Machine Learning, 45, 147–170, 2001

# Robust Learning with Missing Data

MARCO RAMONI marco_ramoni@harvard.edu
*Children's Hospital Informatics Program, Harvard Medical School, Boston, MA 02115, USA*

PAOLA SEBASTIANI sebas@math.umass.edu
*Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01002, USA*

**Editor:** Pat Langley

**Abstract.** This paper introduces a new method, called the *robust Bayesian estimator* (RBE), to learn conditional probability distributions from incomplete data sets. The intuition behind the RBE is that, when no information about the pattern of missing data is available, an incomplete database constrains the set of all possible estimates and this paper provides a characterization of these constraints. An experimental comparison with two popular methods to estimate conditional probability distributions from incomplete data—Gibbs sampling and the EM algorithm—shows a gain in robustness. An application of the RBE to quantify a naive Bayesian classifier from an incomplete data set illustrates its practical relevance.

**Keywords:** Bayesian learning, Bayesian networks, Bayesian classifiers, probability intervals, missing data

## 1. Introduction

The Bayesian estimation of conditional probabilities from a data is a task relevant to a variety of machine learning applications, such as classification (Langley, Iba, & Thompson, 1992) and clustering (Cheeseman & Stutz, 1996). When no entry is missing in the database, these conditional probabilities can be efficiently estimated using standard Bayesian analysis (Good, 1968). Unfortunately, when the database is incomplete, i.e., some entries are reported as unknown, the simplicity and efficiency of this analysis are lost. Exact Bayesian analysis requires that one estimates the conditional probability distributions in each database that can be completed by replacing the missing entries with some value and then computes their average estimate. As the number of the completed databases increases exponentially with the number of missing entries, this exact analysis is computationally intractable.

During the past few years, several methods have been proposed for learning conditional probabilities from incomplete data sets. The two most popular methods are the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977) and Gibbs sampling (Geman & Geman, 1984). Both methods make the simplifying assumption that data are missing at random (Rubin, 1976). Under this assumption, the probability that an entry is not reported is independent of the missing entries in the data set and, in this situation, the missing values can be inferred from the available data. However, there is no way to verify this assumption on a database and, when this assumption is violated, all these methods can suffer of a dramatic decrease in accuracy (Spiegelhalter & Cowell, 1992). This situation motivated the recent development of a deterministic method, called *Bound* and *Collapse* (Ramoni & Sebastiani, 1998), that does not rely, per se, on a particular assumption about the missing data mechanism but allows the user to specify one, including the missing at random assumption. However, it still requires the specification of a particular missing data pattern, and this information may not be readily available. Approximate methods and simplifying assumptions make the task of learning from incomplete data feasible, but they prompt one to question the reliability of the estimates obtained in this way.

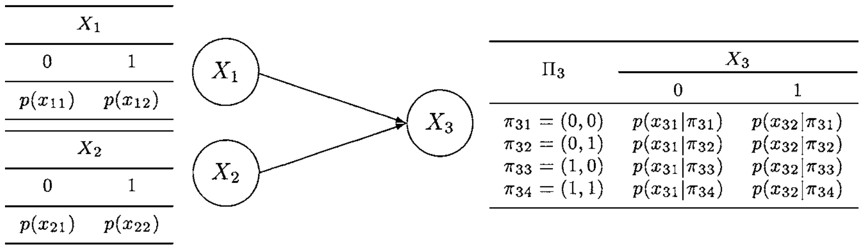
A typical solution to this problem is to measure this reliability by estimating the conditional probabilities under different assumptions about the missing data mechanism and by assessing the sensitivity of the estimates to these assumptions. A drawback of this approach is that each missing data mechanism explored requires a new estimation process and the choice of the mechanisms to consider in this sensitivity analysis is left entirely to the analyst. The rationale behind the approach presented in this paper is closely related to this idea and it can be regarded as an automated method for sensitivity analysis.

This paper introduces the *robust Bayesian estimator* (RBE) to learn conditional probability distributions from incomplete data sets without making any assumption about the missing data mechanism. The major feature of the RBE is to produce probability estimates that are robust with respect to different types of missing data. This robustness is achieved by providing *probability intervals* containing the estimates that can be learned from all completed data sets. The width of these intervals is a monotonically increasing function of the information available in the data set and thus provides a measure of the information conveyed by the data. We will focus on Bayesian networks, although the RBE can be used for the general task of learning conditional probabilities. During the past few years, there has been an increasing interest in algorithms that propagate probability intervals during inference in Bayesian networks (Fertig & Breese, 1993; Ramoni, 1995) but, to our knowledge, no effort has been made to apply the same interval-based approach to the task of learning such networks. The method presented in this paper can be therefore regarded as the learning counterpart of this research on reasoning methods, and we will show how the networks learned with the RBE can be used for classification and inference by means of these propagation methods.

The reminder of this paper describes our approach. Section 2 establishes some notation and reviews the background and motivation of the research. Section 3 describes the theoretical framework of the RBE and the use of probability intervals for inference, while Section 4 outlines the algorithms required for the implementation. Section 5 compares the RBE to expectation and maximization and Gibbs sampling in a controlled experiment, and Section 6 applies the RBE to an incomplete data set in a classification task.

## 2. Learning conditional probabilities

A Bayesian network is defined by a set of variables $\mathcal{X} = \{X_1, \ldots, X_v\}$ and a network structure $S$ represented by a graph of conditional dependencies among the variables in $\mathcal{X}$. A conditional dependency links a *child* variable $X_i$ to a set of *parent* variables $\Pi_i$ and, since we consider discrete variables only, is quantified by the table of conditional distributions of the child variable given each combination $\pi_{ij}$ of values of the parent variables $\Pi_i$. As a



| $X_1$ | |
|---|---|
| 0 | 1 |
| $p(x_{11})$ | $p(x_{12})$ |

| $X_2$ | |
|---|---|
| 0 | 1 |
| $p(x_{21})$ | $p(x_{22})$ |

| $\Pi_3$ | $X_3$ | |
|---|---|---|
| | 0 | 1 |
| $\pi_{31} = (0,0)$ | $p(x_{31}\|\pi_{31})$ | $p(x_{32}\|\pi_{31})$ |
| $\pi_{32} = (0,1)$ | $p(x_{31}\|\pi_{32})$ | $p(x_{32}\|\pi_{32})$ |
| $\pi_{33} = (1,0)$ | $p(x_{31}\|\pi_{33})$ | $p(x_{32}\|\pi_{33})$ |
| $\pi_{34} = (1,1)$ | $p(x_{31}\|\pi_{34})$ | $p(x_{32}\|\pi_{34})$ |

*Figure 1.* The graphical structure of a Bayesian network with variables $X_1$, $X_2$, and $X_3$, and the associated conditional probability distributions.

shorthand, we denote a single variable value $X_i = x_{ik}$ as $x_{ik}$ and a combination of values $\Pi_i = \pi_{ij}$ of the parent variables $\Pi_i$ as $\pi_{ij}$. We denote the number of states of the variable $X_i$ and the number of possible states of the parent variables $\Pi_i$ by $s_i$ and $q_i$, respectively. We also denote the conditional distribution of $X_i$ given the combination $\pi_{ij}$ of parent variables by $(p(x_{i1} \mid \pi_{ij}), \ldots, p(x_{is_i} \mid \pi_{ij}))$. Figure 1 shows an example in which $\mathcal{X} = \{X_1, X_2, X_3\}$ and each variable $X_i$ takes on one of the two values 0 and 1. The variables $X_1$ and $X_2$ have no parents, and they are both parents of $X_3$ so that $\Pi_3 = \{X_1, X_2\}$. The three tables of conditional probabilities give the distributions of $X_1$ and $X_2$ and the conditional distributions of $X_3$ given the four combinations of values of the parent variables.

Suppose we have a data set of $n$ cases $\mathcal{D} = \{c_1, \ldots, c_n\}$, where one case is $c_k = \{x_{1k}, \ldots, x_{uk}\}$, and a graphical structure $S$ over the set of variables $\{X_1, \ldots, X_v\}$. Our task is to estimate the conditional probability tables quantifying the dependence of each variable $X_i$ on its parent variables $\Pi_i$.

When the database is complete, that is, all entries are known, this estimation is straightforward. Let $n(x_{ik} \mid \pi_{ij})$ be the frequency of cases in which the variable $X_i$ appears in its state $x_{ik}$ together with the combination $\pi_{ij}$ of its parents $\Pi_i$. We denote by $n(\pi_{ij}) = \sum_k n(x_{ik} \mid \pi_{ij})$ the frequency of the parents' combination $\pi_{ij}$. Dividing each frequency $n(x_{ik} \mid \pi_{ij})$ by $n(\pi_{ij})$, we obtain the "classical" estimate of the conditional probability $p(x_{ik} \mid \pi_{ij})$, which is

$$\tilde{p}(x_{ik} \mid \pi_{ij}) = \frac{n(x_{ik} \mid \pi_{ij})}{n(\pi_{ij})}. \tag{1}$$

The set of estimates $\{\tilde{p}(x_{ik} \mid \pi_{ij})\}$, for all $i$, $j$ and $k$, quantifies the Bayesian network. These estimates maximize the joint probability of the data, $\Pi_{k=1}^{n} \Pi_{i=1}^{v} p(x_{ik} \mid \pi_{ij})$, also called the *likelihood function*, where the quantity $\Pi_{i=1}^{v} p(x_{ik} \mid \pi_{ij})$ is the joint probability of one case $c_k$.

The estimates $\{\tilde{p}(x_{ik} \mid \pi_{ij})\}$ are only a function of the data, but in some situations we may also want the estimation process to take into account some information, such as an expert's opinion. The Bayesian estimate of $p(x_{ik} \mid \pi_{ij})$ modifies the classical estimate $\tilde{p}(x_{ik} \mid \pi_{ij})$ by augmenting the observed frequencies $n(x_{ik} \mid \pi_{ij})$ by some quantities $\alpha_{ijk}$ that encode the external information in terms of *imaginary* frequencies of sample of size $\alpha$, where $\alpha_{ij}$ is $\sum_k \alpha_{ijk}$. The *Bayesian estimate* of $p(x_{ik} \mid \pi_{ij})$ is computed by applying Eq. (1) to the frequencies $n(x_{ik} \mid \pi_{ij})$, augmented by the quantities $\alpha_{ijk}$, and is

$$\hat{p}(x_{ik} \mid \pi_{ij}) = \frac{\alpha_{ijk} + n(x_{ik} \mid \pi_{ij})}{\alpha_{ij} + n(\pi_{ij})}. \tag{2}$$

By writing Eq. (2) as

$$\hat{p}(x_{ik} \mid \pi_{ij}) = \frac{\alpha_{ijk}}{\alpha_{ij}} \cdot \frac{\alpha_{ij}}{\alpha_{ij} + n(\pi_{ij})} + \frac{n(x_{ik} \mid \pi_{ij})}{n(\pi_{ij})} \cdot \frac{n(\pi_{ij})}{\alpha_{ij} + n(\pi_{ij})},$$

we observe that $\hat{p}(x_{ik} \mid \pi_{ij})$ is an average of the classical estimate $\tilde{p}(x_{ik} \mid \pi_{ij})$ and of the quantity $\alpha_{ijk}/\alpha_{ij}$. The latter is the estimate of $p(x_{ik} \mid \pi_{ij})$ when the data set consists of the imaginary counts $\alpha_{ijk}$ only (i.e., $n(x_{ijk} \mid \pi_{ij}) = 0$, for all $i$, $j$, and $k$) and it is therefore called the *prior probability* of $(x_{ijk} \mid \pi_{ij})$, while $\hat{p}(x_{ik} \mid \pi_{ij})$ is called the *posterior probability*. Note that $\alpha$ is the size of the imaginary sample upon which we base the formulation of the prior probability. As such, $\alpha$ represents a confidence measure of our prior probabilities and, therefore, it is called *prior precision* (Good, 1968; Ramoni & Sebastiani, 1999).

Unfortunately, the simplicity and efficiency of this closed form solution are lost when the database is incomplete, that is, some entries are reported as unknown. The issues involved in the estimating the probabilities from an incomplete data set $\mathcal{D}$ are better explained if we regard $\mathcal{D}$ as the result of a deletion process applied to a complete but unknown database $\mathcal{D}_c$. We define a *consistent completion* of $\mathcal{D}$ to be any complete database $\mathcal{D}_c$ from which we can obtain $\mathcal{D}$ using some deletion process. The set of consistent completions $\{\mathcal{D}_c\}$ is given by all databases in which the unknown entries are replaced by one of the possible values of the unobserved variables. The exact analysis, in this case, consists of applying Eq. (2) to each consistent completion $\mathcal{D}_c$, to yield a *consistent estimate* of the probability $p(x_{ik} \mid \pi_{ij})$, and then to average the consistent estimates. However, as the size of the set $\{\mathcal{D}_c\}$ grows exponentially with the number of missing entries, the exact analysis is computationally intractable. A typical solution is to make simplifying assumptions about the mechanism that causes missing data and to invoke approximate methods. Rubin (1976) classifies missing data mechanisms into three categories:

- *missing completely at random* (MCAR): the probability that an entry will be missing is independent of both observed and unobserved values in the data set;
- *missing at random* (MAR): the probability that an entry will be missing is a function of the observed values in the data set;
- *informatively missing* (IM): the probability that an entry will be missing depends on both observed and unobserved values in the data set.

These models are characterized by associating a dummy variable $R_i$ with each variable $X_i$. For each case in the data set, the variable $R_i$ takes on one of the two values 0 and 1 denoting, respectively, that the entry $x_{ik}$ is observed or not. The probability distribution for each variable $R_i$ specifies the missing data mechanism. Data are MCAR if the probability distribution of each $R_i$ is independent of $\{X_1, \ldots, X_v\}$. When the probability distribution of each $R_i$ is a function of the observed values in the data set, data are MAR, whereas data are IM when the probability distribution of each $R_i$ is a function of the observed and unobserved entries.

When data are either MCAR or MAR, the deletion mechanism is said to be *ignorable* because we can infer the missing entries from the observed ones. The two most popular solutions to handle incomplete data sets—the *expectation maximization* (EM) algorithm (Dempster, Laird, & Rubin, 1977) and Gibbs sampling (Geman & Geman, 1984)—rely on the assumption that data are MAR. The EM algorithm is an iterative method that approximates the estimate in either Eqs. (1) or (2) when data are incomplete and the likelihood function, as defined in the previous section, becomes a mixture of likelihood functions, one for each consistent completion of the data set. EM alternates an expectation step, in which unknown quantities depending on the missing entries are replaced by their expectation in the likelihood function, with a maximization step, in which the likelihood is maximized with respect to the set of unknown probabilities $\{p(x_{ik} \mid \pi_{ij})\}$. The estimates computed by the maximization step are then used to replace unknown quantities by their expectation in the next step, and the whole process is repeated until the difference between successive estimates is smaller than a fixed threshold. The EM algorithm produces an approximation of the estimate $(\alpha_{ijk} + n(x_{ik} \mid \pi_{ij}) - 1)/(\alpha_{ij} + n(\pi_{ij}) - s_i)$, the so called *maximum a posteriori* (Heckerman, Geiger, & Chickering, 1995). However, by setting $\alpha'_{ijk} = \alpha_{ijk} + 1$, the estimate $(\alpha'_{ijk} + n(x_{ik} \mid \pi_{ij}) - 1)/(\alpha'_{ij} + n(\pi_{ij}) - s_i)$ becomes exactly that given in Eq. (2). The convergence rate of this process can be slow and several modifications have been proposed to increase it (Lauritzen, 1995; Zhang, 1996; Russell et al., 1995; Friedman, 1977).

In contrast to the EM algorithm, which is iterative but deterministic, Gibbs sampling is a stochastic method that produces a sample of values for the probabilities $\{p(x_{ik} \mid \pi_{ij})\}$ from which one can compute $\{\hat{p}(x_{ik} \mid \pi_{ij})\}$ as sample means. The method works by generating a Markov chain whose equilibrium distribution is the distribution generating the posterior probabilities $\{\hat{p}(x_{ik} \mid \pi_{ij})\}$. In practice, the algorithm iterates a number of times—called the *burn in*—to reach stability and then takes a final sample from the equilibrium distribution (Thomas, Spiegelhalter, & Gilks, 1992). The advantage of Gibbs sampling over EM is that the simulated sample provides empirical estimates of the variance, as well as *credible intervals*, that is, intervals that contain the $p(x_{ik} \mid \pi_{ij})$ values with a given probability. Gibbs sampling treats missing data as unknown quantities to be estimated so that, as the number of missing entries in the data set increases, the convergence rate of the method decreases very rapidly.

When data are neither MAR nor MCAR, the accuracy of both the EM algorithm and Gibbs sampling can dramatically decrease (Spiegelhalter & Cowell, 1992). This finding raises questions about the reliability of the estimates produced under these assumptions. One solution is to perform a sensitivity analysis to assess the robustness of the estimates with respect to different missing data mechanisms. This consists of repeating the estimation process under different assumptions about the missing data mechanism and evaluating the changes in the estimates. The drawback of sensitivity analysis is that it requires a new estimation process for each missing data mechanism, but it leads naturally to the idea of the Robust Bayesian Estimator (RBE), which we introduce in the next section.

## 3. Learning and reasoning with probability intervals

The solution we propose is based on the idea that, even with no information on the missing data mechanism, an incomplete data set $\mathcal{D}$ constrains the set of estimates that can be induced from its consistent completions. Following this principle, we introduce the *Robust Bayesian Estimator* to learn the conditional probabilities $\{p(x_{ik} \mid \pi_{ij})\}$ with no assumptions about the missing data mechanism. The RBE returns, for each conditional probability $p(x_{ik} \mid \pi_{ij})$, an interval containing all the consistent estimates of $p(x_{ik} \mid \pi_{ij})$ and proceeds by refining this set as more information becomes available. This section shows how to estimate these intervals from an incomplete database and how to reason on the basis of these intervals. The first step of the estimation with the RBE is the definition of *virtual frequencies* that are used to find the extreme points of the probability intervals.

### *3.1. Virtual frequencies*

Suppose that we wish to estimate the conditional probability $p(x_{ik} \mid \pi_{ij})$ from an incomplete database $\mathcal{D}$ in which some entries of the variable $X_i$ and of its parent variables $\Pi_i$ are unknown. These unknown entries give rise to three types of incomplete cases that are relevant to the estimation of $p(x_{ik} \mid \pi_{ij})$:

- The variable $X_i$ takes value $x_{ik}$, the value of $\Pi_i$ is not fully observed, and it can be *consistently* completed as $\pi_{ij}$ (that is, the observed incomplete value of $\Pi_i$ is obtained from $\Pi_i = \pi_{ij}$ using some deletion process).
- The parent variables $\Pi_i$ take value $\pi_{ij}$ and the value of $X_i$ is missing.
- Both values of $X_i$ and $\Pi_i$ are unknown, and the value of $\Pi_i$ can be consistently completed as $\pi_{ij}$.

We denote the frequency of these cases, respectively, by $n(x_{ik} \mid ?)$, $n(? \mid \pi_{ij})$, and $n(? \mid ?)$. Consider, for example, the Bayesian network in figure 1 and suppose that we wish to estimate the conditional probability of $X_3 = 0$ (termed $x_{31}$ in figure 1), given the parent variables configurations $\Pi_3 = (1,\ 0)$ (which we called $\pi_{33}$ in figure 1), from the incomplete data set in Table 1. The cases $c_1$ and $c_6$ are complete and determine $n(x_{31} \mid \pi_{33}) = 2$. All cases $c_3$, $c_5$, $c_7$, $c_8$, and $c_{10}$ can be consistently completed as $X_3 = 0 \mid \Pi_3 = (1,\ 0)$. The case $c_3$ determines $n(? \mid \pi_{33}) = 1$. The cases $c_7$ and $c_{10}$ determines $n(x_{31} \mid ?) = 2$, while the two cases $c_5$ and $c_8$ determine $n(? \mid ?) = 2$.

By completing the cases $c_3$, $c_5$, $c_7$, $c_8$, and $c_{10}$ as $X_3 = 0 \mid \Pi_3 = (1,\ 0)$, we create a particular consistent completion of the data set, in which the event $X_3 = 0 \mid \Pi_3 = (1,\ 0)$ occurs the largest number of times. This idea is the intuition behind the definition of the virtual frequency $\bar{n}(x_{ik} \mid \pi_{ij})$. The quantity $\bar{n}(x_{ik} \mid \pi_{ij})$ is the maximum number of incomplete cases $(X_i, \Pi_i)$ that can be consistently completed as $(x_{ik},\ \pi_{ij})$ and is defined by

$$\bar{n}(x_{ik} \mid \pi_{ij}) = n(? \mid \pi_{ij}) + n(x_{ik} \mid ?) + n(? \mid ?). \tag{3}$$

*Table 1.* An incomplete data set used to describe the virtual frequencies.

| Case | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $c_1$ | 1 | 0 | 0 |
| $c_2$ | 0 | ? | 1 |
| $c_3$ | 1 | 0 | ? |
| $c_4$ | ? | ? | 1 |
| $c_5$ | 1 | ? | ? |
| $c_6$ | 1 | 0 | 0 |
| $c_7$ | ? | 0 | 0 |
| $c_8$ | ? | ? | ? |
| $c_9$ | ? | 0 | 1 |
| $c_{10}$ | ? | 0 | 0 |

Now note that, if we complete the cases $c_3$, $c_4$, $c_5$, $c_8$, and $c_9$ as $X_3 = 1 \,|\, \Pi_3 = (1,\ 0)$, we create a consistent completion of the data set in which the event $X_3 = 0 \,|\, \Pi_3 = (1,\ 0)$ occurs the minimum number of times. We then define the virtual frequency $\underline{n}x_{ik} \,|\, \pi_{ij}$ as the maximum number of incomplete cases $(X_i,\ \Pi_i)$ that can be ascribed to $\pi_{ij}$ without increasing the frequency $n(x_{ik} \,|\, \pi_{ij})$, which is equivalent to

$$\underline{n}(x_{ik} \,|\, \pi_{ij}) = n(? \,|\, \pi_{ij}) + \sum_{h \neq k} n(x_{ih} \,|\, ?) + n(? \,|\, ?). \tag{4}$$

In the next section, we use these virtual frequencies to find the minimum and maximum estimate of the probability $p(x_{ik} \,|\, \pi_{ij})$.

### 3.2. *The robust Bayesian estimator*

We define the robust Bayesian estimator (RBE) as the estimator that returns the probability interval $[\underline{p}(x_{ik} \,|\, \pi_{ij}\,)\bar{p}(x_{ik} \,|\, \pi_{ij})]$ containing the set of consistent estimates of $p(x_{ik} \,|\, \pi_{ij})$. The values $\underline{p}(x_{ik} \,|\, \pi_{ij})$ and $\bar{p}(x_{ik} \,|\, \pi_{ij})$ are, respectively, the minimum and the maximum estimate of $p(x_{ik} \,|\, \pi_{ij})$ that can be found in the consistent completions $\mathcal{D}_c$ of $\mathcal{D}$. Next, Theorem 1 gives a closed form solution for the minimum and maximum value of $\hat{p}(x_{ik} \,|\, \pi_{ij})$.

**Theorem 1.** *Let $\mathcal{D}$ be an incomplete database. The minimum and maximum Bayesian estimate of $p(x_{ik} \,|\, \pi_{ij})$ are, respectively,*

$$\underline{p}(x_{ik} \,|\, \pi_{ij}) = \frac{\alpha_{ijk} + n(x_{ik} \,|\, \pi_{ij})}{\alpha_{ij} + n(\pi_{ij}) + \underline{n}(x_{ik} \,|\, \pi_{ij})} \tag{5}$$

$$\bar{p}(x_{ik} \,|\, \pi_{ij}) = \frac{\alpha_{ijk} + n(x_{ik} \,|\, \pi_{ij}) + \bar{n}(x_{ik} \,|\, \pi_{ij})}{\alpha_{ij} + n(\pi_{ij}) + \bar{n}(x_{ik} \,|\, \pi_{ij})}. \tag{6}$$

**Proof:** Let $y_{ijk}$ be the unknown frequency of $(x_{ik},\ \pi_{ij})$, so that $y_{ij} = y_{ijk} + y_{ij\backslash k}$, where $y_{ij\backslash k} = \sum_{h \neq k} y_{ijk}$ is the known frequency of $\pi_{ij}$. The Bayesian estimate of $p(x_{ik} \mid \pi_{ij})$ that would be computed from the complete data set—if known—is

$$f(y_{ijk},\ y_{ij\backslash k}) = \frac{\alpha_{ijk} + n(x_{ik} \mid \pi_{ij}) + y_{ijk}}{\alpha_{ij} + n(\pi_{ij}) + y_{ijk} + y_{ij\backslash k}}.$$

The information conveyed by the incomplete cases impose three constraints on the variables $y_{ijk}$ and $y_{ij\backslash k}$:

$$\begin{aligned}
&0 \leq y_{ijk} \leq \bar{n}(x_{ik} \mid \pi_{ij})\\
&0 \leq y_{ij\backslash k} \leq \underline{n}(x_{ik} \mid \pi_{ij})\\
&n(? \mid \pi_{ij}) \leq y_{ijk} + y_{ij\backslash k} \leq n(? \mid \pi_{ij}) + \sum_{k} n(x_{ik} \mid ?) + n(? \mid ?).
\end{aligned}$$

This system of inequalities identifies the constraint region displayed in figure 2. One can show that $f(y_{ijk},\ y_{ij\backslash k})$ is an increasing function of $y_{ijk}$ and a decreasing function of $y_{ij\backslash k}$, so that it is maximized when $y_{ijk} = \bar{n}(x_{ik} \mid \pi_{ij})$ and $y_{ij\backslash k} = 0$, and it is minimized when $y_{ijk} = 0$ and $y_{ij\backslash k} = \underline{n}(x_{ik} \mid \pi_{ij})$. Since these points are both in the constraint region, the proof is complete. □

When $X_i$ is a Boolean variable taking on one of the two values $x_{i1}$ and $x_{i2}$, then the probability intervals returned by the RBE are such that

$$\begin{aligned}
\bar{p}(x_{i1} \mid \pi_{ij}) &= 1 - \underline{p}(x_{i2} \mid \pi_{ij})\\
\bar{p}(x_{i2} \mid \pi_{ij}) &= 1 - \underline{p}(x_{i1} \mid \pi_{ij}).
\end{aligned}$$

The RBE does not make any assumption about the missing data model and thus provides a framework for sensitivity analysis, as we will show in the next section.

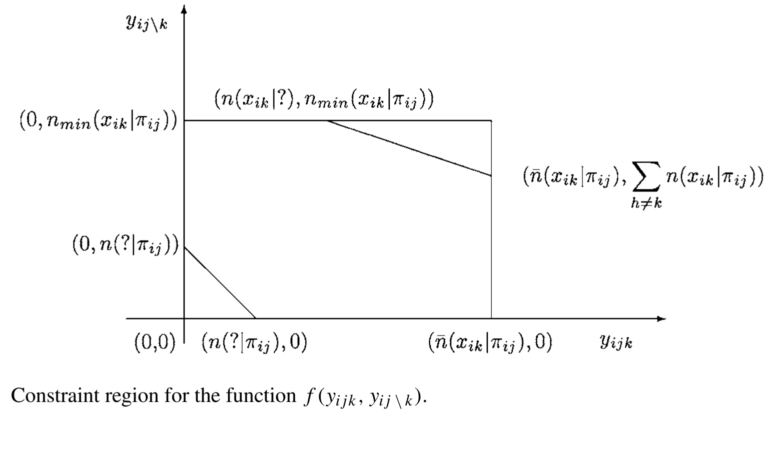

$y_{ij\backslash k}$
$(n(x_{ik}|?), n_{min}(x_{ik}|\pi_{ij}))$
$(0, n_{min}(x_{ik}|\pi_{ij}))$
$(\bar{n}(x_{ik}|\pi_{ij}), \sum_{h \neq k} n(x_{ik}|\pi_{ij}))$
$(0, n(?|\pi_{ij}))$
(0,0)
$(n(?|\pi_{ij}), 0)$
$(\bar{n}(x_{ik}|\pi_{ij}), 0)$
$y_{ijk}$

*Figure 2.* Constraint region for the function $f(y_{ijk}, y_{ij\backslash k})$.

### 3.3. *Interval-based inference*

The probability intervals computed with the RBE can be used, as they are, to perform classification and inference, as well as to provide a form of *sensitivity analysis* for the conclusions achieved by other methods. We first show that the RBE produces intervals that can be used to evaluate the reliability of the estimates provided by the EM algorithm and Gibbs sampling.

Both EM and Gibbs sampling compute Bayesian estimates of the conditional probabilities $p(x_{ik} \mid \pi_{ij})$ under the assumption that data are MAR. However, they do not provide any measure of the impact of the assumed model for the missing data on the estimates. Gibbs sampling returns, for each estimate $\hat{p}(x_{ik} \mid \pi_{ij})$, a credibility interval that represents the uncertainty about the estimate. As these credibility intervals are computed assuming that data are MAR, they do not tell us anything about the sensitivity of each estimate to the MAR assumption.

It is straightforward to show that the width $w(x_{ik} \mid \pi_{ij})$ of the probability interval computed by the RBE as an estimate of $p(x_{ik} \mid \pi_{ij})$ is a monotonic increasing function of the number of incomplete cases. Thus, the wider the interval, the greater the uncertainty due to the incompleteness of the data, and the less reliable the point estimate returned by either the EM algorithm or Gibbs sampling. In this way, we can use the value $1 - w(x_{ik} \mid \pi_{ij})$ as a *local measure of reliability* for each point estimate, to account for the missing data. The average width $\bar{w}$ of all interval estimates provides a *global measure of reliability* of the estimates as $1 - \bar{w}$. Both values $1 - w(x_{ik} \mid \pi_{ij})$ and $1 - \bar{w}$ vary between 0 and 1, with small values denoting lack of reliability and values near 1 denoting high reliability.

Suppose now that we wish to use a Bayesian network, quantified with estimates computed from an incomplete data set, to predict the value of the variable $X_i$, given that we observe the values of a subset of the other variables in the network. The set of variable values observed is called *evidence*, which we denote by $e$. The solution is to compute the probability distribution of $X_i$ given the evidence $e$—using some standard propagation algorithm (Pearl, 1988; Castillo, Gutierrez, & Hadi, 1997)—and then to select the value of $X_i$ with the largest probability, given $e$. We can similarly propagate the probability intervals computed by the RBE with one of the existing propagation algorithms for interval-based Bayesian networks and calculate a probability interval $[\underline{p}(x_{ik} \mid e)\ \bar{p}(x_{ik} \mid e)]$ for each value $p(x_{ik} \mid e)$. Such intervals can be used to make a prediction that does not rely on any particular assumption about the model for the missing data. We accomplish this task by choosing a criterion upon which to base the selection of the $X_i$ value. The *stochastic dominance* criterion (Kyburg, 1983) selects the value $x_{ik}$ of $X_i$ if the minimum probability $\underline{p}(x_{ik} \mid e)$ is larger than the maximum probability $\bar{p}(x_{ih} \mid e)$, for any $h \neq k$. Stochastic dominance is the safest and most conservative criterion since the prediction is independent of the distribution of missing data.

When the probability intervals are overlapping, the stochastic dominance criterion is not applicable and we face a situation of undecidability. In this case, we can rank the probability intervals $[\underline{p}(x_{ik} \mid e)\ \bar{p}(x_{ik} \mid e)]$ by assigning, to each of them, a predictive score, and the decision criterion is to select the value $x_{ik}$ of $X_i$ associated with the interval $[\underline{p}(x_{ik} \mid e)\ \bar{p}(x_{ik} \mid e)]$ receiving the highest score. Let $q(x_{ik})$ be the probability that an unknown value of $X_i$ must be completed as $x_{ik}$. We define the *predictive score* of $x_{ik} \mid e$ by

$$s_q(x_{ik} \mid e) = \underline{p}(x_{ik} \mid e)(1 - q(x_{ik})) + \bar{p}(x_{ik} \mid e)q(x_{ik}).$$

The score $s_q(x_{ik} \mid e)$ falls in the interval $[\underline{p}(x_{ik} \mid e)\ \bar{p}(x_{ik} \mid e)]$ and approaches the maximum when the missing values of $X_i$ are supposed to be all $x_{ik}$, while $s_q(x_{ik} \mid e)$ approaches the minimum when the missing values of $X_i$ are supposed to be different from $x_{ik}$. If we do not want to commit ourselves to any particular missing data mechanism, we can assume that all mechanisms are equally likely, so that the probability that $X_i$ takes on one of the values $x_{ik}$, when the entry in the data set is unknown, is the uniform probability $1/s_i$, where $s_i$ is the number of states of $X_i$ and

$$s_u(x_{ik} \mid e) = \frac{\underline{p}(x_{ik} \mid e)(s_i - 1)}{s_i} + \frac{\bar{p}(x_{ik} \mid e)}{s_i}.$$

When we believe that data are MAR or MCAR, so that the probability that an entry is missing is not a function of the unknown values, we can estimate the distribution $q(x_{ik})$ from the data available as $q(x_{ik}) = n(x_{ik})/\sum_{k=1}^{s_i} n(x_{ik})$.

Stochastic dominance is a special case of this criterion in which $q(x_{ik}) = 0$, and hence we term this criterion, in which we summarize the prediction interval by the point $s_q(x_{ik} \mid e)$, *weak dominance*.

## 4. Implementation of the robust Bayes estimator

This section outlines the algorithm that implements the method described in Section 3. We first describe the estimation procedure in a Bayesian network, and then we analyze the computational complexity of the algorithm.

### 4.1. Estimation procedure

The RBE can be regarded as a batch procedure that parses the data set, stores the observations about the variables, and then computes the conditional probabilities needed to quantify a Bayesian network from these observations. The procedure takes as input a data set $\mathcal{D}$ and a network structure $S$, identified by a set of conditional dependencies $\{d_{X_1}, \ldots, d_{X_I}\}$ associated with each variable in $\mathcal{X}$. A dependency $d_{X_i}$ is an ordered tuple $(X_i, \Pi_i)$, where $\Pi_i$ are parents of $X_i$.

The learning procedure parses each case in $\mathcal{D}$ using the dependencies defining the network structure $S$. Therefore, for each entry in the case, the procedure recalls the dependency within which the variable appears as a child, and identifies the states of its parent variables recorded in the case. For each combination of states $(x_{ik},\ \pi_{ij})$, the procedure maintains the two counters $n(x_{ik} \mid \pi_{ij})$ and $\bar{n}(x_{ik} \mid \pi_{ij})$. For each child variable $X_i$, it also keeps track of the unobserved entries for $X_i$ in a third counter $n_{mis}(x_i)$. These three counters are sufficient to compute the quantity $\underline{n}(x_{ik} \mid \pi_{ij})$ by noting that, since the marginal frequency of cases in which the value of $X_i$ is unknown is $n_{mis}(x_i) = n(? \mid \pi_{ij}) + n(? \mid ?)$, the quantity $\underline{n}(x_{ik} \mid \pi_{ij})$

can be written as a function of $n_{mis}(x_i)$ and $\bar{n}(x_{ik} \mid \pi_{ij})$, giving

$$\underline{n}((x_{ik} \mid \pi_{ij})) = \sum_{h \neq k, h=1}^{c_i} \bar{n}(x_{ih} \mid \pi_{ij}) - (c_i - 2) n_{mis}(x_i).$$

When the detected configuration does not contain any missing data, the first counter $n(x_{ik} \mid \pi_{ij})$ is increased by one. When the value of one or more variables in the combination is missing, the procedure increases the second counter $\bar{n}(x_{ik} \mid \pi_{ij})$ by one, for each configuration of states of the variables with missing entries. When an observation is missing for the child variable, the counter $n_{mis}(x_i)$ is also increased by one.

The procedure for storing the counters plays a crucial role in determining the efficiency of the algorithm. The current implementation uses *discrimination trees* to store the counters, following a slightly modified version of the method proposed by Ramoni et al. (1995) that implements an idea originally due to Cooper and Herskovitz (1992). In this approach, the states of each variable of the network are associated with a discrimination tree whose levels are defined by the possible states of a parent variable. Each path in the discrimination tree represents a possible value of parent variables for that state. In this way, each path is associated with a single conditional probability in the network. Each leaf of the discrimination tree holds the pair of counters $n(x_{ik} \mid \pi_{ij})$ and $\bar{n}(x_{ik} \mid \pi_{ij})$. For each observed entry, the procedure just needs to follow a path in the discrimination tree to identify the counters to be updated. In order to save memory, the discrimination trees are built incrementally: each branch in the tree is created the first time the procedure needs to walk through it. Once the data set has been parsed, the procedure has only to collect the counters $n(x_{ik} \mid \pi_{ij})$, $\bar{n}(x_{ik} \mid \pi_{ij})$, and $n_{mis}(x_i)$ for each variable and compute minimum and maximum using Eqs. (5) and (6).

### 4.2. *Computational complexity*

The RBE procedure takes advantage of the modular nature of Bayesian networks and partitions the search space using the dependencies in the network. The algorithm starts by scanning the data set $\mathcal{D}$ and, for each element in a row, it scans the row again to identify the patterns of the dependencies. If the data set $\mathcal{D}$ contains $n$ rows, one for each case, and $v$ columns, corresponding to the $v$ variables in the network, the upper bound of the execution time for this part of the algorithm is $O(gnv^2)$, where $g$ is the maximum number of parents for a variable in the network. Finally, the procedure scans the generated discrimination trees to compute the virtual counters. The number of such trees created during the learning process is the total number of values that the variables $X_1, \ldots, X_v$ take on. The number of leaves of each discrimination tree associated with a state of each variable $X_i$ equals to the number of combinations of parents states $\pi_{ij}$, and this is the minimum number of conditional distributions required to define a conditional dependency.

## 5. Experimental evaluation

This section reports the results of three experimental comparisons based on natural data. The aim of these experiments was twofold: to evaluate the effectiveness of the RBE estimates

*Table 2*. Description of variables of the network displayed in figure 3.

| Name | Description | States |
|---|---|---|
| $X_1$ | Family anamnesis of coronary heart disease | neg, pos |
| $X_2$ | Strenuous mental work | no, yes |
| $X_3$ | Ratio of beta and alpha lipoproteins | $< 3, \geq 3$ |
| $X_4$ | Strenuous physical work | no, yes |
| $X_5$ | Smoking | no, yes |
| $X_6$ | Systolic blood pressure | $< 140, \geq 140$ |

for inference and to show that the RBE returns probability intervals that provide a reliability measure of the point estimates computed using either the EM algorithm or Gibbs sampling. We begin by describing the data set used for the experiment and the procedure that we used to remove data. The criteria used to evaluate the experimental results are described in Section 5.2, while the results of the experiments are discussed in Section 5.3.

Whittaker (1990, p. 261) reports a data set that involves six Boolean risk factors $X_1, \ldots, X_6$ observed in a sample of 1841 employees of a Czech car factory. Table 2 describes the variables and their values. The data set is complete and we used the K2 algorithm (Cooper & Herskovitz, 1992) to extract the most probable structure, reported in figure 3. The K2 algorithm extracts the most probable network consistent with a partial order among the variables in the data set. We chose $X_1 \leq X_2 \leq X_3 \leq X_4 \leq X_5 \leq X_6$ as the initial order, where $X_i \leq X_j$ means that $X_i$ cannot be a child of $X_j$. We then estimated the fifteen conditional probabilities that quantify this network using Eq. (2), with $\alpha_{ij} = 8/q_i$ and $\alpha_{ijk} = 8/(s_i q_i)$, where $s_i$ and $q_i$ denote, respectively, the number of states of the variable $X_i$ and its parent variables $\Pi_i$.

### 5.1. Procedures used to remove data

We generated three groups of experimental data sets by removing, in each group, values using one of the three missing data mechanisms described in Section 2. In this way, we created three sets of incomplete databases in which, respectively, at most 25%, 50%, and 75% of entries—of some or all the variables in the data set—are missing. We describe each of the deletion procedures in turn.

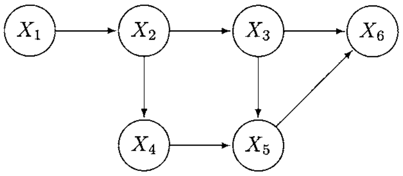


*Figure 3*. The network structure extracted from the data set used for the experimental evaluation. The description of variables is detailed in Table 2.

*Group 1: Missing completely at random.* In this group, all variables in the data set were subject to a deletion process. We associated each variable $X_i$ with a dummy variable $R_i$ that took on one of the two values 0 and 1 with some probability. The original network of figure 3 was augmented by the six variables $R_1, \ldots, R_6$, marginally independent of $X_1, \ldots, X_6$, as shown in figure 4. For each case in the original data set, we generated a combination of values of the six variables $R_i$ and removed the entry of the variable $X_i$ if the value of $R_i$ was 1. Thus, data removed with this process were MCAR. To obtain data sets with different proportions of missing data, this process was repeated with three different sets of probability values $(p(R_1 = 1), \ldots, p(R_6 = 1))$, independently generated from uniform distributions in the intervals [0 0.25], [0.25 0.5] and [0.5 0.75]. For each set of probability values, we generated ten incomplete data sets, so that this group consists of thirty data sets with an average proportion of missing entries 15%, 36%, and 64%.

*Group 2: Missing at random.* In this group, only the variables $X_3$, $X_5$, and $X_6$ were subject to a deletion process. We associated these variables with dummy variables $R_3$, $R_5$, and $R_6$ that took on one of the two values 0 and 1 and, for each case in the original data set, we generated a combination of values for the three variables $R_i$ and removed the entry of the variable $X_i$ if the value of $R_i$ was 1. The distribution for each of the variables $R_3$, $R_5$, and $R_6$ was a function of the variables $X_1$, $X_2$, and $X_4$, as shown in figure 4. Thus, since $X_1$, $X_2$, and $X_4$ are fully observed and the distribution of $R_3$, $R_5$, and $R_6$ is only dependent on the values observed in the incomplete data set, data removed with this process are MAR. We repeated this deletion process with three different sets of probability values that were generated from uniform distributions in the intervals [0 0.25], [0.25 0.5], and [0.5 0.75]. Again, for each set of probability values, we generated ten data sets in which the average proportion of missing entries were 10%, 20%, and 30%.

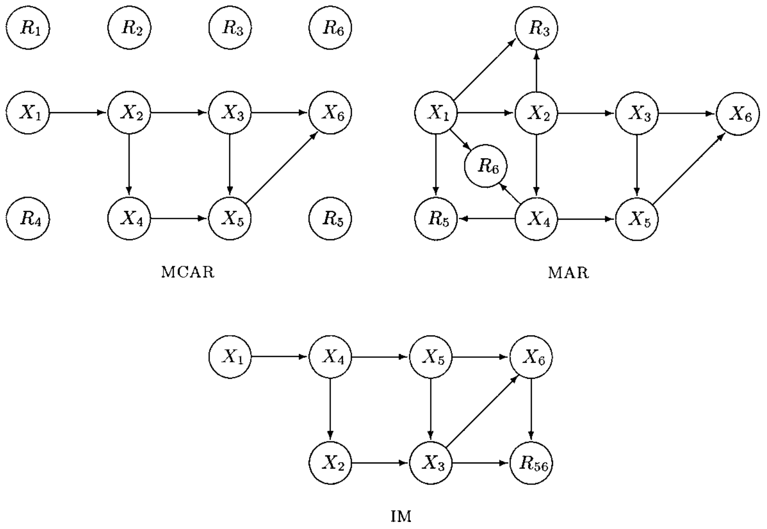


*Figure 4.* Graphical representation of the missing data mechanisms used in the experiments.

*Group 3: Informatively missing.* In this group, only the variables $X_5$ and $X_6$ were subject to a deletion process. We associated the variables $X_5$ and $X_6$ with a dummy variable $R_{56}$ that took on one of the two values 0 and 1 and, for each case in the original data set, we generated a value of the variable $R_{56}$ and removed the entry of both variables $X_5$ and $X_6$ if the value of $R_{56}$ was 1. The distribution of $R_{56}$ is a function of $X_5$ and $X_6$ as follows. The probabilities $p(R_{56} = 1 \mid X_5 = \text{no},\ X_6 < 140)$ and $p(R_{56} = 1 \mid X_5 = \text{no},\ X_6 \geq 140)$ are both zero, so that none of the pair of entries $(X_5 = \text{no},\ X_6 < 140)$ and $(X_5 = \text{no},\ X_6 \geq 140)$ were removed from the data set, while the probabilities $p(R_{56} = 1 \mid X_5 = \text{yes},\ X_6 < 140)$ and $p(R_{56} = 1 \mid X_5 = \text{yes},\ X_6 \geq 140)$ were both randomly generated. Figure 4 depicts the missing data model. Since the distribution of $R_{56}$ depends of the unobserved values in the data set, values removed with this process are IM. The deletion process was repeated with three different sets of probability values that were generated from uniform distributions in the intervals [0 0.25], [0.25 0.5], and [0.5 0.75], and, for each set of probability values, we generated ten incomplete data sets in which the average proportions of missing entries were 1.5%, 5%, and 10%.

### 5.2. *Evaluation criteria*

The deletion procedures described in the previous section produced 90 data sets. From each of these incomplete data sets, we estimated the conditional probability tables using EM (with $\alpha'_{ijk} = 8 / (c_i q_i) + 1$), Gibbs sampling, and the RBE (both with $\alpha_{ijk} = 8 / (c_i q_i)$), and we then evaluated the reliability of the estimates computed with the first two methods using the measures of local and global reliability defined in Section 3.3. We then used the Bayesian networks quantified with these estimates to compute the predictive probabilities of the event $X_6 < 140$ given the 43 relevant evidences in the five risk factors.[1] We also computed prediction intervals for the event $X_6 < 140$ by exact propagation of the probability intervals found with the RBE. We then compared both point-based and interval-based predictions to the values predicted from the Bayesian network, quantified with the complete data set using two performance measures that evaluate, respectively, the predictive accuracy and precision.

The first performance measure compares the number of correct predictions when the criterion selects the value of $X_6$ with the largest probability, given the evidence. For interval-based predictions, we used both the stochastic dominance criterion, which selects the value $X_6 < 140$ if $\underline{p}(X_6 < 140 \mid e) > \bar{p}(X_6 \geq 140 \mid e)$, and the weak dominance criterion—described in Section 3.3—which selects the value of $X_6$ with the highest score. The two scores for $X_6 \mid e$ were computed as

$$\begin{aligned} s_q(X_6 < 140 \mid e) = {} & \underline{p}(X_6 < 140 \mid e)(1 - q(X_6 < 140)) \\ & + \bar{p}(X_6 < 140 \mid e) q(X_6 < 140) \end{aligned}$$

and

$$\begin{aligned} s_q(X_6 \geq 140 \mid e) = {} & \underline{p}(X_6 \geq 140 \mid e)(1 - q(X_6 \geq 140)) \\ & + \bar{p}(X_6 \geq 140 \mid e) q(X_6 \geq 140), \end{aligned}$$

where $q(X_6 < 140)$ is the probability that an unknown value of $X_6$ in the data set can be ascribed to a value smaller than 140, and $q(X_6 \geq 140) = 1 - q(X_6 < 140)$. Since both the EM algorithm and Gibbs sampling assume that data are MAR, we encoded the same assumption by setting $q(X_6 < 140) = n(X_6 < 140)/n$. The average estimated probabilities $q(X_6 < 140)$ were 0.57 in the first two groups of data sets and 0.57, 0.56, and 0.55 in the last group, in which data were informatively missing.

The second performance measure compares the cumulative Kullback-Liebler distance between the distributions of $X_6 \mid e_j$—conditional on the 43 evidences $e_j$—computed from the network quantified with complete data, and the conditional distribution of $X_6 \mid e_j$ computed from the network quantified with incomplete data using EM, Gibbs sampling, and the RBE. Since the variable $X_6$ is Boolean, one can easily show that $s_q(X_6 < 140 \mid e_j) = 1 - s_q(X_6 \geq 140 \mid e_j)$, so the two scores define a conditional distribution for $X_6 \mid e_j$ that was used to evaluate the predictive precision of the RBE. Denote by $p(x_{6k} \mid e_j)$ the distribution of $X_6$, given the evidence $e_j$, induced from the complete data, and by $\hat{p}(x_{6k} \mid e_j)$ the distribution of $X_6$, given $e_j$, induced from the incomplete data set using the various methods. This performance measure can be computed as

$$\sum_{j=1}^{43} \sum_{k=1}^{2} p(x_{6k} \mid e_j) \log \frac{p(x_{6k} \mid e_j)}{\hat{p}(x_{6k} \mid e_j)}.$$

Note that this quantity is zero whenever $p(x_{6k} \mid e_j) = \hat{p}(x_{6k} \mid e_j t)$ for all $j = 1, \ldots, 43$, and that it increases as the accuracy of the approximation decreases. We compared the results obtained from the RBE estimates with the results obtained from an implementation of the accelerated EM algorithm in GAMES (Thiesson, 1995) and the implementation of Gibbs sampling called BUGS (Thomas, Spiegelhalter, & Gilks, 1992).

### *5.3. Results and discussion*

We begin by describing the use of the probability intervals computed by the RBE as a measure of global reliability for the estimates found with EM and Gibbs sampling. Figure 5 plots the average lengths $\bar{w}$ of the intervals computed with the RBE versus different proportions of missing data in the three groups of incomplete data sets. The plot shows that $\bar{w}$ is an increasing function of the proportion of missing entries in the data set. From the plot, we can deduce the global reliability $1 - \bar{w}$ of the estimates induced from the incomplete data sets. When data are MCAR, the global reliability of the estimates is about 0.65 when 25% of data are missing, but this decreases to less than 0.1 when 75% of data are missing. When data are MAR, the reliability decreases to 0.45 when less than 75% of the entries of $X_3$, $X_5$, and $X_6$ are missing, but when data are IM, the reliability is 0.75 in the worst case, in which, on average, 10% of the entries in the data set are missing.

The high uncertainty about the estimates must be taken into account when the network induced from data is used for inference. Figure 6 displays the average lengths of the prediction intervals computed by propagating the estimates returned by the RBE. The plot shows that, when data are either MCAR or MAR, the reliability of the prediction can be as small as 0.05—given by $1 - 0.95$—when 75% of data are missing. We note that the average

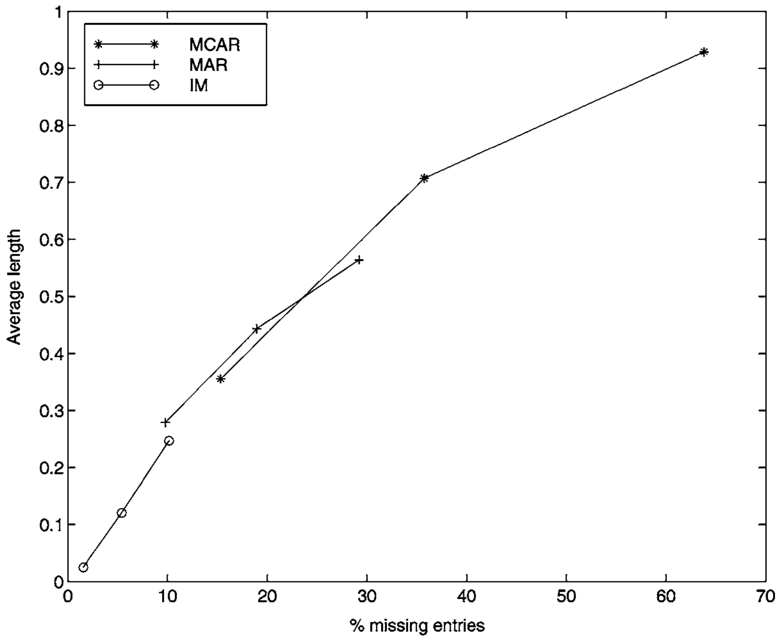


*Figure 5.* Average lengths of the probability intervals estimated with the RBE for different proportion of missing entries and missing data mechanisms.

width of the prediction intervals depends on the amount of missing data per variable. For example, when about 20% of data of all the variables are MCAR, the average width of the prediction intervals is smaller than the average width of the same prediction intervals when an equivalent amount of data is missing only for three variables.

The predictive accuracy of the networks quantified from the incomplete data sets is affected by the increasing uncertainty. Table 3 reports the first performance measure, which, as described in the previous section, is the average number of correct predictions in the nine groups of data sets. Gibbs sampling and the EM algorithm lead to the same results, so we report the common values under the heading GS-EM. RBEn is the average number of correct predictions in the networks quantified with the RBE when one adopts the weak dominance criterion with $q(X_6 < 140 \mid e) = n(X_6 < 140) / n(X_6)$. This criterion, when coupled with the MAR assumption made by both EM and Gibbs sampling, has 100% of correct predictions, which is superior to both methods. The stochastic dominance criterion leads to undecidability in all cases in the first two groups of incomplete data sets. When 25% of the entries of $X_5$ and $X_6$ are IM, the criterion yields the correct prediction in 30 cases, but the other 13 are undecidable.

Figure 7 plots median values for the second performance measure, which evaluates the precision of the predictive probabilities of $X_6$ computed with the weak dominance criterion and those induced by the networks quantified with EM and Gibbs sampling. When data are MAR, both methods determine more accurate predictive distributions compared to the weak dominance. However, this gain of accuracy is lost when data are MCAR and the proportion

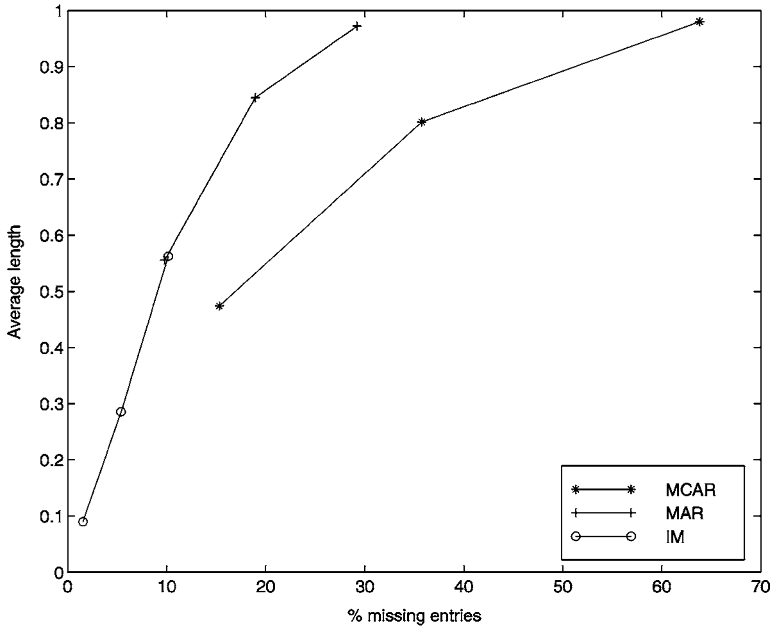


*Figure 6.* Average lengths of prediction intervals for different proportion of missing entries and different missing data mechanisms.

of missing data is either small or large. The probabilities of $X_6$, computed using the weak dominance criterion, become more accurate when data are IM, and all methods work under the assumption that data are MAR.

The experimental results point out the advantages and disadvantages of the RBE. When data are MCAR or MAR, both the EM algorithm and Gibbs sampling can use the information available in the data to compute accurate estimates of the conditional probabilities, while the RBE—provided with the same information about the missing data mechanism—cannot reach, in most cases, the same predictive precision, although its predictive accuracy appears

*Table 3.* Performance on the first measure for the three groups of conditions.

| Proportion missing | MCAR | | MAR | | IM | |
|---|---|---|---|---|---|---|
| | GS-EM | RBEn | GS-EM | RBEn | GS-EM | RBEn |
| 25% | 42.7 | 43.0 | 42.7 | 43.0 | 43.0 | 43.0 |
| 50% | 41.1 | 43.0 | 40.3 | 43.0 | 43.0 | 43.0 |
| 75% | 41.2 | 43.0 | 40.1 | 43.0 | 40.0 | 43.0 |

GS-EM is the average number of correct predictions in the networks quantified with Gibbs sampling and EM. RBEn is average number of correct prediction in the networks quantified with the RBE when one adopts the weak dominance criterion with $q(X_6 < 140 \mid e) = n(X_6 < 140)/n(X_6)$.

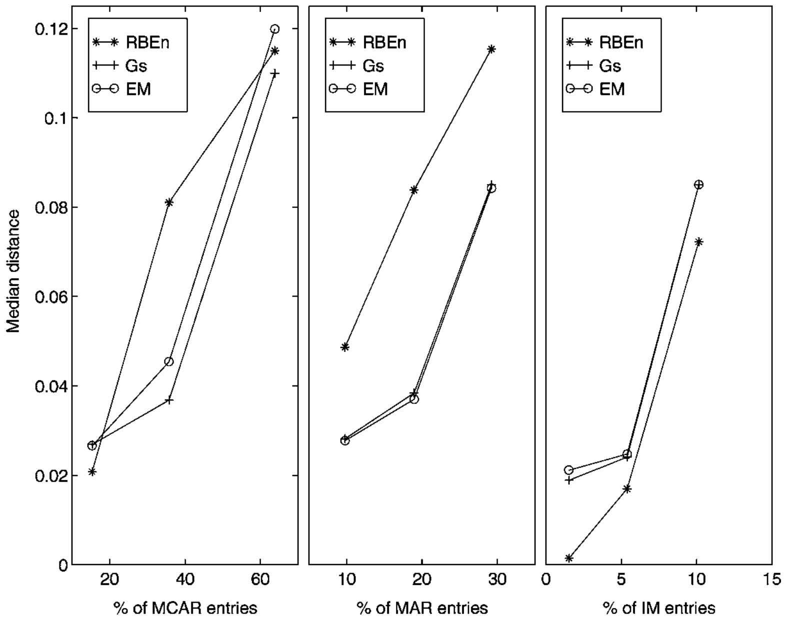


*Figure 7.* Median values of the cumulative Kullback-Liebler distance between values predicted by the RBE, Gibbs sampling, and EMA versus the average percentage of missing data.

to be superior. However, when data are IM and all methods make the same wrong assumption about the missing data mechanism, the RBE performs better in terms of both predictive accuracy and precision. The application that we describe in the next section will shed further light on the properties and usefulness of the RBE.

## 6. An application to a classification task

This section illustrates an application of the RBE to a classification task to show the gain of classification accuracy and coverage achieved by a classifier trained, with the RBE, on a naturally incomplete data set, and tested using classification rules that do not rely on assumptions about the missing data mechanism. The standard procedure, in this case, would be to not update the counts when a value is missing (Domingos & Pazzani, 1997; Friedman, Geiger, & Goldszmidt, 1997) or to assign the unknown entries to a dummy value (Quinlan, 1993). We describe, first, the data set and the statistical model used in this application, and we then show the advantages of using the RBE in a supervised learning task.

### *6.1. The data set*

We used data on Congressional Voting Records, available from the Machine Learning Repository at the University of California, Irvine (Blake, Keogh, & Merz, 1998). The data set describes votes for each of the 435 member of the US House of Representative on the 16 key issues during the 1984. Hence, the data set consists of 435 cases on 16 binary attributes and two classes that represent the party affiliation. There are 289 values reported as unknown. Although these missing entries amount to 4% of the data set, the number of incomplete cases is 203, more than 45% of the total. A feature of this data set is that the unknown entries, and hence what members of the US House of Representative did not vote on, can be predictive. Therefore, it makes sense to treat the missing entries as real values.

### 6.2. *The classification model*

The classification model we used was the naive Bayesian classifier (Langley, Iba, & Thompson, 1992), shown in figure 8, in which the 16 key votes are represented as Boolean attributes $\{A_1, \ldots, A_{16}\}$ that take on values $a_{ik}$ either *yes* or *no*. All attributes $A_i$ are treated as conditionally independent given the class variable $C$, which represents the party affiliation: either *Republican* or *Democratic*. With complete data, the training step consists of estimating the conditional probability distributions for each attribute $A_i$, given the class membership $C = c_j$, using Eq. (2). As shorthand, we write $A_i = a_{ik}$ as $a_{ik}$ and $C = c_j$ as $c_j$. Once the classifier is trained, we can use it to classify new cases. In this example, the classification step consists of the identification of the party affiliation of a Congressman given the set of his/her votes on each of the 16 issues. A standard application of Bayes' theorem lets us calculate the posterior probability of the party affiliation $C = c_j$ given the set of attribute values $e_k = \{A_1 = a_{1k}, \ldots, A_{16} = a_{16k}\}$ as

$$p(c_j \mid e_k) = \frac{\prod_{k=1}^{16} p(a_{ik} \mid c_j) p(c_j)}{\sum_{h=1}^{2} \prod_{k=1}^{16} p(a_{ik} \mid c_h) p(c_h)}, \tag{7}$$

and the case is assigned to the party with the highest posterior probability. Since the data set contains unknown attributes values, we used the RBE to train the classifier. As the estimates of

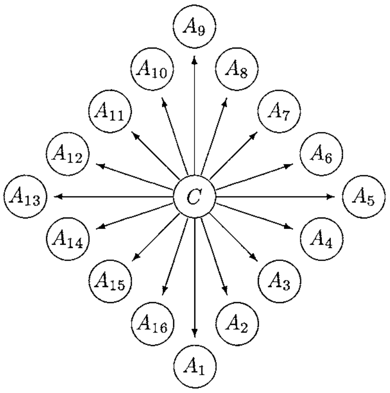


*Figure 8.* Structure of the Bayesian classifier used on the "voting records" data set.

the conditional probabilities $p(a_{ik} \mid c_j)$ are the probability intervals $[\underline{p}(a_{ik} \mid c_j)\bar{p}(a_{ik} \mid c_j)]$, we cannot simply apply Eq. (7) to compute the posterior probability $p(c_j \mid e)$. However, we can still iteratively compute the probability interval $[\underline{p}(c_j \mid e_k)\bar{p}(c_j \mid e_k)]$ containing the posterior probability $p(c_j \mid e_k)$. Since the class variable is Boolean, it is sufficient to compute $\bar{p}(c_1 \mid e_k)$ and $\bar{p}(c_2 \mid e_k)$, from which we can compute $\underline{p}(c_1 \mid e_k) = 1 - \bar{p}(c_2 \mid e_k)$ and $\underline{p}(c_2 \mid e_k) = 1 - \bar{p}(c_1 \mid e_k)$, as noted in Section 3.2. The initialization step sets

$$\bar{p}(c_j \mid a_{1k} = \frac{\bar{p}(a_{1k} \mid c_j)p(c_j)}{\bar{p}(a_{1k} \mid c_j)p(c_j) + \underline{p}(a_{1k} \mid c_h)p(c_h)} \tag{8}$$

for $h \neq j$. Thus, after the first incorporation of evidence, the point-valued probabilities $p(c_j)$ are replaced by interval probabilities and the next iteration steps set

$$\begin{aligned}&\bar{p}(c_j \mid a_{1k}, \ldots, a_{ik})\\ &\quad = \frac{\bar{p}(a_{ik} \mid c_j)\bar{p}(c_j \mid a_{1k}, \ldots, a_{ik-1})}{\bar{p}(a_{ik} \mid c_j)\bar{p}(c_j \mid a_{1k}, \ldots, a_{ik-1}) + \underline{p}(a_{ik} \mid c_h)\underline{p}(c_h \mid a_{1k}, \ldots, a_{ik-1})}\end{aligned}$$

for $h \neq j$ and each attribute value. Since Formula 8 can be applied iteratively, the classifier retains the propagation properties of a standard Bayesian classifier and enjoys the same low time and memory requirements. When attributes are not binary or more than two classes are involved, however, more general methods must be used to apply Bayes' theorem to probability intervals (Fertig & Breese, 1993; Snow, 1991).

A further difference between a standard Bayesian classifier and one trained with the RBE lies in the class assignment criterion. A standard scheme assigns a case to the class with the highest posterior probability. A classifier trained using the RBE would assign, unequivocally, a case to the class $c_j$ only when the stochastic dominance criterion is met, and hence $\underline{p}(c_j \mid e_k) > \bar{p}(c_h \mid e_k)$ for $h \neq j$. When this condition does not hold, we can resort to the weak dominance criterion described Section 3.3. In this case, we assign the score $s_q(c_j \mid e_k) = \underline{p}(c_j \mid e_k)(1 - q(c_j)) + \bar{p}(c_j \mid e_k)q(c_j)$ to $c_j \mid e_k$ and then select the class with the highest score. If we do not want to commit ourselves to any particular assumption about the missing data mechanism, we can use the uniform distribution $q(c_1) = q(c_2) = 0.5$, yielding the score $s_u(c_j \mid e_k)$.

### 6.3. *Evaluation of the method*

We assessed the accuracy of the learned classifier by running 20 replicates of a five-fold *cross validation* experiment. We divided the data set $\mathcal{D}$ in five mutually exclusive data sets $\mathcal{D}_1, \ldots, \mathcal{D}_5$ of approximately the same size. For each data set $\mathcal{D}_i$, we trained the classifier on $\mathcal{D}$ with the cases in $\mathcal{D}_i$removed, then tested it on $\mathcal{D}_i$. The test step was the classification of cases in $\mathcal{D}_i$, in which the classifier returned the class selected, using either the stochastic dominance criterion or the score $s_u$. We computed the *classification accuracy* as the average number of cases that were classified correctly in the five test sets. We calculated the *classification coverage* as the ratio between the number of cases classified and the total number of cases. We used both measures were used to evaluate the gain of classification accuracy and coverage compared to the standard Bayesian classifier, in which the missing entries are either ignored or assigned to a dummy value. We repeated the whole procedure 20 times and averaged the outcomes.

### 6.4. *Results and discussion*

We measured the classification accuracy and coverage obtained by a classifier trained with the RBE and tested using both the stochastic dominance criterion (NBCs) and the weak dominance criterion (NBCw). For comparison, we collected the same measure of a classifier trained by assigning the missing entries to dummy values (NBC*) and by disregarding the missing entries when updating the counts (NBCm). Table 4 summarizes the results. The average width of conditional probability intervals estimated by the RBE is 0.053, which let the stochastic dominance criterion achieve 95% coverage. The weak dominance criterion left none of the cases unclassified, which raises the classification coverage to 100%. The average classification accuracy of the classifier, under the stochastic dominance criterion, is 92.05 ± 1.67%, and it is bounded by the two extreme situations in which the unclassified cases are all regarded as predictive failures (87.56% accuracy) or success (93.09% accuracy). If we use the classification score under the weak dominance criterion, the classifier reaches 90.21 ± 1.7% accuracy. This result is similar to the accuracy of the classifier trained by assigning the missing entries to dummy values and it is slightly superior to the accuracy of the classifier trained by disregarding the missing entries.

Strong dominance achieves the highest accuracy at the price of the lowest coverage and identifies the group of 95% of cases on which there is no classification ambiguity due to the missing values in the database. In particular, all other classifiers achieve the same accuracy on this 95% of cases. The classification of the remaining 5% of cases, by using the weak dominance criterion or the other two classifiers NBC* and NBCm, raises the coverage to 100% at the price of decreasing the accuracy. Now note that the classification accuracy $\theta$ of the NBC*, the NBCm, or the NBCw can be written as

$$\theta = \theta_s \gamma_s + \theta_l (1 - \gamma_s),$$

where $\theta_s$ and $\gamma_s$ are accuracy and coverage of the classifier NBCs and $\theta_l$ is the accuracy of the other classifiers on the cases left unclassified by the NBCs. Thus, the quantity $\theta_l$ gives

*Table 4*. Classification coverage and accuracy of the naive Bayesian classifier on the Congressional voting data.

| Classification | NBCm | NBC* | NBCw | NBCs |
|---|---|---|---|---|
| Accuracy | 90.02 ± 1.05% | 90.21 ± 1.04% | 90.21 ± 1.04% | 92.05 ± 1.67% |
| Coverage | 100% | 100% | 100% | 95% |

NBCm is the classifier trained by disregarding the missing entries during the calculation of the counts. NBC* is the classifier trained by assigning the missing entries to dummy values. NBCw and NBCs are, respectively, the classifier trained with the RBE and tested using the weak dominance criterion and stochastic dominance criterion.

a measure of the classification accuracy achieved by the other classifiers when one relaxes the strong dominance criterion to increase the classification coverage. By using the values in Table 4, we compute $\theta_l = 0.55$ for both the NBC* and the NBCw and $\theta_l = 0.51$ for the NBCm. Hence, disregarding the missing entries leads essentially to randomly classifying those cases that were left unclassified by the NBCs, while both the NBC* and NBCw are slightly more accurate than a simple random assignment to classes. This last finding would suggest evidence of an informative pattern of missing data and discourage the enforcement of an inappropriate assumption on the missing data mechanism to merely increase the classification coverage. In summary, strong dominance identifies the subset of cases that are unequivocally classified, independently of the missing data mechanism, and the subset of cases whose classification is directly influenced by the assumed missing data mechanism. One can then use the accuracy and coverage of the classification based on the strong dominance criterion to derive an estimate of the classification accuracy obtained under the assumed missing data mechanism and therefore evaluate the impact of this assumption on classification.

## 7. Conclusions

Real-world data sets are often incomplete, and machine learning methods must be able to handle incomplete data before they can be widely applied. A key issue for such learning methods is the reliability of the knowledge they generate. This paper introduced a robust method, the RBE, to estimate conditional probabilities. Compared to traditional Bayesian estimation methods, the RBE relies on a new strategy: rather than guessing the value of missing data on the basis of the available information, it bounds the set of all estimates consistent with the data. The estimates computed by the RBE are, therefore, probability intervals containing all possible estimates that could be computed from the consistent completions of the database and, as such, they are robust with respect to the distribution of missing data. Hence, one feature of the RBE is to provide an automated method to analytically perform sensitivity analysis, at a low computational cost, with respect to different assumptions on the missing data mechanism.

Efficient sensitivity analysis is not, however, the only feature of the RBE. The probability intervals it computes can be used, as they are, to draw robust inferences. Furthermore, these intervals provide a measure of the information conveyed by the data and can be used to assess the reliability of the point-valued estimates computed by other learning methods, like EM and Gibbs sampling, that are based on specific assumptions about the missing data mechanism. The experimental evaluation presented in this paper showed that when data are missing at random or completely at random, both EM and Gibbs sampling can exploit the information provided by the observed data to return more precise inferences, although the predictive accuracy of the RBE appears to be superior. However, when the deletion process produces data that are informatively missing, the RBE is more reliable in terms of both accuracy and precision.

The application described in Section 6 showed another potential use of the method introduced in this paper. Interval-based classification breaks the database in two sets: the set of those cases that can be classified independently of any assumption made about the missing data mechanism, and the set of cases that can be classified only by assuming a particular mechanism. One can then use the accuracy and coverage of the interval-based classification to derive an estimate of the classification accuracy obtained under the assumed missing data mechanism and therefore evaluate the impact of this assumption on the overall classification task. Interval-based classification seems to be a promising area of application of the RBE, and a systematic investigation of this problem could lead to the development of a robust Bayes classifier, able to couple the computational efficiency of standard naive Bayes with the ability to handle incomplete databases with no assumption about the missing data mechanism. A more ambitious avenue would involve applying the same interval-based approach to other machine learning tasks, such as clustering, dependency discovery, and the identification of hidden variables.

## Acknowledgments

The authors thank Greg Cooper, Bo Thiesson, and Zdenek Zdrahal, and the anonymous referees for their useful suggestions. They are grateful to Pat Langley for the painstaking efforts spent on the earlier versions of this paper. This research was supported, in part, by the ESPRIT program of the Commission of the European Communities under contract EP29105 and by equipment grants from Apple Computer and Sun Microsystems.

## Note

1. One can show that, given the structure of conditional independence in the network, there are only 43 combinations of values of the variables $X_1, \ldots, X_5$ that yield different conditional probabilities of $X_6 < 140$.

## References

Blake, C., Keogh, E., & Merz, C. J. (1998). *UCI Repository of machine learning databases*. Department of Information and Computer Sciences, University of California, Irvine, CA.

Castillo, E., Gutierrez, J. M., & Hadi, A. S. (1997). *Expert systems and probabilistic network models*. New York, NY: Springer.

Cheeseman, P. & Stutz, J. (1996). Bayesian classification (AutoClass): Theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 153–180). Cambridge, MA: MIT Press.

Cooper, G. F. & Herskovitz, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309–347.

Dempster, A. P., Laird, D., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, B, 39*, 1–38.

Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103–130.

Fertig, K. W. & Breese, J. S. (1993). Probability intervals over influence diagrams. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15*, 280–286.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* (pp. 1277–1284). San Francisco, CA: Morgan Kaufmann.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*, 131–163.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721–741.

Good, I. J. (1968). *The estimation of probability: An essay on modern bayesian methods*. Cambridge, MA: MIT Press.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combinations of knowledge and statistical data. *Machine Learning, 20*, 197–243.

Kyburg, H. E. (1983). Rational belief. *Behavioral and Brain Sciences, 6*, 231–273.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. In *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 223–228). Menlo Park, CA: AAAI Press.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis, 19*, 191–201.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Ramoni, M. (1995). Ignorant influence diagrams. In *Proceedings of the Fortheenth International Joint Conference on Artificial Intelligence* (pp. 1808–1814). San Francisco, CA: Morgan Kaufmann.

Ramoni, M., Riva, A., Stefanelli, M., & Patel, V. (1995). An ignorant belief network to forecast glucose concentration from clinical databases. *Artificial Intelligence in Medicine, 7*, 541–559.

Ramoni, M. & Sebastiani, P. (1998). Parameter estimation in Bayesian networks from incomplete databases. *Intelligent Data Analysis Journal, 2*, 139–160.

Ramoni, M. & Sebastiani, P. (1999). Bayesian methods. In M. Berthold & D. J. Hand (Eds.), *Intelligent data analysis. An introduction* (pp. 29–166). New York, NY: Springer.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.

Russell, S., Binder, J., Koller, D., & Kanazawa, K. (1995). Local learning in probabilistic networks with hidden variables. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (pp. 1146–1151). San Francisco, CA: Morgan Kaufmann.

Snow, P. (1991). Improved posterior probability estimates from prior and linear constraint system. *IEEE Transactions on Systems, Man, and Cybernetics, 21*, 464–469.

Spiegelhalter, D. J. & Cowell, R. G. (1992). Learning in probabilistic expert systems. In *Bayesian statistics 4* (pp. 447–466). Oxford, UK: Oxford University Press.

Thiesson, B. (1995). Accelerated quantification of Bayesian networks with incomplete data. In *Proceedings of the First ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 306–311). New York, NY: ACM Press.

Thomas, A., Spiegelhalter, D. J., & Gilks, W. R. (1992). Bugs: A program to perform Bayesian inference using Gibbs Sampling. In *Bayesian statistics 4* (pp. 837–42). Oxford, UK: Oxford University Press.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. New York, NY: Wiley.

Zhang, N. L. (1996). Irrelevance and parameter learning in Bayesian networks. *Artificial Intelligence, 88*, 359–373.

Received
Revised
Accepted June 1, 2000
Final Manuscript September 5, 2000