KMi

Knowledge Media Institute

Robust Parameter Learning in Bayesian Networks with Missing Data

Marco Ramoni   Paola Sebastiani

KMI-TR-29

July 1996 (revised October 1996)

The Open University

Robust Parameter Learning in Bayesian Networks with Missing Data

Marco Ramoni
Knowledge Media Institute
The Open University

Paola Sebastiani
Department of Actuarial Science and Statistics
City University

Abstract

Bayesian Belief Networks (BBNs) are a powerful formalism for knowledge representation and reasoning under uncertainty. During the past few years, Artificial Intelligence met Statistics in the quest to develop effective methods to learn BBNs directly from databases. Unfortunately, real-world databases include missing and/or unreported data whose presence challenges traditional learning techniques, from both the theoretical and computational point of view. This paper introduces a new method to learn the probabilities defining a BBNs from databases with missing data. The intuition behind this method is close to the robust sensitivity analysis interpretation of probability: the method computes the extreme points of the set of possible distributions consistent with the available information and proceeds by refining this set as more information becomes available. This paper outlines the description of this method and presents some experimental results comparing this approach to the Gibbs Samplings.

Keywords: Bayesian Belief Networks; Bayesian Learning; Robustness; Dirichlet Distribution; Probability Intervals; Missing Data; Gibbs Sampling.

Reference: KMi Technical Report KMI-TR-29, June 1997 (revised October 1996). Also in *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, Fort Laurendale, FL, 1997.

Address: 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

Bayesian Belief Networks (BBNs) provide a powerful formalism to reason under uncertainty. Although in their original concept BBNs were mainly designed to encode the knowledge of human experts, their statistical roots soon prompted for the development of methods to learn them directly from databases of cases rather than from the insight of human domain experts [2, 1, 4]. This choice can be extremely rewarding when the domain of applications 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 domain experts.

A common assumption made by the current learning methods is that the database at hand is complete and does not contain any missing datum. Unfortunately, real-world databases are rarely complete: unreported, lost, and corrupted data are a distinguished feature of databases. In order to move on real-world applications, methods to learn BBNs have to face the challenge of learning from databases with missing data and this challenge has been accepted by a number of researchers during the past few years, producing different methods to cope with this problem.

This paper introduces a new method to learn conditional probabilities in BBNs from incomplete databases. The assumption of this method is that the BBNs generated by the learning process should enable the problem solver to reason on the basis of the available information and thus requires the learning method to return results whose precision is a monotonic increasing function of the available information.

2. Background

A BBN is a direct acyclic graph in which nodes represent stochastic variables and arcs represent conditional dependencies among variables. We shall limit our attention to discrete variables taking a finite number of values associated to mutually exclusive and exhaustive events. A conditional dependency links a *child* variable to a set of *parent* variables, and it is defined by the set of conditional probabilities of each state of the child variable given each combination of states of the parent variables in the dependency.

More formally, a BBN is defined by a set of *variables* $\mathcal{X} = \{X_1, \ldots, X_I\}$ and a structure S defining a graph of conditional dependencies among the elements of $\mathcal{X}$. The structure S allows us to decompose the joint probability of a particular set of values of the variables in $\mathcal{X}$, say $x_j = \{x_{1j}, \ldots x_{Ij}\}$, as

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

where $pa(X_i)$ are the parent nodes of X_i, and $x_{pa(X_i)j}$ denotes the states of $pa(X_i)$ in x_j. Clearly, $p(X_i = x_{ij}|pa(X_i) = x_{pa(X_i)j}) = p(X_i = x_{ij})$ for a root node X_i. In the following we will denote $X_i = x_{ij}$ as x_{ij}, and $pa(X_i) = x_{pa(X_i)j}$ as $pa(x_i)$. We shall consider the conditional probabilities defining the BBN as being generated by parameters

$\theta = \{\theta_1, \dots, \theta_K\}$, so that the joint probability of a case x_j is

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

where θ_i parameterizes the probability of x_{ij} given the parent configuration $pa(x_i)$.

Suppose we are given a database of J independent cases $\mathcal{D} = \{C_1, \dots, C_J\}$, each case $C_J = \{X_1 = x_{1j}, \dots, X_I = x_{Ij}\}$ being a set of entries. Given a network structure S, the task here is to learn the conditional probabilities defining the dependencies in the BBN from $\mathcal{D}$.

The most common approach to learn the parameter vector θ is Maximum Likelihood. With a database of J independent cases the likelihood function is

$$l(\theta) = \prod_{j=1}^{J} p(C_j|\theta),$$

and if the database is complete, the Maximum Likelihood estimates of the conditional probabilities are the observed frequencies of the relevant cases in the database.

The Bayesian approach extends the standard parameter estimation techniques by regarding the parameters θ as random variables, whose prior distribution represents the observer's belief about the parameters before observing any data. Given the information in the database, the prior density $\pi(\theta)$ is updated in the posterior density using Bayes' theorem, and hence

$$\pi(\theta|\mathcal{D}) = \frac{\pi(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})} \quad \text{where} \quad p(\mathcal{D}) = \int_{\mathcal{R}^K} \pi(\theta)p(\mathcal{D}|\theta)d\theta.$$

The Bayesian estimate of θ is then the posterior expectation $E(\theta|\mathcal{D})$ of θ.

Common assumptions of the Bayesian approach to learn BBNs are that the parameters are (i) mutually independent , and (ii) have a Dirichlet distribution. Under (i) the joint prior density of θ can be decomposed as

$$\pi(\theta) = \prod_{k=1}^{K} \pi(\theta_k),$$

thus allowing "local computations", while (ii) facilitates the computation of the posterior density by taking advantages of conjugate analysis. Consider for instance the variable X_i, taking values $\{x_{i1}, \dots, x_{iM}\}$, and the parameters $\theta_i = \{\theta_{i1}, \dots, \theta_{iM}\}$ associated to the conditional probabilities $p(x_{im}|pa(x_i))$, $m = 1, \dots, M$, so that $\sum_m \theta_{im} = 1$. A Dirichlet prior for θ_i, denoted as $D(\alpha_{i1}, \dots, \alpha_{iM})$, $\alpha_{im} > 0$, is a continuous multivariate distribution with density function proportional to

$$\prod_m \theta_{im}^{\alpha_{im}-1}.$$

The hyper-parameters α_{im}s have the following interpretation: $\alpha_{i+} = \sum_m \alpha_{im}$ can be regarded as an imaginary sample size needed to formulate this prior information about θ_i, and the mean of θ_{im} is α_{im}/α_{i+}, $m = 1, \ldots, M$. Note that the prior mean is the marginal probability of $x_{im}|pa(x_i)$. For instance, a uniform prior with $\alpha_{im} = 1$, for all m, would assign uniform probabilities to each $x_{im}|pa(x_i)$.

With complete data, the posterior distribution of the parameters can be computed exactly using standard conjugate analysis:

$$\pi(\theta_i|\mathcal{D}) = D(\alpha_{i1} + n(x_{i1}|pa(x_i)), \ldots, \alpha_{iM} + n(x_{iM}|pa(x_i))),$$

where $n(x_{im}|pa(x_i))$ is the frequency of cases in the database with $x_{im}|pa(x_i)$. The Bayes estimate of the conditional probability of $x_{im}|pa(x_i)$, given the information in the database, i.e. the posterior mean of θ_{im}, is then

$$\frac{\alpha_{im} + n(x_{im}|pa(x_i))}{\alpha_{i+} + n(pa(x_i))},$$

where $n(pa(x_i)) = \sum_m n(x_{im}|pa(x_i))$.

Unfortunately, the situation is quite different when some entries in the database are missing. When a datum is missing, there is a set of possible *complete* databases, one for each possible value of the variable for which the observation is missing. Exact analysis would require the computation of the joint posterior distribution of the parameters given each possible completion of the database, and then mix these over all possible completions. This is apparently infeasible.

A deterministic method, proposed by [7] and further developed by [3], provides a way to approximate the exact posterior distribution by processing data sequentially. An alternative approach is a stochastic approximation of the posterior distribution using for instance Markov Chain Monte Carlo (MCMC) methods, such as the Gibbs Sampling. These solutions share a common strategy known as *imputation*: they try to complete the database by inferring 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 so that the incomplete database is a representative sample of the complete one. Unfortunately this is unrealistic as the complete database assumption: there is often a reason for data to be missing in real databases.

When the "Missing at Random" assumption is violated — such as when data are systematically missing so that the probability of a missing entry depends on the state of the corresponding variable — even the most reliable of these methods suffers of dramatic decreases in accuracy. The completion of the database using the available information in the database itself leads the learning system to ascribe the missing data to known values in the database and, in the case of systematically missing data, to twist the estimate of probabilities in the database. It is apparent that this behavior can prevent the applicability of learning methods based on imputation to learn BBNs because, for the general case, they can produce strongly biased estimates of the conditional probabilities of the variables in the database and therefore unreliable BBNs.

3. Method

The solution we propose is a method to learn parameters in a BBN which is *robust* with respect to the distribution of missing data. This method computes the set of possible posterior distributions consistent with the available information in the database and proceeds by refining this set as more information becomes available. Instead of summarizing this information *somehow*, we then represent it via intervals, whose extreme points are the minimum and the maximum Bayes estimate that would have been inferred from all possible completions of the database. Such extreme estimates can be easily computed from the frequencies of incomplete cases in the database. Full details can be found in [6], we report here only the main result.

Let X_i be a variable in the BBN and $n^\bullet(x_{im}|pa(x_i))$ be the frequency of cases with $X_i = x_{im}$, given the parent configuration $pa(x_i)$, which have been obtained by completing incomplete cases. Note that 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. Suppose further that we start from total ignorance, thus the parameter θ_{im} which is associated to $p(x_{im}|pa(x_i))$ is assigned a uniform prior. Then the Bayes estimate $E(\theta_{im}|\mathcal{D})$, that would have been computed from the complete database satisfies

$$E(\theta_{im}|\mathcal{D}) = p(x_{im}|pa(x_i), \mathcal{D}) \geq \frac{1 + n(x_{im}|pa(x_i))}{k + \sum_m n(x_{im}|pa(x_i)) + \max_{h \neq m} n^\bullet(x_{ih}|pa(x_i))} \tag{1}$$

and

$$E(\theta_{im}|\mathcal{D}) = p(x_{im}|pa(x_i), \mathcal{D}) \leq \frac{1 + n(x_{im}|pa(x_i)) + n^\bullet(x_{im}|pa(x_i))}{k + \sum_m n(x_{im}|pa(x_i)) + n^\bullet(x_{im}|pa(x_i))}. \tag{2}$$

Note that the sum of the maximum posterior probability $p(x_{im}|pa(x_i))$ and the minimum posterior probabilities $p(x_{ih}|pa(x_i))$, with $h \neq m = 1, ..., M$, is one. It is worth noting that these bounds depend only on the frequencies of complete entries in the database and the "artificial" frequencies of the completed entries, so that they can be computed in batch mode.

When X_i is a binary variable, taking for instance values 1 and 0, (1) and (2) simplify to

$$E(\theta_{im}|\mathcal{D}) = p(X_i = 1|pa(x_i), \mathcal{D}) \geq \frac{1 + n(1|pa(x_i))}{2 + n(1|pa(x_i)) + n(0|pa(x_i)) + n^\bullet(0|pa(x_i))} \tag{3}$$

and

$$E(\theta_{im}|\mathcal{D}) = p(X_i = 1|pa(x_i)) \leq \frac{1 + n(1|pa(x_i)) + n^\bullet(1|pa(x_i))}{2 + n(1|pa(x_i)) + n(0|pa(x_i)) + n^\bullet(1|pa(x_i))}. \tag{4}$$

The main feature of this method is its robustness with respect to the distribution of missing data: it does not relies on the assumption that data are missing at random because it does not try to infer them from the available entries in the database. The basic intuition behind our method is that we are better off if, rather than trying to complete the database by guessing the value of missing data, we regard the available information as a set of constraints on the possible distributions in the database and we reason on the basis of the set of probability distributions consistent with the database at hand.

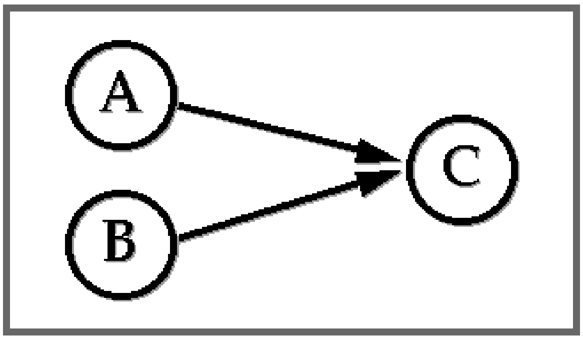

Figure 1: The simple network structure of the BBN used for the experimental evaluation.

4. 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. The focus of these experiments was mainly to compare the robustness of the two methods when data are systematically missing in the database.

We compared an implementation of our method to the implementation of the Gibbs Sampling provided by the program BUGS [8]. In the following experiments, we used the implementation of BUGS version 0.5 running on a Sun Sparc 5 under SunOS 5.5 and the an implementation of our method written in Common Lisp running on the same platform under CLISP version 1996/10/10.

Figure 1 shows the graphical structure of the simple BBN — defined by three binary variables A, B, and C — used for this comparison. We generated a database of 100 random cases from the following probability distribution:

$$\begin{array}{lcl} p(A=1) & = & 0.5 \\ p(B=1) & = & 0.4 \\ p(C=1|A=1,B=1) & = & 0.95 \\ p(C=1|A=1,B=2) & = & 0.05 \\ p(C=1|A=2,B=1) & = & 0.1 \\ p(C=1|A=2,B=2) & = & 0.8 \end{array}$$

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 iteratively deleting 2% of the database by systematically removing the entries reporting the value $B = 1$ and, each time, we ran the two learning algorithms. 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. This procedure was iterated until no entry $B = 1$ was reported in the database.

Figure 2 plots the parameter estimates given by the two systems against the percentage of entries still present in the database. Stars represent the point estimates given by the

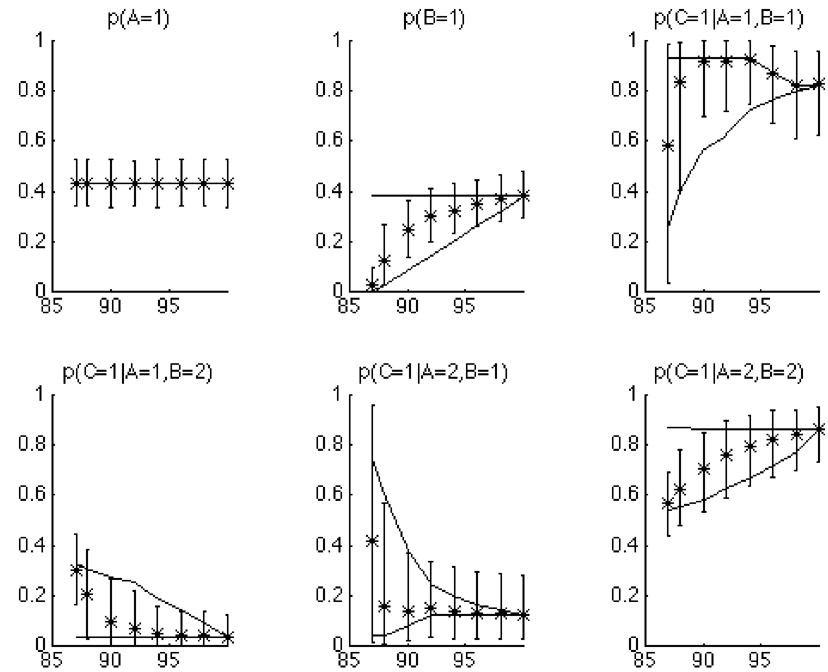

Figure 2: Estimates of the parameters defining the BBN depicted in Figure 1 against percentage of entries in the database. Sample of 100 cases.

Gibbs Sampling and errorbars indicates 95% confidence intervals. Solid lines report the lower and upper bounds of the probability intervals inferred by our method.

The bias of the Gibbs Sampling is absolutely clear. For instance, when 87% of the entries is available in the database but all the entries of $B = 1$ are missing, the estimate given by the Gibbs Sampling for $p(B = 1)$ lies on the lower extreme of the interval estimated by our method and the value computed in the complete database is excluded by the 95% confidence interval. The same effect can be noted in the estimates of $p(C = 1|A = 1, B = 2)$ and $p(C = 1|A = 2, B = 2)$, where the final value remains excluded up to the 88% of the entries in the database.

The bias is made even more remarkable as the size of the database increases. Figure 3 shows the results of the same experiment based on a database of 500 cases. The point estimates are comparable with the estimates extracted from the database of 100 cases. However, the estimates for $B = 1$ shows that, when all the entries $B = 1$ are missing, the sample size tights up the confidence interval around the estimate 0.0045 so the estimate 0.425 obtained from the complete database, is definitely excluded, with an error overpassing the 40%. The bias is more evident than in the previous experiment also for the estimates of $p(C = 1|A = 1, B = 2)$ and $p(C = 1|A = 2, B = 2)$, where the value estimated from the complete database remains excluded by the confidence intervals up to the 92% of the entries. In all cases, our method returned intervals which always contain the estimates obtained by the complete database. The width of the interval accounts for the amount of information available in the database 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.

As a matter of facts, the effect of the strong bias of the Gibbs Sampling is remarkable in the predictive performance of the BBN, and no measure of its reliability can be derived by the point-valued probability and the confidence interval. Suppose that $A = 2$ and $B = 2$ are observed and we want to predict the value of C. Since in this case $p(C = 1)$ reduces to $p(C = 1|A = 2, B = 2)$, the bottom right plot in Figure 3 reports the behavior of the marginal probability of $p(C = 1)$ as well. Suppose that we use the estimates learned by the Gibbs Sampling with 85% of the complete data, when all the entries $B = 1$ are missing. The prediction of the Gibbs Sampling is $p(C = 1) = p(C = 1|A = 2, B = 2) = 0.508$ with a confidence interval of $[0.443, 0.571]$ against the value 0.776 inferred from the complete database. Instead, our method returns the probability interval $[0.51, 0.8]$, thus including the value predicted using the complete database and providing a measure of the reliability of the prediction through the width of the interval.

A further difference between the performances of the two systems has been the execution time: in the worse case, Gibbs Sampling took over 4 minutes to run to completion the learning process of a single databases, while our system ran to completion the same task in less than 20 milliseconds. Further experimental results, comparing the two methods when data are missing at random, using different network topologies and larger databases, are reported in [6].

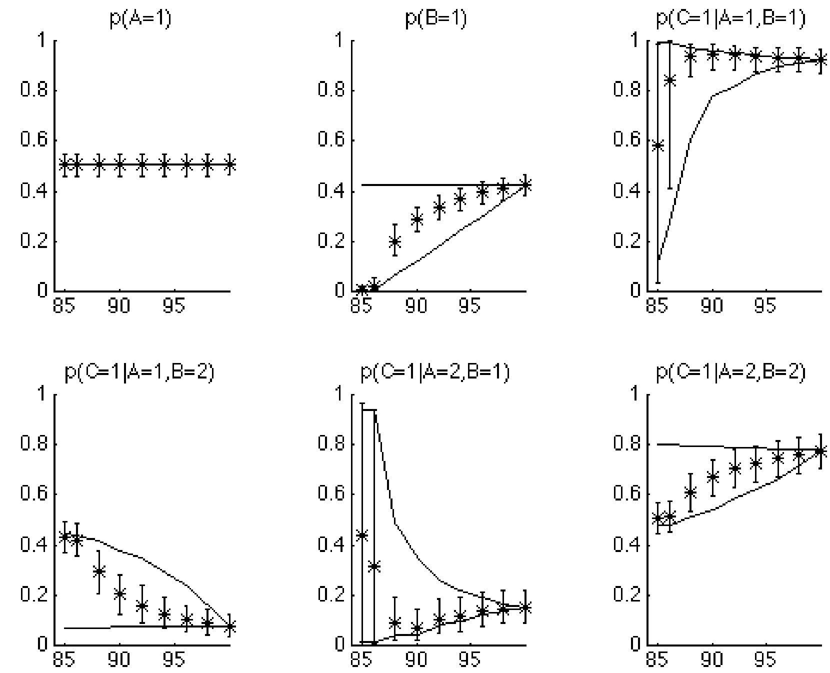

Figure 3: Estimates of the parameters defining the BBN depicted in Figure 1 against percentage of entries in the database. Sample of 500 cases.

5. Conclusions

This paper introduced a new method to learn conditional probabilities in a BBN from a database. The main feature of this method is its robustness with respect to the distribution of missing data because it does not try to infer them from the available information. The basic intuition behind our method is that we are better off if, rather than trying to complete the database by guessing the value of missing data, we regard the available information as a set of constraints on the possible distributions in the database. In this way, our learning algorithm returns probability intervals which account for the reliability of the information available in the database. These intervals can be then propagated using current techniques, such as [5]. An experimental comparison between our method and a stochastic method shows a remarkable difference in accuracy between the two methods and the computational advantages of our deterministic method with respect to the stochastic one.

Acknowledgments

Authors thank Greg Cooper, Pat Langley, Paul Snow, and Zdenek Zdrahal for their useful suggestions during the development of this research. Equipment has been provided by generous donations from Apple Computers 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 Cowel, 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, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combinations of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[5] M. Ramoni. Ignorant influence diagrams. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1808–1814, S. Mateo, CA, 1995. Morgan Kaufman.

[6] M. Ramoni and P. Sebastiani. Robust learning with missing data. Technical Report KMi-TR-28, Knowledge Media Institute, The Open University, 1997.

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

[8] 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.