Learning Bayesian Network Models from Incomplete Data using Importance Sampling

Carsten Riggelsen and Ad Feelders
Institute of Information & Computing Sciences
Utrecht University
P.O. Box 80098, 3508TB Utrecht
The Netherlands

Abstract

We propose a Bayesian approach to learning Bayesian network models from incomplete data. The objective is to obtain the posterior distribution of models, given the observed part of the data. We describe a new algorithm, called eMC^4, to simulate draws from this posterior distribution. One of the new ideas in our algorithm is to use importance sampling to approximate the posterior distribution of models given the observed data and the current imputation model. The importance sampler is constructed by defining an approximate predictive distribution for the unobserved part of the data. In this way existing (heuristic) imputation methods can be used that don't require exact inference for generating imputations.

We illustrate eMC^4 by its application to modeling the risk factors of coronary heart disease. In the experiments we consider different missing data mechanisms and different fractions of missing data.

1 Introduction

Bayesian networks are probabilistic models that can represent complex interrelationships between random variables. It is an intuitively appealing formalism for reasoning with probabilities that can be employed for diagnosis and prediction purposes. Furthermore, learning Bayesian networks from data may provide valuable insight into the (in)dependences between the variables.

In the last decade, learning Bayesian networks from data has received considerable attention in the research community. Most learning algorithms work under the assumption that *complete* data is available. In practical learning problems one frequently has to deal with missing values however. The presence of incomplete data leads to analytical intractability and high computational complexity compared to the complete data case. It is very tempting to "make the problem go away" either by deleting observations with missing values or using ad-hoc methods to fill in (impute) the missing data. Such procedures may however lead to biased results, and, in case of imputing a single value for the missing data, to an overconfidence in the results of the analysis.

We avoid such ad-hoc approaches and use a method that takes *all* observed data into account, and correctly reflects the increased uncertainty due to missing data. We do assume however that the missing data mechanism is *ignorable* as defined by Little and Rubin (1987). Essentially this means that the probability that some component is missing may depend on observed components, but not on unobserved components.

Our approach is Bayesian in the sense that we are not aiming for a single best model, but want to obtain (draws from) a posterior distribution over possible models. We show how to perform model averaging over Bayesian network models, or alternatively, how to get a range of good models, when we have incomplete data. We develop a method that can handle a broad range of imputation methods without violating the validity of the models returned. Our approach is not restricted to any imputation technique in particular, and therefore allows for imputation methods that do not require expensive inference in a Bayesian network.

This paper is organised as follows. In section 2 we briefly review previous research in this area and show how our work fits in. In section 3 we describe model learning from complete data. In sections 4 and 5 we introduce a new algorithm, called eMC^4, for Bayesian network model learning from incomplete data. We performed a number of experiments to test eMC^4 using

real life data. The results of those experiments are reported in section 6. Finally, we summarize our work and draw conclusions.

2 Previous research

Here we briefly review relevant literature on learning Bayesian networks from incomplete data. Two popular iterative approaches for learning parameters are Expectation-Maximization (EM) by Dempster et al. (1977) and a simulation based Gibbs sampler (Geman and Geman, 1984) called Data Augmentation (DA) introduced by Tanner and Wong (1987). For Bayesian networks EM was studied by Lauritzen (1995). The Expectation step (E-step) involves the performance of inference in order to obtain sufficient statistics. The E-step is followed by a Maximization step (M-step) in which the Maximum Likelihood (ML) estimates are computed from the sufficient statistics. These two steps are iterated until the parameter estimates converge.

Data Augmentation (DA) is quite similar but is nondeterministic. Instead of calculating expected statistics, a value is drawn from a predictive distribution and imputed. Similarly, instead of calculating the ML estimates, one draws from the posterior distribution on the parameter space (conditioned on the sufficient statistics of the most recent imputed data set). Based on Markov chain Monte Carlo theory this will eventually return realizations from the posterior parameter distribution. There are also EM derivatives that include a stochastic element quite similar to DA (see McLachlan and Krishnan, 1997).

Bound and Collapse (BC) introduced by Ramoni and Sebastiani (2001) is a two-phase algorithm. The bound phase considers possible completions of the data sample, and based on that computes an interval for each parameter estimate of the Bayesian network. The collapse phase computes a convex combination of the interval bounds, where the weights in the convex combination are computed from the available cases. The collapse phase seems to work quite well for particular missing data mechanisms but unfortunately is not guaranteed to give valid results for ignorable mechanisms in general.

Learning models from incomplete data so to speak adds a layer on top of the parameter learning methods described above. For EM, Friedman (1998) showed that doing a model selection search *within* EM will result in the best model in the limit accoring to some model scoring criterion. The Structural EM (SEM) algorithm is in essence similar to EM, but instead of computing expected sufficient statistics from the same Bayesian network model throughout the iterations, a model selection step is employed. To select the next model, a model search is performed, using the expected sufficient statistics obtained from the current model and current parameter values.

Ramoni and Sebastiani (1997) describe how BC can be used in a model selection setting. As remarked before however, BC is not guaranteed to give valid results for ignorable mechanisms in general, and the risk of obtaining invalid results unfortunately increases when the model structure is not fixed.

In contrast to SEM, our aim is not to select a single model, but to obtain a posterior probability distribution over models that correctly reflects uncertainty, including uncertainty due to missing data. Therefore our approach is more related to the simulation based DA described above.

3 Learning from complete data

In this section we discuss the Bayesian approach to learning Bayesian networks from complete data. First we introduce some notation. Capital letters denote discrete random variables, and lower case denotes a state. Boldface denote random vectors and vector states. We use $\Pr(\cdot)$ to denote probability distributions (or densities) and probabilities. $\mathcal{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_c)$ denotes the multinomial data sample with c i.i.d. cases. A Bayesian network (BN) for $\boldsymbol{X} = (X^1, \ldots, X^p)$ represents a joint probability distribution. It consists of a directed acyclic graph (DAG) m, called the model, where every vertex corresponds to a variable X^i, and a vector of conditional probabilities $\boldsymbol{\theta}$, called the parameter, corresponding to that model. The joint distribution factors recursively according to m as $\Pr(\boldsymbol{X}|m, \boldsymbol{\theta}) = \prod_{i=1}^{p} \Pr(X^i|\boldsymbol{\Pi}(X^i), \boldsymbol{\theta})$, where $\boldsymbol{\Pi}(X^i)$ is the parent set of X^i in m.

Since we learn BNs from a Bayesian point of view, model and parameter are treated as random variables M and $\boldsymbol{\Theta}$. We define distributions on parameter space $\Pr^{\boldsymbol{\Theta}}(\cdot)$ and model space $\Pr^{M}(\cdot)$. The superscript is omitted and we simply write $\Pr(\cdot)$ for both. The distribution on the parameter space is a product Dirichlet distribution which is conjugate for the multinomial sample $\mathcal{D}$, i.e. Bayesian updating is easy because the posterior once $\mathcal{D}$ has been taken into consideration is again Dirichlet, but with updated hyper parameters. The MAP model is found by maximizing with respect to M

$$\Pr(M|\mathcal{D}) \propto \Pr(\mathcal{D}|M) \cdot \Pr(M) \tag{1}$$

where $\Pr(\mathcal{D}|M)$ is the normalizing term in Bayes theorem when calculating the posterior Dirichlet

$$\Pr(\mathcal{D}|M) = \int \Pr(\mathcal{D}|M, \boldsymbol{\Theta}) \Pr(\boldsymbol{\Theta}|M) d\boldsymbol{\Theta} \tag{2}$$

where $\Pr(\mathcal{D}|M,\boldsymbol{\Theta})$ is the likelihood, and $\Pr(\boldsymbol{\Theta}|M)$ is the product Dirichlet prior. In Cooper and Herskovits (1992) a closed formula is derived for (2) as a function of the sufficient statistics for $\mathcal{D}$ and prior hyper parameters of the Dirichlet. This score can be written as a product of terms each of which is a function of a vertex and its parents. This decomposability allows local changes of the model to take place without having to recompute the score for the parts that stay unaltered, that is, only the score for vertices whose parents set has changes needs to be recomputed.

Instead of the MAP model, we may be interested in the expectation of some quantity Δ of models using $\Pr(M|\mathcal{D})$ as a measure of uncertainty over all models

$$\mathrm{E}[\Delta] = \sum_M \Delta_M \cdot \Pr(M|\mathcal{D}) \approx \frac{1}{q}\sum_{i=1}^{q} \Delta_{m^i} \qquad (3)$$

where the Monte Carlo approximation is obtained by sampling $\{m^i\}_{i=1}^q$ from $\Pr(M|\mathcal{D})$.

It is infeasible to calculate the normalizing factor $\Pr(\mathcal{D})$ required to obtain equality in equation (1). Madigan and York (1995) and Giudici and Castelo (2003) propose to use (enhanced) Markov chain Monte Carlo Model Composition (eMC^3) for drawing models from this distribution leaving the calculation of the normalizing term implicit. It is a sampling technique based on Markov chain Monte Carlo Metropolis-Hastings sampling summarized in the following iterated steps. At entrance assume model m^t:

1. Draw model m^{t+1} from a proposal distribution $\Pr(M|m^t)$ resulting in a slightly modified model compared to m^t (addition, reversal or removal of an arc).
2. The proposed model m^{t+1} is accepted with probability

   $$\alpha(m^{t+1}, m^t) = \min\left\{1, \frac{\Pr(\mathcal{D}|m^{t+1})\Pr(m^{t+1})}{\Pr(\mathcal{D}|m^t)\Pr(m^t)}\right\},$$

   otherwise the proposed model is rejected and $m^{t+1} \stackrel{\text{def}}{=} m^t$.

For $t \to \infty$ the models can be considered samples from the invariant distribution $\Pr(M|\mathcal{D})$. Note that in step two the normalizing factor $\Pr(\mathcal{D})$ has been eliminated. For enhanced MC^3 a third step is required, *Repeated Covered Arc Reversals* (RCAR) which simulates the neighbourhood of equivalent DAG models. We refer to Kocka and Castelo (2001) for details.

4 Learning from incomplete data

Running standard eMC^3 can be quite slow, especially for large models and data sets. In the presence of missing data, a prediction 'engine' (predicting missing components) so to speak has to be wrapped around eMC^3. Obtaining a prediction engine which will always make the correct predictions is infeasible to construct, and when the engine itself has to adapt to the ever changing model this becomes even worse. An approximate predictive engine is usually easier to construct, but will obviously sometimes make slightly wrong predictions. In this section we show how approximation can be used together with eMC^3 to obtain realizations from the posterior model distribution such that prediction errors are corrected for.

Our goal is to compute (3) when we have missing data. To be more precise, if we write $\mathcal{D} = (\mathcal{O},\mathcal{U})$ to denote the observed part $\mathcal{O}$ and the unobserved part $\mathcal{U}$, our goal is to get draws from $\Pr(M|\mathcal{O})$ such that we can use the approximation in (3). Due to incompleteness the integral in (2) no longer has a tractable solution. Our approach is instead to rewrite the posterior model distribution such that U can be "summed out" by way of "filling in". Note that the desired model posterior can be written as

$$\Pr(M|\mathcal{O}) = \sum_U \Pr(M|\mathcal{O},U)\Pr(U|\mathcal{O}). \qquad (4)$$

The first term is the distribution given in (1) involving the prior and the marginal likelihood (2). The second part is the predictive distribution which can be considered the predictor of the missing data based on the observed part. We explicitly model the predictive distribution as a BN with model M', the *imputation* model. We therefore write

$$\Pr(M|\mathcal{O},M') = \sum_U \Pr(M|\mathcal{O},U,M')\Pr(U|\mathcal{O},M') \quad (5)$$

We assume that M is independent of M' given O and U, i.e. once we are presented with complete data, the imputation model has become irrelevant

$$\Pr(M|O,U,M') = \Pr(M|O,U).$$

The Monte Carlo approximation of (5) is calculated as

$$\Pr(M|\mathcal{O},M') \approx \frac{1}{n}\sum_{i=1}^{n} \Pr(M|\mathcal{O},\mathcal{U}^i)$$

where $\mathcal{U}^i \sim \Pr(U|\mathcal{O},M')$ for $i = 1,\ldots,n$. So, if we could compute realizations from this predictive distribution we could approximate $\Pr(M|\mathcal{O},M')$. Unfortunately we can not use simple sequential Bayesian updating (Spiegelhalter and Lauritzen, 1990) for determining $\Pr(U|\mathcal{O},M')$. Instead of sampling from the true predictive distribution, we define an *approximate predictive distribution* which can act as a proposal distribution for suggesting imputations. The predictive

distribution can be rewritten such that it can act as a quality measure for a proposed imputation. This is accomplished by using *importance sampling*. Denote the approximate predictive distribution $\Pr^*(U)$ and rewrite (5)

$$\Pr(M|\mathcal{O},M') = \sum_U \Pr(M|\mathcal{O},U)\frac{\Pr(U|\mathcal{O},M')}{\Pr^*(U)}\Pr^*(U)$$

Sample $\mathcal{U}^i \sim \Pr^*(U)$ for $i = 1,\ldots,n$ and use that sample in the importance sampling approximation

$$\Pr(M|\mathcal{O},M') \approx \frac{1}{W}\sum_{i=1}^{n} \underbrace{\frac{\Pr(\mathcal{U}^i|\mathcal{O},M')}{\Pr^*(\mathcal{U}^i)}}_{w_i} \Pr(M|\mathcal{O},\mathcal{U}^i) \quad (6)$$

where $W = \sum_{i=1}^{n} w_i$ is the normalizing constant. Now rewrite the predictive distribution

$$\Pr(\mathcal{U}^i|\mathcal{O},M') = \Pr(\mathcal{U}^i,\mathcal{O}|M')\frac{1}{\Pr(\mathcal{O}|M')}$$

where the term $\Pr(\mathcal{U}^i,\mathcal{O}|M')$ is given by (2). Through normalization the denominator $\Pr(\mathcal{O}|M')$ disappears as it is independent of U. It therefore suffices to calculate the weights as

$$w_i = \frac{\Pr(\mathcal{U}^i,\mathcal{O}|M')}{\Pr^*(\mathcal{U}^i)} \quad (7)$$

where the numerator can be computed efficiently and the denominator is the probability of the proposal. The marginal likelihood in the numerator is in this context not used as a scoring criterion for models. Instead we use it as a scoring criterion for *imputations*. The denominator compensates for the bias introduced by drawing from $\Pr^*(U)$ rather than the correct predictive distribution.

Given a set $\{\mathcal{U}^i\}_{i=1}^n$ that has been sampled from $\Pr^*(U)$, sampling from the mixture approximation (6) of $\Pr(M|\mathcal{O},M')$ is done as follows:

1. The probability of selecting sample $\mathcal{U}^i$ augmenting $\mathcal{O}$ is proportional to the importance weight w_i.
2. The now complete sample $(\mathcal{O},\mathcal{U}^i)$ is used for sampling models from $\Pr(M|\mathcal{O},\mathcal{U}^i)$ using eMC^3.

We can now draw from $\Pr(M|\mathcal{O},M')$, but our goal was to obtain draws from $\Pr(M|\mathcal{O})$, i.e. we need samples from the posterior model distribution given observed data *without* conditioning on the imputation model M'. The desired distribution is obtained by Gibbs sampling. Given an imputation model m^l draw the following

$$\begin{aligned} m^{l+1} &\sim \Pr(M|\mathcal{O},m^l) \\ &\vdots \end{aligned}$$

Algorithm e**MC**$^4(n,q,k)$

```
1  m^0 ← G = (V = X, E = ∅)
2  r ← 0
3  for l ← 0 to k
4      W ← 0
5      U^0 ← U^r
6      for i ← 0 to n
7          w_i ← Pr(O,U^i|m^l) / Pr*(U^i)
8          W ← W + w_i
9          if i ≠ n then draw U^{i+1} ~ Pr*(U)
10     draw r ~ Pr(i) = w_i/W
11     m^0 ← m^l
12     for t ← 0 to q
13         m^t ← RCAR(m^t)
14         draw m^{t+1} ~ Pr(M|m^t)
15         B ← Pr(O,U^r|m^{t+1}) / Pr(O,U^r|m^t)
16         draw α ~ Bernoulli(min{1, B})
17         if α ≠ 1 then m^{t+1} ← m^t
18     m^{l+1} ← m^{q+1}
19     LogToFile(m^{l+1})
```

Figure 1: The eMC^4 algorithm

which for $l \to \infty$ results in a chain of realizations from $\Pr(M|\mathcal{O})$. This in effect allows us to calculate (3).

When n is large, the mixture approximation is close to the real distribution $\Pr(M|\mathcal{O},M')$. However, the invariant model distribution is reached for any value assigned to n if we make sure that one of the imputation proposals is the *current* imputation, i.e. the augmented data sample that was selected at the last mixture draw before entering the eMC^3 loop. By this overlap, n is indirectly increased every time the mixture is set up. From a practical point of view however, n does have an impact on how well the model Markov chain mixes. Small n implies slow mixing depending on how far the approximate predictive distribution is from the real predictive distribution.

We do not discuss parameter estimation in this paper, but merely mention that using the importance sampler presented above, it is also possible to approximate the posterior *parameter* distribution. In (6) simply plug in $\Pr(\boldsymbol{\Theta}|M',\mathcal{O},\mathcal{U}^i)$ instead of $\Pr(M|\mathcal{O},\mathcal{U}^i)$ to obtain the required posterior.

To summarize, figure 1 contains the pseudo-code for the algorithm called *enhanced Markov Chain Monte Carlo Model Composition with Missing Components*, or for short, eMC^4. In line 1 an initial empty graph is defined. In lines 6–9 the imputations take place and the importance weights are calculated. Line 10 is the first step in drawing from the mixture. Lines

12–17 perform the eMC^3 algorithm based on the augmented sample selected. In the absence of relevant prior knowledge, a uniform model prior is assumed in line 15. Angelopoulos and Cussens (2001) discuss the construction of informative model priors. The choice of k (number of iterations of the Gibbs model sampler) depends on when the Markov chain of models has converged. Monitoring the average number of edges is one method for doing so suggested by Giudici and Green (1999). Once this average stabilizes the chain has converged.

5 Proposal distribution $\Pr^*(U)$

We can choose freely the approximate predictive distribution from which samples are drawn for (6), but of course some choices are better than others. Ideally, the appoximate predictive distribution should be close to the real predictive distribution, because otherwise n is required to be large to obtain enough samples from the region where the mass of the real distribution is located. Existing imputation techniques can be used, as long as they can be cast in the form of a distribution from which imputations can be drawn. Naturally it is also a requirement that $\Pr^*(U)$ has support when $\Pr(U|\mathcal{O}, M')$ has support. A uniform proposal distribution is probably unwise unless a very small fraction of data is missing. On the other hand, a distribution based on M' with parameters estimated using EM is not needed. Employing the BC algorithm for parameter estimation using M' could be interesting, since it is fast and is reported to often give reasonable results. Alternatively, a simple *available cases analysis* may prove to be good enough.

It is not a requirement to use the actual imputation model M' as the basis for $\Pr^*(U)$. In fact $\Pr^*(U)$ need not be modeled as a BN at all. However, an imputation method that does not take the independences portrayed by M' into account will have a hard time proposing qualified imputations, because the degree of freedom is simply too high (assuming that nothing is known about the missing data mechanism). Similarly, the parameter does not need to be derived from the data; however, since the missing data mechanism is assumed to be ignorable, all information we need to impute (predict) is contained (indirectly) in $\mathcal{O}$, and therefore predictions should at least depend on observed values. We propose to model $\Pr^*(U) \stackrel{\text{def}}{=} \Pr(U|\mathcal{O}, M', \boldsymbol{\theta})$ as a BN with model M' and parameter $\boldsymbol{\theta} = \mathrm{E}[\boldsymbol{\Theta}|\mathcal{O}, \mathcal{U}^{M'}]$, where expectation is with respect to $\Pr(\boldsymbol{\Theta}|M')$, and $(\mathcal{O}, \mathcal{U}^{M'})$ is the augmented sample from which M' was learned. The latter makes sense because $(\mathcal{O}, \mathcal{U}^{M'})$ is the sample from which the model was learned and therefore reflects the most appropriate sample on which to base the parameter.

In order do draw multivariates we propose the following method based on Gibbs sampling, where realizations are drawn on a univariate level. Denote a case j in $\mathcal{D}$ by $\boldsymbol{d}_j = (x_j^1, \ldots, x_j^p) = (\boldsymbol{o}_j, \boldsymbol{u}_j) = (\boldsymbol{o}_j, (u_j^1, \ldots, u_j^{r(j)}))$, where $\boldsymbol{o}_j$ and $\boldsymbol{u}_j$ refer to the observed and unobserved part of the case. The j'th case for $\mathcal{U}^t$ is sampled as follows

$$\begin{aligned} u_j^{1,t} &\sim \Pr(U_j^1 | u_j^{2,t-1}, \ldots, u_j^{r(j),t-1}, \boldsymbol{o}_j, M', \boldsymbol{\theta}) \\ &\vdots \\ u_j^{r(j),t} &\sim \Pr(U_j^{r(j)} | u_j^{1,t}, \ldots, u_j^{r(j)-1,t}, \boldsymbol{o}_j, M', \boldsymbol{\theta}). \end{aligned}$$

Based on Markov chain Monte Carlo theory, correlated multivariate realizations $\boldsymbol{u}_j \sim \Pr(\boldsymbol{U}_j|\boldsymbol{o}_j, M', \boldsymbol{\theta})$ are obtained when $t \to \infty$. Since each draw in the Gibbs sampler is univariate, and the entire Markov blanket of variable U^i has evidence, inference does not require any advanced techniques.

With the suggested Gibbs sampler we effectively collect *all realizations* including samples in the burn-in phase. The idea is to let the importance sampler decide on the quality of the proposed imputations.

We can calculate the importance weights efficiently without explicitely knowing the actual probabilities $\Pr(\mathcal{U}^i|\mathcal{O}, M', \boldsymbol{\theta})$. This can be seen by rewriting the importance weights from (7):

$$\begin{aligned} w_i &= \frac{\Pr(\mathcal{U}^i, \mathcal{O}|M')}{\Pr(\mathcal{U}^i|\mathcal{O}, M', \boldsymbol{\theta})} \\ &= \frac{\Pr(\mathcal{U}^i, \mathcal{O}|M') \cdot \Pr(\mathcal{O}|M', \boldsymbol{\theta})}{\Pr(\mathcal{U}^i, \mathcal{O}|M', \boldsymbol{\theta})}. \end{aligned}$$

By normalization of these weights, $\Pr(\mathcal{O}|M', \boldsymbol{\theta})$ cancels out, and it suffices to calculate the weights as

$$w_i = \frac{\Pr(\mathcal{U}^i, \mathcal{O}|M')}{\Pr(\mathcal{U}^i, \mathcal{O}|M', \boldsymbol{\theta})},$$

the ratio of the marginal likelihood over the likelihood given $\boldsymbol{\theta}$. This ratio is easily obtained because both probabilites can be computed efficiently in closed form. In summary, we can propose imputations efficiently *and* we can compute the "quality" of such a propsal efficiently as well.

When the structural difference between the imputation model M' and the exit model M is kept relatively small (dependent on q), we can make the following observations when $\boldsymbol{\theta}$ is assigned $\mathrm{E}[\boldsymbol{\Theta}|\mathcal{O}, \mathcal{U}^{M'}]$:

(1) Multivariate *predictions* based on the two models are correlated. Hence n need not be large in order to compensate for the predictive difference. However, we may still need many realizations $\mathcal{U}^i$ in order to

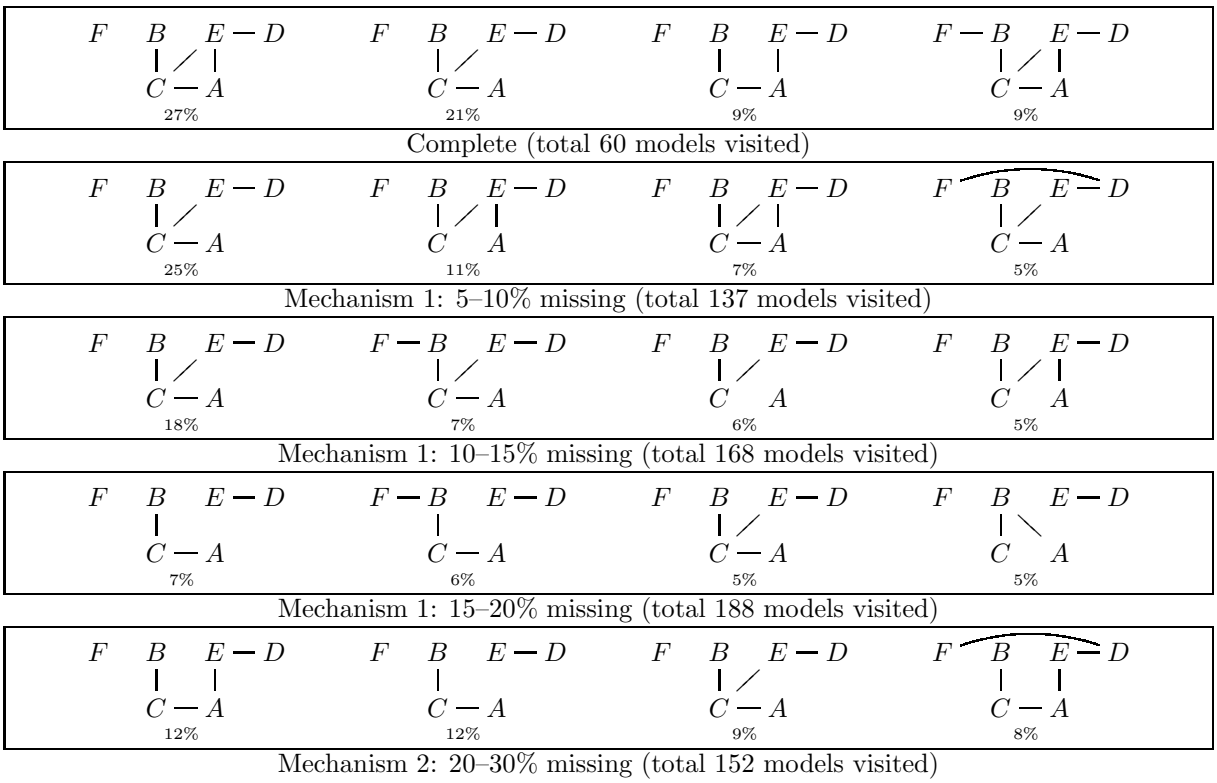

Figure 2: Top four visited models. Note that all edges are reversible.

capture the entire distribution. Correlation between multivariate predictions allows us to keep n relatively small and only move slightly in a direction towards a 'better' prediction.

(2) Because models are correlated and as a consequence also the predictions, the first predictions obtained by running the Markov blanket Gibbs sampler may be good. The Gibbs sampler so to speak picks up from where it left the last time and continues imputation using the new model. This means that the there is a fair chance that an initial imputation is actually selected.

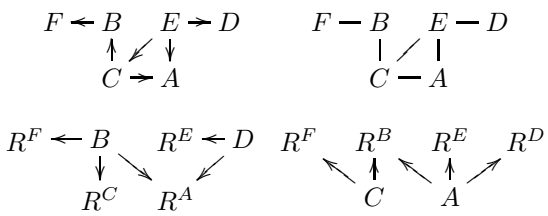

Figure 3: *Top*: Generating model (left), Essential graph (right). *Bottom*: Mechanism 1 (left), Mechanism 2 (right).

6 Experimental evaluation

In this section we perform a small experimental evaluation of eMC^4 and briefly discuss the results. Because our approach is Bayesian, comparison of the results with model *selection* methods for incomplete data such as SEM is not very useful.

We used a data set from Edwards and Havránek (1985) about probable risk factors of coronary heart disease. The data set consists of 1841 records and 6 binary variables, A: smoking, B: strenuous mental work, C: strenuous physical work, D: blood pressure under 140, E: ratio β to α proteins less than 3, F: family history of coronary heart disease. Because several DAGs encode the same set of assumptions about independence, we depict results as *essential graphs*, a canonical representation of an equivalence class (see Chickering, 1995; Kocka and Castelo, 2001).

Based on an eMC^3 run using the 1841 complete records, the 1st model in figure 3 is a highly probable model (although edge $F-B$ is not strongly supported). The 2nd model in the figure is the corresponding essential graph of the DAG. The parameter corresponding to the DAG model was determined based on the aforementioned data set, and 1800 new records were sampled from the BN. Incomplete sets were generated by

applying missingness mechanism one in figure 3 on the complete sample. This graph explicitly defines how response R^i of variable i depends on observed variables. Since for all i, R^i only depends on completely observed variables, the missingness mechanism is clearly ignorable. Three incomplete sets were generated with 5–10%, 10–15% and 15–20% missing components. The probability of non-response of variable i conditional on a parent configuration of R^i was selected from the specified interval.

On the basis of the generating model and the missingness mechanism, we would expect the following results. Since response of C only depends on B and the association $C - B$ is strong, a big fraction of components can be deleted for C without destroying support for the edge in the data. Association $D - E$ is also strong so discarding components for E will probably not have a major impact on the edge either. Association $E - A$ is influenced by B and D because the response is determined by those variables. Values for E and A may be absent often and therefore information about the association might have changed. This may also be the case for the edges $C - E$ and $C - A$.

We ran eMC^4 using each incomplete data set. Parameter q (number of eMC^3 iterations) was set to 150 and n to 25 (number of imputations). It took about 15 minutes on a 2 GHz machine before the Markov chain appeared to have converged.

In figure 2 the top four models are depicted along with their sampling frequencies. Notice the presence of the strong associations $C - B$ and $D - E$ everywhere, as expected. When the fraction of missing components for two associated variables increases it has a big impact on the support of such an association. Indeed, from the figure we see that the support for associations between variables A, C and E has changed. The sample frequencies and the number of visited models also suggest that the variance of the posterior distribution becomes bigger when more components are deleted. There is no longer a pronounced 'best' model.

The plot in figure 4 shows this more clearly. Here the cumulative frequencies are plotted against models (sorted on frequency in descending order). A steep plot indicates a small variance. For complete data the 10 best models account for 90% of the distribution whereas for 15–20% missing components only 50% of the distribution is accounted for by the best 10 models. To investigate the similarity of the models between the three incomplete sets, we used equation (3) to calculate Δ, where Δ_M is set to 1 when there is an edge between two vertices of interest in M. This results in the expected probability of the presence of edges as seen in figure 4. We can see, as we would expect,

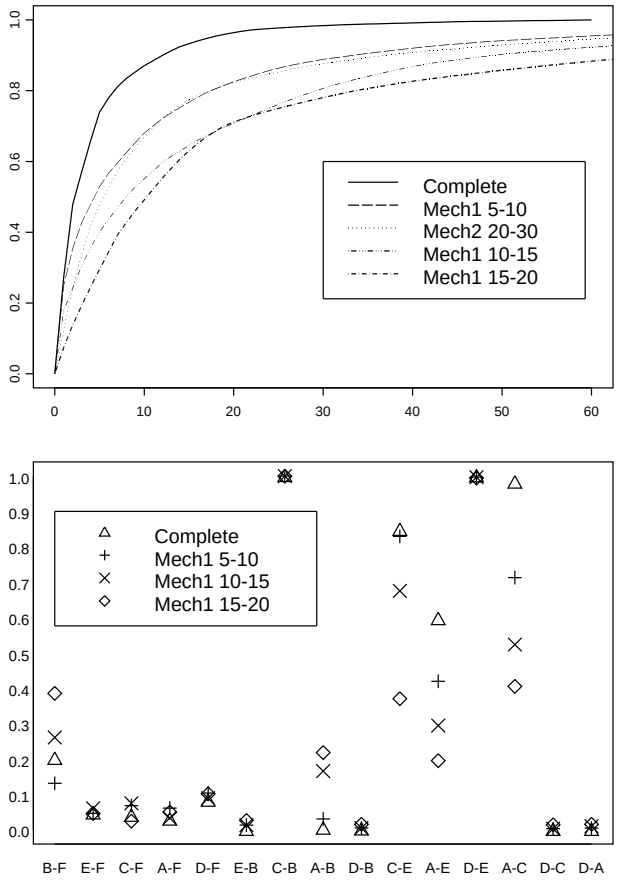

Figure 4: *Top*: Cumulative frequencies. *Bottom*: Expected probability of edges.

that the distance between points of complete data and incomplete data is dependent on the fraction of missing components. Diamonds (15–20%) have the biggest distance to triangles (complete), and plusses (5–10%) the smallest.

As we saw for mechanism one, discarding components for two associated variables can have a big impact on the presence of the corresponding edge in sampled models. For strongly associated variables the impact is less pronounced. We created another incomplete data set using mechanism two in figure 3. For the associated variables C, E and A the mechanism only discards components that we think will not severely impact these associations. For the strong association $E - D$ discarding components on both should not matter. We expect that we are able to remove a substantial fraction of components and still obtain reasonable models. We selected the fraction of missing components in the interval 20–30%. In the last row in figure 2 we see that although a substantial fraction of components were deleted, the models learned are quite similar to the models from the complete set.

To illustrate that it is not the fraction of missing components that determines the variance but rather the fraction of missing information (Little and Rubin, 1987), we plotted the cumulative frequency in figure 4. The variance of the posterior distribution is similar to the variance of the posterior for mechanism one with 5–10% missing components. This means that although the fraction of missing components is much higher than 5–10%, the uncertainty due to missing data has not changed substantially.

7 Conclusion

We have presented eMC^4 for simulating draws from the posterior distribution of BN models given incomplete data. In contrast to existing methods for BN model learning with incomplete data, we take a Bayesian approach and approximate the posterior model distribution given the observed data. Different imputation methods may be used, and specifically we describe a method that does not require exact inference in a BN. By using importance sampling we give all multivariate realizations of the Markov chain a 'chance' of being selected rather than just returning the last realization as in traditional Gibbs sampling. Importance sampling makes it possible to exploit qualified, yet not perfect imputation proposals. From a computational point of view specifying an approximate distribution is cheaper than a perfect one.

Valuable insight is gained when sampling models from the posterior; an illustration of the kind of information one can derive from posterior realizations is given in section 6. A posterior distribution is more informative than just a single model. This is especially true in the case of incomplete data, since the increased uncertainty due to missing data is reflected in the probability distribution.

References

N. Angelopoulos and J. Cussens. Markov chain Monte Carlo using trees-based priors on model structure. In J. Breese and D. Koller, editors, *Proc. of the Conf. on Uncertainty in AI*, pages 16–23, 2001.

D. Chickering. A transformational characterization of equivalent Bayesian networks. In P. Besnard and S. Hanks, editors, *Proc. of the Conf. on Uncertainty in AI*, pages 87–98, 1995.

G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society, Series B*, 34:1–38, 1977.

D. Edwards and T. Havránek. A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72(2):339–351, 1985.

N. Friedman. The Bayesian structural EM algorithm. In G. F. Cooper and S. Moral, editors, *Proc. of the Conf. on Uncertainty in AI*, pages 129–138, 1998.

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

P. Giudici and R. Castelo. Improving Markov chain Monte Carlo model search for data mining. *Machine Learning*, 50(1):127–158, 2003.

P. Giudici and P. Green. Decomposable graphical gaussian model determination. *Biometrika*, 86(4): 785–801, 1999.

T. Kocka and R. Castelo. Improved learning of Bayesian networks. In D. Koller and J. Breese, editors, *Proc. of the Conf. on Uncertainty in AI*, pages 269–276, 2001.

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

R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley and Sons, 1987.

D. Madigan and J. York. Bayesian graphical models for discrete data. *Intl. Statistical Review*, 63:215–232, 1995.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

M. Ramoni and P. Sebastiani. Learning Bayesian networks from incomplete databases. In D. Geiger and P. Shenoy, editors, *Proc. of the Conf. on Uncertainty in AI*, pages 401–408, 1997.

M. Ramoni and P. Sebastiani. Robust learning with missing data. *Machine Learning*, 45(2):147–170, 2001.

D. J. Spiegelhalter and S. L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1990.

M. Tanner and W. Wong. The calculation of posterior distributions by data augmentation. *J. of the Am. Stat. Assoc.*, 82(398):528–540, 1987.