KNOWLEDGE MEDIA INSTITUTE

Parameter Estimation in Bayesian Networks from Incomplete Databases

Marco Ramoni *Paola Sebastiani*

KMI-TR-57

November 1997

The Open University

Parameter Estimation in Bayesian Networks from Incomplete Databases

Marco Ramoni
Knowledge Media Institute
The Open University

Paola Sebastiani
Department of Actuarial Science and Statistics
City University

Abstract

Current methods to learn Bayesian Networks from incomplete databases share the common assumption that the unreported data are missing at random. This paper describes a method — called *Bound* and *Collapse* (BC) — to learn Bayesian Networks from incomplete databases which allows the analyst to efficiently integrate information provided by the observed data and exogenous knowledge about the pattern of missing data. BC starts by *bounding* the set of estimates consistent with the available information and then *collapses* the resulting set to a point estimate via a convex combination of the extreme points, with weights depending on the assumed pattern of missing data. Experiments comparing BC to Gibbs Samplings are provided.

Keywords: Bayesian Inference; Bayesian Networks; Gibbs Sampling; Missing Data.

Reference: KMi Technical Report KMi-TR-57, November 1997.

Address: Marco Ramoni. Knowledge Media Institute, The Open University, 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

Bayesian Belief Networks (BBNs) provide a powerful formalism to reason under uncertainty. A BBN is a direct acyclic graph where nodes represent stochastic variables and direct arcs identify dependencies between a set of *parent* nodes and a *child* node. In the original development of BBNs, domain experts were supposed to be their main source of information: their independence assumptions, when coupled with the subjective assessment of the conditional dependencies among the variables, produces a sound and compact probabilistic representation of the domain knowledge. However, the statistical roots of BBNs soon prompted for the use of learning methods to extract them directly from databases of cases rather than from the insight of human domain experts [2, 1, 8]. This choice can be extremely rewarding when the domain of application generates large amounts of statistical information and aspects of the domain knowledge are still unknown or controversial, or too complex to be encoded as subjective probabilities of few experts.

When the database is complete, the induction of the conditional probabilities quantifying the BBN can be efficiently performed using a standard statistical technique known as Bayesian Conjugate analysis. Unfortunately, when the database is incomplete, i.e. some data are reported as unknown, simplicity and efficiency of conjugate analysis is lost, and exact Bayesian updating becomes infeasible: its complexity is exponential in the number of missing data [2]. During the past few years, several methods have been proposed to learn conditional probabilities from incomplete databases, such as sequential updating [15, 3], the EM algorithm [4], or Gibbs Sampling (GS) [7]. GS is a stochastic, iterative method and it is currently becoming the most popular in the statistical community, although its limitations are well known: it is highly resource demanding, the convergence rate may be slow, and the execution time heavily depends on the number of missing data.

According to the classification of missing data mechanisms proposed by [13] and further developed by [10] and [6], when the database consists of values of a single variable X, unreported data are said to be "Missing at Random" (MAR) if the probability that an observation on X is missing does not depend on X; if this probability depends on X, then the missing data mechanism is said to be "Not Ignorable" (NI). In the first case, both observed and unknown entries are generated by the same mechanism, so that the available data are representative of the complete but unknown database. In the second case, the observed entries may no longer be representative and exogenous information about the missing data mechanism is required. This classification generalizes to more complex structures in which a set of variables $\mathcal{X} = \{X_1, \ldots, X_I\}$, related by some association model, is observed. This is the case of direct graphical models, which are the basic component of BBNs. Spelling out the previous classification of missing data mechanisms within this framework, we have now three possible situations. Data are MAR when the probability that a variable X_i in $\mathcal{X}$ is not observed depends at most on the configuration of its parent variables. A special case is when this probability is neither dependent on X_i nor on its parents. In this case, data are said to be "Missing Completely at Random" (MCAR). When the probability that X_i is missing depends on X_i and/or its parents, the missing data mechanism is NI. When data are MAR or MCAR, the observed database contains all we need to know to make inference since no other information is required to distribute the missing data. Under a NI missing data mechanism, on the other hand, the amount of information needed about patterns of missing data to have reliable estimates can be large so that, often, data are supposed to be MAR, or incomplete cases are removed from the database. Current methods, such as Sequential Updating, EM algorithm and GS, implicitly assume that unreported data are MAR and we shall see, in the reminder of this paper, that this assumption can have a dramatic effect on the propriety of inference.

This paper will present a new method to estimate conditional distributions defining a BBN from an incomplete database, which, with a minimum amount of information about the missing data process, yields accurate estimates. Ramoni and Sebastiani [11] introduce a deterministic method to learn conditional probabilities from incomplete databases which does not assume any information on the missing data mechanism. The method *bounds* the set of possible estimates consistent with the available information by computing the minimum and the maximum Bayesian estimate that would be obtained from all possible completions of the database. This process returns probability intervals containing all possible estimates consistent with the available information. In this paper, we present a technique to use exogenous knowledge about the pattern of missing data to collapse these bounds into a unique value via a convex combination of the extreme points with weights depending on the assumed pattern of missing data. Because of its strategy, we call this method *Bound* and *Collapse* (BC). The use of BC allows the encoding of prior knowledge of the pattern of missing data, such as the MAR assumption. Experimental evaluations show that the estimates provided by BC are very close to those provided by GS, when data are MAR, but can be more robust when the missing data mechanism is NI. Furthermore, for each parent configuration, BC reduces the cost of estimating each conditional distribution of X_i to the cost of one exact Bayesian updating and one convex combination for each state of X_i.

The structure of the paper is as follows. In Section 2 we review background material on estimation of conditional probabilities in BBNs. Section 3 illustrates BC and provides an approximation of the variance of the estimates, from which approximate posterior distributions are derived. An experimental comparison with GS is given in Section 4.

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 distributions of X_i given the configurations of the parent variables. We shall consider discrete variables only and, therefore, each variable X_i will bear a finite set of c_i states. We will denote $X_i = x_{ik}$ by x_{ik}. The structure $\mathcal{M}$ yields a factorization of the joint probability of a set of values $x_k = \{x_{1k}, \ldots, x_{Ik}\}$ of the variables in $\mathcal{X}$ as

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

where π_{ij} denotes the state of Π_i in x_k. We will denote $\Pi_i = \pi_{ij}$ by π_{ij} and the number of states of Π_i by q_i.

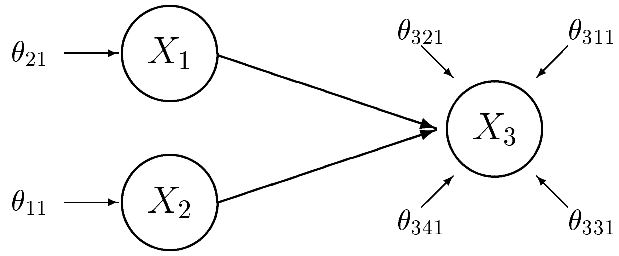

Figure 1: A Simple BBN

Suppose we are given a database of n cases $\mathcal{D} = \{x_1, \ldots, x_n\}$ and a graphical model $\mathcal{M}$ specifying the dependencies among variables in $\mathcal{X}$. Our task is to estimate, from $\mathcal{D}$, the conditional probabilities quantifying the dependencies in the graph. These conditional probabilities will be regarded as unknown parameters $\theta = (\theta_{ijk})$, where $\theta_{ijk} = p(x_{ik}|\pi_{ij}, \theta)$, and $\theta_{ij} = (\theta_{ij1}, \ldots, \theta_{ijc_i})$ is the parameter vector associated to the conditional distribution of $X_i|\pi_{ij}$, to be inferred from $\mathcal{D}$. Figure 1 shows an example of a simple BBN in which the set $\mathcal{X}$ is $\{X_1, X_2, X_3\}$ and $c_i = 2$ for $i = 1, 2, 3$. The graph encodes the marginal independence of X_1 and X_2 and they are both parents of X_3. Thus, Π_3 takes four values π_{ij} corresponding to the four combinations of states of X_1 and X_2, that we will denote as $\pi_{31} = (1, 1)$, $\pi_{32} = (1, 2)$, $\pi_{33} = (2, 1)$ and $\pi_{34} = (2, 2)$. Six independent parameters $\theta = (\theta_1, \theta_2, \theta_{31}, \theta_{32}, \theta_{33}, \theta_{34})$ quantify the BBN, where $\theta_{11} = p(x_{11}|\theta)$, $\theta_{21} = p(x_{21}|\theta)$, and $\theta_{3j1} = p(x_{31}|\pi_{3j}, \theta)$ for $j = 1, 2, 3, 4$. From these parameters, we obtain the parameter vectors $\theta_1 = (\theta_{11}, 1 - \theta_{11})$, $\theta_2 = (\theta_{21}, 1 - \theta_{21})$, $\theta_{31} = (\theta_{311}, 1 - \theta_{311})$, $\theta_{32} = (\theta_{321}, 1 - \theta_{321})$ $\theta_{33} = (\theta_{331}, 1 - \theta_{331})$, $\theta_{34} = (\theta_{341}, 1 - \theta_{341})$.

Several techniques are available to estimate the unknown θ_{ijk}. Let $n(x_{ik}|\pi_{ij})$ be the frequency of (x_{ik}, π_{ij}) in the database, and let $n(\pi_{ij}) = \sum_k n(x_{ik}|\pi_{ij})$ be the frequency of π_{ij}. The joint probability of a case x_k can be written as a function of the unknown θ_{ijk} as $p(x_k|\theta) = \prod_{i=1}^{I} \theta_{ijk}$, and hence, if cases are independent, the joint probability of the database $\mathcal{D}$ is

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

This quantity, called the likelihood function $l(\theta)$ when regarded as a function of θ alone, plays a crucial role in many statistical methods. A popular approach, called Maximum Likelihood (ML), calculates the parameter values that maximize $l(\theta)$. It is well known [9] that the ML estimate of θ_{ijk} is the relative frequency of relevant cases in the database: $\hat{\theta}_{ijk} = n(x_{ik}|\pi_{ij})/n(\pi_{ij})$. A drawback of ML is that $\hat{\theta}_{ijk} = 0$ whenever $n(x_{ik}|\pi_{ij}) = 0$, so that the estimate can be too extreme when the database is sparse.

Compared to ML, the Bayesian approach allows us to incorporate prior information about the parameters θ_{ijk} in the estimation process. The crucial aspect of the Bayesian approach is to regard θ_{ijk} as a random variable, whose *prior* distribution represents the observer's belief about the conditional probability, before data are seen. Bayesian learning means to use $\mathcal{D}$ to update the prior belief on the parameters using Bayes' theorem. In this way, probability estimates will not be 0 even for unobserved parent-child configurations, as long as the prior belief will not allow it. Using Bayes' theorem, the prior density $p(\theta)$ is updated in the *posterior* density

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

where $p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$ is the *marginal probability* of $\mathcal{D}$. The standard Bayesian estimate of θ is the *posterior expectation* of θ: $E(\theta|\mathcal{D})$. A common assumption, which leads to local computations, is that the parameter vectors θ_{ij} and $\theta_{i'j'}$ are independent for $i \neq i'$ (global independence) and for $j \neq j'$ (local independence) [15]. Global and local independence yield a factorization of the joint prior density $p(\theta) = \prod_{ij} p(\theta_{ij})$ which, if the database $\mathcal{D}$ is complete, induces an equivalent factorization of the posterior density of θ:

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

and this allows us to independently update the distribution of θ_{ij}, for all i, j. A further saving in computation is achieved if, for all i and j, the prior distribution of θ_{ij} is a *Dirichlet* distribution with *hyperparameters* $\{\alpha_{ij1}, \cdots, \alpha_{ijc_i}\}$, that is $\theta_{ij} \sim D(\alpha_{ij1}, \ldots, \alpha_{ijc_i})$ with $\alpha_{ijk} > 0$. In this case, the prior density of θ_{ij} is

$$p(\theta_{ij}) = \prod_k \frac{\Gamma(\sum_k \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \theta_{ijk}^{\alpha_{ijk}}.$$

The prior hyperparameters α_{ijk} encode the observer's prior belief and can be regarded as frequencies of imaginary cases needed to formulate the prior distribution. As a matter of fact, the marginal probability of $(x_{ik}|\pi_{ij})$ is α_{ijk}/α_{ij}, $\alpha_{ij} = \sum_{k=1}^{c_i} \alpha_{ijk}$ is the *prior precision* on θ_{ij}, and $\alpha = \sum_{ij} \alpha_{ij}$ is the size of the imaginary database. The situation of initial ignorance can be represented by assuming $\alpha_{ijk} = \alpha/(c_i q_i)$ for all i, j and k, so that the prior probability of $(x_{ik}|\pi_{ij})$ is simply $1/c_i$ [5]. [15] shows that parameter independence and prior Dirichlet distributions imply that the posterior density of θ is still a product of Dirichlet densities and

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

so that local and global independence are retained after the updating. Furthermore, the Bayesian estimate of the probability of $(x_{ik}|\pi_{ij})$ is the posterior expectation of θ_{ijk}:

$$E(\theta_{ijk}|\mathcal{D}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij})}.$$

Unfortunately, this simplicity is lost when the database is incomplete. Exact analysis requires the computation of the joint posterior distribution of the parameters by considering all possible completions of incomplete cases. Suppose, for instance, that the case x_k is incomplete, with entries on X_i and/or its parents missing whilst other variables are observed. Let $\mathcal{D} = \mathcal{D}_o \cup x_k$, where $\mathcal{D}_o$ denotes the part of the database with no missing data. The posterior distribution of the parameters associated to the conditional distribution of $X_i|\pi_{ij}$ turns out to be the mixture:

$$\sum_k D(\alpha_{ij1} + n(x_{i1}|\pi_{ij}) + \delta_1(k), \ldots, \alpha_{ijc_i} + n(x_{ic_i}|\pi_{ij}) + \delta_{c_i}(k))p(x_{ik}, \pi_{ij}|\mathcal{D}_o)$$
$$+D(\alpha_{ij1} + n(x_{i1}|\pi_{ij}), ..., \alpha_{ijc_i} + n(x_{ic_i}|\pi_{ij}))(1 - p(\pi_{ij}|\mathcal{D}_o))$$

where $\delta_i(k) = 1$ if $i = k$, $\delta_i(k) = 0$ otherwise, and the frequencies are based on $\mathcal{D}_o$ [15]. The first term in the mixture computes the possible completions of the child variable X_i given the parent configuration π_{ij}. The last term is conditioned on completing the parent configuration to a state different from π_{ij}, so that the distribution of θ_{ij} is not updated. When variables in different configurations are not observed, the likelihood function $l(\theta)$ does not factorize any longer, and hence the advantages of local computations are lost. As the number of incomplete cases increases, exact updating becomes apparently infeasible: its complexity is in fact exponential in the number of missing data [2], and approximate methods are therefore required.

GS is currently the most popular method for Bayesian estimation in complex problems such us inference from incomplete samples. GS is an iterative, stochastic method in which missing data are treated as unknown parameters. Let $\beta = \{\beta_1, \cdots, \beta_p\}$ be the augmented parameter vector, which is given by θ and new parameters representing missing entries, and let $\beta^{(0)}$ be initial values. Suppose that the conditional distributions of $\beta_i|(\beta_1, \cdots, \beta_{i-1}, \beta_{i+1}, \cdots, \beta_p)$ are known for each i. This assumption is easily satisfied in BBNs with discrete data and Dirichlet priors [16]. The first sample is simulated as follows: $\beta_1^{(1)}$ is sampled from the conditional distribution of $\beta_1|(\beta_2^{(0)}, \cdots, \beta_p^{(0)}, \mathcal{D})$; then $\beta_2^{(1)}$ is sampled from the conditional distribution of $\beta_2|(\beta_1^{(1)}, \beta_3^{(0)}, \cdots, \beta_p^{(0)}, \mathcal{D})$ and the process is iterated to completion of $\beta^{(1)}$. This process is repeated several times and it is known [16] that, under broad conditions, it provides a sample from the joint posterior distribution of β, from which a sample from the posterior distribution of θ can be extracted. This sample is used to compute empirical estimates of the posterior means and any other function of the parameters. In practical applications, the algorithm iterates a number of times (*burn-in*) to reach stability and then a final sample from the joint posterior distribution of the parameters is taken. The underlying assumption is that unreported data are MAR, so that the probability of an entry being missing does not depend on the state of the corresponding variable, but it may depend on the parent configuration π_{ij}. Thus, the incomplete samples within parent configurations are representative of the complete but unknown ones.

3. Method

This section introduces a deterministic method, called BC, to estimate parameters from an incomplete database. This method first *bounds* the possible estimates consistent with the

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	?	?

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

Figure 2: Virtual frequencies $n^\bullet(x_{3k}|x_1, x_2)$ consistent with the incomplete database.

available data using a technique given in [11], and then *collapses* the resulting interval to a point via a convex combination of the extreme estimates using information on the pattern of missing data. The basic intuition behind BC is that an incomplete database is still able to constrain the possible estimates within a set and that, when exogenous information is available on the pattern of missing data, this can be used to select a point estimate within the set of possible ones.

3.1 Bound

Let X_i be a variable in $\mathcal{X}$ with parent variable Π_i. Denote by $n(?|\pi_{ij})$ the frequency of cases in which only the entry on the child variable is missing, by $n(x_{ik}|?)$ the frequency of cases in which only the parent configuration is unknown but it can be completed as π_{ij}, and by $n(?|?)$ the frequency of cases in which the entries X_i, Π_i are unknown and they can be completed as (x_{ik}, π_{ij}). As in [11], we define *virtual* frequencies:

$$\begin{aligned} n^\bullet(x_{ik}|\pi_{ij}) &= n(?|\pi_{ij}) + n(x_{ik}|?) + n(?|?) \\ n_\bullet(x_{ik}|\pi_{ij}) &= n(?|\pi_{ij}) + \sum_{h \neq k=1}^{c_i} n(x_{ih}|?) + n(?|?). \end{aligned}$$

Thus, $n^\bullet(x_{ik}|\pi_{ij})$ is the maximum achievable frequency of (x_{ik}, π_{ij}) in the incomplete database that is used to compute the maximum of $p(x_{ik}|\pi_{ij})$. The virtual frequency $n_\bullet(x_{ik}|\pi_{ij})$ is used to compute the minimum estimate of $p(x_{ik}|\pi_{ij})$. Note that, if X_i is binary, $n^\bullet(x_{i1}|\pi_{ij}) = n_\bullet(x_{i2}|\pi_{ij})$, and $n^\bullet(x_{ik}|\pi_{ij}) = n^\bullet_{ij}$ and $n_\bullet(x_{ik}|\pi_{ij}) = n_{\bullet ij}$, for all k, when $n(x_{ik}|?)$ are all identical. When data are missing only on the child variable, then $n^\bullet(x_{ik}|\pi_{ij}) = n_\bullet(x_{ik}|\pi_{ij}) = n_{\bullet ij}$. An example is given in Figure 2 for the BBN depicted in Figure 1. For instance, $n(x_{31}|?) = 1$ from case x_4, $n(?|(1,1)) = 0$ and $n(?|?) = 1$ from x_5, from which $n^\bullet(x_{31}|(1,1)) = 2$ is obtained. Furthermore, $n(x_{32}|?) = 1$ from x_3, so that $n_\bullet(x_{31}|(1,1)) = 2$ and $n^\bullet(x_{32}|(1,1)) = n_\bullet(x_{31}|(1,1))$. Ramoni and Sebastiani [11] show that the maximum Bayesian estimate of $p(x_{ik}|\pi_{ij})$ is:

$$p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij}) + n^\bullet(x_{ik}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij}) + n^\bullet(x_{ik}|\pi_{ij})} \tag{1}$$

and the minimum Bayesian estimate is

$$p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij}) + n_\bullet(x_{ik}|\pi_{ij})}. \tag{2}$$

The probability interval defined by $(p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D})\ p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}))$ contains all possible estimates consistent with $\mathcal{D}$, therefore, it is sound and it is the *tightest* estimable interval. When X_i is binary, then $p_\bullet(x_{i1}|\pi_{ij}, \mathcal{D}) = 1 - p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$.

The minimum estimate of θ_{ijk} is achieved when the missing data mechanism is such that, with probability 1, all cases $(X_i =?, \pi_{ij})$ and $(X_i =?, \Pi_i =?)$ are completed as (x_{ih}, π_{ij}), for $h \neq k$, all cases $(X_i = x_{ih}, \Pi_i =?)$ are completed as (x_{ih}, π_{ij}) if $h \neq k$, and, if $h = k$, as $(x_{ik}, \pi_{ij'})$, for some $j' \neq j$. Thus, $n_\bullet(x_{ik}|\pi_{ij})$ is the total virtual frequency of cases with $\Pi_i = \pi_{ij}$, for $h \neq k$, and the frequency of (x_{ik}, π_{ij}) is not augmented.

The maximum estimate of θ_{ijk} is achieved when the missing data mechanism is such that, with probability 1, all cases $(X_i =?, \pi_{ij})$, $(X_i =?, \Pi_i =?)$ and $(X_i = x_{ik}, \Pi_i =?)$ are completed as (x_{ik}, π_{ij}) and, for $h \neq k$, $(X_i = x_{ih}, \Pi_i =?)$ are completed as $(x_{ih}, \pi_{ij'})$, for some $j' \neq j$. Thus, $n^\bullet(x_{ik}|\pi_{ij})$ is the virtual frequency of (x_{ik}, π_{ij}) as well as of π_{ij}, while the frequency of other cases with the same parent configuration is not augmented.

The main feature of this method is its independence of the distribution of missing data because it does not try to infer them: with no information on the missing data mechanism, an incomplete database can only induce bounds on the possible estimates that could be learned. It is worth noting that, along this approach, a complete database is just a special case, within which available data are enough to constrain the set of possible estimates to a single point. A further advantage of this method is that the width of each interval accounts for the amount of information available in $\mathcal{D}$ about the parameter to be estimated, and it represents a measure of the quality of probabilistic information conveyed by the database about a parameter: the wider the interval, the greater the uncertainty due to the incompleteness of the database. In this way, intervals provide an explicit representation of the reliability of the estimates, which can be taken into account when the extracted BBN is used to perform a particular task.

3.2 Collapse

The second step of BC collapses the intervals estimated in the *bound* step into point estimates using a convex combination of the extreme estimates. This convex combination can be computed either by using of external information about the pattern of missing data or by a dynamic estimation of this pattern from the available information in the database, when data are assumed to be MAR.

3.2.1 Using Exogenous Knowledge

Suppose that some external information is available on the pattern of missing data. The analyst can encode this information as a probability distribution describing, for each variable in the database, the probability of a datum being missing as:

$$p(x_{ik}|\pi_{ij}, X_i =?) = \phi_{ijk} \tag{3}$$

where $k = 1, ..., c_i$ and $\sum_k \phi_{ijk} = 1$. Note that this is only a part of the information required about the distribution of missing data. We have seen that incomplete cases can be either $(x_{ik}, \Pi_i = ?)$, or $(X_i = ?, \pi_{ij})$ or $(X_i = ?, \Pi_i = ?)$. Thus, a full description of the pattern of missing data requires the probabilities $p(\pi_{ij}|\Pi_i = ?)$ and $p(x_{ik}|\pi_{ij}, \Pi_i = ?)$, as well as ϕ_{ijk}. We will show that the probabilities ϕ_{ijk} can be used to obtain accurate estimates of θ_{ijk}, if we exclude, amongst the possible patterns of missing data, those extreme mechanisms that yield the lower bounds $p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$. These mechanisms imply that all incomplete cases $(x_{ik}, \Pi_i = ?)$, except at least one, can be completed as (x_{ik}, π_{ij}), so that at most one minimum probability in all conditional distributions quantifying a parent-child dependency can be obtained. In order to be able to limit the amount of information about the process underlying missing data, we will assume that if all incomplete case $(x_{ik}, \Pi_i = ?)$ are completed as (x_{ik}, π_{ij}), then for all $k \neq h$, $(x_{ih}, \Pi_i = ?)$ cannot be completed as (x_{ih}, π_{ij}). This assumption allows us to derive new local lower bounds from the maximum probabilities as follows.

Each maximum probability $p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$ is obtained when all incomplete cases that could be completed as (x_{ik}, π_{ij}) are attributed to (x_{ik}, π_{ij}), and the observed frequencies of the other states of X_i given π_{ij} are not augmented. This is equivalent to assuming that the posterior distribution of θ_{ij} is a Dirichlet in which only the kth hyperparameter is updated by taking into account all possible completions, that is:

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

This distribution identifies a single probability for each other state of the variable X_i given π_{ij} as:

$$p_{k\bullet}(x_{il}|\pi_{ij}, \mathcal{D}) = \frac{\alpha_{ijl} + n(x_{il}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij}) + n^\bullet(x_{ik}|\pi_{ij})}$$

for $l \neq k$. Thus, the maximum probabilities induce c_i extreme probability distributions:

$$\{p^\bullet(x_{ik}|\pi_{ij}), p_{k\bullet}(x_{ih}|\pi_{ij}, \mathcal{D}), \quad k \neq h\},$$

for $k = 1, \ldots, c_i$. Therefore, the set of $c_i - 1$ local minima for $E(\theta_{ijk}|\mathcal{D})$ is $\{p_{h\bullet}(x_{ik}|\pi_{ij}, \mathcal{D})\}$, $h \neq k = 1, \ldots, c_i$ and the global minimum, in this set, is:

$$p^l_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) = \min_h \{p_{h\bullet}(x_{ik}|\pi_{ij}, \mathcal{D})\}.$$

The local minimum of $E(\theta_{ijk}|\mathcal{D})$ can also be written as:

$$p^l_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij}) + \max_{h \neq k} n^\bullet(x_{ih}|\pi_{ij})}, \tag{4}$$

which shows that the difference between $p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$ and $p^l_\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$ depends only on those cases in which the state of the child variable is known and the parent configuration is not. Note that $p^l_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) = p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$ if either X_i is binary, or $n(x_{ik}|?) = 0$ for all k, or if there is no missing data mechanism consistent with $n_\bullet(x_{ik}|\pi_{ij})$. This is the assumption made above: the completion of all cases $(x_{ik}, \Pi_i =?)$ as (x_{ik}, π_{ij}) implies that, for $h \neq k$, $(x_{ih}, \Pi_i =?)$ must be completed as $(x_{ih}, \pi_{ij'})$ for some $j' \neq j$. Note that, the difference $p^l_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) - p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$ is positive and it is easy to show that it is maximized when $n(x_{ik}|?) = m$, for all k, so that it approaches 0 at the same rate of m, and it becomes negligible as the database increases. Furthermore, if we happen to know that, for some missing data mechanism, the estimate of $p(x_{ik}|\pi_{ij}) = p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$ for some k, then, for all $h \neq k$ and j, $p^l_\bullet(x_{ih}|\pi_{ij}, \mathcal{D}) < p(x_{ih}|\pi_{ij}) < p^\bullet(x_{ih}|\pi_{ij}, \mathcal{D})$, so that all the other estimates are bounded below by $p^l_\bullet(x_{ih}|\pi_{ij}, \mathcal{D})$. Thus, at most one minimum can be achieved, as noted above.

The distribution of missing entries in terms of ϕ_{ijk} can now be used to identify a point estimate within the interval $[p^l_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}), p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D})]$ via a convex combination of the extreme probabilities:

$$\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}, \phi_{ijk}) = \sum_{l \neq k} \phi_{ijl} p_{l\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}) + \phi_{ijk} p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}). \tag{5}$$

The intuition behind (5) is that the upper bound of $p(x_{ik}, \pi_{ij})$ is obtained when all incomplete cases are completed as (x_{ik}, π_{ij}). Thus, if $p(x_{ik}|\pi_{ij}, X_i =?) = 1$ for a particular k, then (5) will return the upper bound of the interval probability as estimate of $p(x_{ik}|\pi_{ij})$, and $p_{k\bullet}(x_{ih}|\pi_{ij}, \mathcal{D})$ as estimates of $p(x_{ih}|\pi_{ij})$, $h \neq k$. This case corresponds to the assumption that data are *systematically* missing about x_{ik}. On the other hand, when no information on the mechanism generating missing data is available, and therefore all patterns of missing data are equally likely, then $\phi_{ijk} = 1/c_i$.

Since we consider only the extreme estimates induced by the maximum probabilities, the information needed in the collapse step is limited to the probabilities ϕ_{ijk}, and there is no need to specify probabilities of completions for cases in which the parent configuration is unknown. This limited amount of information about the process underlying missing data implies that $\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}, \phi_{ijk}) \geq p^l_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) \geq p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$. Thus, $\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}, \phi_{ijk})$ cannot be equal to $p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$ unless $p^l_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) = p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$. However, as noted above, this error would be limited to at most one conditional probability for each parent-child configuration, and it becomes negligible as the database increases, for we have shown above that $p^l_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) - p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) \to 0$.

As the number of missing entries decreases, $p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D})$ and $p_{h\bullet}(x_{ik}|\pi_{ij}, \mathcal{D})$ approach $(\alpha_{ijk} + n(x_{ijk}|\pi_{ij}))/(\alpha_{ij} + n(\pi_{ij}))$, so that, when the database is complete, (5) returns the exact estimate $E(\theta_{ijk}|\mathcal{D})$. As the number of missing entries increases then $p_{h\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}) \to 0$, for all l, and $p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) \to 1$, so that the estimate (5) approaches the prior probability ϕ_{ijk}, and coherently nothing is learned from a database in which all entries on (X_i, π_{ij}) are missing.

Furthermore, the estimates so found define a probability distribution since

$$\sum_{k=1}^{c_i} \hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}, \phi_{ijk}) \quad = \quad \sum_{k=1}^{c_i} \phi_{ijk} p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}) + \sum_{k=1}^{c_i} \sum_{h \neq k} \phi_{ijh} p_{h\bullet}(x_{ik}|\pi_{ij}, \mathcal{D})$$

$$= \sum_{k=1}^{c_i} \phi_{ijk}\{p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}_i) + \sum_{h \neq k} p_{k\bullet}(x_{ih}|\pi_{ij}, \mathcal{D}_i)\}$$
$$= \sum_{k=1}^{c_i} \phi_{ijk} = 1.$$

Finally, if $n^\bullet(x_{ik}|\pi_{ij}) = n^\bullet_{ij}$ then (5) simplifies to

$$\frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij}) + n^\bullet_{ij}\phi_{ijk}}{\alpha_{ij} + n(\pi_{ij}) + n^\bullet_{ij}} \tag{6}$$

so that the incomplete cases are distributed across the states of X_i according to the prior knowledge on the pattern of missing data. Note that (6) is the *expected Bayesian estimate*, given the assumed pattern of missing data.

3.2.2 Using Available Information

Suppose now that data are MAR, so incomplete samples within parent configurations are *representative samples* of the complete but unknown ones. In this case, the probability of a completion is $\phi_{ijk} = p(x_{ik}|\pi_{ij})$, and it can be estimated from the available data as

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

Then $\hat{\phi}_{ijk}$ can be used to compute (5). Thus, when unreported data are MAR, BC estimates are corrections of the estimates computed from the observed data. In this case, as the number of missing entries increases, the estimate (5) still approaches the prior probability α_{ijk}/α_{ij}, so that again we have a coherent estimate and no updating is performed when data are totally missing. In particular, if $n^\bullet(x_{ik}|\pi_{ij}) = n^\bullet_{ij}$ then (6) becomes

$$\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij}) + n^\bullet_{ij}\hat{\phi}_{ijk}}{\alpha_{ij} + n(\pi_{ij}) + n^\bullet_{ij}} = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij})} \tag{7}$$

which is a consistent estimate of θ_{ijk}, since $\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D})$ is a generalized version of the Maximum Likelihood Estimate of θ_{ijk}. This particular case is considered in [14]. If $\alpha_{ijk} = 0$, then (7) is the classical Maximum Likelihood Estimate of θ_{ijk} [10].

GS would complete the database by guessing the missing data from the current estimate of $p(x_{ik}|\pi_{ij})$, use the completed database to learn the parameters and iterate the procedure until convergence is reached. Clearly, the estimates of the conditional probabilities compute by (5) with ϕ_{ijk} replaced by $\hat{\phi}_{ijk}$ are the expected estimates and, as the database increases, they will be the same estimates returned by GS. The advantage of BC is that, as a deterministic method, it does not pose problems of convergence detection and monitoring, and it reduces the cost of estimating each conditional distribution of $X_i|\pi_{ij}$ to the cost of one exact Bayesian updating and one convex combination for each state of X_i.

3.3 Variance

Bounds (1) and (2) contain all estimates that could be obtained from all possible completed databases, and hence give an overall measure of the information available. The value in (5) is an estimate of the posterior expectation of θ_{ijk} that would be obtained from the complete database $\mathcal{D}$, before losing some of the entries. With a complete database, the exact posterior distribution would be $D(\alpha_{ij1}+n(x_{i1}|\pi_{ij}),\ldots,\alpha_{ijc_i}+n(x_{ic_i}|\pi_{ij}))$, so that the posterior variance of θ_{ijk} would be

$$V(\theta_{ijk}|\mathcal{D}) = \frac{(\alpha_{ijk}+n(x_{ik}|\pi_{ij}))(\alpha_{ij}-\alpha_{ijk}+n(\pi_{ij})-n(x_{ik}|\pi_{ij}))}{(\alpha_{ij}+n(\pi_{ij}))^2(\alpha_{ij}+n(\pi_{ij})+1)}.$$

Note that

$$V(\theta_{ijk}|\mathcal{D}) = \frac{E(\theta_{ijk}|\mathcal{D})(1-E(\theta_{ijk}|\mathcal{D}))}{\alpha_{ij}+n(\pi_{ij})+1}.$$

Since (5) is an estimate of $E(\theta_{ijk}|\mathcal{D})$, the posterior variance can be estimated as

$$\hat{V}(\theta_{ijk}|\mathcal{D}) = \frac{\hat{p}(x_{ik}|\pi_{ij},\mathcal{D},\phi_{ijk})(1-\hat{p}(x_{ik}|\pi_{ij},\mathcal{D},\phi_{ijk}))}{\alpha_{ij}+n(\pi_{ij})+1} \tag{8}$$

where $n(\pi_{ij})$ are the observed frequencies. Since $n(\pi_{ij})$ is smaller than or at most equal to the frequencies in $\mathcal{D}$, (8) will be an upper bound of $V(\theta_{ijk}|\mathcal{D})$, and approaches the exact posterior variance as the number of missing entries decreases. If we want to approximate the marginal posterior distribution of θ_{ijk}, we can use a moment-matching approximation [3], such as

$$\theta_{ijk}|\mathcal{D},\phi_{ijk} \sim D(\tilde{\alpha}_{ijk1},\tilde{\alpha}_{ijk2}),$$

where $\tilde{\alpha}_{ijk1},\tilde{\alpha}_{ijk2}$ are such that

$$\begin{aligned}
\hat{p}(x_{ik}|\pi_{ij},\mathcal{D},\phi_{ijk}) &= \frac{\tilde{\alpha}_{ijk1}}{\tilde{\alpha}_{ijk1}+\tilde{\alpha}_{ijk2}} \\
\hat{V}(\theta_{ijk}|\mathcal{D}) &= \frac{\tilde{\alpha}_{ijk1}\tilde{\alpha}_{ijk2}}{(\tilde{\alpha}_{ijk1}+\tilde{\alpha}_{ijk2})^2(\tilde{\alpha}_{ijk1}+\tilde{\alpha}_{ijk2}+1)}.
\end{aligned}$$

From these two equations it is easy to derive

$$\begin{aligned}
\tilde{\alpha}_{ijk1} &= \frac{\hat{p}(x_{ik}|\pi_{ij},\mathcal{D},\phi_{ijk})^2(1-\hat{p}(x_{ik}|\pi_{ij},\mathcal{D},\phi_{ijk}))}{\hat{V}(\theta_{ijk}|\mathcal{D})} - \hat{p}(x_{ik}|\pi_{ij},\mathcal{D},\phi_{ijk}) \\
\tilde{\alpha}_{ijk2} &= \frac{\hat{p}(x_{ik}|\pi_{ij},\mathcal{D},\phi_{ijk})(1-\hat{p}(x_{ik}|\pi_{ij},\mathcal{D},\phi_{ijk}))^2}{\hat{V}(\theta_{ijk}|\mathcal{D})} + \hat{p}(x_{ik}|\pi_{ij},\mathcal{D},\phi_{ijk}) - 1.
\end{aligned}$$

An alternative approximation could be based on the use of BC to estimate the expected posterior precision, as described in [12].

$p(X_1 = 1) = 0.5$	$p(X_2 = 1) = 0.4$
$p(X_3 = 1 \mid (X_1 = 1, X_2 = 1)) = 0.9$	$p(X_3 = 2 \mid (X_1 = 1, X_2 = 1)) = 0.05$
$p(X_3 = 1 \mid (X_1 = 1, X_2 = 2)) = 0.05$	$p(X_3 = 2 \mid (X_1 = 1, X_2 = 2)) = 0.5$
$p(X_3 = 1 \mid (X_1 = 2, X_2 = 1)) = 0.1$	$p(X_3 = 2 \mid (X_1 = 2, X_2 = 1)) = 0.6$
$p(X_3 = 1 \mid (X_1 = 2, X_2 = 2)) = 0.8$	$p(X_3 = 2 \mid (X_1 = 2, X_2 = 2)) = 0.01$

Table 1: Probabilities used to generate the random sample

4. Experimental Evaluation

GS is currently considered the most appropriate solution to the problem of learning BBNs from incomplete databases, although its limitations are well-known: its convergence rate is slow and it is resource consuming. The aim of these experiments is to compare the accuracy of the parameter estimates provided by GS and our method as the available information in the database decreases. Since GS assumes that unreported data are MAR, in order to make the comparisons meaningful, we had BC to make the same assumption. However, it is worth recalling that, while this assumption is embedded in the methodological nature of GS, BC can easily encode different patterns of missing data, if known.

4.1 Materials

We compared the implementation of BC provided by BKD [12] to the implementation of GS provided by the BUGS version 0.5 [16] on a Sun Sparc 5 under SunOS 5.5. The graphical structure of the BBN used in this experimental evaluation is the same as the one used in our previous description in Section 2 and depicted in Figure 1. The only difference between the two BBNs is that, while the variables X_1 and X_2 are still binary, X_3 is here ternary, so that the joint probability distribution specifying the BBN of our experiments is defined by 10 parameters. We generated a database of 1000 random cases from the probabilities reported in Table 1. The conditional probabilities were then estimated from the simulated sample. The prior distribution of the unknown parameters was a product of Dirichlet distributions defined as $\theta_1, \theta_2 \sim D(6,6)$, and $\theta_{3j} \sim D(1,1,1)$, $j = 1,2,3,4$. This is equivalent to assuming uniform prior probabilities in the BBN, and 12 as initial precision. Results are given in Figures 3, 4, 5, 6, 7 and 8, in which the parent configurations are coded as $\pi_{31} = (X_1 = 1, X_2 = 1)$, $\pi_{32} = (X_1 = 1, X_2 = 2)$, $\pi_{33} = (X_1 = 2, X_2 = 1)$ and $\pi_{34} = (X_1 = 2, X_2 = 2)$.

4.2 Methods

We ran 4 different tests, using 4 different missing data mechanisms. In all cases, the estimates computed with GS are based on a first burn-in of 2000 runs, that was enough to reach stability, and a final sample of 2000 cases. Estimates were then computed as sample means. The prior distribution of the parameters was the same used with complete data.

$p(X_1 =?) = 0.8$	$p(X_2 =?) = 0.07$
$p(X_3 =?\|(X_1 = 1, X_2 = 1)) = 0.76$	$p(X_3 =?\|(X_1 = 1, X_2 = 2)) = 0.04$
$p(X_3 =?\|(X_1 = 2, X_2 = 1)) = 0.50$	$p(X_3 =?\|(X_1 = 2, X_2 = 2)) = 0.58$

Table 2: Probabilities used to generate the incomplete sample in Test 1

Sample Parsed	$n(x_{21})$	Missing Entries
0%	404	0.0%
25%	320	3.0%
50%	242	5.0%
75%	162	8.0%
100%	80	11

Table 3: Proportion of missing entries in Test 2. Figures in the last column are the proportions of missing entries in the 4 incomplete samples.

4.2.1 Test 1: Missing at Random (MAR)

In the first experiment, data were randomly deleted in the following way. A vector β of 6 numbers in [0,1] was randomly generated, and elements of β were taken as the probabilities of deleting the occurrence of variable X_i, independently of its value, given the parent configuration, as shown in Table 2. The complete sample was randomized, that is original cases were randomly permuted, and then we run through each case in the sample and deleted the value of $X_i|\pi_{ij}$ with probability $p(X_i =?|\pi_{ij})$. This process generated 4 incomplete samples in which 10%, 20%, 30% and 40% of entries were removed. The deletion process ensures that the missing data mechanism is MAR. From each incomplete samples we computed BC estimates and stochastic estimates using GS.

4.2.2 Test 2: Informative Deletion of $X_2 = 1$ (INFX2)

In the second experiment, we deleted entries with $(X_2 = 1)$ with probabilities 0.8 in 25%, 50%, 75% and 100% of the complete sample. Thus, we generated 4 incomplete samples, in which only entries with $(X_2 = 1)$ were missing. The frequencies of remaining cases with $(X_2 = 1)$ in the 4 incomplete samples are given in Table 3, from which the proportion of missing entries reported in the last column is easily derived. Thus, at most 11% of entries are removed, and this corresponds to having at most 33% of incomplete cases in the sample. This missing data mechanism is NI because the probability of X_2 being missing depends on the state of X_2. As in the first test, from each incomplete samples we computed BC estimates and stochastic estimates.

4.2.3 Test 3: Informative Deletion of $X_2 = 1, X_3 = 3$ (INFX2,X3)

In the third experiment we deleted entries $(X_2 = 1, X_3 = 3)$ with probabilities 0.9 in 25%, 50%, 75% and 100% of the complete sample. Hence, as in the previous cases, we

Sample Parsed	$n(x_{33}\|\Pi_3 = (1,1))$	$n(x_{33}\|\Pi_3 = (2,1))$	Missing Entries
0%	7	84	0.0%
25%	3	67	0.7%
50%	3	44	1.5%
75%	1	28	2.0%
100%	1	7	2.8%

Table 4: Proportion of missing entries in Test 3. Figures in the last column are the proportions of missing entries in the 4 incomplete samples.

generated 4 incomplete samples, in which only entries with $(X_2 = 1, X_3 = 1)$ were missing. Frequencies of remaining $(X_2 = 1, X_3 = 3)$ and proportion of missing entries in the 4 incomplete samples are given in Table 4. Note that in this case, the proportion of missing entries in the 4 incomplete samples is very small. As in Test 2, the probability of X_2 and X_3 being missing depends on their values, so that the missing data mechanism here is NI. To conclude the test, we computed BC estimates and stochastic estimates using GS in all 4 incomplete samples.

4.2.4 Test 4: Systematic Deletion of $X_2 = 1$ (SYSX2)

In the last experiment, we systematically deleted entries $(X_2 = 1)$ in 25%, 50%, 75% and 100% of the complete sample, so that in the last sample no entry with $(X_2 = 1)$ is left. Since, in the complete sample, there are 404 entries $(X_2 = 1)$, the proportions of missing entries in the 4 samples are easily found to be 3.4%; 6.7%; 10.1%; 13.4%. As in all other cases, we computed BC estimates and stochastic estimates using GS in all incomplete samples.

4.3 Results

Figures 3, 4, 5, 6, 7 and 8 report estimates of the conditional probabilities quantifying the BBN used in this experiment against proportions of missing entries when data are MAR, and proportions of sample subject to missing data for the results of Tests 2, 3 and 4. Dotted lines are the lower and upper bounds reported by BC, and solid lines are BC estimates, computed under the assumption that data are MAR. Thus, weights used in the convex combination are computed from the observed entries in the database, as described in Section 3.2.2. Dashed lines display 95% credibility intervals computed using the approximate posterior distribution described in Section 3.3. Stochastic estimates computed with GS and relative 95% empirical credibility intervals are reported as $\times$ and error bars whose extreme points are 2.5% and 97.5% empirical quantiles. Figure 3 reports estimates of $p(x_{11})$ computed in the 4 incomplete samples generated in Test 1. In the other 3 tests, values of X_1 are always observed and, therefore, are not reported.

Apparent results for estimates computed when data are MAR are the increasing width of probability intervals computed in the bound step, and the overall accuracy of both BC and stochastic estimates until 70% of entries are left in the sample. This is not surprising since both methods are provided with the correct information about the missing data

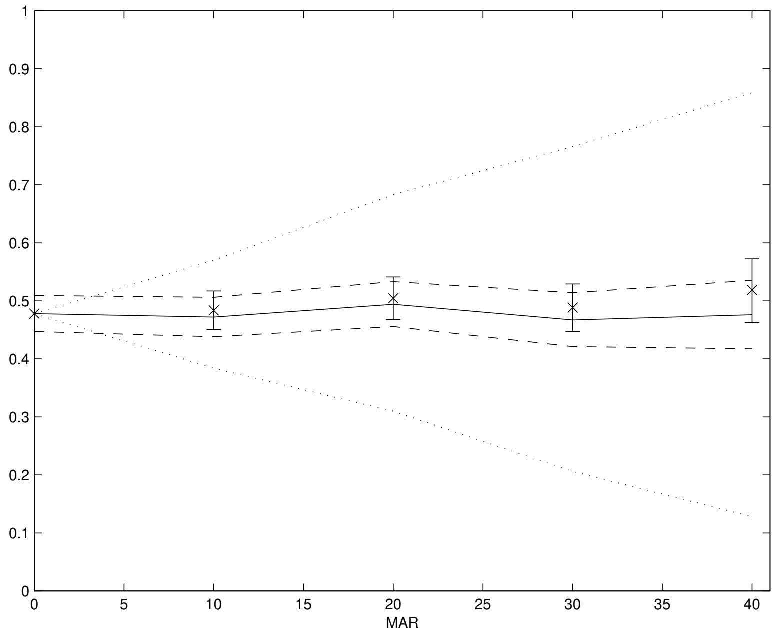

Figure 3: Estimates of $p(x_{11})$ when unreported data are MAR.

mechanism. Furthermore, probability intervals contain all possible estimates consistent with the available information, so that when data are heavily missing, the uncertainty about the estimates increases. When 40% of entries are missing, there is some bias in both BC and stochastic estimates of the conditional distributions of $X_3|\Pi_3 = (1,1)$ (Figure 5), $X_3|\Pi_3 = (2,1)$ (Figure 7) and $X_3|\Pi_3 = (2,2)$ (Figure 8). BC estimates $p(x_{3k}|(1,1))$ $(k = 1,2,3)$ as $(0.78, 0.16, 0.06)$ with 95% credibility intervals (0.614 0.900), (0.057 0.308) and (0.008 0.172). GS returns $(0.65, 0.19, 0.16)$ with 95% credibility intervals (0.524 0.788), (0.076 0.320) and (0.047 0.290). Compared to $(0.89, 0.07, 0.04)$ found in the complete sample, the largest bias is 0.24, and all estimates of θ_{31k} are ruled out from the 95% credibility intervals reported by GS. The reason for such a bias can be the slight overestimation of $p(x_{11})$ (0.52 instead of 0.48 in the complete sample.) With 40% of missing data, the observations on the variable X_1 are heavily missing, so that a small distortion in the estimate of $p(x_{11})$ can lead GS to inflate the sample of cases relevant to the estimation of $p(x_{3k}|(1,1))$ $(k = 1,2,3)$. Similarly, stochastic estimates of $p(x_{3k}|(2,2))$ $(0.7, 0.046, 0.24)$ exhibit a larger bias than those computed by BC $(0.86, 0.019, 0.117)$, since values computed in the complete sample are $(0.792, 0.0130.195)$. However, credibility intervals contain the estimates found in the complete database. Opposite results arise for the estimates of the conditional distribution of $X_3|\Pi_3 = (2,1)$. BC estimates $p(x_{3k}|(2,1))$ $(k = 1,2,3)$ as $(0.09, 0.66, 0.25)$ compared to the estimates $(0.10, 0.51, 0.39)$ in the complete sample, whilst GS returns $(0.05, 0.59, 0.35)$, so that the largest bias of the BC estimates is 0.15. However, credibility intervals found by BC (0.026 0.19), (0.51 0.79) and (0.13 0.39) contain the estimates found in the complete sample. Note however, that such intervals are not comparable to the probability intervals returned in the bound step. The latter contain all possible estimates consistent with the

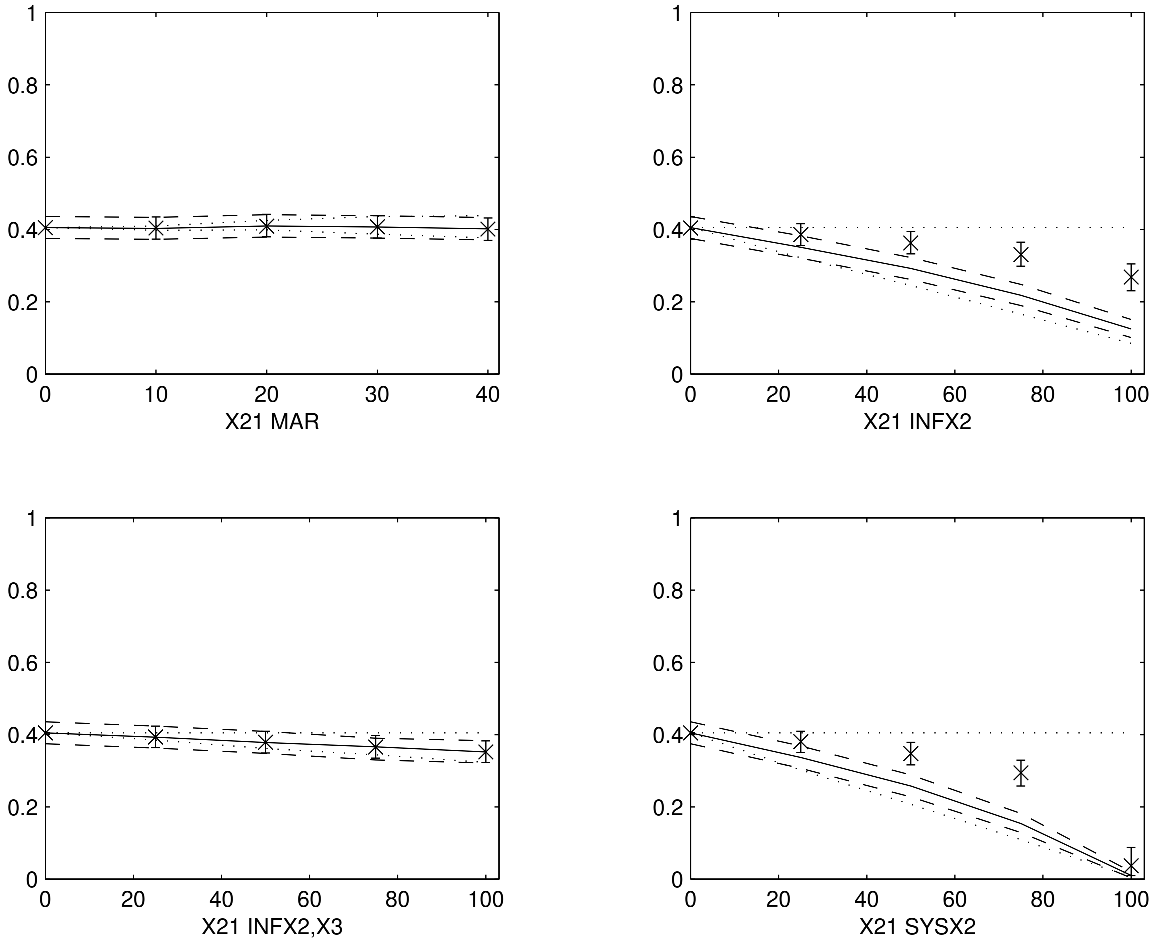

Figure 4: Estimates of $p(x_{21})$ in the 4 tests, that are represented from left to right, top to bottom.

available data, and hence are a measure of the quality of information conveyed by the sample about a parameter. Estimates given by GS, when the database is incomplete, are based on a sample of the most likely reconstructions of the missing entries, and this relies on the prior belief about the parameters, on the observed data, but most of all on the assumption that the missing data mechanism is MAR. Associated 95% credibility intervals represent posterior uncertainty about the parameters, given the assumed missing data mechanism. Thus, a tight interval denotes a high confidence in the posterior estimate, and the tightness is generally an increasing function of the sample size. However, the width of credibility intervals computed by GS does not seem to take into account the fact that the sample is incomplete, since there are no significant changes as the number of missing entries increases. Instead, the width of credibility intervals returned by BC increases as the number of missing data does. This cautious behavior of BC translates into more robust results. This effect becomes more evident when the missing data mechanism is NI and both methods continue

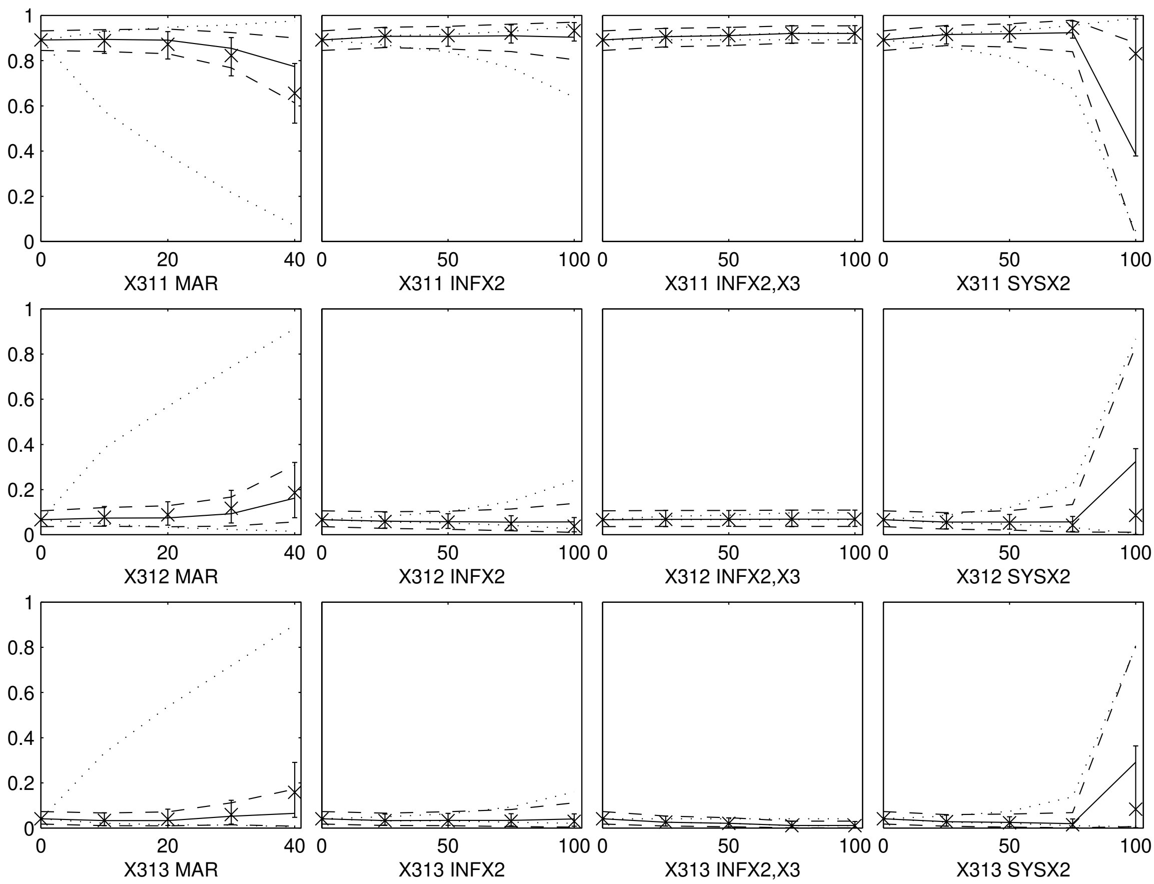

Figure 5: Estimates of $p(X_3 = x_{3k}|\Pi_3 = (1,1))$, $k = 1,2,3$ in the 4 tests.

to assume that unreported data are MAR. When data are missing on the variable X_2 alone with probability 0.9 (Test 2), we have comparable results for the estimates of θ_{31k}, but BC finds larger credibility intervals than GS, so that it is again aware of the uncertainty due to missing data in the database. Some bias is evident in the estimates of θ_2 and all other conditional probabilities. GS returns more accurate estimates of θ_2 compared to BC, particularly when data are heavily missing. However, the results of GS are rather surprising. Consider, for instance, the last sample which is totally subject to missing data. The frequencies of complete cases left after the deletion process are $n(X_2 = 1) = 80$, $n(X_2 = 2) = 496$, from which $\hat{\phi}_{21} = (6 + 80)/(12 + 80 + 496) = 0.15$. Thus if unreported data were MAR, so that the observed cases would be a representative sample of the complete one, we would expect to distribute the 424 cases with $X_2 =?$ as 64$\approx 0.15 \times 424$ with $(X_2 = 1)$ and the remaining 360 with $(X_2 = 2)$, and hence the expected estimate is $\hat{p}(x_{21}|\mathcal{D}) = 0.15$. BC returns 0.13. Thus data available, coupled with the MAR assumption are a natural explanation of the results computed by BC. On the other hand, GS guesses missing data from the available information, so that the final estimates are based on a sample of most likely reconstructions, and in this process it accounts somehow of the joint information conveyed by X_1, X_2 and

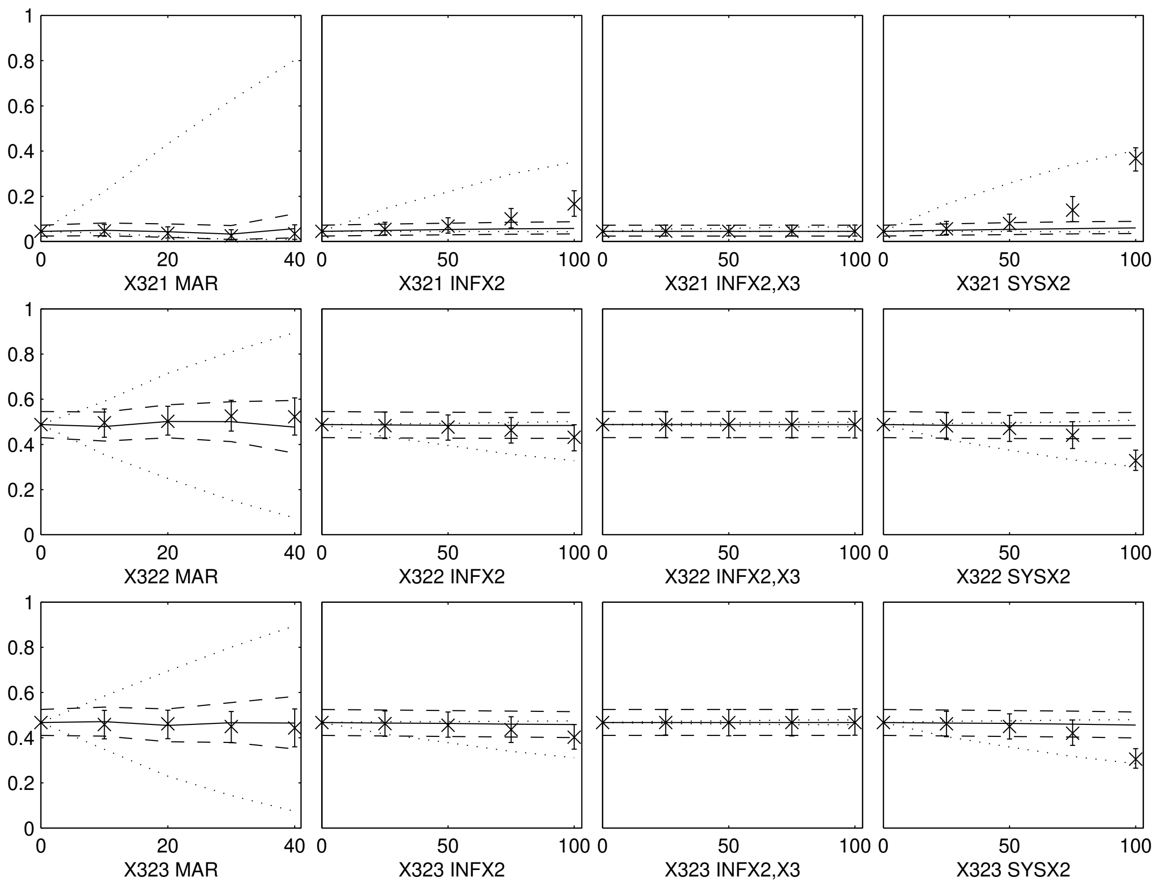

Figure 6: Estimates of $p(X_3 = x_{3k}|\Pi_3 = (1,2))$, $k = 1,2,3$ in the 4 tests.

X_3. If this strategy is rewarding for estimation of θ_2, there are several biased estimates of the conditional probabilities (Figures 6, 7 and 8) and in 17 cases the estimates of the conditional probabilities computed in the original database are ruled out by the 95% credibility intervals returned by GS.

Results of Test 3 are overall comparable, both methods return biased estimates of θ_{33k}, $k = 1,2,3$, whilst other estimates are extremely robust. In this case, the frequency of complete cases $(x_{3k}, (2,1))$ left in the sample after the deletion process are $(20,108,67)$ when 25% of the sample is parsed, $(20,108,44)$ when 50% is parsed, $(20,108,28)$ when 75% is parsed, and $(20,108,7)$ when the whole sample is subject to missing data. The MAR assumption yields estimates of θ_{333} that go to 0, and estimates of θ_{331} and θ_{332} that are pushed up.

Results of the last test are rather interesting. In this case, both methods are extremely sensitive to the MAR assumption in the estimation of θ_2 even though GS produces more accurate estimates than BC. However, as in Test 2, the way incomplete data are processed can be a winning strategy for estimation of θ_2 but it becomes dramatic for estimation of the other parameters. Consider first the estimates of θ_{31k}. Until 25% of the sample is

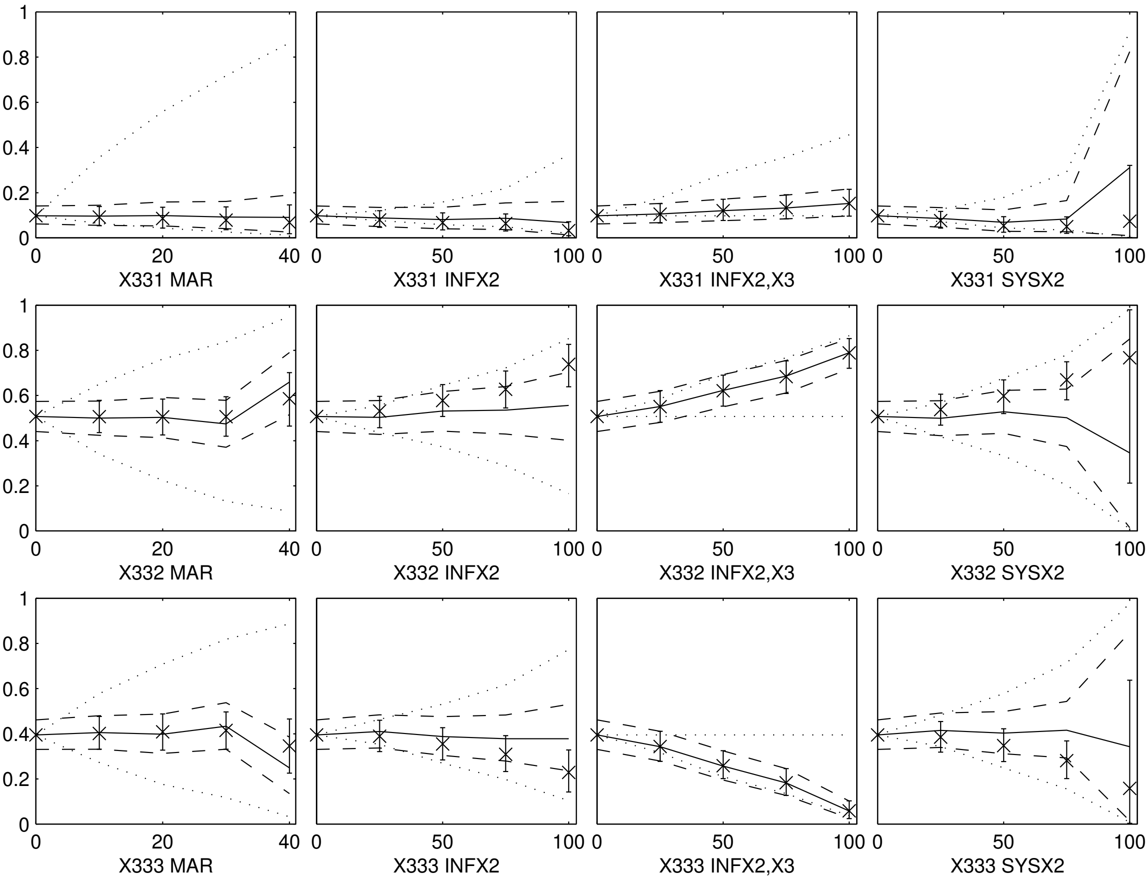

Figure 7: Estimates of $p(X_3 = x_{3k}|\Pi_3 = (2,1))$, $k = 1,2,3$ in the 4 tests.

complete, both stochastic and BC estimates are quite robust. In the final sample, in which all entries with $(X_2 = 1)$ have been removed, GS returns very accurate estimates, whilst BC returns estimates that are nearly uniform and large credibility intervals. Note that, in this case, the frequencies of $n(x_{3k}|(1,1))$ are all 0, so that BC estimates simply uses the prior information and the large credibility intervals associated are a further sign of the large uncertainty. This behavior is very reasonable, since there are no relevant data in the sample. On the other hand, it is a little bit mysterious the way in which GS process the information available from the data. This lack of controllability becomes even clearer with estimation of θ_{33k}. Stochastic estimates are now extremely biased, even when only 50% of the sample is subject to deletion of $(X_2 = 1)$. BC instead, returns very reliable estimates until 75% of cases $(X_2 = 1)$ are removed, and then coherently uniform estimates are returned in the last sample, in which frequency of $(x_{3k}, (2,1))$ are all 0. The bias of GS is even more apparent for the estimation of θ_{32k} and θ_{34k} when more than 50% of the sample is subject to missing data. For instance, estimates of θ_{32k} in the final sample are $(0.17, 0.43, 0.40)$ with associated 95% credibility intervals (0.11 0.22), (0.37 0.49) and (0.35 0.46), so that the estimates computed in the original complete database $(0.045, 0.498, 0.467)$ are all ruled out. BC estimates, on the

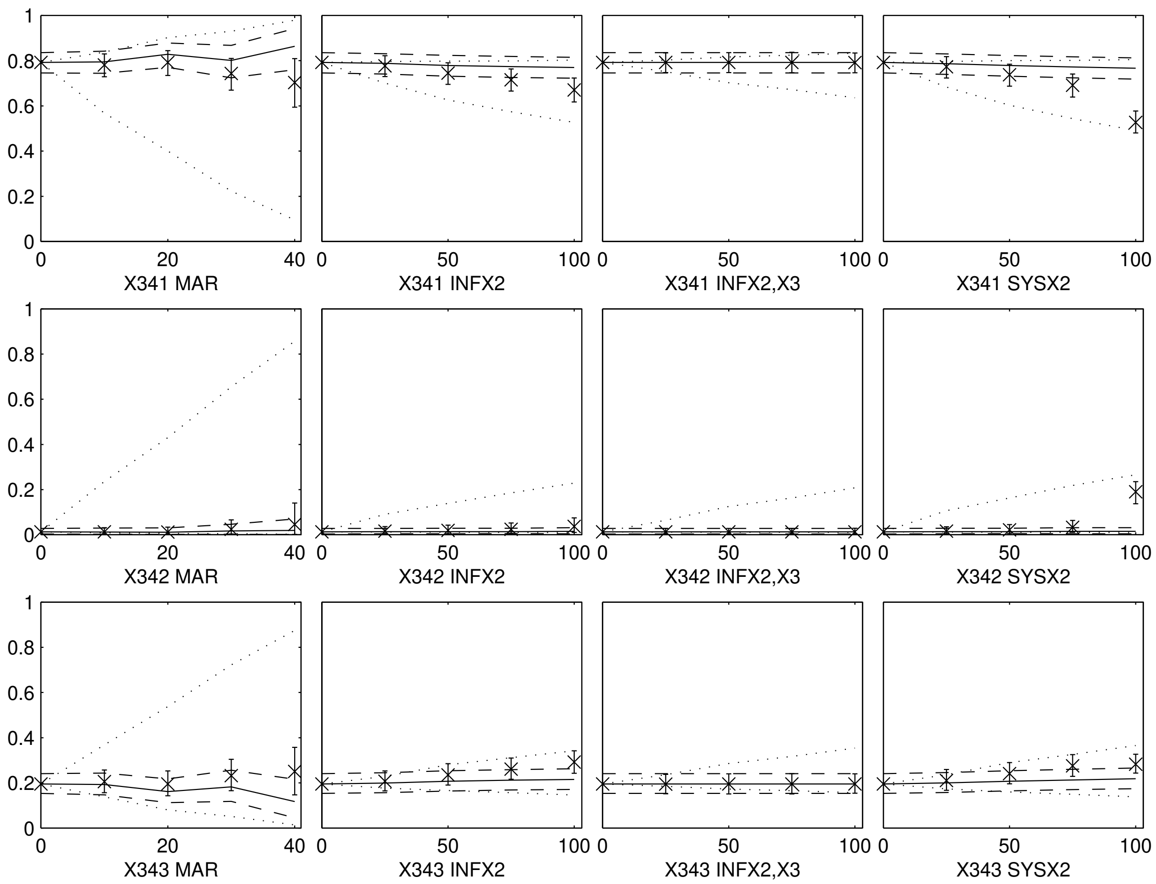

Figure 8: Estimates of $p(X_3 = x_{3k}|\Pi_3 = (2,2))$, $k = 1,2,3$ in the 4 tests.

other hand, do not seem to be affected by missing data. Note that, in this case, frequencies of $(x_{3k},(1,2))$ are $(12,140,134)$ as in the complete sample, and frequencies of incomplete cases $(X_1 = 1, X_2 =?)$, $(X_1 = 2, X_2 =?)$ are respectively $(137,8,4)$ and $(17,88,70)$. These frequencies are used to compute bounds on the possible estimates of θ_{32k}, but the effect of $\hat{\phi}_{32k} = (0.045,0.488,0.467)$ is negligible, since $\hat{p}(x_{3k}|(1,2),\mathcal{D}) = (0.058,0.484,0.458)$. Thus, when some information is available, BC seems to be able to exploit this information more effectively than GS. The strategy of GS of guessing missing data on the basis of the available information can be extremely biased, because final estimates can be based on inflated samples. Furthermore, the tight credibility intervals associated to the estimates, can give the analyst a wrong confidence in the conclusions. BC is more cautious and controllable, and indeed, the number of errors cumulated in the 4 tests is by far smaller than the number of errors of GS (Table 5.) Both methods, however, show the risk of imputing a wrong missing data mechanism that can yield severe bias. However, BC provides probability intervals, which can make the analyst aware of the range of possible estimates, and hence of the quality of information on which inference is based. A further difference between the performances of the 2 systems has been the execution time: in the worse case, GS took over 12 minutes

	BC					GS				
Test	s_1	s_2	s_3	s_4	Tot	s_1	s_2	s_3	s_4	Tot
MAR	0	0	0	0	0	0	0	0	3	3
INFX2	1	1	1	1	4	0	2	6	9	17
INFX2,X3	0	0	4	4	8	0	2	4	4	10
SYSX2	1	1	1	2	5	0	6	9	7	22
Tot	2	2	6	7	17	0	10	19	23	52

Table 5: Estimation errors cumulated by BC and GS in the 4 samples s_i generated in each test. An error is made when the 95% credibility interval associated to the estimate that does not contain the estimate computed in the complete sample

to analyze a sample, while BC ran to completion in less than 2 seconds, independently of the number of missing entries. This means that, with a minimum effort, other processes of missing data could be tested, and the sensitivity of the estimates on the assumed missing data mechanism can be quickly examined.

5. Conclusion

BC provides a method to learn conditional probabilities in a BBN from incomplete databases. Contrary to the current methods, BC does not try to complete the database, but it computes the extreme points of the possible distributions consistent with the database and then collapses them into a point estimate using the available information about the pattern of missing data. This feature allows the encoding of exogenous knowledge about the pattern of missing data, and subsumes methods using the MAR assumption as special cases, providing performances equal or even higher in accuracy. Furthermore, the probability intervals used by BC provide a specific measure of the *quality* of information conveyed by the database and an explicit representation of the impact of the assumption made on the pattern of missing data. From a computational standpoint, BC has all the advantages of a deterministic method: it provides a stopping rule — while GS is guaranteed to converge only asymptotically — and its computational cost is reduced to the cost of one exact updating and one convex combination for each state of $X_i|\pi_{ij}$. A final remark is due: although this paper focuses on BBNs, BC is a general method to learn conditional probabilities and may be applied to any data analysis task.

Acknowledgments

Authors thank Greg Cooper, Paul Snow, and Zdenek Zdrahal for their helpful suggestions during the development of this research, and the anonymous referees for their useful comments on the first version of this paper. This research was partially supported by equipment grants from Apple Computer and Sun Microsystems.

References

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

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

[3] R.G. Cowell, A.P. Dawid, and P. Sebastiani. A comparison of sequential learning methods for incomplete data. In *Bayesian Statistics 5*, pages 533–542. Clarendon Press, Oxford, 1996.

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

[5] D Geiger and D Heckerman. A characterization of Dirichlet distributions through local and global independence. *Ann. Statist.*, 25:1344–1368, 1997.

[6] A Gelman, J B Carlin, H S Stern, and D B Rubin. *Bayesian Data Analysis.* Chapman and Hall, London, 1995.

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

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

[9] S L Lauritzen. *Graphical Models.* Clarendon Press, Oxford, 1996.

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

[11] 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.

[12] M. Ramoni and P. Sebastiani. Learning Bayesian networks from incomplete databases. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, San Mateo, CA, 1997. Morgan Kaufmann.

[13] D B Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[14] P. Sebastiani and M. Ramoni. Bayesian inference from data subject to non response using bound and collapse. Technical Report KMi-TR-48, Knowledge Media Institute, The Open University, 1997. Available at http://kmi.open.ac.uk/techreports/KMi-TR-48.

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

[16] A. Thomas, D.J. Spiegelhalter, and W.R. Gilks. Bugs: A program to perform Bayesian inference using Gibbs Sampling. In *Bayesian Statistics 4*, pages 837–42. Clarendon Press, Oxford, 1992.