KMi

KNOWLEDGE MEDIA INSTITUTE

Learning Bayesian Networks from Incomplete Databases

Marco Ramoni   *Paola Sebastiani*

KMI-TR-43

February 1997

The Open University

Learning Bayesian Networks from Incomplete Databases

Marco Ramoni
Knowledge Media Institute
The Open University

Paola Sebastiani
Department of Actuarial Science and Statistics
City University

Abstract

Bayesian approaches to learn the graphical structure of Bayesian Belief Networks (BBNs) from databases share the assumption that the database is complete, that is, no entry is reported as unknown. Attempts to relax this assumption often involve the use of expensive iterative methods to discriminate among different structures. This paper introduces a deterministic method to learn the graphical structure of a BBN from a possibly incomplete database. Experimental evaluations show a significant robustness of this method and a remarkable independence of its execution time from the number of missing data.

Keywords: UNCERTAINTY AND METHODS FOR LEARNING AND DATA MINING: Bayesian Belief Networks, Bayesian Learning, Missing Data, Model Selection.

Reference: KMi Technical Report KMi-TR-43 (February 1997).

Address: Marco Ramoni, Knowledge Media Institute. The Open University. Walton Hall, Milton Keynes, United Kingdom MK7 6AA. PHONE: +44 (1908) 655721, FAX: +44 (1908) 653169, EMAIL: `M.Ramoni@open.ac.uk`, URL: `http://kmi.open.ac.uk/~marco`.

1. Introduction

A Bayesian Belief Network (BBN) [7] is direct acyclic graph where nodes represent stochastic variables and arcs represent conditional dependencies among these variables. A conditional dependency links a *child* variable to a set of *parent* variables and it is defined by the set of conditional distribution of the child variable given a combination of states of its parent variables. Although in their original concept BBNs were designed to rely on human experts to define the graphical structure and assess the conditional probabilities defining a BBNs, during the past few years, an increasing number of efforts has been addressed toward the development of methods able to directly construct BBNs from databases of cases, rather than from the insight of human experts. Early results in this quest toward an efficient method to learn BBNs from databases were based on non Bayesian approaches [11], but almost immediately a seminal paper by Cooper and Herskovitz [4] gave rise to a stream of research within a Bayesian framework [1, 5]. Along this second approach, there are two main tasks involved in the learning process of a BBN from a database: the induction of the graphical model best fitting the database at hand, and the extraction of the conditional probabilities defining the dependencies in a given graphical model.

Methods to perform the first of these tasks, known as *model selection*, typically involve two components: a search procedure to explore the space of possible graphical models and a score metric to assess the goodness-of-fit of a particular model. Current approaches exploit heuristics to reduce the search space and use the scoring metric to drive the search process. Although the process of extracting BBNs from databases in known to be NP-Hard for the general case [2], under certain assumptions, these methods are able to extract quite large BBNs from databases of thousands of cases. One of these assumptions is that the database is *complete*, that is, no entry in the database is unknown. The presence of *hidden variables* in the model is a special case of this situation, where all the observations about a particular variable are not reported in the database. The reason for this assumption is that a key step in the Bayesian learning process is the computation of the marginal likelihood of the database given a graphical model. This computation can be performed efficiently when the database is complete using exact Bayesian updating, but it becomes intractable when data are missing in the database, and therefore, methods to approximate the marginal likelihood of data have to be used. Well-known methods typically involve the use of the EM algorithm or Markof Chain Monte Carlo methods, such as Gibbs sampling [3]. The basic strategy underlying these methods is based on the *Missing Information Principle* [6]: fill in the missing observations on the basis of the available information. EM performs this task by replacing the missing entries with the maximum likelihood estimates extracted from the available data and proceeds by iteratively estimating and replacing until stability is reached within a certain threshold. On the other hand, the Gibbs Sampling generates a value for the missing data from some conditional distributions and provides a stochastic estimation of the posterior probabilities. Unfortunately, these processes are usually highly resource demanding, their convergence rates may be slow, and their execution time heavily depends on the number of missing data.

Ramoni and Sebastiani [9] introduced a deterministic method to estimate the conditional probabilities defining the dependencies in a BBN which does not rely on the Missing Information Principle. This method, called *Bound* and *Collapse* (BC), starts by *bounding* the set of possible estimates consistent with the available observations in the database and then *collapses* the resulting interval estimate to a point estimate via a convex combination of the extreme estimates with weights depending on the assumed pattern of missing data. The pattern of missing data may be either provided by an external source of information or may be estimated from the available information under the assumption that data are missing at random. Experimental evaluations show clearly that the estimates provided by BC are equivalent to the ones provided by the Gibbs Sampling, when data are missing at random, and they are more robust to departure from the true pattern of missing data. On the other hand, the computational cost of BC is reduced to the cost of two exact Bayesian updating — one for each extreme distribution — plus the cost of a convex combination for each parameter in the BBN. Experiments comparing BC to the Gibbs Sampling have shown that the use of BC reduces the execution time of several order of magnitudes.

This paper describes how BC can be used to estimate the marginal likelihood of a database given a model and how the principle underlying BC can be extended from the task of learning the conditional probabilities of a BBN to the task of extracting the graphical model from an incomplete database. The reminder of this paper is structured as follows: Section 2 introduces the technical *background* of the research, Section 3 describes the new *method*, Section 4 outlines the description of an implemented *system* based on this method, Section 5 reports some some results from a preliminary *experimental evaluation* of this system, and finally Section 6 summarizes the relevant results.

2. Background

A BBN is defined by a set of *variables* $\mathcal{X} = \{X_1, \ldots, X_I\}$ and a direct acyclic graph defining a model $\mathcal{M}$ of conditional dependencies among the elements of $\mathcal{X}$. A conditional dependency links a *child* variable X_i to a set of *parent* variables Π_i, and it is defined by the conditional distribution of the child variable given a configuration of the parent variables. We shall consider discrete variables only, and denote by c_i the number of states of X_i, and q_i the number of states of Π_i. The model $\mathcal{M}$ yields a factorization of the joint probability of a particular set of values of the variables in $\mathcal{X}$, say $x_k = \{x_{1k}, \ldots x_{Ik}\}$, as

$$p(\mathcal{X} = x_k|\mathcal{M}) = \prod_{i=1}^{I} p(X_i = x_{ik}|\Pi_i = \pi_{ij}, \mathcal{M}), \tag{1}$$

where π_{ij} denotes the state of Π_i in x_k. In the following we will denote $X_i = x_{ik}$ as x_{ik}, and $\Pi_i = \pi_{ij}$ as π_{ij}.

Suppose we are given a database of n cases $\mathcal{D} = \{x_1, \ldots, x_n\}$ from which we wish to select a model $\mathcal{M}$ of conditional dependencies among the variables in the database. We adopt a Bayesian approach, so that if $p(\mathcal{M})$ is our prior belief about a particular model $\mathcal{M}$, we can use the information in the database $\mathcal{D}$ to compute the posterior probability of $\mathcal{M}$ given the data:

$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{M}, \mathcal{D})}{p(\mathcal{D})},$$

and then we choose the model which has the highest posterior probability. When the comparison is between two rival models $\mathcal{M}_1$ and $\mathcal{M}_2$, this is equivalent to choose $\mathcal{M}_1$ if the Bayes factor:

$$\frac{p(\mathcal{M}_1, \mathcal{D})}{p(\mathcal{M}_2, \mathcal{D})},$$

is greater than one. Clearly the valuation can be done independently of the marginal probability $p(\mathcal{D})$, by just considering the joint probability of the model and the data: $p(\mathcal{M}, \mathcal{D})$. It is well known [4], that $p(\mathcal{M}, \mathcal{D})$ can be easily computed if the conditional probabilities defining $\mathcal{M}$ are regarded as random variables θ_{ijk} whose prior distribution represents the observer's beliefs before seen any data. The joint probability of a case x_k can then be written in terms of the random vector $\theta = \{\theta_{ijk}\}$ as:

$$p(x_k|\theta) = \prod_{i=1}^{I} \theta_{ijk}.$$

This parameterization of the probabilities defining $\mathcal{M}$ allows us to write:

$$p(\mathcal{M}, \mathcal{D}) = \int p(\mathcal{M}, \mathcal{D}, \theta) d\theta = p(\mathcal{M}) \int p(\theta|\mathcal{M}) p(\mathcal{D}|\theta) d\theta \qquad (2)$$

where $p(\theta|\mathcal{M})$ is the prior density of θ, and $p(\mathcal{D}|\theta)$ is the sampling model. The integral (2) is defined on the parameter space whose dimension depends on the complexity of the network. A solution in closed form exists if:

1. *The database is complete;*
2. *The cases are independent, given the parameter θ associated to $\mathcal{M}$;*
3. *The prior distribution of the parameters is conjugate to the sampling model $p(\mathcal{D}|\theta)$;*
4. *The parameters are marginally independent.*

Details of the calculations can be found for instance in [4]. Let $n(x_{ik}|\pi_{ij})$, $i = 1, ..., I; j = 1, ..., q_i; k = 1, ..., c_i$, be the frequency of cases in the database with $x_{ik}|\pi_{ij}$, so that $n(\pi_{ij}) = \sum_{k=1}^{c_i} n(x_{ik}|\pi_{ij})$ is the frequency of cases with π_{ij}. Assumptions 1 and 2 lead to

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{I} \prod_{j=1}^{q_i} \prod_{k=1}^{c_i} \theta_{ijk}^{n(x_{ik}|\pi_{ij})}.$$

A prior distribution on the parameters that satisfies 3 and 4 is a product of Dirichlet distributions. Thus if we denote by $\theta_{ij} = (\theta_{ij1}, ..., \theta_{ijc_i})$ the vector of parameters associated to the conditional distribution of $X_i|\pi_{ij}$, we have

$$\theta_{ij} \sim D(\alpha_{ij1}, ..., \alpha_{ijc_i}).$$

The prior hyper-parameters α_{ijk}s can be regarded as frequencies of imaginary cases needed to formulate the prior distribution, in fact the marginal probability of $x_{ik}|\pi_{ij}$ is α_{ijk}/α_{ij}, and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ is the prior precision on θ_{ij}. It is well known [10] that, under the assumptions 1 – 4, the posterior distribution of θ is still a product of Dirichlet distributions, and

$$\theta_{ij}|\mathcal{D} \sim D(\alpha_{ij1} + n(x_{i1}|\pi_{ij}), ..., \alpha_{ijc_i} + n(x_{ic_i}|\pi_{ij})).$$

Thus an estimate of $p(x_{ik}|\pi_{ij})$ is the posterior expectation of θ_{ijk}:

$$\frac{\alpha_{ijk} + n(x_{ijk}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij})},$$

and the posterior precision on θ_{ij} is $\alpha_{ij} + n(\pi_{ij})$. Furthermore, the integral (2) has the solution:

$$p(\mathcal{M}, \mathcal{D}) = p(\mathcal{M}) \prod_{i=1}^{I} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n(\pi_{ij}))} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ijk} + n(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}, \tag{3}$$

where

$$p(\mathcal{D}|\mathcal{M} = \prod_{i=1}^{I} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n(\pi_{ij}))} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ijk} + n(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}$$

is the marginal likelihood of $\mathcal{D}$ given $\mathcal{M}$. Thus, $p(\mathcal{D}|\mathcal{M})$ depends on the updated hyper-parameters of $\theta_{ij}|\mathcal{D}$, and the posterior precision on θ_{ij}. If we assume a uniform prior on the parameters, so that $\alpha_{ijk} = 1$ for all ijk, then

$$p(\mathcal{M}, \mathcal{D}) = p(\mathcal{M}) \prod_{i=1}^{I} \prod_{j=1}^{q_i} \frac{(c_i - 1)!}{(c_i + n(\pi_{ij}) - 1)!} \prod_{k=1}^{c_i} n(x_{ik}|\pi_{ij})!. \tag{4}$$

Note that

$$\prod_{i=1}^{I} \prod_{j=1}^{q_i} \frac{(c_i + n(\pi_{ij}) - 1)!}{(c_i - 1)! \prod_{k=1}^{c_i} n(x_{ik}|\pi_{ij})!} \tag{5}$$

is the number of ways in which the frequencies could have been observed by changing the order of the entries in the database, so that $p(\mathcal{M}, \mathcal{D})$ is proportional to the reciprocal of the number of possible databases that could give rise to the observed counts.

The probability (3) is the base for an algorithm proposed by [4] to induce the model from a database. Suppose we have a partial order on the variables so that $X_i \prec X_j$ if X_i cannot be parent of X_j. Let $\mathcal{P}_i$ be the set of current parents of X_i, thus $\mathcal{P}_i$ is the empty set if X_i is a root node. Then the local contribution of a node X_i and its parents Π_{ij} to the joint probability of $(\mathcal{M}, \mathcal{D})$ is measured by

$$g(X_i, \mathcal{P}_i) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n(\pi_{ij}))} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ijk} + n(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}. \tag{6}$$

The algorithm proceeds by adding a parent at a time and computing $g(X_i, \mathcal{P}_i)$. The set $\mathcal{P}_i$ is expanded to include the parent nodes that give the largest contribution to $g(X_i, \mathcal{P}_i)$, and stops if the probability does not increase any longer. When the database is complete, the joint probability of $(\mathcal{M}, \mathcal{D})$, as given in (3), depends on the counts $n(x_{ik}|\pi_{ij})$ and the hyper-parameters α_{ijk}. Suppose now that we are given an incomplete database, say $\mathcal{D}_{inc} = \mathcal{D}_{obs} \cup \mathcal{D}_{mis}$, where $\mathcal{D}_{mis}$ denotes the part of $\mathcal{D}_{inc}$ with missing entry. The exact probability of $(\mathcal{M}, \mathcal{D}_{inc})$ is

$$p(\mathcal{M}, \mathcal{D}_{inc}) = \sum_{com} p(\mathcal{M}, \mathcal{D}_{inc}, \mathcal{D}_{com}) = \sum_{com} p(\mathcal{D}_{inc}) p(\mathcal{M}, \mathcal{D}_{com}|\mathcal{D}_{inc})$$

where the sum is over all possible complete databases $\mathcal{D}_{com}$ consistent with the available data. Clearly, as the number of missing entries increases, the exact calculation of $p(\mathcal{M}, \mathcal{D}_{inc})$ is infeasible, and some approximation is needed.

3. Method

In this section we will show that it is possible to approximate the hyper-parameters of the posterior distributions of θ_{ij}, from which an estimate of (3) can be obtained. The method consists of two steps:

Step 1: Estimate the posterior expectation of θ_{ijk}, for all ijk, using BC.

Step 2: Estimate the posterior precision of θ_{ij}, for all ij.

Let $\hat{p}_{ijk}$ be an estimate of the posterior expectation of θ_{ijk}, and let $\hat{\alpha}_{ij}$ be an estimate of the posterior precision of θ_{ij}. Then the distribution $D(\hat{\alpha}_{ij}\hat{p}_{ij1}, ..., \hat{\alpha}_{ij}\hat{p}_{ijc_i})$ will have precision $\hat{\alpha}_{ij}$ and expectations $\hat{p}_{ijk}$, so that the posterior distribution of θ_{ij} can be approximated to

$$\theta_{ij}|\mathcal{D} \sim D(\hat{\alpha}_{ij}\hat{p}_{ij1}, ..., \hat{\alpha}_{ij}\hat{p}_{ijc_i}). \tag{7}$$

From (7) we can then derive an estimate of (3):

$$\hat{p}(\mathcal{M}, \mathcal{D}) = p(\mathcal{M}) \prod_{i=1}^{I} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\hat{\alpha}_{ij})} \prod_{k=1}^{c_i} \frac{\Gamma(\hat{\alpha}_{ij}\hat{p}(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}, \tag{8}$$

which can be also used to extend the algorithm in [4] to the induction of Bayesian networks from incomplete databases. In fact the local contribution of a node X_i and its parents Π_{ij} to the joint probability of $(\mathcal{M}, \mathcal{D})$ defined for the complete databases by 6 can be estimated as

$$\hat{g}(X_i, \mathcal{P}_i) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\hat{\alpha}_{ij})} \prod_{k=1}^{c_i} \frac{\Gamma(\hat{\alpha}_{ij}\hat{p}(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}. \tag{9}$$

3.1 Step 1: Posterior Expectation

The BC method proposed by [9] is a technique to estimate the conditional probabilities defining a BBN from an incomplete database. Suppose we have a model $\mathcal{M}$ of conditional dependencies, specifying for each X_i the parent variable Π_{ij}. The method starts by bounding the set of possible posterior distributions of θ_{ij} consistent with the database, and then

case	X_1	X_2	X_3
x_1	1	2	2
x_2	2	?	1
x_3	?	1	2
x_4	?	?	1
x_5	1	?	?

$$\Rightarrow \quad \begin{array}{ll} n^\bullet(x_{31}|(1,1)) = 2 & n^\bullet(x_{31}|(1,2)) = 2 \\ n^\bullet(x_{31}|(2,1)) = 2 & n^\bullet(x_{31}|(2,2)) = 2 \\ n^\bullet(x_{32}|(1,1)) = 2 & n^\bullet(x_{32}|(1,2)) = 1 \\ n^\bullet(x_{32}|(2,1)) = 1 & n^\bullet(x_{32}|(2,2)) = 0 \end{array}$$

Table 1: Completions $n^\bullet(x_{3k}|x_1, x_2)$ consistent with the incomplete database.

collapses the extreme distributions in one single Dirichlet by taking into account the assumed pattern of missing data.

Let $n^\bullet(x_{ik}|\pi_{ij})$ be the frequency of cases with $X_i = x_{ik}$, given the parent configuration π_{ij}, which have been obtained by completing incomplete cases. These completions can be due either to an incomplete observation of the parent configuration, or to an incomplete observation of the variable X_i itself, or both. An example is given in Table 1 for the model

$$\begin{array}{ccc} X_1 & & \\ & \searrow & \\ X_2 & \longrightarrow & X_3 \end{array} \qquad X_i \quad \text{binary}, \quad i = 1, 2, 3.$$

For each incomplete case in the database, let ϕ_{ijk} be the probability of a completion:

$$\phi_{ijk} = p(x_{ik}|\pi_{ij}, X_i =?). \tag{10}$$

Suppose further that data are missing at random, so that $\mathcal{D}_{obs}$ is a *representative sample* of the complete database $\mathcal{D}$. Thus the probability of a completion can be estimated from $\mathcal{D}_{obs}$ as

$$\hat{\phi}_{ijk} = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij})}.$$

Then an estimate of $E(\theta_{ik}|\mathcal{D}_{inc})$ is given by:

$$\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijk}) = \sum_{l \neq k} \hat{\phi}_{ijl} p_{l\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}) + \hat{\phi}_{ijk} p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}), \tag{11}$$

where

$$p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij}) + n^\bullet(x_{ik}|\pi_{ij})}{\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij}) + n^\bullet(x_{ik}|\pi_{ij})} \tag{12}$$

and

$$p_{l\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij}) + n^\bullet(x_{il}|\pi_{ij})}. \tag{13}$$

The value in (12) is the upper bound of $p(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$, which is achieved when all incomplete cases in the database which could be completed as $x_{ik}|\pi_{ij}$ are assigned to $x_{ik}|\pi_{ij}$, and the other incomplete cases are assigned to $x_{ih}|\pi_{il}$, any h, and $l \neq j$. Thus each maximum probability $p^{\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$ is obtained from a Dirichlet distribution

$$D_k(\alpha_{ij1} + n(x_{i1}|\pi_{ij}), \ldots, \alpha_{ijk} + n(x_{ik}|\pi_{ij}) + n^{\bullet}(x_{ik}|\pi_{ij}), \ldots, \alpha_{ijc_i} + n(x_{ic_i}|\pi_{ij}))$$

which identifies a unique probability $p_{k\bullet}(x_{il}|\pi_{ij}, \mathcal{D}_{inc})$ for the other states of the variable X_i given π_{ij} from which $p_{l\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$ is obtained. The estimates $\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijk})$, $k = 1, ..., c_i$, so found define a probability distribution since $\sum_{k=1}^{c_i} \hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijk}) = 1$.

As the number of missing entries in $\mathcal{D}_{inc}$ decreases, $p^{\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$ and $p_{l\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$ approach $(\alpha_{ijk} + n(x_{ijk}|\pi_{ij}))/(\alpha_{ij} + n(\pi_{ij}))$ so that, when the database is complete, (11) returns the exact estimate $E(\theta_{ijk}|\mathcal{D}_{inc})$. As the number of missing entries increases then both $\hat{\phi}_{ijk}$ and the estimate in (11) approaches the prior probability α_{ijk}/α_{ij}, so that the estimation method is coherent: no updating is performed when data are totally missing.

If $n^{\bullet}(x_{ik}|\pi_{ij}) = n^{\bullet}_{ij}$, as for instance when data are missing only on the child variable, then (11) simplifies to

$$\frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij}) + \hat{\phi}_{ijk} n^{\bullet}_{ij}}{\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij}) + n^{\bullet}_{ij}}, \tag{14}$$

which is an estimate of the expected Bayesian estimate given $\mathcal{D}_{obs}$ and the assumption that data are missing at random, that is:

$$\frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij}) + n^{\bullet}_{ij} p(x_{ik}|\pi_{ij}, \theta)}{\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij}) + n^{\bullet}_{ij}}.$$

A remarkable property is that, if $\alpha_{ijk} = 0$, then 14 is the maximum likelihood estimate of θ_{ijk} [6]. The method is not limited to the assumption that data are missing at random. For instance, when no information on the mechanism generating the missing data is available and therefore any pattern of missing data is equally likely, then $\hat{\phi}_{ijk} = 1/c_i$. Experimental comparisons [9] have shown that, when data are missing at random, the estimates computed by the BC method are equivalent to those obtained using stochastic methods based on the Missing Information Principle, as the Gibbs Sampling, are are more robust to departures from the true pattern of missing data.

Furthermore, note that the extreme probabilities (13) lead to a lower bound of $p(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$, that is $p_{\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}) = \min_l\{p_{l\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})\}$, which is then given by:

$$p_{\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij}) + \max_{l \neq k} n^{\bullet}(x_{il}|\pi_{ij})}. \tag{15}$$

The interval $[p_{\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}), p^{\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})]$ contains all posterior estimates of θ_{ijk} that would be obtained from the possible completions of the database and therefore it provides a measure of the quality of information conveyed by $\mathcal{D}_{inc}$ about θ_{ijk} [8].

3.2 Step 2: Posterior Precision

The value in (11) is an estimate of $E(\theta_{ijk}|\mathcal{D}_{inc})$. We now derive an estimate of the posterior precision of θ_{ij}. Suppose we have $n(\pi_{ij})$ cases completely observed on π_{ij}, so that $n -$

$\sum_j n(\pi_{ij})$ is the number of cases partially observed on the parent variable Π_i. Let $\theta_i = (\theta_{i1}, ..., \theta_{iq_i})$ be the parameters associated to the joint probability distribution of Π_i, and let $D(\beta_{i1}, ..., \beta_{iq_i})$ be the prior distribution, so that $\beta_i = \sum_j \beta_{ij}$ is the prior precision. Suppose that data are missing at random. If we knew the probability distribution of π_{ij} we could distribute the incomplete cases across the states of Π_i, so that the expected precision of the posterior distribution of θ_{ij} would be $\alpha_{ij} + n(\pi_{ij}) + p(\pi_{ij})(n - \sum_j n(\pi_{ij}))$. Thus if $\hat{p}(\pi_{ij}|\mathcal{D}_{inc})$ is an estimate of $p(\pi_{ij})$, an estimate of the posterior precision is

$$\hat{\alpha}_{ij} = \alpha_{ij} + n(\pi_{ij}) + \hat{p}(\pi_{ij}|\mathcal{D}_{inc})(n - \sum_j n(\pi_{ij})). \tag{16}$$

Clearly, $\hat{\alpha}_{ij}$ is the exact posterior precision when the database is complete, and as the number of missing entries increases then $\hat{\alpha}_{ij}$ depends heavily on $\hat{p}(\pi_{ij}|\mathcal{D}_{inc})$. We can apply the BC method used in Section 3.1 to obtain the estimate $\hat{p}(\pi_{ij}|\mathcal{D}_{inc})$. If data are missing at random, an estimate of $\phi_{ij} = p(\Pi_i = \pi_{ij}|\Pi =?)$, $j = 1, ..., q_i$, is

$$\hat{\phi}_{ij} = \frac{\beta_{ij} + n(\pi_{ij})}{\beta_i + \sum_h n(\pi_{ih})}.$$

We can then apply (11) to obtain

$$\hat{p}(\pi_{ij}) = \sum_{l \neq j = 1}^{q_i} \hat{\phi}_{il} p_{l\bullet}(\pi_{ij}|\mathcal{D}_{inc}) + \hat{\phi}_{ij} p^{\bullet}(\pi_{ij}|\mathcal{D}_{inc})$$

where

$$p^{\bullet}(\pi_{ij}|\mathcal{D}_{inc}) = \frac{\beta_{ij} + n(\pi_{ij}) + n^{\bullet}(\pi_{ij})}{\beta_i + \sum_h n(\pi_{ih}) + n^{\bullet}(\pi_{ij})}$$

$$p_{l\bullet}(\pi_{ij}|\mathcal{D}_{inc}) = \frac{\beta_{ij} + n(\pi_{ij})}{\beta_i + \sum_h n(\pi_{ih}) + n^{\bullet}(\pi_{il})},$$

with $n^{\bullet}(\pi_{ij})$ denoting the number of possible completions of incomplete cases on π_{ij}. We thus see that as the number of missing entries increases, the estimate $\hat{\alpha}_{ij}$ tends to $\alpha_{ij} + (\beta_{ij}/\beta_i)n$ so that the cases are distributed according to the prior belief about the parameters defining the BBN.

4. The System

The method described in the previous section has been implemented in a system able to learn BBNs from (possibly) incomplete databases. The system takes as input a database and a partial order on the variables occurring in it and returns a BBN. The system performs two tasks: *i)* extracts a graphical model from the available information in the database using the method described in Section 3 and *ii)* assesses the conditional probabilities for the extracted graphical model using the BC method.

The first task uses the search procedure devised by [4], and replaces the measure 6 to estimate the local contribution of a node X_i and its parents Π_{ij} to the joint probability of

	Generating Structure	Variables	n
$\mathcal{M}_1$	$X_1 \longrightarrow X_2 \longrightarrow X_3$	all binary	1,000
$\mathcal{M}_2$	$X_1 \longrightarrow X_2 \longrightarrow X_3$	X_1, X_2 binary, X_3 ternary	1,000
$\mathcal{M}_3$	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & & X_4 \\ & \nwarrow & & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	all binary	5,000
$\mathcal{M}_4$	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & & X_4 \\ & \nwarrow & & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	X_1, X_2, X_3, X_4 ternary X_5 quaternary	10,000

Table 2: Generating structures used for the experimental evaluations.

	Generating Structure $\mathcal{M}_1$	$\log(\hat{p}(\mathcal{D}_{inc}\|\mathcal{M}))$	Run Time (sec)
100	$X_1 \longrightarrow X_2 \longrightarrow X_3$	-1437.64	12
80	$\begin{array}{ccc} X_2 & \longrightarrow & X_3 \\ \uparrow & \nearrow & \\ X_1 & & \end{array}$	-1426.48	13
60	$X_1 \quad X_2 \longrightarrow X_3$	-1446.51	11
40	$X_1 \longrightarrow X_2 \longrightarrow X_3$	-1447.19	12
20	$X_1 \longrightarrow X_2 \longrightarrow X_3$	-1414.63	12

Table 3: Models induced from the database generated from $\mathcal{M}_1$ for different percentages of complete data.

	$p(X_1 = 1)$	$p(X_2 = 1)$	$p(X_3 = 1)$
100	0.11	0.78	0.56
80	0.11	0.78	0.57
60	0.12	0.79	0.56
40	0.11	0.79	0.57
20	0.10	0.79	0.60

Table 4: Marginal probabilities in the networks induced from the database generated from $\mathcal{M}_1$ for different percentages of complete data.

	100	80	60	40	20
$\log(g(X_1))$	-356.43	-352.92	-366.62	-349.84	-323.95
$\log(g(X_2))$	-531.59	-525.65	-519.15	-512.21	-505.97
$\log(g(X_3))$	-689.87	-691.72	-689.42	-689.62	-678.11
$\log(g(X_2, X_1))$	-519.15	-512.47	-519.87	-511.20	-483.45
$\log(g(X_3, X_1))$	-690.99	-692.12	-692.38	-683.91	-661.60
$\log(g(X_3, X_2))$	-562.06	-560.36	-560.74	-586.15	-607.22
$\log(g(X_3, (X_1, X_2)))$	-564.40	-553.69	-566.18	-593.53	-609.02

Table 5: Estimate of $\log(g(X_i, \Pi))$ for several percentages of available entries in the database generated from $\mathcal{M}_1$.

$(\mathcal{M}, \mathcal{D})$ with a measure able to cope with missing data. Therefore, the choice on whether adding a parent node to a variable X_i is done by evaluating the estimate

$$\hat{g}(X_i, \mathcal{P}_i) = \prod_{j=1}^{q_i} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ij})\Gamma(\hat{\alpha}_{ij}\hat{p}(x_{ik}|\pi_{ij}))}{\Gamma(\hat{\alpha}_{ij})\Gamma(\alpha_{ijk})}.$$

The second task is performed using the method described in [9] and it is implemented using a slightly modified version of the algorithms described in [8]. The system has been developed in Common Lisp and CLOS under Machintosh Common Lisp v.4. Being entirely CLtL2 compliant, the system should be easily portable to any platform where a CLtL2 development environment is available.

5. Experimental Evaluation

The aim of the experiments described in this Section is to evaluate the accuracy of the estimate (8) as the number of missing entries in the database increases.

5.1 Materials and Methods

We considered four different models described in Table 2. From each of these models we generated a random sample of n cases, and applied the algorithm for the induction of the model from the data, using an initial order which was consistent with the generating struc-

	100	80	60	40	20
$X_1 \quad X_3 \quad X_2$	-1577.88	-1570.29	-1575.19	-1551.66	-1508.03
$X_1 \to X_3 \quad X_2$	-1579.00	-1570.69	-1578.15	-1545.95	-1491.51
$X_1 \quad X_2 \to X_3$	-1450.07	-1438.93	-1446.51	-1448.20	-1437.14
$X_1 \to X_3 \leftarrow X_2$	-1452.41	-1432.26	-1451.95	-1455.57	-1438.93
$X_1 \to X_2 \quad X_3$	-1565.45	-1557.11	-1575.91	-1550.65	-1485.51
$X_2 \leftarrow X_1 \to X_2$	-1566.57	-1557.51	-1578.87	-1544.94	-1469.00
$X_1 \to X_2 \to X_3$	-1437.64	-1425.76	-1447.23	-1447.19	-1414.62
$\begin{array}{ccc} X_1 & \to & X_2 \\ \downarrow & \swarrow & \\ X_3 & & \end{array}$	-1439.98	-1419.08	-1452.67	-1454.56	-1416.42

Table 6: $\log(\hat{p}(\mathcal{D}_{inc}|\mathcal{M}))$ for all possible models consistent with $X_3 \prec X_2 \prec X_1$, for different percentages of available data generated from $\mathcal{M}_1$.

ture, and assuming uniform prior distributions on the parameters. We then sequentially deleted 20% of the sample at random, until the database was empty. On each incomplete database we run our system to induce the model from the data. The learning process was then completed by computing the estimate of the probability $\hat{p}(\mathcal{D}_{inc}|\mathcal{M})$ and the BC estimates of the conditional probabilities defining the induced model. From the estimates of the conditional probabilities the marginal probabilities were computed. These experiments were performed on a Machintosh PowerBook 5300 using the system described in Section 4.

5.2 Results and Discussion

Tables 3 and 7 show the models induced from the databases generated from the two chain models $\mathcal{M}_1$ and $\mathcal{M}_2$, the estimates of $\log(p(\mathcal{D}_{inc}|\mathcal{M})$ for several percentages of available entries, and the total run time taken by the whole learning process to extract the graphical model and estimate the parameters of the BBN, excluding the time of reading the database from the disk. The marginal probabilities are displayed in Tables 4 and 8. The initial order on the variables was in both cases $X_3 \prec X_2 \prec X_1$.

The models learned from the database generated from $\mathcal{M}_1$ are the correct ones when 40% and 20% of the entries in the database are available, and coherently the model of independence is induced from the empty database. With 60% and 80% of the original data the models induced from the data differ from the generating structure in one link. Run times show a remarkable independence from the percentage of missing data in the database.

Table 5 gives the estimates $\log(\hat{g}(X_i, \Pi_{ij}))$ computed in each step of the algorithm. When 80% of the entries are available, $\log(\hat{g}(X_3, (X_1, X_2))) = -553.6903$ and $\log(\hat{g}(X_3, X_2)) = -560.3615$, so that the model induced from the incomplete database is $\exp(-553.6903 + 560.3615) = 789.34$ times more likely than the generating structure, if we assume that the prior distribution on the eight possible models consistent with the order $X_3 \prec X_2 \prec X_1$ is uniform. The strong evidence against the model used to generate the database can be due to the fact that $p(X_3 = 1|X_2 = 2) = 0.1$ and $p(X_2 = 2) = 0.77$ in the generating structure. In the complete database $n(X_3 = 1|X_2 = 2) = 22$ which becomes 11 when

	Generating Structure $\mathcal{M}_2$	$\log(\hat{p}(\mathcal{D}_{inc} \mid \mathcal{M}))$	Run Time (sec)
100	$X_1 \longrightarrow X_2 \longrightarrow X_3$	-1869.86	12
80	$\begin{array}{ccc} X_2 & \longrightarrow & X_3 \\ \uparrow & \nearrow & \\ X_1 & & \end{array}$	-1855.40	13
60	$X_1 \longrightarrow X_2 \longrightarrow X_3$	-1865.62	11
40	$\begin{array}{ccc} X_2 & \longrightarrow & X_3 \\ \uparrow & \nearrow & \\ X_1 & & \end{array}$	-1825.61	12
20	$X_1 \longrightarrow X_2 \longrightarrow X_3$	-1770.44	12

Table 7: Models induced from the database generated from $\mathcal{M}_2$ for different percentages of available data.

20% of entries are deleted, so that the small number of counts may cause the imprecision of the estimate $\log(\hat{p}(\mathcal{D}_{inc} \mid \mathcal{M}))$. The conditional probabilities estimated for the model selected are $p(X_3 = 1 \mid (X_1 = 1, X_2 = 1) = 0.77$ and $p(X_3 = 1 \mid (X_1 = 2, X_2 = 1) = 0.70$, $p(X_3 = 1 \mid (X_1 = 1, X_2 = 2) = 0.12$ and $p(X_3 = 1 \mid (X_1 = 2, X_2 = 2) = 0.11$, so that the estimate of the marginal probability of $X_3 = 1$ differs from the estimate obtained from the complete database by 1%. When 60% of the entries are available $\log(\hat{g}(X_2)) = -519.1485$ and $\log(\hat{g}(X_2, X_1)) = -519.8694$ so that the model induced from the data is only 2 times more likely than the generating structure. Again the marginal probabilities computed from the induced network are very similar to the marginal probabilities found in the model induced from the complete database: thus the choice of a slightly different model has little effect on the predicting power of the network.

Table 6 gives the estimate $\log(\hat{p}(\mathcal{D}_{inc} \mid \mathcal{M}))$ for the eight possible models consistent with the initial ordering of the variables. Such estimates can be computed from the values in Table 5 by adding relevant terms. The estimates are very accurate until 40% of the entries in the original database are retained. When only 20% of the entries is available the error of the estimate increases, but nevertheless the model induced from the database is equal to the generating structure. If we assume that the set of possible models is limited to the eight models consistent with the order $X_3 \prec X_2 \prec X_1$, and that they are a priori equally likely, then from the values in Table 6 we can compute the marginal probability of $\mathcal{D}$ and of the four incomplete databases $\mathcal{D}_{inc}$ from which we can compute the posterior probabilities of all possible models. The posterior probability of the model induced from the database with 80% of the entries is 0.9987, against a probability 0.0012 for the generating structure. The other models have posterior probabilities near 0. With 60% of the entries, the posterior probability of the induced model is 0.6699, against 0.3258 for the generating structure.

Similar results are found for the models induced from the database generated from $\mathcal{M}_2$. The models induced from the databases with 80% and 40% of the entries differs from the generating structure in one link, and they are respectively $\exp(-979.8927 + 985.2711) = 216.68$ and $\exp(-992.9008 + 994.7959) = 6.65$ more likely than the generating structure, under the assumptions that a priori the possible models are equally likely. The estimates of the marginal probabilities are very similar to those obtained in the complete database, again

	$p(X_1 = 1)$	$p(X_2 = 1)$	$p(X_3 = 1)$	$p(X_3 = 2)$
100	0.11	0.78	0.25	0.30
80	0.12	0.78	0.23	0.30
60	0.12	0.79	0.23	0.29
40	0.12	0.79	0.23	0.28
20	0.10	0.81	0.19	0.30

Table 8: Marginal probabilities in the networks induced from the database generated from $\mathcal{M}_2$ for different percentages of complete data.

	Generating Structure $\mathcal{M}_3$	$\log(\hat{p}(\mathcal{D}_{inc}\|\mathcal{M}))$	Run Time (sec)
100	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & & X_4 \\ & \nwarrow & & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	-25024.09	183
80	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & \longrightarrow & X_4 \\ & \nwarrow\!\!\!\!\nearrow & & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	-24673.57	191
60	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & & X_4 \\ & \nwarrow\!\!\!\!\nearrow & & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	-24870.98	187
40	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & \longrightarrow & X_4 \\ & \nwarrow\!\!\!\!\nearrow & & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	-24814.57	188
20	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & & X_4 \\ \uparrow & \nwarrow\!\!\!\!\nearrow & & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	-25112.2423	185

Table 9: Models induced from the database generated from $\mathcal{M}_3$ for different percentages of available data.

showing that the consequence of a slightly different model has little effect on the reasoning process. The estimate of $\log(p(\mathcal{D}_{inc}|\mathcal{M}))$ is very accurate until the database contains 40% of the original entries. The total run times make even clear that the source of complexity is the search space and the performances of the method remain insensitive to the number of missing data. This result is not surprising when we realize that the computational cost of BC amounts to just two exact Bayesian updating and a convex combination for each parameter, and therefore it does not depend on the number of missing data. The number of missing data affects only the storage procedure described in [8] but its effect is limited by taking advantage of the local independence of the BBN and by using discrimination trees to store the counters of observed data and keep track of the possible completions.

The models induced from the databases generated from $\mathcal{M}_3$ and $\mathcal{M}_4$ are given in Table 9 and 11. Table 10 displays the marginal probabilities computed by the networks induced from the various incomplete databases. The initial order on the variables was in both

	$X_1 = 1$	$X_2 = 1$	$X_3 = 1$	$X_4 = 1$	$X_5 = 1$
100	0.20	0.70	0.39	0.30	0.52
80	0.20	0.71	0.38	0.29	0.53
60	0.21	0.70	0.39	0.29	0.53
40	0.21	0.70	0.39	0.30	0.53
20	0.21	0.69	0.40	0.31	0.54

Table 10: Marginal probabilities in the networks induced from the database generated from $\mathcal{M}_3$ for different percentages of complete data.

cases $X_5 \prec X_4 \prec X_3 \prec X_2 \prec X_1$. The model induced from the complete database are equal to the generating structure, and coherently the empty structure is induced when data are totally missing. As the number of entries available decreases, at most two extra dependencies are induced from the database. The only exception is the model induced from the database generated from $\mathcal{M}_4$ with 80% of the entries available. In this case 4 extra dependencies are learned compared to the generating structure, and the Bayes factor of the induced model against the generating structure is e^{13}. However the conditional probabilities learned are only slightly different, so that the estimates of the marginal probabilities are extremely robust thus limiting the effect in the subsequent reasoning process. The estimates of $\log(p(\mathcal{D}_{inc}|\mathcal{M}))$ are again extremely accurate.

6. Conclusions

Missing data represent a challenge for learning methods because they may affect their use in real-world applications, where databases are often incomplete. Current methods to learn BBNs from incomplete databases rely on iterative methods, such as EM or the Gibbs Sampling, to obtain an approximate estimate of the marginal likelihood of the database given a graphical model, a fundamental step in the process of extracting the graphical structure of a BBN from a database of cases. This paper introduced a deterministic method able to provide such an estimation and to extract the graphical structure of a BBN from an incomplete database. When coupled with a method able to assess the parameters of a BBN given a graphical model from an incomplete database, this method gives rise to systems able to extract a complete BBN from an incomplete database, such as the system outlined in Section 4. Preliminary experimental evaluations show a significant robustness of this method and a remarkable independence of its execution time from the number of missing data. Information about the experiments, including databases, generated BBNs, and log files of the program, are available at the URL http://kmi.open.ac.uk/~marco/projects/kdd/experiments/uai97. The prototype program used for the experimental evaluations is available at http://kmi.open.ac.uk/~marco/projects/kdd/software/uai97.

References

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

	Generating Structure $\mathcal{M}_4$	$\log(\hat{p}(\mathcal{D}_{inc} \mid \mathcal{M}))$	Run Time (sec)
100	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & & X_4 \\ & \nwarrow & & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	-40125.37	275
80	$\begin{array}{ccc} X_5 & \longleftarrow & X_3 \\ \uparrow & \nwarrow\!\!\!\!\!\nearrow & \uparrow \\ X_1 & \longrightarrow & X_2 \\ & \searrow & \downarrow \\ & & X_4 \end{array}$	-39369.00	296
60	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & & X_4 \\ & \nwarrow & \uparrow & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	-42145.68	278
40	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & & X_4 \\ & \nwarrow\!\!\!\!\!\nearrow & & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	-40131.97	289
20	$\begin{array}{ccccc} X_5 & \longleftarrow & X_3 & & X_4 \\ & \nwarrow & \uparrow & \nearrow & \\ X_1 & \longrightarrow & X_2 & & \end{array}$	-39952.47	285

Table 11: Models induced from the database generated from $\mathcal{M}_4$ for different percentages of available data.

[2] D. M. Chickering and D. Heckerman. Learning Bayesian networks is NP-hard. Technical Report MSR-TR-94-17, Microsoft Research, Microsoft Corporation, 1994.

[3] D. M. Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network. Technical Report MSR-TR-96-08, Microsoft Research, Microsoft Corporation, 1996.

[4] G.F. Cooper and E. Herskovitz. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[5] D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combinations of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[6] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* Wiley, New York, 1987.

[7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[8] M. Ramoni and P. Sebastiani. Robust learning with missing data. Technical Report KMi-TR-28, Knowledge Media Institute, The Open University, 1996. Available at http://kmi.open.ac.uk/techreports/KMi-TR-28.

[9] M. Ramoni and P. Sebastiani. The use of exogenous knowledge to learn Bayesian networks from incomplete databases. Technical Report KMi-TR-

44, Knowledge Media Institute, The Open University, 1996. Available at http://kmi.open.ac.uk/techreports/KMi-TR-44.

[10] D.J. Spiegelhalter and S.L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:157–224, 1990.

[11] P. Sprites, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Springer Verlag, New York, 1993.