatic form, 19 show increased total serum calcium and 5 do not. Of those patients in which the metastatic form was not observed, 19 show increased total serum calcium and 77 do not, etc. These and the remaining frequencies are contained in Table 1. The main purpose of a Bayesian belief network is to perform probabilistic inference. If for a patient one or more of the variables are observed and are known for certain, this affects the probabilities of the neighboring states in the graph. Evidence and updated probabilities *propagate* through the network. Various kinds of probabilistic inference like medical diagnosis, prediction, or explanation are special cases of propagating probabilities in a Bayesian network. Bayesian belief networks belong to the class of graphical probabilistic models ([6,11,15,30,43]; for tutorials and related work on uncertainty in artificial intelligence see the `http://www.auai.org` page and the references given there).

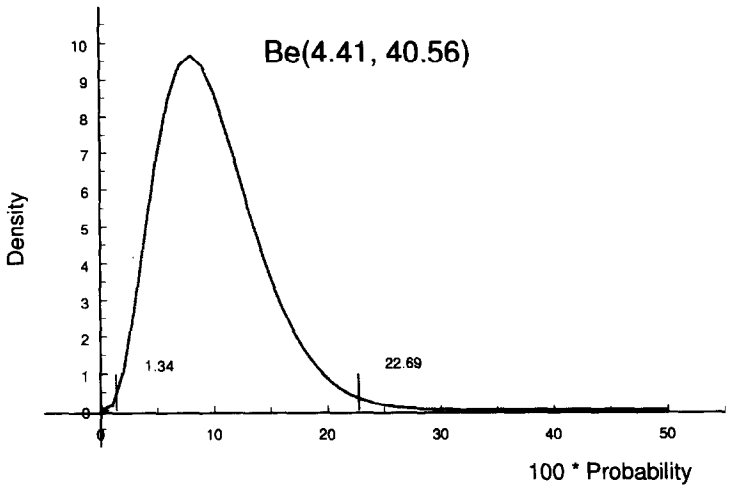

Fig. 2. The beta distribution $Be(4.41, 40.56)$ together with a 99% highest density interval for Cooper's example.

Usually, the probabilities in Bayesian networks are treated as though they were *known precisely*. In the present paper we analyse Bayesian networks in which the probabilities are *not known precisely*. Experts often cannot provide exact point probabilities, providing intervals instead. Probability estimates derived from empirical data are often based on small sample sizes. In such cases the probabilities in a Bayesian network cannot be considered to be precise point values. In the literature, several proposals have been made how to handle imprecision in dependency structures, such as lower and upper bounds [9, 12, 42], propagation of variances [8, 31, 38], and second order distributions [17, 21, 25, 26, 33, 39, 40]. A tutorial is provided in [16].

We treat probabilities that are not known precisely in the same way they are treated in Bayesian statistics [3], as *uncertain quantities* to which a (second order) probability density function is attached. The distributions express the imprecision. If little is known about the uncertain quantity, the distribution is flat and its variance large. If much is known, the distribution is tight and its variance small. The use of a second order probability distribution is a standard procedure in Bayesian statistics and there is nothing especially exciting about it. The procedure actually goes back to Thomas Bayes. He was one of the first who plotted a continuous probability density function, a beta distribution (upside-down) over the unit interval.

The method proposed in this paper allows the derivation of the following inferences: If a patient does not intermittently fall into comas ($\neg d$) but suffers from severe headaches (e), then the probability of a metastatic cancer (a) is 0.098. However, there is an appreciable imprecision associated with this estimate. We can be 99% sure that the true probability lies in the interval 0.0134 and 0.227. The standard deviation of the estimate is 0.0436. The imprecision may be expressed by the beta distribution $[a|\neg d, e] = Be(4.41, 40.56)$, where the brackets are used as a shorthand notation for "the probability density function of the parameter corresponding to the probability of a given $\neg d$ and

e". While the full example is based on a total sample size of 120 cases the precision of the present inference corresponds to a sample size of 45 cases only. Fig. 2 shows the beta distribution together with a 99% highest density interval (the shortest interval with probability content 0.99). Further analysis shows that the severe headaches are not really essential for inferences about the metastatic cancer. Conditioning on $\neg d$ alone leads to the distribution $[a|\neg d] = Be(4.50, 47.44)$ with mean 0.087 and standard deviation 0.0387. The precision even slightly increases when D is instantiated only as compared when both D and E are instantiated. We will come back to this at first sight counterintuitive property.

2. Basic model

We consider a set of vertices (nodes, variables) V and a set of directed edges E (arcs, probabilistic dependencies between variables) defined on $V \times V$. The vertices and the edges are represented by a graph $G = (V, E)$. If the arcs do not contain cycles, the graph is a directed acyclic graph (DAG). With each node $X \in V$ and the set of its parents $\mathrm{pa}(X) = \{U_1, \ldots, U_m\}$ we associate a weight table $\langle X, U_1, \ldots, U_m \rangle$. If $\langle X, U \rangle$ is a two-dimensional weight table, then we denote the marginal of X along U by $\langle X, U \rangle^{\downarrow U}$. More generally, if $W = \langle X_1, X_2, \ldots, X_n \rangle$ is an n-dimensional weight table, we denote the marginal along the subset $\{Z_1, \ldots, Z_m\} \subseteq \{X_1, \ldots, X_n\}$ by $\langle X_1, X_2, \ldots, X_n \rangle^{\downarrow \{Z_1, \ldots, Z_m\}}$.

We follow [14] and denote a probability density function (pdf) by brackets. Joint, conditional, and marginal distributions are written as $[X, Y]$, $[X|Y]$, and $[X]$, respectively. The product of densities is denoted by $*$, e.g., $[X, Y] = [X|Y] * [Y]$ etc. The weight tables define (second order) pdfs $[X|\mathrm{pa}(x)]$ for each variable X. We conceive the weights in a table as the shape parameters of Dirichlet distributions. Dirichlet distributions and their special versions for binary cases, the beta distributions, are defined as follows:

Definition 1 (*Dirichlet distribution*). Let $(Y_1, \ldots, Y_d)$ be a random vector on the simplex $\mathcal{S}^d = \{(y_1, \ldots, y_d): y_i \leqslant 0,\ i = 1, \ldots, d; \sum_{i=1}^{d} y_i \leqslant 1\}$ and $(\nu_1, \ldots, \nu_D)$ a vector of reals with $(\nu_1 > 0, \ldots, \nu_D > 0)$, where $d = D - 1$. If the density of the random vector is given by

$$p(y_1, \ldots, y_d) = \frac{\Gamma(\nu_1 + \cdots + \nu_D)}{\Gamma(\nu_1) \cdots \Gamma(\nu_D)} y_1^{\nu_1 - 1} \cdots y_d^{\nu_d - 1} \left(1 - \sum_{i=1}^{d} y_i\right)^{\nu_D - 1}, \tag{1}$$

we say that $(Y_1, \ldots, Y_d)$ follows a d-variate Dirichlet distribution. We use the shorthand $[Y_1, \ldots, Y_d] = Di(\nu_1, \ldots, \nu_D)$.

Definition 2 (*Beta distribution*). Let Y be a random variable on the unit interval. If its density is given by

$$p(y) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} y^{\nu_1 - 1}(1 - y)^{\nu_2 - 1}, \tag{2}$$

Y is beta distributed with shape parameters ν_1 and ν_2. We write for short $[Y] = Be(\nu_1, \nu_2)$. Its mean and variance are given by

$$\mathrm{E}(Y) = \frac{\nu_1}{\nu_1 + \nu_2}, \tag{3}$$

$$\mathrm{var}(Y) = \frac{\nu_1 \nu_2}{(\nu_1 + \nu_2)^2(\nu_1 + \nu_2 + 1)}. \tag{4}$$

In a beta distribution we interpret the sum $\nu_1 + \nu_2$ as the total amount of evidence available about the point probability $\nu_1/(\nu_1 + \nu_2)$. ν_1 is the weight in favor of an event, a proposition, or a hypothesis, and ν_2 the weight against it. Weights of evidence were extensively discussed by Keynes [20].

The beta or Dirichlet distributions implement a system of second order pdfs on the probability parameters underlying the network. If a node X has no parents, then the pdf is a marginal distribution. If the node has n parents the weight table has $n + 1$ dimensions. If the number of possible values of the node under consideration is m_0 and the number of possible values of its parents is $m_1, \ldots, m_n$ then the weight table is of order $m_0 \times m_1 \times \cdots \times m_n$.

The probability parameters underlying the network are not directly observable but *hidden* random variables. Of course, in the graph of a Bayesian network also the hidden variables should be represented by nodes. To each discrete propositional random variable with D possible values we should attach a parent that represents the $(D-1)$-dimensional continuous probability vector $(\pi_1, \ldots, \pi_{D-1})$ (one dimension is lost because the probabilities add up to one). The (second order) probability distribution of the vector is a Dirichlet distribution. It is specified by the numbers contained in the weight table of the node. The relationship between the hidden nodes and their children, though, is redundant: the conditional probabilities of the discrete states of the child nodes (propositional variables) are equal to the values of the hidden variables: $P(x_i|\pi_i) = \pi_i$. Because every propositional variable has a twin hidden variable and because the relationship between the hidden variable and the propositional variable is redundant the hidden variables are not drawn in the graph of a weighted Bayesian network. Drawing the twin nodes would unneccessarily complicate graph.

The space defined by the hidden probability variables is a subspace of all possible Bayesian belief networks for the domain of propositional variables under consideration. It is a subspace, and not the full space, because it respects the conditional independencies in the network. The Dirichlet distributions are treated as Bayesian *posterior* distributions. The qualitative independence/dependence structure (that is visually represented in the graph), is taken for sure. The numerical specifications of the underlying (conditional) probabilities are taken as uncertain quantities that are not known for sure. We investigate the propagation of posterior densities in Bayesian networks with given structure but uncertain parameters. Methods how to learn such structures from prior knowledge and data were described by Heckerman, Geiger and Chickering [17].

If all the weight tables are obtained from frequency counts in one big *complete* database with no missing cases, then the joint distribution over the domain of all variables and all conditional distributions are Dirichlet [44]. The propagation of probabilities

can be performed by one of the usual methods of propagating point probabilities; the weights of the Dirichlet distributions are obtained by simply multiplying the resulting probabilities by the total sample size, i.e., by taking expected frequencies. Completeness, though, is a strong assumption. It is often violated in practical applications. A database may have been combined from several sources, it may contain objective data and subjective expertize etc. We next introduce concepts that are helpful to find an approximate solution for incomplete data.

3. Natural children and natural neighbors

In a complete contingency table the marginals along any of its dimensions are equal to the sums of the corresponding cell counts. A similar relationship may or may not hold for the weights in a weighted network. The case of a perfectly additive relationship between the cells and the marginals allows an easy mathematical treatment of the member distribution. With additive weights the member distribution is also a beta or Dirichlet distribution and we thus stay within the same family of probability distributions. What shall we do when the cell counts and the marginals do not add up? In inferential statistics this case occurs when some of the data are missing [29]. The treatment of the missing data can be related to the complete case by calculating the weighted averages of complete solutions, where averaging is performed over the space of the missing data. This results in probability mixtures [26]. For computational purposes the solutions are too complicated. Incomplete data are usually analyzed by *expectation maximization* (EM) algorithms [10]. EM is an iterative procedure providing maximum likelihood estimates in the presence of missing data. The precision (variance) of the estimates can be approximated [29]. EM has several disadvantages. In large networks the iterative algorithm is slow. Furthermore, in large networks there are multiple local maxima and it is difficult to find out that a global maximum has been found [11].

We combine local noniterative estimation with Gibbs sampling. The probabilities at each node in the network depend only upon the states of the neighbor nodes (Markov blanket). The estimation of conditional probabilities in a Markov blanket with missing data can be done without iteration. We use the δ method to estimate means and variances of the conditional probabilities. We next give a definition of the additive relationships between cell weights and marginals. We introduce the concepts of a natural child, natural parents, and natural neighbors.

Definition 3 (*Natural child*). Let Y have the parents $\mathrm{pa}(Y) = \{X_1, \ldots, X_m\}$ and let $X_k \in \mathrm{pa}(Y)$ have the parents $\mathrm{pa}(X_k) = \{U_1, \ldots, U_m\}$. Y is a natural child of X_k if the marginal weights of X_k in the table $\langle Y, \mathrm{pa}(Y)\rangle$ along $\{Y, \mathrm{pa}(Y) \setminus X_k\}$ and in the table $\langle X_k, \mathrm{pa}(X_k)\rangle$ along $\mathrm{pa}(X_k)$ are identical:

$$\langle Y, \mathrm{pa}(Y)\rangle^{\downarrow\{Y,\mathrm{pa}(Y)\setminus X_k\}} = \langle X_k, \mathrm{pa}(X_k)\rangle^{\downarrow\mathrm{pa}(X_k)}. \tag{5}$$

If X has no parents, then Y is a natural child of X if $\langle Y, X\rangle^{\downarrow Y} = \langle X\rangle$.

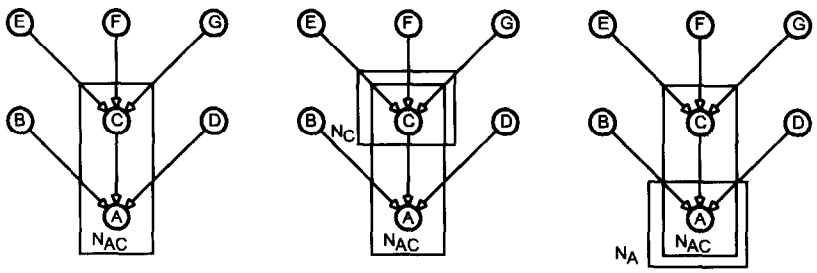

Fig. 3. The plates drawn around A and C indicating that A is a natural child of C (left), more is known about C (middle), and more is known about A.

Visually, we represent a parent together with its natural child in a *plate*. A plate is a rectangle drawn around a set of nodes with a repetition number N written in its left lower corner. Plates were introduced by Buntine [5] (who gives credit to Spiegelhalter). Plates indicate a data set of the same kind. The set of nodes is instantiated simultaneously or repeatedly by N observations. A plate shows a complete set of N data. Often N is a sample size, or it is the total sum of weights in a weight table. In the left panel of Fig. 3 a plate with the repetition number N_{AC} is drawn around the nodes A and C. A node Y is a nonnatural child of the parent X_k if the marginal weights of X_k within the table containing X_k and the parents of X_k are (i) larger or (ii) smaller than those within the table containing Y and the parents of Y. We say that in the first case, we have more information about the parents, and that in the second case, we have more information about the child. This is the case if, for example, in a sample of observations some cases are missing or if additional cases are available. In the middle panel of Fig. 3 an extra plate is drawn around node C to indicate N_C additional observed cases on C only. More is known about C than about A. In the right panel an extra plate is drawn around A. More is known about A than about C.

If a parent has two or more natural children, then their repetition numbers must be identical. It follows that the parent and the natural children can be put into a single plate. We call a parent natural if all its children—taking also the parents of these children into account—are natural. If a parent has two or more children that are natural in respect to *all* their parents, then their repetition numbers must be identical. It follows that the parent and the children can be put into a single plate. In the left panel of Fig. 4 node E has the natural children A, B, and C. The three children are also natural in respect to their parents D and F. A, B, C, D, E, and F can thus be put into a single plate. The parents of a node, its children and the parents of these children are called the neighbors or the Markov blanket of the node. The probability distribution of the states of the node depends on the state of nodes in the Markov blanket and on these only. The condition in which the Markov blanket builds a plate is important:

Definition 4 (*Natural neighbors*). A node has natural neighbors (Markov blanket) if all its children are natural in respect to all their parents.

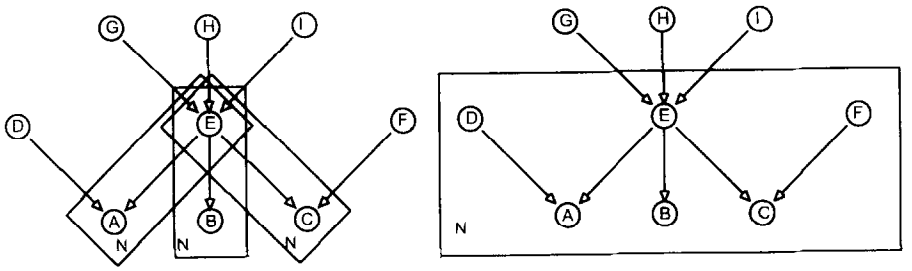

Fig. 4. The plates drawn around E and its natural children A, B, and C can be represented by a single plate drawn around A, B, C, D, E, and F.

4. Member distribution

To propagate the second order pdf in a weighted graph we use stochastic simulation. It turns out that in stochastic simulation Bayes' theorem, or what is called "generalized Bayes' theorem" [34], plays a central role. We have investigated second order pdfs for a Bayes' parameter in [21,24,26]. We will make use of the previous results obtained for this Bayes' parameter and extend them to the slightly more general situation corresponding to the generalized Bayes' theorem.

4.1. Member parameter

Consider a disease A that can be present or absent, and a symptom B that also can be present or absent. Assume the marginal or base rate probability of A being present is α, and the conditional symptom probability of the symptom B given the disease is present is β_1, and the conditional symptom probability of the symptom given the disease is not present is β_2. If we observe a patient showing symptom B the probability that the patient suffers from disease A is given by Bayes' theorem

$$\mu = \frac{\alpha\beta_1}{\alpha\beta_1 + (1-\alpha)\beta_2}. \tag{6}$$

We call μ the *member parameter*. It is just Bayes' theorem in a parametric form. If the probabilities α, β_1, and β_2 are imprecise and the imprecision is expressed by the three beta distributions

$$\alpha \sim Be(a_1, a_2), \qquad \beta_1 \sim Be(b_{11}, b_{12}), \qquad \beta_2 \sim Be(b_{21}, b_{22}), \tag{7}$$

we want to infer the distribution of the member parameter μ given B and the weights of evidence a_1, a_2, b_{11}, b_{12}, b_{21}, and b_{22}. We call this distribution the *member distribution*. The member distribution is a second order pdf defined for Bayes' formula. It tells us how precise our posterior probability is.

4.2. Beta and mixed beta member distributions

Let A and B be two binary random variables. We have shown the following theorem [23,26]:

Theorem 5 (Natural child).

(1) *If* α, β_1, β_2, *and* μ *are the probability parameters underlying the propositional variables* A, $B|A$, $B|\neg A$, *and* $A|b$, *respectively, and if the second order pdfs of the first three parameters are* $\alpha \sim Be(a_1, a_2)$, $\beta_1 \sim Be(b_{11}, b_{12})$, *and* $\beta_2 \sim Be(b_{21}, b_{22})$, *and*

(2) *if the three parameters* α, β_1, *and* β_2 *are independent, and*

(3) *if* B *is a natural child of* A, *i.e.,* $a_1 = b_{11} + b_{12}$ *and* $a_2 = b_{21} + b_{22}$,

then the pdf of the Bayes' parameter μ *is given by* $\mu \sim Be(b_{11}, b_{21})$.

That is if B is a natural child of A then the member distribution is a beta distribution and its parameters can directly be read off from B's weight table. Note that the weight table of A is not needed.

The theorem directly generalizes to natural parents, the notation, however, becomes more cumbersom. We assume that the nodes are binary and use upper case characters "A", "B", "$\neg A$", "$\neg B$", etc. to denote node variables. We use lower case characters "a", "$\neg a$", "b", "$\neg b$", etc. to denote instantiated nodes. We use conjunctions like "AB", "$A\neg B$", etc. to locate cells in the weight tables. We finally denote the corresponding cell weights by "$\langle A\rangle$", "$\langle \neg A\rangle$", "$\langle a\rangle$", "$\langle \neg a\rangle$". Using this notation Theorem 5 reads: If $[A] = Be(\langle A\rangle, \langle \neg A\rangle)$, $[B|A] = Be(\langle AB\rangle, \langle A\neg B\rangle)$, and $[B|\neg A] = Be(\langle \neg AB\rangle, \langle \neg A\neg B\rangle)$, then $[A|b] = Be(\langle Ab\rangle, \langle \neg Ab\rangle)$. In the following theorem $\langle Ab_1 f(b_1)\rangle$ denotes the weight in the cell indexed by the variable node value A, the instantiated node value b_1 and the node values of the parents of b_1, etc.

Theorem 6 (Natural neighbors). *If* A *has natural neighbors, if its children are* $B_1, \ldots, B_m$, *and if the underlying probability parameters are independent and Dirichlet distributed, then*

$$[A|b_1, \ldots, b_m] = Be(\langle Ab_1 f(b_1) \cdots Ab_m f(b_m)\rangle, \langle \neg Ab_1 f(b_1) \cdots \neg Ab_m f(b_m)\rangle). \tag{8}$$

In the case in which the weights of evidence are not additive, the member distribution is a *mixture of beta distributions* [26]. The mixing weights follow a Polya–Eggenberger probability distribution. Two cases must be distinguished: (i) the case in which more is known about the marginals than about the conditional probabilities, and (ii) the case in which less is known about the marginals than about the conditional probabilities. In the first case we may know more about the presence (for example) of the disease than about the conditional symptom probabilities, that is $a_1 > b_{11} + b_{12}$. The difference $D = a_1 - (b_{11} + b_{12})$ is positive. D may be conceived as the number of *missing data*, i.e., as the number of cases for which we know the disease to be present but do not know the symptom. In the second case we may know less about the presence (for example) of the disease than about the conditional symptom probabilities, that is $a_1 < b_{11} + b_{12}$. In

inferential statistics this situation may arise if in a contingency table the sampling of the marginals is random for a_1 cases but fixed by the experimenter for the remaining $D = b_{11} + b_{12} - a_1$ cases. For inferences about the marginal probabilities only the a_1 cases can be used in the statistical analysis. For inferences about the conditional symptom probabilities all $b_{11} + b_{12}$ cases can be used. In the first and in the second case the missing data can be predicted probabilistically. In statistics the probability distribution of a future sample given an observed one is called a *predictive distribution* [1,2]. It may be shown that the predictive distribution in both our cases is a Polya–Eggenberger distribution [19]. The member distribution turns out to be a probability mixture of beta distributions where the mixing weights are Polya–Eggenberger probabilities [26]:

Theorem 7 (Nonnatural neighbors).

(1) *If* α, β_1, β_2, *and* μ *are the probability parameters underlying the propositional variables* A, $B|A$, $B|\neg A$, *and* $A|b$, *respectively, and if the second order pdfs of the first three parameters are* $\alpha \sim Be(a_1, a_2)$, $\beta_1 \sim Be(b_{11}, b_{12})$, *and* $\beta_2 \sim Be(b_{21}, b_{22})$, *and*

(2) *if* α, β_1, *and* β_2 *are independent, then the pdf of the Bayes' parameter* μ *is given by*

$$\mu = \sum_{d_1=\min(d_1)}^{\max(d_1)} \sum_{d_2=\min(d_2)}^{\max(d_2)} PE(D_1, b_{11} + s_1 d_1, b_{12} + D_1 - s_1 d_1) \times PE(D_2, b_{21} + s_2 d_2, b_{22} + D_2 - s_2 d_2) \times Be(b_{11} + s_1 d_1, b_{21} + s_2 d_2). \tag{9}$$

Let $B_i = b_{i1} + b_{i2}$ *and* $D_i = |a_i - B_i|$. *Then the range of* d_1 *and* d_2 *is constrained by*

$$\min(d_i) = \begin{cases} 0, & \text{if } a_i \geqslant B_i, \\ \max(0, b_{i1} + D_i - B_i), & \text{if } a_i < B_i, \end{cases}$$

$$\max(d_i) = \begin{cases} D_i, & \text{if } a_i > B_i, \\ \min(b_{i1}, D_i), & \text{if } a_i \leqslant B_i, \end{cases}$$

$$s_i = \begin{cases} 1, & \text{if } a_i > B_i, \\ 0, & \text{if } a_i = B_i, \\ -1, & \text{if } a_i < B_i. \end{cases}$$

The Polya–Eggenberger distribution is defined as follows:

Definition 8 (*Polya–Eggenberger distribution*). Let Y be a discrete random variable. If its distribution is given by

$$P(y|n, g, h) = \binom{n}{y} \frac{g(g+1s)(g+2s)\cdots(g+(y-1)s)}{(g+h)(g+h+1s)\cdots(g+h+(n-1)s)} \times h(h+1s)(h+2s)\cdots(h+(n-y-1)s), \tag{10}$$

it is a Polya–Eggenberger distribution and we write $Y \sim PE(n, g, h)$.

For $s = 1$ the Polya–Eggenberger distribution is equivalent to a beta-binomial distribution (see, e.g., [2]), for $s = 0$ to a binomial, and for $s = -1$ to a hypergeometric distribution. For more details we refer to [26]. In a more general structure the constaints may be obtained by linear programming. In a large network containing many missing observations, though, the calculation of the exact beta mixtures becomes cumbersom. Below, we employ an approximation based on the δ method.

5. Stochastic simulation

The use of stochastic simulation in Bayesian networks was proposed by Pearl [34]. Hrycej [18] has shown that the stochastic simulation in a Bayesian network is a special case of Gibbs sampling. It has extensively been employed to Bayesian networks [13, 36, 39, 41].

At the start each instantiated node is clamped to its constant value and each non-instantiated node is set to an arbitrary value. Then, iteratively, the following steps are performed:

(1) Select a nonclamped node, e.g., in the alphabetical order of the node names.

(2) Compute the means $m(\mu)$ and the variances $\mathrm{var}(\mu)$ of the member distributions for each of the values of this node, given the current values of the neighbor nodes. In a Bayesian network containing point probabilities, the local probabilities are obtained by the *generalized Bayes' theorem*

$$P(X|\mathrm{rest}(x)) = KP(X|\mathrm{pa}(x)) \prod_{j=1}^{m} P(y_j|f_j(x)). \tag{11}$$

The upper case letters refer to random nodes, the lower case letters to instantiated nodes. K is a normalizing factor, $\mathrm{pa}(x)$ represents the parents of X, y_j the children of X, $f_j(x)$ the parents of y_j, and $\mathrm{rest}(x)$ represents all variables except X. For the second order pdfs we use a completely analogous formula to determine the means of the distributions at each node. The variances are calculated by the δ method that is described below.

(3) Determine a new value for the node by selecting a random number. The probability for each value is equal to the mean of the member distribution of this value.

(4) Build the sum of the means and variances of the member distribution for each value of the current node.

The sums are averaged, and finally beta distributions $Be(p, q)$ are fitted to the means $m(\mu)$ and variances $\mathrm{var}(\mu)$; p and q are obtained from:

$$N = \frac{m(\mu)(1 - m(\mu))}{\mathrm{var}(\mu)} - 1 \quad \text{and} \quad p = m(\mu)N, \quad q = N - p. \tag{12}$$

It is interesting to note the central role Bayes' theorem plays in stochastic simulation. We turn to the determination of the variances in step (2).

5.1. The δ method

Gibbs sampling allows the propagation of first order probabilities in Bayesian networks with incomplete data. Principally, it is possible to employ a Gibbs sampler also to obtain second order densities. At the hidden nodes we would have to generate random probabilities according to a distribution law, a beta distribution, for example. The random probabilities, in turn, would determine the state probabilities at the associated child nodes. This would lead to a computationally very expensive two-level sampling process. We will use a shortcut instead. We directly employ variance estimates obtained at each node to calculate the precision of the second order distributions. In this section we describe the method of how to obtain the variances of the second order distribution at each node given its natural or nonnatural neighbors. Nonnatural neighbors correspond to incomplete data.

The member parameter μ as introduced in Eq. (6) is a nonlinear function g of the variables α, β_1, and β_2. For each of these parameters the pdf is known. Can we derive the mean and the variance of the member parameter? For linear functions g of a random variable X we have $\mathrm{E}[g(X)] = g(\mathrm{E}[X])$. This is not true if g is not linear. In many cases, though, the mean and the variance of $g(X)$ can be approximated by the δ method [4, 32]:

Definition 9 (δ *rule*). Let $(X_1, \ldots, X_n)$ be independent random variables with means $(E_1, \ldots, E_n)$ and variances $(V_1, \ldots, V_n)$. If $f(X_1, \ldots, X_n)$ is a function of the variables that can be partially differentiated at $f(E_1, \ldots, E_n)$ with respect to $E_1, \ldots, E_n$, then $f(X_1, \ldots, X_n)$ is asymptotically normal with mean

$$E_{f(X_1,\ldots,X_n)} = f(E_1, \ldots, E_n) \tag{13}$$

and variance

$$\mathrm{var}_{f(X_1,\ldots,X_n)} = \sum_{i=1}^{n} V_i \left(\frac{\partial f(E_1, \ldots, E_n)}{\partial E_i} \right)^2. \tag{14}$$

The mean of each variable X in a Bayesian network conditioned on the state of all other variables can be approximated by applying the formula for the expectation (13) upon the generalized Bayes' theorem (11). We only need to rewrite the generalized Bayes' theorem in the form of expectations:

$$\mathrm{E}[X|\mathrm{rest}(x)] = K\mathrm{E}[X|\mathrm{pa}(x)] \prod_{j=1}^{m} \mathrm{E}[y_j|f_j(x)]. \tag{15}$$

The expectations are estimates of the underlying probability parameters. The variance of the member distribution is obtained from Eq. (14) as a sum of variance components:

$$\mathrm{var}[X|\mathrm{rest}(x)] = \mathrm{var}[X|\mathrm{pa}(x)] + \sum_{j} \mathrm{var}[y_j|f_j(x)], \tag{16}$$

i.e., from the sum of the variance components of the jointly instantiated parents and the sums of the variance components of each of the children under proper consideration of the instantiated values of their parents. The variance component due to the parents is obtained by applying the δ rule to one variable only

$$\begin{aligned}&\mathrm{var}[X|\mathrm{pa}(x)]\\&\approx V[X|\mathrm{pa}(x)]\left(\frac{\partial f(\mathrm{E}[X|\mathrm{pa}(x)],\mathrm{E}[y_1|f_1(x)],\ldots,\mathrm{E}[y_m|f_m(x)])}{\partial \mathrm{E}[X|\mathrm{pa}(x)]}\right)^2\end{aligned} \tag{17}$$

and similarily the variance components due to the children are obtained from

$$\begin{aligned}&\mathrm{var}[y_j|f_j(x)]\\&\approx V[y_j|f_j(x)]\left(\frac{\partial f(\mathrm{E}[X|\mathrm{pa}(x)],\mathrm{E}[y_1|f_1(x)],\ldots,\mathrm{E}[y_m|f_m(x)])}{\partial \mathrm{E}[y_j|f_j(x)]}\right)^2.\end{aligned} \tag{18}$$

Building partial derivatives and collecting terms finally leads to the following expressions

$$\begin{aligned}&\mathrm{var}[X|\mathrm{rest}(x)]\\&=\left(\frac{A\prod_{j=1}^{m}B_j}{D^2}\right)^2\left[\frac{V[X|\mathrm{pa}(x)]}{A^2}+\sum_{j=1}^{n}\sum_{k=j}^{m}\frac{V[y_{jk}|f_j(x)]}{(\mathrm{E}[y_{jk}|f_j(x)])^2}\right]\end{aligned} \tag{19}$$

where n is the number of children and m the number of possible values of X, and

$$A=\mathrm{E}[X|\mathrm{pa}(x)](1-\mathrm{E}[X|\mathrm{pa}(x)]), \tag{20}$$

$$B_j=\mathrm{E}[y_j|f_j(x)](1-\mathrm{E}[y_j|f_j(x)]), \tag{21}$$

$$\begin{aligned}D=&\mathrm{E}[X|\mathrm{pa}(x)]\prod_{j=1}^{n}\mathrm{E}[y_j|f_j(x)]\\&+(1-\mathrm{E}[X|\mathrm{pa}(x)])\prod_{j=1}^{n}(1-\mathrm{E}[y_j|f_j(x)]).\end{aligned} \tag{22}$$

The $\mathrm{E}[\cdot|\cdot]$ terms are easily obtained by dividing cell weights by marginals, e.g., $b_{11}/(b_{11}+b_{12})$. The $V[\cdot|\cdot]$ are variances of beta distributions, e.g., $b_{11}b_{12}/[(b_{11}+b_{12})^2(b_{11}+b_{12}+1)]$. The square bracket notation is used to avoid too many greek symbols and to stay as close as possible to the symbols used in the literature for point probabilities. $\mathrm{E}[X|\mathrm{rest}(x)]$ stands for the member parameter μ at node X given all other nodes except X.

5.2. Program

The propagation is performed by a program written in C. The navigation through the graph is supported by the Raima Database Manager [35]. This database is network oriented and supports the definition and processing of directed graphs by pointers. In that

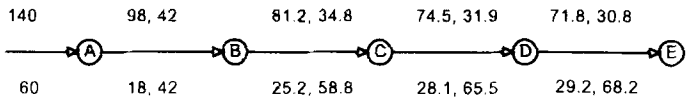

Fig. 5. Simple chain with base rate weights $\langle a \rangle = 140$, $\langle \neg a \rangle = 60$, and conditional weights $\langle b|a \rangle = 98$, $\langle \neg b|a \rangle = 42$, $\langle b|\neg a \rangle = 18$, and $\langle \neg a|\neg a \rangle = 42$, etc.

respect it is different from relational databases. For the numerical examples stochastic simulation was performed with 1000 iterations.

Stochastic simulation may not work well when the estimated probabilities in the network are close to zero or one [7]. We intend to replace the δ method by a method that is closer to beta mixtures. Especially, the mean and the variances at each node given all its neigbhors may be obtained directly from the mixtures. But even the calculation of the means and variances of the mixtures requires the determination of many Γ-terms in the Polya–Eggenberger weights and the direct programming of the formulas does not look promising.

6. Examples

6.1. A simple chain

Consider the chain in Fig. 5. Denote the full graph consisting of all five nodes by G_5. Denote any subgraph consisting of 1, 2, 3, or 4 nodes by G_1, G_2, G_3, and G_4, respectively. Assume 200 cases were observed under natural sampling conditions, 140 $A = a$ cases, 60 $A = \neg a$ cases etc. At each node the conditional probabilities are 0.7/0.3 and 0.3/0.7, respectively. Without any nodes instantiated the probability of a is distributed as $[a|G_5] = Be(79, 32)$ with mean 0.71. Note that the distribution is much flatter than the marginal distribution of a without B, C, D, and E being included in the system, which of course is $[a|G_1] = Be(140, 60)$. When $B = b$ is clamped we obtain $[a|b, G_5] = Be(98, 18)$ and the distribution remains the same when, additionally, C, D, and E are clamped. If we let the *reference system grow* in which inferences are made we obtain $[a|b, G_2] = Be(98, 18)$ which due to the Markov property is identical to $[a|b, G_5]$.

There is very slow learning at the beginning as we instantiated bottom-up: $[a|e, G_5] = Be(81, 31)$, $[a|d, G_5] = Be(84, 30)$, $[a|c, G_5] = Be(89, 27)$, and finally $[a|b, G_5] = Be(98, 18)$ again. The situation of the top-down inferences is different: The marginal of e in G_5 is $[e|G_5] = Be(61, 58)$, and then we obtain $[e|a, G_5] = Be(61, 58)$, $[e|b, G_5] = Be(64, 55)$, $[e|c, G_5] = Be(68, 48)$, $[e|d, G_5] = Be(72, 31)$. Of course, $[e|d, G_5]$ is equivalent to the distribution directly provided to the system.

The "long distance" forward inference from A to E is noninformative in respect to the first order probability which is $61/(61 + 58) = 0.51$, a value that is practically equal to 0.5. The value is equal to the base rate of e. The "long distance" backward inference from E to A is also noninformative; it results in the probability $81/(81 + 31) = 0.72$

which is practically identical to the base rate of a. Prediction (forward inference) and diagnosis (backward inference) at worst results in the base rate probabilities. However, in both cases the precision in the 5-node system containing the nodes A, B, C, D, and E, is much worse than in the 1-node system containing only A or E, respectively.

We recognize the Markov property in the Bayesian network: The distributions at any node of the chain depends only upon its parent and its child—the grandparent, or any other predecessors, and grandchild, or any other decendent, do not provide additional information. The distributions $[a|b,G_5]$, $[a|b,G_4]$, $[a|b,G_3]$, $[a|b,G_2]$ are equal because of the independence structure in the chain.

The weight tables of the chain have natural children only. The local member distributions in the stochastic simulation process may be determined approximately by the δ method or exactly by Theorem 5 or 6 for natural sampling. There were practically no differences between both methods.

If we change the marginal weights of a from 140/60 to 35/15 the distribution of a given b changes from $Be(98,18)$ to $Be(47,9)$. That is, if we in the present example divide the marginal weights by 4 we have to divide the conditional ones by about 2.

6.2. *A simple triangle*

Consider the network in Fig. 6. Denote the subgraph consisting of A and its marginal weights only by G_1, denote the subgraph consisting of A and B by G_2 and the graph consisting of all three nodes by G_3. The marginal distribution of A is different in all three structures: $[a|G_1] = Be(140,60)$, $[a|G_2] = Be(77,13)$, and $[a|G_3] = Be(37,16)$. Accordingly, the conditional distributions of A given B in the two graphs G_2 and G_3 are $[a|b,G_2] = Be(98,16)$ and $[a|b,G_3] = Be(49,9)$. The distribution of $[a|b,c,G_3]$ is $Be(69,13)$. It is important to note that the distributions are not invariant with respect to the *reference system* in which they are determined. Generally we observe that the larger the system the smaller the resulting weights of evidence. The limiting condition occurs when the inferences are independent of the structural extension. Then their first and second order distributions just remain the same. This happened in the chain but not in the triangle structure. In highly connected structures, such as large cliques, e.g., the loss in imprecision by system extensions will be larger than in systems in which many variables are independent.

If we reduce the marginal weights of A from 140/60 to 35/15 the marginal distribution of A in G_3 becomes $Be(22,10)$. The distribution of $[a|b,G_3]$ is $Be(31.5,6)$, and the distribution of $[a|b,c,G_3]$ is $Be(39,7)$. The precision of the inferences about A—or α, to be more precise—has decreased appreciably.

We have stated [22] that the imprecision of probabilistic inferences—under otherwise comparable circumstances—increases as the systems in which they are embedded get more complex. We should therefore strive to keep the inferential systems simple. The trade-off between complexity and accuracy has recently been studied in Bayesian networks by the minimum description length criterion [27]. From our viewpoint the trade-off may be illustrated by an example that at first looks terribly counterintuitive. Can more data make us more uncertain about our inferences? Consider the following problem:

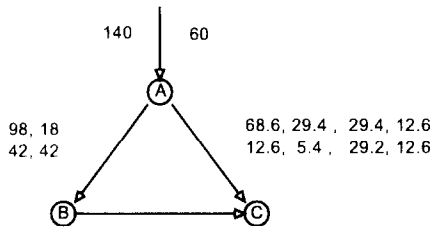

Fig. 6. Simple triangle with base rate weights $\langle a\rangle = 140$, $\langle \neg a\rangle = 60$ and conditional weights $\langle b|a\rangle = 98$, $\langle b|\neg a\rangle = 18$, etc. and $\langle \neg b|\neg a\rangle = 42$, and weights $\langle c|a, b\rangle = 68.6$, $\langle c|a, \neg b\rangle = 29.4$, $\langle c|\neg a, b\rangle = 29.4$, $\langle c|\neg a, \neg b\rangle = 12.6$, etc.

7. Bad news

Imagine that you are a doctor on a remote island. One of the residents is brought in to see you. After a careful investigation you suspect the patient is suffering from disease A. However, a definite diagnosis can only be made after laboratory blood tests and you do not have the expensive technical equipment.

Since you have arrived on the island, you have investigated 40 similar cases. Later (after careful laboratory checking) you found out that 30 cases actually had A, and 10 did not.

(1) What is the probability that your patient is suffering from A?
The probability is ....

(2) As your experience is limited to 40 cases only, your estimate cannot be absolutely precise. Give a confidence interval for your estimate!
I am 90% sure that the true value of the probability lies between ... and ....

For some time you thought that the diagnostic sign B might be relevant for the diagnosis of A. You found out that in the 30 cases suffering from A, only 9 showed the symptom and 21 did not. Of the 10 cases not suffering from A, 3 showed symptom B and 7 did not.

You realize that your patient is showing the diagnostic sign B.

(1) What is the probability that your patient is suffering from A given that the patient shows B?
The probability is ....

(2) Give a confidence interval for your estimate!
I am 90% sure that the true value lies between ... and ....

Of course, both the first and the second point probability estimates should be 0.75. Your second estimate, though, should be be less precise than your first one, and the second interval should therefore be wider than the first one. Assuming the uncertainty is expressed by beta distributions the first distribution is $Be(30, 10)$. The mean of the distribution is 0.75 and the 90% confidence interval is $(0.64, 0.86)$. The posterior is $Be(9, 3)$ with the same mean of course and the confidence interval $(0.56, 0.94)$. The second distribution is flatter than the first one. Its variance is larger.

In the example the observed data is nondiagnostic. The probabilistic conditioning on the nondiagnostic data seems to make things worse—which is counterintuitive. We

assume that additional new information can never be bad for the quality of our inferences. We need a principle that protects us from considering irrelevant data. Nondiagnostic data is not necessarily neutral to our arguments. It increases the imprecision of an argument. Inferences become more noisy.

7.1. The resolution of the paradox

In the first part the cover story describes a reference system G_1 consisting of only one node and its associated frequencies, i.e., the disease A together with the counts $a_1 = 30$ and $a_2 = 10$. If we assume the improper prior $Be(0,0)$, then $[a|G_1]$ is distributed as

$$[a|G_1] = [a|a_1, a_2] = Be(a_1, a_2) = Be(30, 10).$$

In the second part the cover story introduces a second node, the symptom B, together with the associated counts $b_{11} = 9$, $b_{12} = 21$, $b_{21} = 3$, and $b_{22} = 7$. The resulting reference system G_2 now has two nodes. Let us investigate the distribution of α in G_2 prior to the information that our case shows symptom B

$$[a|G_2] = [a|a_1, a_2; b_{11}, b_{12}, b_{21}, b_{22}].$$

The first order marginal probability of B may be estimated by the compound probability estimates

$$P(B) = P(A)P(B|A) + P(\neg A)P(B|\neg A) = 0.3.$$

Let's assume for a moment that this value would be known exactly. If with probability 0.3 we observe B then with this probability we will observe the member distribution $Be(9,3)$ which has the mean $9/12$ and the variance 0.01442. Similarily, with the probability $P(\neg B) = 0.7$ we will observe the member distribution $Be(21,7)$ which has the mean $21/28$ and the variance 0.00647. The expected variance is therefore $0.3 \times 0.01442 + 0.7 \times 0.00647 = 0.00885$. If we fit a beta distribution we obtain

$$[a|G_2] = Be(14.13, 6.06).$$

This is the expected posterior distribution in G_2 resulting from a preposterior analysis. As $P(B)$ is not known exactly there will actually be some more variability in the distribution. Note that the mean of this distribution is again 0.7, i.e., it is the mean of the marginal distribution of A in G_1. Its precision though corresponds to only $14.13 + 6.06 = 20.19$ cases. This is only half of 40, the total number of cases effective in G_1. Observing symptom B or $\neg B$ in G_2 leads to the member distributions

$$[a|b, G_2] = Be(9,3) \quad \text{and} \quad [a|\neg b, G_2] = Be(2,7),$$

respectively. These distributions are the posterior distributions in the system G_2 after having clamped symptom b and $\neg b$, respectively. The distribution of $[a|G_1]$ and $[a|G_2]$ should not be interpreted as prior and posterior distributions because they do not belong to the same reference system.

What makes this problem counterintuitive? Intuitively we discard the additional information as soon as we have realized that it does not change the probability estimate.

Thus, the situation reduces to G_1 and the accuracy does not change, of course. We do not conditionalize on irrelevant information. We perform an elementary pruning process. When we discuss the problem we tend to compare distributions in G_1 and G_2. We compare the information about $[a|G_1]$ in G_1 with the information about $[a|G_2]$ in G_2. This may be misleading. The conditioning is not based on the same information. The prior and the posterior do not belong to the same system. Intuitively additional information is associated with an improved state of knowledge. We believe in the principle of "monotone information" in the sense that more information is equivalent to better knowledge and less uncertainty. The example shows that additional data can decrease the precision.

We have to protect inferential systems from variables with low diagnosticity. Such variables have low positive impact upon the first order probabilities but may have considerable negative impact upon the precision of the system's inferences. There is a trade-off between the improvement in first order probabilities and the loss in second order precision.

References

[1] J. Aitchison, *The Statistical Analysis of Compositional Data* (Chapman and Hall, London, 1986).

[2] J. Aitchison and I.R. Dunsmore, *Statistical Prediction Analysis* (Cambridge University Press, Cambridge, 1975).

[3] J.M. Bernardo and A.F.M. Smith, *Bayesian Theory* (Wiley, Chichester, 1994).

[4] Y.M. Bishop, S.E. Fienberg and P.W. Holland, *Discrete Multivariate Analysis: Theory and Practice* (MIT Press, Cambridge, MA, 1975).

[5] W.L. Buntine, Operations for learning with graphical models, *J. Artif. Intell. Res.* **2** (1994) 159–225.

[6] W. Buntine, A guide to the literature on learning probabilistic networks from data, *IEEE Trans. Knowledge Data Eng.* (to appear).

[7] R.M. Chavez and G.F. Cooper, A randomized approximation algorithm for probabilistic inference on Bayesian belief networks, *Networks* **20** (1990) 661–685.

[8] P. Che, R.E. Neapolitan, J. Kenevan and M. Evens, An implementation of a method for computing the uncertainty in inferred probabilities in belief networks, in: D. Heckerman and A. Mamdami, eds., *Uncertainty in Artificial Intelligence* (Kaufmann, San Mateo, CA, 1993) 292–300.

[9] G. Coletti, A. Gilio and R. Scozzafava, Conditional events with vague information in expert systems, in: B. Bouchon-Meunier and R.R. Yager, eds., *Uncertainty in Knowledge Bases* (Springer, Berlin, 1991) 106–114.

[10] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *J. Roy. Stat. Soc. Ser. B* **39** (1977) 1–38.

[11] C. Edwards, *Introduction to Graphical Modelling* (Springer, New York, 1995).

[12] K.W. Fertig and J.S. Breese, Interval influence diagrams, in: H. Henrion, R.D. Shachter, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **5** (North-Holland, Amsterdam, 1990) 149–161.

[13] A. Gammerman, Z. Luo, C.G.G. Aitken and M.J. Brewer, Exact and approximate algorithms and their implementations in mixed graphical models, in: A. Gammerman, ed., *Probabilistic Reasoning and Bayesian Belief Networks* (Alfred Waller, Henley-on-Thames, 1995) 33–53.

[14] A.E. Gelfand and A.F.M. Smith, Sampling-based approaches to calculating marginal densities, *J. Am. Stat. Assoc.* **85** (1990) 398–409.

[15] P. Hájek, T. Havránek and R. Jiroušek, *Uncertain Information Processing in Expert Systems* (CRC Press, Boca Raton, FL, 1992).

[16] D. Heckerman, A tutorial on learning Bayesian networks, Microsoft Research, Advanced Technology Division, Microsoft Corporation, Redmond, WA (1995) (heckerma@microsoft.com).

[17] D. Heckerman, D. Geiger and D. Chickering, Learning Bayesian networks: the combination of knowledge and statistical data, *Mach. Learn.* **20** (1995) 197–243.
[18] T. Hrycej, Gibbs sampling in Bayesian networks, *Artif. Intell.* **46** (1990) 351–363.
[19] N.L. Johnson and S. Kotz, *Urn Models and their Application* (Wiley, New York, 1977).
[20] J.M. Keynes, *A Treatise on Probability* (MacMillan, London, 1921).
[21] G.D. Kleiter, Bayesian diagnosis in artificial intelligence, *Artif. Intell.* **54** (1992) 1–32.
[22] G.D. Kleiter, Properties of probabilistic imprecision, in: B. Buchon-Meunier, L. Valverde and R.R. Yager, eds., *Uncertainty in Intelligent Systems* (North-Holland, Amsterdam, 1993) 155–170.
[23] G.D. Kleiter, Natural sampling: rationality without base rates, in: G.H. Fischer and D. Laming, eds., *Contributions to Mathematical Psychology, Psychometrics, and Methodology* (Springer, New York, 1994) 375–388.
[24] G.D. Kleiter, The precision of Bayesian classification: the multivariate normal case, *Int. J. General Syst.* **22** (1994) 139–157.
[25] G.D. Kleiter, Expressing imprecision in probabilistic knowledge, *J. Italian Stat. Soc.* **2** (1994) 213–232.
[26] G.D. Kleiter and M. Kardinal, A Bayesian approach to imprecision in belief nets, in: V. Mammitzsch and H. Schneeweiß, eds., *Symposia Gaussiana, Proceedings of the 2nd Gauss Symposium, Conference B: Statistical Sciences* (De Gruyter, Berlin, 1995) 91–105.
[27] W. Lam and F. Bacchus, Learning Bayesian belief networks: an approach based on the MDL principle, *Comput. Intell.* **10** (1994) 269–293.
[28] S.L. Lauritzen and N. Wermuth, Graphical models for associations between variables, some of which are qualitative and some quantitative, *Ann. Stat.* **17** (1989) 31–57.
[29] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data* (Wiley, New York, 1987).
[30] R.E. Neapolitan, *Probabilistic Reasoning in Expert Systems* (Wiley, New York, 1990).
[31] R.E. Neapolitan and J.R. Kenevan, Investigations of variances in belief networks, in: *Uncertainty in Artificial Intelligence* (North-Holland, Amsterdam, 1991) 232–240.
[32] G.W. Oehlert, A note on the delta method, *Am. Stat.* **46** (1992) 27–29.
[33] G. Paaß, Second order probabilities for uncertain and conflicting evidence, in: P.P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **6** (North-Holland, Amsterdam, 1991) 447–456.
[34] J. Pearl, *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufman, San Mateo, CA, 1988).
[35] Raima Database Manager, 1605 NW Sammamish Rd. Suite 200, Isaaquah, WA, 98027, USA (1994).
[36] A. Runnalls, A survey of sampling methods for inference on directed graphs, in: P. Cheeseman and R.W. Oldford, eds., *Selecting Models from Data* (Springer, New York, 1994) 153–162.
[37] D.J. Spiegelhalter, Probabilistic reasoning in predictive expert systems, in: L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* (North-Holland, Amsterdam, 1986) 47–68.
[38] D.J. Spiegelhalter, A unified approach to imprecision and sensitivity of beliefs in expert systems, MCR Biostatistics Unit, Cambridge (1991).
[39] D. Spiegelhalter, A. Dawid, S. Lauritzen and R. Cowell, Bayesian analysis in expert systems, *Stat. Sci.* (1993) 219–283.
[40] D.J. Spiegelhalter, R.C.G. Franklin and K. Bull, Assessment, criticism and improvement of imprecise subjective probabilities for a medical expert system, in: M. Henrion, R.D. Shachter, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **5** (North-Holland, Amsterdam, 1990) 285–294.
[41] A. Thomas, D. Spiegelhalter and W. Gilks, BUGS: a program to perform Bayesian inference using Gibbs sampling, in: J. Bernardo, J. Berger, A. Dawid and A.F.M. Smith, eds., *Bayesian Statistics* **4** (Oxford University Press, Oxford, 1992) 837–842.
[42] P. Walley, *Statistical Reasoning with Imprecise Probabilities* (Chapman and Hall, London, 1991).
[43] J. Whittaker, *Graphical Models in Applied Multivariate Statistics* (Wiley, Chichester, 1990).
[44] S.S. Wilks, *Mathematical Statistics* (Wiley, New York, 1962).