KMi

KNOWLEDGE MEDIA INSTITUTE

The Use of Exogenous Knowledge to Learn Bayesian Networks from Incomplete Databases

Marco Ramoni Paola Sebastiani

KMI-TR-44

February 1997

The Open University

The Use of Exogenous Knowledge to Learn 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 the information provided by the database and the 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 the Gibbs Samplings are also provided.

Keywords: Applications: *Information extraction;* Theory and General Principles: *Uncertainty and noise in data;* Algorithms and Techniques: *Bayesian inference.*

Reference: KMi Technical Report KMi-TR-44 (February1997).

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) is defined by a set of *variables* $\mathcal{X} = \{X_1, \ldots, X_I\}$ and a direct acyclic graph defining a structure S 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 state of the parent variables. We shall consider discrete variables only. The structure S yields a factorization of the joint probability of a particular set of values of the variables in $\mathcal{X}$, say $x_j = \{x_{1j}, \ldots x_{Ij}\}$, as

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

where π_{ij} denotes the state of Π_i in x_j. In the following we will denote $X_i = x_{ij}$ as x_{ij}, and $\Pi_i = \pi_{ij}$ as π_{ij}. Several efforts have been addressed to develop Bayesian methods to learn BBNs from databases [2, 1, 5]. Along this approach, we shall consider the conditional probabilities defining the BBN as being generated by parameters $\theta = \{\theta_1, \ldots, \theta_K\}$, so that the joint probability of a case x_j is

$$p(x_j|\theta) = \prod_{i=1}^{I} p(x_{ij}|\pi_{ij}, \theta_i),$$

where θ_i parameterizes the probability of x_{ij} given the parent configuration π_{ij}. We assume that the parameters θ_k are random variables, whose prior distribution represents the observer's belief about the conditional probabilities specifying the BBN and that we are given a database of J cases $\mathcal{D} = \{x_1, \ldots, x_J\}$ from which we want to update the prior knowledge on the parameters. Given the database, the prior density $p(\theta)$ is updated in the posterior density by using Bayes' theorem, and hence $p(\theta|\mathcal{D}) = p(\theta)p(\mathcal{D}|\theta)/p(\mathcal{D})$, where $p(\mathcal{D})$ is the marginal probability of the database, and $p(\mathcal{D}|\theta) = \prod_j p(x_j|\theta)$, if different cases are independent given θ. The Bayesian estimate of θ is then the posterior expectation $E(\theta|\mathcal{D})$ of θ. We shall assume that the parameters are mutually independent and have a Dirichlet distribution. Thus, by taking advantage of local computation and conjugate analysis [9], if the prior distribution of the parameter vector $\theta_i = \{\theta_{i1}, \ldots, \theta_{iM}\}$ associated to the conditional probabilities $p(x_{im}|\pi_{ij})$, $m = 1, \ldots, M$, is a Dirichlet $D(\alpha_{i1}, \ldots, \alpha_{iM})$, $\alpha_{im} > 0$, then the posterior distribution is still a Dirichlet $D(\alpha_{i1} + n(x_{i1}|\pi_{ij})), \ldots, \alpha_{iM} + n(x_{iM}|\pi_{ij})))$, where $n(x_{im}|\pi_{ij})$ is the frequency of cases in the database with $x_{im}|\pi_{ij}$. The Bayes estimate of the conditional probability of $x_{im}|\pi_{ij}$, given the information in the database, i.e. the posterior mean of θ_{im}, is then

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

where $n(\pi_{ij}) = \sum_m n(x_{im}|\pi_{ij})$. Unfortunately, the situation is quite different when some entries in the database are missing. Suppose for instance that the case x_j is incomplete, with entries on X_i and/or its parents missing, so that $\mathcal{D}_{inc} = \mathcal{D}_{obs} \cup x_j$, with $\mathcal{D}_{obs}$ denoting the part of the database with complete entries. Exact analysis would require the computation of the joint posterior distribution of the parameters by considering all possible completions of the incomplete case. Assuming that properties of local and global independence are retained, the exact updating can be performed locally, so that the posterior distribution of the parameters associated to the conditional distribution of $X_i|\pi_{ij}$ turns out to be the mixture:

$$\sum_m D(\alpha_{i1} + n(x_{i1}|\pi_{ij}) + \delta_1(m), \ldots, \alpha_{iM} + n(x_{iM}|\pi_{ij}) + \delta_M(m))p(x_{im}, \pi_{ij}|\mathcal{D}_{obs})$$
$$+D(\alpha_1 + n(x_{i1}|\pi_{ij}), ..., \alpha_M + n(x_{iM}|\pi_{ij}))(1 - p(\pi_{ij}|\mathcal{D}_{obs})$$

where $\delta_i(j) = 1$ if $i = j$ and 0 otherwise, and the counts are based on $\mathcal{D}_{obs}$. The first term in the mixture depends on the possible completion of the child variable X_i given the state π_{ij} of the parents variables. The last term takes into account the completions that would lead to the updating of different conditional distributions. As the number of completions increase, the exact local updating is apparently infeasible: its complexity is in fact exponential in the number of missing data in the database [2], and approximate methods therefore are required.

Deterministic methods, such as fractional updating or matched moments [9, 3] are a possible solution. However, the need to process data sequentially, and the fact that the results of the learning process are sensitive to the order of cases in the database are serious limitations that can impair their applicability. An alternative approach, very popular nowadays, is the use of a stochastic approximation of the posterior distribution using Markov Chain Monte Carlo (MCMC) methods, such as the Gibbs Sampling, see [4] for a recent review. When some of the entries in the database are missing, the Gibbs Sampling treats the missing data as unknown parameters, so that in each iteration, for each missing entry, a value is sampled from the conditional distribution of the corresponding variables, given the available data, all the parameters and the values simulated for the other missing entries. The algorithm is iterated to reach stability, and then a sample from the joint posterior distribution is taken which can be used to provide empirical estimates of the posterior means [10]. Convergence rate and convergence detection are open issues of this stochastic method. These methods share a common strategy: they complete the database by inferring somehow the missing data from the available information and then learn from the completed database. The underlying assumption is that the unreported data are Missing at Random (MAR) so that $\mathcal{D}_{obs}$ is a representative sample of $\mathcal{D}$. Experimental evidences [8, 7] show that, when this assumption is violated, these methods are not robust with respect to the pattern of missing data.

Ramoni and Sebastiani [7] introduced an efficient method to learn conditional probabilities from incomplete databases which does not rely on this assumption. The method *bounds* the set of possible distributions consistent with the available information by computing the minimum and the maximum Bayes estimate that would be obtained from the possible completions of the database. This paper presents a technique to use prior knowledge about the pattern of missing data to identify a unique distribution within these bounds. This new method *collapses* the set of possible distributions into a single one via a convex combination of the extreme points with weights depending on the assumed pattern of missing data. An approximation of the variance of the estimates is also provided. 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, without any need to guess the missing data. Experimental valuations show that the estimates provided by

BC are equivalent to the ones provided by the Gibbs Sampling, when data are MAR, and they are more robust to departure from the true pattern of missing data, thus making the computational effort of stochastic methods vane.

2. Method

In [7] it was shown that the missing entries in the database lead to bounds on the possible estimates that could be obtained from the complete database. Thus the *bound* step of the BC method computes, for each parameter, a probability interval whose extreme points are the minimum and the maximum Bayes estimate that would be inferred from all possible completions of the database.

Let X_i be a variable in the BBN and $n^\bullet(x_{im}|\pi_{ij})$ be the frequency of cases with $X_i = x_{im}$, 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. Clearly $n^\bullet(x_{im}|\pi_{ij}) = n^\bullet_{ij}$ for all m when data are missing only on the child variable. Suppose further that we start from total ignorance, so that the parameter vector θ_i which is associated to $p(x_i|\pi_{ij})$ is assigned a uniform prior. Then, the Bayes estimate $E(\theta_{im}|\mathcal{D})$, that would be computed from the complete database, is bounded above by

$$p^\bullet(x_{im}|\pi_{ij}, \mathcal{D}_{inc}) = \frac{1 + n(x_{im}|\pi_{ij}) + n^\bullet(x_{im}|\pi_{ij})}{M + \sum_m n(x_{im}|\pi_{ij}) + n^\bullet(x_{im}|\pi_{ij})} \qquad (1)$$

and below by

$$p_\bullet(x_{im}|\pi_{ij}, \mathcal{D}_{inc}) = \frac{1 + n(x_{im}|\pi_{ij})}{M + \sum_m n(x_{im}|\pi_{ij}) + \max_{l \neq m} n^\bullet(x_{il}|\pi_{ij})} \qquad (2)$$

Each maximum probability $p^\bullet(x_{im}|\pi_{ij}, \mathcal{D}_{inc})$ is obtained from a Dirichlet distribution $D_m(1 + n(x_{i1}|\pi_{ij}), \ldots, 1 + n(x_{im}|\pi_{ij}) + n^\bullet(x_{im}|\pi_{ij}), \ldots, 1 + n(x_{iM}|\pi_{ij}))$ which identifies a unique probability for the other states of the variable X_i given π_{ij}, say $p_{m\bullet}(x_{il}|\pi_{ij}, \mathcal{D}_{inc})$, and $p_\bullet(x_{im}|\pi_{ij}, \mathcal{D}_{inc}) = \min_l\{p_{l\bullet}(x_{im}|\pi_{ij}, \mathcal{D}_{inc})\}$. These bounds depend only on the frequencies of complete entries in the database and the "artificial" frequencies of the completed entries, so that the whole data can be processed at once. This method is implemented using an algorithm based on discrimination trees to efficiently store information about the parameters [6].

The main feature of this method is that it does not rely on any assumption on the distribution of the missing data because it does not try to infer them: the available information can only give rise to constraints on the possible estimates that could be learned from the database. Furthermore, the use of this robust method provide BC with the a new measure of information: the width of the interval accounts for the amount of information available in $\mathcal{D}_{inc}$ about the parameter to be estimated, and represents a measure of the quality of the probabilistic information conveyed by the database about a parameter. 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.

The second step of BC method collapses the interval estimated in the *bound* step into a point estimate using a convex combination of the distributions D_m induced by the maximum probabilities. This convex combination can be achieved either by the use of external information about the pattern of missing data or by a dynamic estimation of this pattern from the available information in the database. Suppose that some information on the pattern of missing data is available, which allow the user to formulate a probability distribution describing, for each incomplete case in the database, the probability of a completion as:

$$p(x_{im}|\pi_{ij}, X_i =?, \theta) = \phi_{ijm} \tag{3}$$

where $m = 1, ..., M$ and $\sum_{m=1}^{M} \phi_{ijm} = 1$. Such information can be used to summarize the interval estimate in a point estimate via a convex combination of the extreme probabilities:

$$\hat{p}(x_{im}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijm}) = \sum_{l \neq m} \phi_{ijl} p_{l\bullet}(x_{im}|\pi_{ij}, \mathcal{D}_{inc}) + \phi_{ijm} p^{\bullet}(x_{im}|\pi_{ij}, \mathcal{D}_{inc}). \tag{4}$$

This point estimate is the expected Bayes estimate that would be obtained from the complete database $\mathcal{D}$, if known, when the mechanism generating the missing data in the database is described by (3). Note that the estimates so found define a probability distribution since $\sum_{m=1}^{M} \hat{p}(x_{im}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijm}) = 1$. The intuition behind (4) is that the lower bound of the estimate of $p(x_{im}|\pi_{ij})$ is achieved when all the incomplete cases are assigned to $x_{ih}|\pi_{ik}$ and $n \neq m$ or $k \neq j$, and the upper bound is achieved when all the incomplete cases are completed as $x_{im}|\pi_{ij}$. Thus if the user specifies that $p(x_{im}|\pi_{ij}, X_i =?, \theta) = 1$ for a particular m, then (4) will return the upper bound of the interval probability as estimate of $p(x_{im}|\pi_{ij})$, and the lower bounds of the other interval probabilities as estimates of $p(x_{ih}|\pi_{ij})$, $h \neq m$. This case corresponds to the assumption that data are *systematically* missing about a particular state. On the other hand, when no information on the mechanism generating missing data is available and therefore any pattern of missing data is equally likely, then $\phi_{ijm} = 1/M$. Note finally that if $n^{\bullet}(x_{im}|\pi_{ij}) = n_{ij}^{\bullet}$ then (4) simplifies to

$$\frac{1 + n(x_{im}|\pi_{ij}) + n_{ij}^{\bullet}\phi_{ijm}}{M + \sum_{m} n(x_{im}|\pi_{ij}) + n_{ij}^{\bullet}} \tag{5}$$

Suppose now that the user has sufficient information to believe that data are MAR, so that $\mathcal{D}_{obs}$ is a *representative sample* of $\mathcal{D}$. Thus the probability of a completion can be estimated from $\mathcal{D}_{obs}$ as

$$\hat{\phi}_{ijm} = \frac{1 + n(x_{im}|\pi_{ij})}{M + \sum_m n(x_{im}|\pi_{ij})},$$

and then $\hat{\phi}$ can be used to compute (4). In particular (5) becomes

$$\hat{p}(x_{im}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijm}) = \frac{1 + n(x_{im}|\pi_{ij}) + n_{ij}^{\bullet}\hat{\phi}_{ijm}}{M + \sum_m n(x_{im}|\pi_{ij}) + n_{ij}^{\bullet}} \tag{6}$$

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

$$\frac{1 + n(x_{im}|\pi_{ij}) + n_{ij}^{\bullet} p(x_{im}|\pi_{ij}, \theta)}{M + \sum_m n(x_{im}|\pi_{ij}) + n_{ij}^{\bullet}}.$$

The Gibbs Sampling would complete the database by guessing the missing data from the current estimate of $p(x_{im}|\pi_{ij}, \theta)$, use the completed database to learn the parameters and iterate the procedure until convergence is reached. Clearly, the estimates of the conditional probabilities returned in (4) with ϕ_{ijm} replaced by $\hat{\phi}_{ijm}$ are the expected estimates, and for large samples they will be the same. On the other hand, the computational cost of BC is reduced to the cost of two exact updating — one for each extreme distribution — plus the cost of a convex combination for each parameter in the BBN.

The bounds associated to the point estimates returned by our method contain the estimates that could be obtained from all possible completed databases, and hence give an overall measure of the information available. The value in (4) is an approximation of the Bayesian estimate of θ_{im} that would be obtained from the complete database $\mathcal{D}$, before losing some of the entries. Such estimate would be the posterior expectation of a Dirichlet distribution with parameters $(1 + n(x_{i1}|\pi_{ij})), \ldots, 1 + n(x_{iM}|\pi_{ij})))$ and counts based on $\mathcal{D}$, so that the posterior variance would be

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

Note that

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

Since (4) is an approximation of $E(\theta_{im}|\mathcal{D})$, we can approximate the posterior variance by

$$\hat{V}(\theta_{im}|\mathcal{D}_{inc}) = \frac{\hat{p}(x_{im}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijm})(1 - \hat{p}(x_{im}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijm}))}{M + n(\pi_{ij}) + 1} \tag{7}$$

where $n(\pi_{ij})$ are the counts observed in $\mathcal{D}_{obs}$. Since $n(\pi_{ij})$ is smaller than or at most equal to the counts in $\mathcal{D}$, (7) will be an upper bound of $V(\theta_{im}|\mathcal{D})$. If we want to approximate the marginal posterior distribution of θ_{im}, we can further assume $\theta_{im}|\mathcal{D}_{inc}, \phi_{ijm} \sim D(\tilde{\alpha}_{m1}, \tilde{\alpha}_{m2})$, where $\tilde{\alpha}_{m1}, \tilde{\alpha}_{m2}$ are such that

$$\begin{aligned} \hat{p}(x_{im}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijm}) &= \tfrac{\tilde{\alpha}_{m1}}{\tilde{\alpha}_{m1} + \tilde{\alpha}_{m2}} \\ \hat{V}(\theta_{im}|\mathcal{D}_{inc}) &= \tfrac{\tilde{\alpha}_{m1}\tilde{\alpha}_{m2}}{(\tilde{\alpha}_{m1} + \tilde{\alpha}_{m2})^2 (\tilde{\alpha}_{m1} + \tilde{\alpha}_{m2} + 1)}. \end{aligned}$$

3. Experimental Evaluation

The Gibbs Sampling is currently considered one of the most feasible solutions to the problem of learning BBNs from databases with missing data, although its limitations are well-known: the convergence rate is slow and resource consuming. The aim of these experiments were to compare the accuracy of the parameter estimates provided by the Gibbs Sampling and our method as the available information in the database decreases. Therefore, in order to make the comparison with Gibbs Sampling meaningful, BC was provided with the MAR assumption.

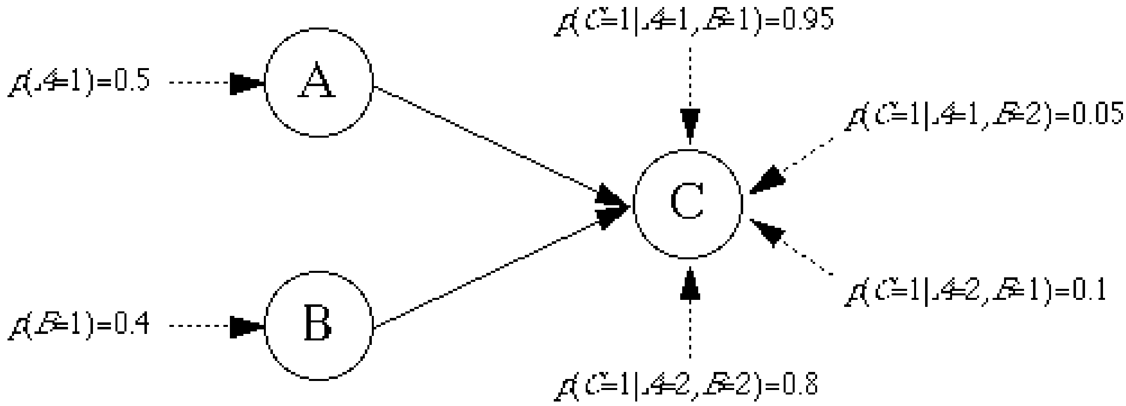

Figure 1: The BBN used for the experimental evaluation.

3.1 Materials and Methods

We compared an implementation of BC written in Common Lisp under CLISP version 1996/10/10 to the implementation of the Gibbs Sampling provided by the BUGS version 0.5 [10] on a Sun Sparc 5 under SunOS 5.5. Figure 1 shows the graphical structure of the BBN — defined by three binary variables A, B, and C — used for this comparison. We generated a database of 1000 random cases from the probability distribution reported in Figure 1.

3.2 Results

Using the materials described above, we ran two different experiments, one fulfilling and one violating the MAR assumption. In the first experiment, we started with a complete database, where all the parameters are independent and uniformly distributed, and ran both our algorithm and the Gibbs Sampling on it. Then, we proceed by randomly deleting 20% of the entries of the database and running the two methods until the database was empty. Each run of the Gibbs Sampling is based on 2,000 iterations, which were sufficient to reach stability, and a final sample of 2,000 cases. Figure 2 plots the parameter estimates given by the two systems against the percentage of entries still present in the database. Dotted lines are the bounds computed by our method. Solid lines links the point estimates returned by our method under the MAR assumption, so that the weights used in the convex combination are computed from the observed entries in the database. Dashed lines display the 95% confidence intervals around the BC estimates, which were computed using the approximate posterior distributions for the parameters described at the end of section 2. Stars represent the point estimates given by the Gibbs Sampling and error bars reports the 95% confidence intervals for these estimates. Both the BC and the Gibbs Sampling confidence intervals are based on the 2.5 % and 97.5 % quantiles. Not surprisingly BC under the MAR assumption returns the accurate estimates of the Gibbs Sampling, with confidence intervals almost identical. The main difference between the performances of the two systems has been the execution time: in the worse case, Gibbs Sampling took over 3

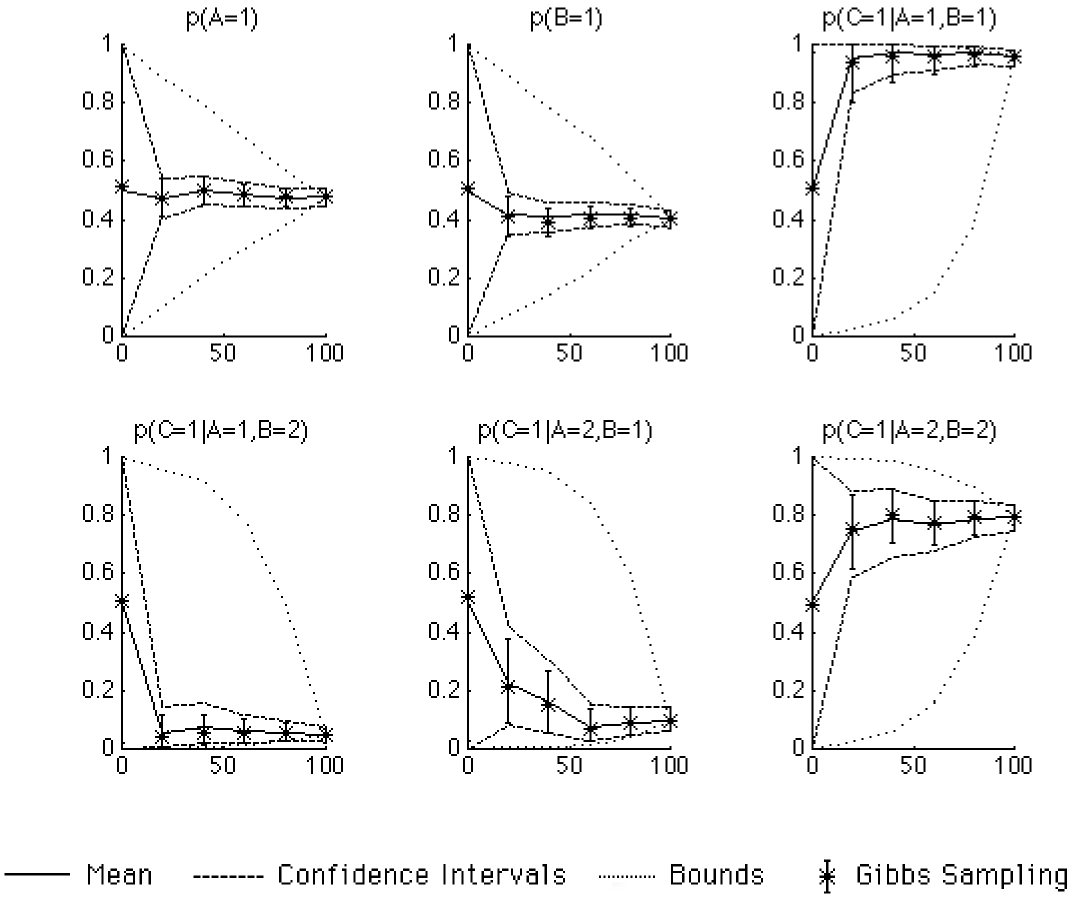

Figure 2: Estimates against percentage of entries. Entries are randomly missing.

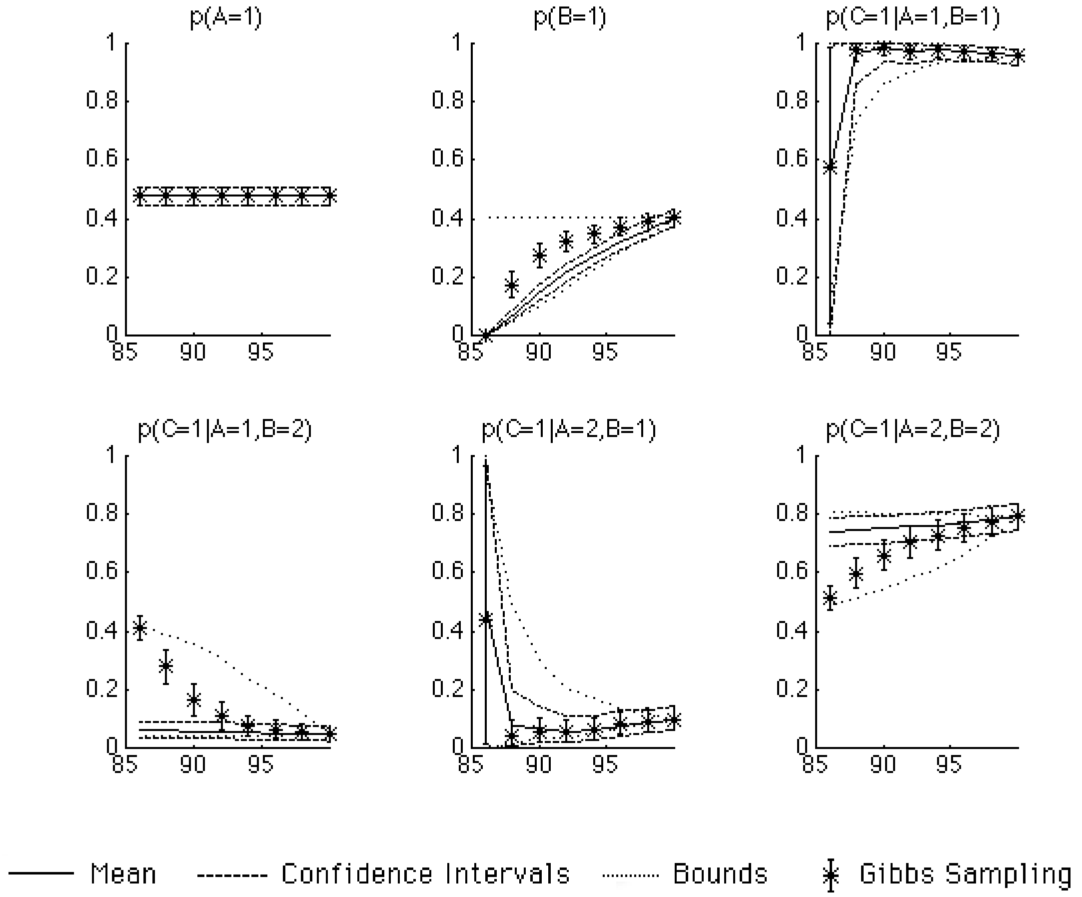

Figure 3: Estimates against percentage of entries. Entries $B = 1$ are removed.

minutes to run to completion, while BC ran to completion the same task in less than 20 milliseconds.

The procedure followed during the second experiment is identical to the procedure used in the first one, except that, at each step, the 2% of the entries $B = 1$ were systematically deleted until no entry $B = 1$ was reported in the database. Results are summarized by Figure 3. As expected, the estimates returned by the Gibbs Sampling and by our method using the weight computed from the entries actually observed in the database are almost identical. An apparently surprising exception occurs in the estimation of $p(C = 1|A = 2, B = 2)$ and $p(C = 1|A = 1, B = 2)$: with the 86% of the database but all the entries $B = 1$ removed, the Gibbs Sampling returns in the first case an estimate of 0.4123 and 0.5876, respectively. Both estimates include an error of almost 40% with respect to the estimates computed for the same parameters with the complete databases: 0.0453 and 0.9546 respectively. On the contrary, BC under the same MAR assumption, catches the correct value even when all the observations $B = 1$ are missing. Again, the difference in computation time was significant: while our implementation of BC took always less than 20 milliseconds to run to completion, the Gibbs Sampler took up to 8 minutes to complete its analysis.

3.3 Discussion

The first important result of our experiment is the overall equivalence between the Gibbs Sampling and the BC estimates, which are almost identical when the MAR assumption is fulfilled by the database. When this assumption is violated, BC seems to be significantly more robust than the Gibbs Sampling: the little gain in the estimates of $p(B=1)$ is a small reward for the appalling error in the estimates of $p(C=1|A=1,B=2)$ and $p(C=1|A=2,B=2)$. The strategy of guessing the missing data from the available information leads the Gibbs Sampler to ascribe most of the missing entries to $B=2$, so that the final estimates of $p(C=1|A=1,B=2)$ and $p(C=1|A=2,B=2)$ are based on an inflated sample. On the other hand, the BC estimates computed under the MAR assumption are based on weights which depend only on the actually observed sample. It is worth noting that, if information about the systematic pattern of missing data were available, this could have been explicitly encoded in BC by setting $\phi(B=2)=1$: the estimations returned by BC would have collapsed on the bounds of the reported intervals and, therefore, even more accurate than the estimates returned under the MAR assumption. The second point is the significant difference in computational cost: milliseconds against minutes.

4. 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 before learning, 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 knowledge external to the database about the pattern of missing data, and subsumes the methods using 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 the Gibbs Sampling is guaranteed to converge only asymptotically — and its computational cost is reduced to the cost of two exact updating. 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 system based on probability theory.

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] D. Heckerman. A tutorial on learning Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Microsoft Corporation, 1995.

[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] 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.

[7] M. Ramoni and P. Sebastiani. Robust parameter learning in Bayesian networks with missing data. In *Proceedings of the Sixth Workshop on Artificial Intelligence and Statistics*, pages 339–406, Fort Lauderdale, FL, 1997.

[8] D J Spiegelhalter and R G Cowell. Learning in probabilistic expert systems. In *Bayesian Statistics 4*, pages 447–466. Clarendon Press, Oxford, 1992.

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

[10] 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.