Learning Bayesian Networks from Incomplete Data: An Efficient Method for Generating Approximate Predictive Distributions

Carsten Riggelsen
Department of Information & Computing Sciences
Universiteit Utrecht
P.O. Box 80.098, 3508TB Utrecht, The Netherlands
carsten@cs.uu.nl

Abstract

We present an efficient method for learning Bayesian network models and parameters from incomplete data. With our approach an approximation is obtained of the predictive distribution. By way of this distribution any learning algorithm that works for complete data can be easily adapted to work for incomplete data as well. Our method exploits the dependence relations between the variables explicitly given by the Bayesian network model to predict missing values. Based on strength of influence and predictive quality, a subset of those predictor variables is selected, from which an approximate predictive distribution is generated by taking the observed part of the data into consideration. The approximate predictive distribution is obtained by traversing the data sample only twice and no iteration is required. Therefore our algorithm is more efficient than iterative algorithms such as EM and SEM. Our experiments show that the method performs well both for parameter learning and model learning compared to EM and SEM.

1 Introduction

Most methods for performing statistical data analysis require complete data samples in order to work or produce valid results. Unfortunately real-life databases are rarely complete. For doing statistical analysis of incomplete data, the standard tools for complete data often don't suffice anymore. Principled data analysis of incomplete data leads to analytical intractability and high computational complexity compared to the complete data scenario.

This paper is concerned with learning Bayesian networks (BN), a formalism built upon statistical principles. BNs are so-called directed graphical models which is a class of statistical models defined by a collection of conditional independences between variables represented by a graph. This graph offers an appealing way of structuring an otherwise confusing number of equations expressing the (in)dependences between variables.

BNs occupy a prominent position in decision support environments where they are used for diagnostic and prediction purposes. Also, in the context of data mining especially the graphical structure (model) of a Bayesian network is an appealing formalism for visualising the relationships between domain variables.

In this paper we show how to learn BNs from incomplete data when the missing data mechanism is *ignorable* as defined by Little & Rubin [15], which entails that data should be missing at random (MAR) or missing completely at random (MCAR). This essentially means that the probability that some entry is missing possibly depends on observed data, but is independent of unobserved data. In the typical MAR missing data mechanism, the probability of occurrence of a missing entry in a variable depends on fully observed covariates only. Without the ignorability assumption it is impossible to develop a fully automated procedure that produces statistically valid results. The reason is that the MAR assumption provides a minimal condition on which valid statistical analysis can be performed without modelling the underlying missing data mechanism. Under the MAR assumption, all information about the missing data, necessary for performing valid statistical analysis, is contained in the observed data, but structured in a way that complicates the analysis [3].

The method we develop has low computational cost compared to most existing incomplete data methods. The price we have to pay for this efficient algorithm is a certain degree of approximation.

Our missing data approach is not directly linked to model learning or parameter estimation *per se*. Instead we focus on the so-called *predictive distribution* which plays a crucial role when we want existing learning methods developed for complete data to work with incomplete data as well.

We proceed as follows: In section 2 we give a short review of previous research on learning Bayesian networks from incomplete data. Section 3 introduces the notation and in section 4 the problem with incomplete data is briefly illustrated. In section 5 methods for learning Bayesian networks are discussed, followed by section 6 and 9 where we develop our method for creating predictive distributions. Section 10 summarises the required steps for generating the approximate predictive distribution. In section 11 we test our method for parameter estimation and model selection, and finally in section 12 we draw conclusions.

2 Previous research

Probably the most well-known technique for learning parameters of statistical models from incomplete data is the Expectation-Maximisation (EM) algorithm by Dempster, Laird & Rubin [7]. Several EM derivatives exists, such as Generalised EM (GEM) and versions with a stochastic element; see for instance McLachlan & Krishnan [16]. EM was studied in the context of graphical models by Lauritzen [14]. In its original form, EM is a deterministic iterative two-step algorithm that converges towards the Maximum-Likelihood (ML) parameter estimates (or the maximum a posteriori (MAP) estimates).

When applied to learning parameters, the E-step in EM involves the performance of inference in order to obtain the values of the expected sufficient statistics. It is followed by an M-step which takes the sufficient statistics from the E-step, and calculates the ML-estimates. These ML-estimates are taken as the parameters of the Bayesian network, and the E-step is performed again, i.e., the expected values of the sufficient statistics are calculated by inference using the ML-parameters calculated in last M-step. Repeated application will finally return the ML-estimates.

Learning Bayesian network models from incomplete data so to speak adds a layer on top of the parameter learning methods described above. For EM, Friedman [10, 9] showed that doing a model selection search *within* EM will result in the best model in the limit according to some model scoring criterion. This 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 values of the sufficient statistics obtained from the current model and current parameter values.

Tanner & Wong [23] introduced a stochastic simulation-based approach for learning parameters called Data Augmentation (DA), which can be considered a Bayesian method for learning. DA is quite similar to EM, but instead of calculating the expected sufficient statistics, a value is drawn from a predictive distribution and filled in (imputed). Similarly, instead of calculating the ML-estimates, a parameter value is drawn from the posterior distribution on the parameter space conditional on the most recent fully imputed data sample. Based on Markov chain Monte Carlo theory this will in the limit return realisations from the posterior parameter distribution conditional on the observed data.

Riggelsen & Feelders [22] describe a Bayesian method for model learning from incomplete data. The method is an imputation-based approach, where possible completions of the data are scored together with the observed part of the data. Using importance sampling, imputations can be re-used as models are sampled from the posterior model distribution. In Riggelsen [21] importance sampling is used for parameter learning, but in contrast to DA, the method described approximates the parameter posterior by a mixture distribution.

Cowell, Dawid & Sebastiani [6] review several methods for representing a parameter posterior through a Bayesian sequential approach. They consider deterministic rules such as fractional updating and matched moments for collapsing components, making the functional form of the posterior a mixture distribution with a tractable number of components.

Another algorithm for parameter estimation and model selection from incomplete data is Bound and Collapse (BC) by Ramoni & Sebastiani [18, 19]. The bound phase considers possible completions of the data sample. The sufficient statistics for these bounds are used for computing an interval in which the actual parameter estimate lies. The collapse phase computes a convex combination of these extreme parameter bounds, where the weights of the convex combination are computed from the observed available cases of the data. BC seems to work for some missing data mechanisms, but unfortunately it can produce rather unpredictable results for others.

Like BC, the method we describe in this paper is a non-iterative procedure. To some degree our method resembles BC, but on crucial points it differs; most notably, we approach the missing data problem from a prediction point of view.

3 Preliminaries

We start by introducing some notation and by defining the concept of a Bayesian network.

Capital letters denote discrete random variables and lower case letters denote states. Boldface denote vectors and vector states. We use $\Pr(\cdot)$ to denote probability distributions (or densities) and probabilities; the context makes it clear what is meant.

A Bayesian network (BN) for the discrete random variables $\boldsymbol{X} = (X^1, \ldots, X^p)$ 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:

$$\Pr(\boldsymbol{X}|m,\boldsymbol{\theta}) = \prod_{i=1}^{p} \Pr(X^i|\boldsymbol{\Pi}^i,\boldsymbol{\theta}) = \prod_{i=1}^{p} \theta_{X^i|\boldsymbol{\Pi}^i},$$

where $\boldsymbol{\Pi}^i$ is the parent set of X^i in m. We use $\boldsymbol{\Lambda}^i$ to denote the set of children of X^i in m.

The Markov blanket of a variable X^i is defined as:

$$\boldsymbol{\Phi}^i = \boldsymbol{\Pi}^i \cup \boldsymbol{\Lambda}^i \cup \{X^j|X^j \in \boldsymbol{\Pi}^k, X^k \in \boldsymbol{\Lambda}^i\} \setminus \{X^i\}.$$

It follows from the BN decomposition according to m that $X^i \perp\!\!\!\perp \boldsymbol{X} \setminus \boldsymbol{\Phi}^i \,|\, \boldsymbol{\Phi}^i$, i.e., that given $\boldsymbol{\Phi}^i$, X^i is independent of all remaining variables.

4 Learning from data

As will become clear in the next section, the likelihood function plays a crucial role in learning BNs, both in the frequentist and the Bayesian approach.

Suppose we are given data in the form of a multinomial sample $\mathcal{D} = (\boldsymbol{d}_1, \dots, \boldsymbol{d}_c)$ consisting of c i.i.d. cases, and every record record $\boldsymbol{d}_l = (x_l^1, \dots, x_l^p)$ is a particular configuration of the variables $\boldsymbol{X} = (X^1, \dots, X^p)$. Given a Bayesian network model, m, the likelihood of $\mathcal{D}$ is:

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D};m) = \prod_{l=1}^{c} \prod_{i=1}^{p} \theta_{x_l^i|\boldsymbol{\pi}_l^i} = \prod_{i=1}^{p} \prod_{x^i} \prod_{\boldsymbol{\pi}^i} \theta_{x^i|\boldsymbol{\pi}^i}^{s(x^i,\boldsymbol{\pi}^i)},$$

where $s(\cdot)$ is the number of cases in the data sample with that particular configuration; the sufficient statistics. We notice that for complete data the likelihood is a simple product of terms.

Most learning algorithms rely on the fact that the functional form of the likelihood is a product of terms. For incomplete data this is unfortunately no longer true. Suppose that we are given a data sample with unobserved items $\mathcal{D} = (\mathcal{O},\mathcal{U}) = (\boldsymbol{d}_1, \dots, \boldsymbol{d}_c)$ consisting of c i.i.d. cases. Each record $\boldsymbol{d}_l = (x_l^1, \dots, x_l^p) = (\boldsymbol{o}_l, (u_l^1, \dots, u_l^{r(l)}))$ is a p-dimensional vector, and has an observed part $\boldsymbol{o}_l$ and an $r(l)$-dimensional unobserved part $\boldsymbol{u}_l$. The likelihood of observed data is:

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{O};m) = \prod_{l=1}^{c} \sum_{\boldsymbol{u}_l} \prod_{i=1}^{p} \theta_{x_l^i|\boldsymbol{\pi}_l^i},$$

that is, the unobserved items are summed out for each record. The likelihood is no longer a simple product, but includes summations as well. For a given record there will for observed variables be as many terms as there are completions of the ancestral variables in m.

Example: Assume we are given the following model m with two binary variables, $X^1 \rightarrow X^2 \rightarrow X^3$, and assume that we are given the following incomplete data sample:

	X^1	X^2	X^3
$\boldsymbol{d}_1$	f	t	t
$\boldsymbol{d}_2$	t	?	t

After seeing $\boldsymbol{d}_1$, the likelihood $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{d}_1;m)$ is still a simple product:

$$(1-\theta_{X^1=t})^1 \theta^1_{X^2=t|X^1=f} \theta^1_{X^3=t|X^2=t},$$

but once $\boldsymbol{d}_2$ is observed, we need to sum out X^2, and the likelihood $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{d}_1,\boldsymbol{d}_2;m)$ becomes (where $1-\theta_{X^i=t|\boldsymbol{\pi}^i} = \theta_{X^i=f|\boldsymbol{\pi}^i}$):

$$\theta^1_{X^1=f}\theta^1_{X^1=t}\theta^1_{X^2=t|X^1=f}\theta^1_{X^2=t|X^1=t}\theta^2_{X^3=t|X^2=t}+$$

$$\theta^1_{X^1=f}\theta^1_{X^1=t}\theta^1_{X^2=t|X^1=f}\theta^1_{X^2=f|X^1=t}\theta^1_{X^3=t|X^2=t}\theta^1_{X^3=t|X^2=f},$$

a summation over likelihoods of complete data, one per completion.

In rare cases the likelihood of incomplete data remains a product, namely if all descendants of missing items are missing as well. Hence, if this holds for all the records, then the resulting likelihood has the same functional form as the likelihood of complete data, i.e., it contains no summations. This resembles the so-called monotone pattern of missingness described in Little & Rubin [15]. In the previous example, if X^3 would be missing in record two, where X^2 is already missing, the likelihood would be the product:

$$(1-\theta_{X^1=t})^1 \theta^1_{X^1=t} \theta^1_{X^2=t|X^1=f} \theta^1_{X^3=t|X^2=t}.$$

5 Learning Bayesian networks

Here we review the methods usually applied when learning BNs. We discuss both the frequentist point of view and the Bayesian approach.

5.1 Parameter estimation In order to learn the parameters of a BN from a frequentist point of view, the Maximum-Likelihood estimates are used. For complete data these are:

$$\hat{\theta}_{x^i|\boldsymbol{\pi}^i} = \frac{s(x^i,\boldsymbol{\pi}^i)}{s(\boldsymbol{\pi}^i)}.$$

With incomplete data, however, finding the parameter values that maximise the likelihood is no trivial task; we then have to turn to numerical optimisation techniques.

In the Bayesian approach to parameter learning, a prior distribution, $\Pr(\boldsymbol{\Theta}|m)$, on the parameter space is defined, and by applying Bayes' rule, we have $\Pr(\boldsymbol{\Theta}|\mathcal{D},m) \propto \mathcal{L}(\boldsymbol{\Theta}|\mathcal{D};m)\cdot\Pr(\boldsymbol{\Theta}|m)$, where $\Pr(\boldsymbol{\Theta}|\mathcal{D},m)$ is the posterior distribution. A natural summary statistic is then the expectation of the parameters with respect to the posterior. For complete data, the known conjugate prior is the (product) Dirichlet distribution for which the 1st moment is:

$$\hat{\theta}_{x^i|\boldsymbol{\pi}^i} = \mathrm{E}[\Theta_{x^i|\boldsymbol{\pi}^i}|\mathcal{D},m] = \frac{\alpha(x^i,\boldsymbol{\pi}^i)+s(x^i,\boldsymbol{\pi}^i)}{\alpha(\boldsymbol{\pi}^i)+s(\boldsymbol{\pi}^i)},$$

where $\alpha(\cdot)$ denotes the prior hyper parameters (counts) of the Dirichlet distribution.

For incomplete data the Dirichlet distribution is not conjugate, and in fact the posterior will in that case be a mixture distribution with as many components as there are summation terms in the likelihood of observed data—exponential in the number of missing items in the sample. From a computational point of view it is entirely intractable to retain all these in memory and to perform statistical analysis.

5.2 Model learning For model learning, the frequentist applies a penalised likelihood scoring criterion, such as the MDL criterion. These kinds of scores are defined as the product of the ML-estimates and a penalty term counteracting overfitting (preventing too many arcs in the model):

$$Score(M|\mathcal{D}) = \prod_{i=1}^{p}\prod_{x^i}\prod_{\boldsymbol{\pi}^i}\hat{\theta}_{x^i|\boldsymbol{\pi}^i}^{s(x^i,\boldsymbol{\pi}^i)}\cdot penalty.$$

Since the ML-estimates are difficult to determine for incomplete data, so is the computation of this penalised scoring metric.

For the Bayesian on the other hand the entire posterior distribution on the model space is of interest, which is found by integration and application of Bayes' rule: $\Pr(M|\mathcal{D}) \propto \int\mathcal{L}(\boldsymbol{\theta}|\mathcal{D};M)\cdot\Pr(\boldsymbol{\theta}|M)d\boldsymbol{\theta}\cdot\Pr(M)$. The posterior is approximated by empirical samples obtained via MCMC, circumventing the need for calculation of the required normalising constant. For model selection the maximum a posteriori model (MAP) is used, in which case the normalising constant can be disregarded entirely. Under the assumption of prior indifference between models, $\Pr(M)$ is the same for all models. In that case for the MAP model all that remains is to maximise $\Pr(\mathcal{D}|M) = \int\mathcal{L}(\boldsymbol{\theta}|\mathcal{D};M)\cdot\Pr(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$, the marginal likelihood.

The marginal likelihood has a closed form under the assumption of complete data, in which case it is the normalising constant for the prior product Dirichlet over the normalising constant for the posterior product Dirichlet (see Riggelsen [20] and Heckerman, Geiger, Chickering [12]):

$$\Pr(\mathcal{D}|M) = \prod_{i=1}^{p}\prod_{\boldsymbol{\pi}^i}\frac{\dfrac{\Gamma\Big(\alpha(\boldsymbol{\pi}^i)\Big)}{\prod_{x^i}\Gamma\Big(\alpha(x^i,\boldsymbol{\pi}^i)\Big)}}{\dfrac{\Gamma\Big(s(\boldsymbol{\pi}^i)+\alpha(\boldsymbol{\pi}^i)\Big)}{\prod_{x^i}\Gamma\Big(s(x^i,\boldsymbol{\pi}^i)+\alpha(x^i,\boldsymbol{\pi}^i)\Big)}}.$$

As seen, both model learning as well as parameter learning are functions of the *sufficient statistics* $s(\cdot)$ from the data sample gathered through the likelihood, which is assumed a product of terms. In order to still be able to apply the learning methods and metrics developed for complete data, we thus need to predict the missing values from the observed part of the data, i.e., make the incomplete data look like it is complete. In terms of sufficient statistics this means that we need to obtain the *expected* sufficient statistics given the observed part of the data.

6 Predicting missing values

The predictive distribution we define as the distribution of missing values given observed values and DAG model m:

$$\Pr(U|\mathcal{O},m).$$

The expected sufficient statistics with respect to this predictive distribution is then:

$$\hat{s}(x^i,\boldsymbol{\pi}^i) = \mathrm{E}[s(x^i,\boldsymbol{\pi}^i)] = \sum_{\mathcal{U}} s(x^i,\boldsymbol{\pi}^i)\Pr(\mathcal{U}|\mathcal{O},m),$$

Intuitively we may think of this equation as a way of "producing" values from the predictive distribution that are "filled in" where data is missing. Hence, each completion gives rise to the sufficient statistics $s(x^i,\boldsymbol{\pi}^i)$.

For incomplete data, we can now use $\hat{s}(\cdot)$ everywhere in the complete data learning methods and metrics. The problem lies in how one should derive or define the predictive distribution.

Unfortunately there is no easy way of obtaining the exact predictive distribution without iteration. EM reaches the "correct" parameterisation for this distribution in the limit, at every step getting closer and closer to the true predictive distribution. In doing so, EM predicts the expected sufficient statistics by performing relatively expensive inference in the BN. DA on the other hand is in the limit guaranteed to produce realisations from the predictive distribution, but several realisations are generally required in order to estimate the expected sufficient statistics.

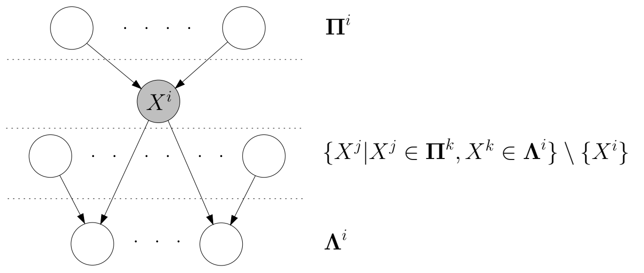
$\mathbf{\Pi}^i$
X^i
$\{X^j|X^j \in \mathbf{\Pi}^k, X^k \in \mathbf{\Lambda}^i\} \setminus \{X^i\}$
$\mathbf{\Lambda}^i$

Figure 1: Markov blanket of X^i. The Markov blanket blocks influence from all other variables.

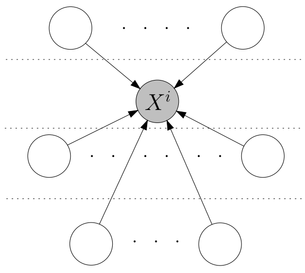

Figure 2: New model m' derived from m. All variables in the Markov blanket of X^i in m are directed towards X^i in m'.

7 The general idea

The basic idea behind our method, is to predict missing values on a per variable basis, by focusing only on those observed variables that directly influence the variables that are missing. The BN model explicitly identifies the relevant predictor variables as those variables that form the Markov blankets. By looking for cases in the data that have similar values on the observed predictor variables, a predictive distribution for the missing data is created. When the number of predictors is too large, a selection is made, and only the best predictors (those that have strongest influence) are considered when generating the predictive distribution.

8 Approximate predictive distributions

In this section we propose how to approximate the predictive distribution. The way the approximation is defined, naturally leads to a non-iterative algorithm. Additionally, there is no need to perform exact inference as in EM, or draw several realisations as in DA.

We assume for any record l the following factorisation:

$$(8.1) \qquad \Pr(\boldsymbol{U}_l|\boldsymbol{o}_l, m) = \prod_{k=1}^{r(l)} \Pr(U_l^k|\boldsymbol{o}_l, m),$$

saying that we are able to predict single missing values independently of each other through separate predictive distributions. This assumption in fact *does* hold given that the missing variables are *d-separated* (see Pearl [17] for details) from each other, i.e., observed variables block influences between unobserved variables.

The dependence on observed variables $\boldsymbol{O}_l$ is explicitly given by model m. If the independence assumptions expressed in equation 8.1 actually do hold, then the dependence of U_l^k on $\boldsymbol{O}_l$ is in fact only on a subset of the observed variables, namely the Markov blanket of U_l^k.

Figure 1 illustrates the Markov blanket of a variable X^i. If $X_l^i = U_l^k$, we can thus predict it if all variables in the Markov blanket are observed:

$$\Pr(X_l^i|\boldsymbol{o}_l, m) = \Pr(X_l^i|\boldsymbol{\phi}_l^i, m).$$

If not all variables of the Markov blanket are observed, then there might be other observed variables outside the blanket that influence the prediction. We return to this issue in section 8.2, and from now on *define* the approximate predictive distribution for record l as:

$$\Pr(\boldsymbol{U}_l|\boldsymbol{o}_l, m) = \prod_{k=1}^{r(l)} \Pr(U_l^k|\boldsymbol{\phi}_l^k, m),$$

where $\Pr(X^i|\boldsymbol{\Phi}^i, m)$ is the univariate predictive distribution for X^i.

The decomposition of the predictive distribution for X^i according to m is then:

$$(8.2) \qquad \Pr(X^i|\boldsymbol{\Phi}^i, m) \propto \theta_{X^i|\mathbf{\Pi}^i} \prod_{X^q \in \mathbf{\Lambda}^i} \theta_{X^q|\mathbf{\Pi}^q}.$$

Hence, in order to predict the missing value, we need first estimate the parameters $\theta_{X^j|\mathbf{\Pi}^j}$. To avoid iteration, we could instead do available cases analysis, or BC for that purpose. The problem with these approaches is that they in one way or another only consider observations in the data pertaining directly to X^j and $\mathbf{\Pi}^j$; they neglect the fact that other variables also influence the estimation when X^j or $\mathbf{\Pi}^j$ are (partly) missing. With incomplete data $\hat{\theta}_{x^j|\boldsymbol{\pi}^j}$ is not only a function of X^j and $\mathbf{\Pi}^j$, but of other variables as well. Disregarding this influence effectively means that valuable information is discarded that otherwise could help in making a better (and perhaps unbiased) prediction of X^i.

We propose to change m such that the parametrisation of the predictive distribution of X^i remains dependent on the Markov blanket of X^i, but in a way that predictions are less sensitive to slightly inaccurate parameter estimates.

Given the variables in the Markov blanket of m, the univariate predictive distribution based on this new model m' captures no extra assumption about independence compared to the predictive distribution based on m. This means that $\Pr(X^i|\mathbf{\Phi}^i, m)$ and $\Pr(X^i|\mathbf{\Phi}^i, m')$ have the same predictive capabilities. Specifically, define m' such that $\mathbf{\Pi}^i_{m'} = \mathbf{\Phi}^i_m$ and $\mathbf{\Lambda}^i_{m'} = \varnothing$, i.e., extend the parent set of X^i to include *all* variables of the Markov blanket of X^i and remove any children. Figure 2 illustrates m' derived from m.[1] The independences *actually* holding in m, are disregarded without further ado for predicting missing values:

$$\Pr(X^i|\mathbf{\Pi}^i, m') = \theta_{X^i|\mathbf{\Pi}^i} \leq_{CI} \Pr(X^i|\mathbf{\Phi}^i, m),$$

where $\leq_{CI}$ means that the distribution on the left is less restrictive in its conditional independence assumptions compared to the distribution on the right.

Obtaining $\hat{\theta}_{X^i|\mathbf{\Pi}^i}$ for m' is now done in a way related to available cases analysis, but all variables highly relevant for determining this parameter of the predictive distribution are considered *jointly.* In m' we thus explicitly consider all relevant variables for the parametrisation *together*, in contrast to m where *separate* child-parent variables are considered for estimating the parameters required according to decomposition m.

In a sense m' allows for a rather direct prediction approach; there is no need to perform inference because the most relevant variables for making a prediction are considered jointly. For the predictive distribution based on m, it is indirectly assumed that the correct parametrisation of the actual BN is given *prior* to predicting the missing values; it is by means of (simple) inference that the missing values are predicted, that is, by applying eq. 8.2. However, the only way of estimating these parameters accurately is to have knowledge about the missing values, but of course these missing values are exactly what is subject to prediction.

Note that across equivalent BN models vertices have the same Markov blankets, i.e., the Markov blanket of vertices does not change for equivalent DAG models. Thus m' is unique across equivalent DAG models. However, the parent set for the vertices within an equivalence class is *not* the same. The estimates computed by the BC algorithm and available case analysis are functions of the parent set $\mathbf{\Pi}^j$ (and X^j). This means that these methods return different results for equivalent DAGs. This is undesirable, because equivalent DAGs are statistically indistinguishable.

8.1 Parameter estimation For the parametrisation of the predictive distribution, a *similar cases* approach is used, based on cases where X^i is observed.

Define $match(\mathbf{\Pi}^i, \boldsymbol{D}_l; X^i)$ as the function that returns the *degree of match* between the configuration of the variables in $\mathbf{\Pi}^i$ and the corresponding variables in record l, *given* that X^i matches the corresponding variable in l. If X^i_l is observed and $x^i = x^i_l$, none of the parent variables in record l are missing and there is a perfect match between the two configurations, then the function returns 1. However, if some parents are missing, but the observed parents all match, then it returns the fraction 1/number-of-possible-configurations of the missing parent(s). The total number of occurrences is then counted in the following way:

$$s^*(x^i, \boldsymbol{\pi}^i) = \sum_{l=1}^{c} match(\boldsymbol{\pi}^i, \boldsymbol{d}_l; x^i),$$

i.e. all records are matched, and the sum is the required statistic.

Example: Assume that we have a model $\{X^2, X^3, X^4\} \rightarrow X^1$, and that the following data is given:

	X^1	X^2	X^3	X^4
$\boldsymbol{d}_1$	f	?	t	?
$\boldsymbol{d}_2$	?	f	t	f
$\boldsymbol{d}_3$	t	?	f	t
$\boldsymbol{d}_4$	t	t	f	t

Suppose we require $s^*(X^1 = t, \mathbf{\Pi}^1 = (X^2 = t, X^3 = f, X^4 = t))$. Any record with $X^1 = t$ is considered, and all those records are counted for which the parent set can be (or is) completed as $\mathbf{\Pi}^1 = (X^2 = t, X^3 = f, X^4 = t)$, resulting in $s^*(X^1 = t, \mathbf{\Pi}^1 = (X^2 = t, X^3 = f, X^4 = t)) = 1.5$. The first perfect match is from record 4. In record 3 there is a partial match, and since X^2 has two possible alternative configurations, 0.5 is returned as the degree of match.

For $s^*(X^1 = f, \mathbf{\Pi}^1 = (X^2 = t, X^3 = t, X^4 = f))$ the only (partial) match is record 1, but since X^2 and X^4 are missing, there are 4 possible completions, so $s^*(X^1 = f, \mathbf{\Pi}^1 = (X^2 = t, X^3 = t, X^4 = f)) = 0.25$.

Using the degree of a match, fully matched observations count more than partial matches; in a partial match several possible configurations “share the single count” between them which is only fair given the fact that the missing values could have been any of the configurations.

Gathering the sufficient statistics based on (a subset of) similar observations from the observed records is valid for a broad range of MAR mechanisms, but is

[1] We note that a m' is defined for each univariate predictive distribution. m' is only concerned with X^i and the variables in $\mathbf{\Phi}^i$ (of m).

not necessarily optimal for arbitrary missing data mechanisms. In contrast to BC and available case analysis, the decomposition according to m' is less sensitive to the actual underlying MAR mechanism because of the larger dependence components captured by considering jointly all variables of the Markov blanket of m.

Obtaining the statistics $s^*(\cdot)$ requires no iteration, and can be done by running through the data sample only once. For each record, all statistics for all vertices X^i are collected using the match-function.

Finally, to estimate $\theta_{x^i|\boldsymbol{\pi}^i}$, we use the ML-estimates (or the expectation of the posterior Dirichlet):

$$\hat{\theta}_{x^i|\boldsymbol{\pi}^i} = \frac{s^*(x^i, \boldsymbol{\pi}^i)}{\sum_{x^i} s^*(x^i, \boldsymbol{\pi}^i)}.$$

8.2 Prediction and missing parents For every single variable in $\boldsymbol{X}$ a predictive distribution can be created the way discussed in the previous sections and applied in all records where X^i is missing and all parents are observed. However, in some records not all $\boldsymbol{\Pi}^i$ in m' are observed, and consequently the absent parent variable(s) can't be used as (a) predictor(s). We therefore define the predictive distribution in slightly different way than in the fully observed parent case. The variables for which there are no values have to be "summed out" such that only the observed variables act as predictors. For instance, if in record l the variable X^i_l is missing and needs to be predicted, and a subset of predictors $\boldsymbol{V}_l \subseteq \boldsymbol{\Pi}^i$ is missing (so $\boldsymbol{V}_l \subset \boldsymbol{U}_l$), the predictive distribution for X^i_l is $\Pr(X^i_l|\boldsymbol{\Pi}^i_l \backslash \boldsymbol{V}_l, m')$. The parameters for m' are obtained in terms of $s^*(\cdot)$ defined earlier on by summing out the missing variables:

$$s^*(x^i_l, \boldsymbol{\pi}^i_l \setminus \boldsymbol{v}_l) = \sum_{\boldsymbol{v}_l} s^*(x^i_l, \boldsymbol{\pi}^i_l).$$

This means that when the parent set is not fully observed the ML-estimates are functions of these marginal statistics rather than the "original" statistics.

In summary, we need to traverse the sample once for obtaining the statistics $s^*(\cdot)$. From these statistics we can create the required univariate predictive distributions for both complete and incomplete parent sets.

9 Predictive quality

As already noted, the predictive distributions based on m and m' differ in the conditional independence assumptions, but in a way that they still have the same predictive capabilities. However, this actually only holds when parameters of the predictive distributions are estimated from a sufficiently large sample.

To estimate $\theta_{x^i|\boldsymbol{\pi}^i}$ we need several "examples" with $\boldsymbol{\pi}^i$ and x^i observed to get an accurate prediction. If the cardinality of $\boldsymbol{\Pi}^i$ is large there are potentially many different configurations; there may not be enough examples of a particular configuration $(x^i, \boldsymbol{\pi}^i)$ to make a reliable estimate of $\theta_{x^i|\boldsymbol{\pi}^i}$ for m'. Consequently the prediction of X^i may suffer.

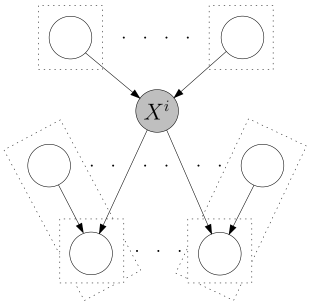

Figure 3: The potential predictors of X^i from m. The dotted boxes indicate the predictors that are checked for their predictive quality.

The problem is thus that predictions for vertices with a large Markov blanket may be unreliable if the sample size is small. Therefore, we propose to reduce the number of predictors. Instead of taking all vertices of a Markov blanket in m as the parents of X^i in m', we suggest to select as parents the best n variables of the Markov blanket of X^i in m; best in terms of predictive power.

9.1 Selecting predictive variables First we need to distinguish between the vertices of the Markov blanket in terms of *direct dependences* and *induced dependences.* We have that $\boldsymbol{\Pi}^i \not\perp\!\!\!\perp X^i$ and $\boldsymbol{\Lambda}^i \not\perp\!\!\!\perp X^i$, i.e. X^i depends on both parents and children; this is a direct dependence because in m we have that $\boldsymbol{\Pi}^i \rightarrow X^i$ and $\boldsymbol{\Lambda}^i \leftarrow X^i$. The induced dependence is related to the *parents* of $\boldsymbol{\Lambda}^i$ in the following way: For every $X^j \in \boldsymbol{\Lambda}^i$ we have that $\boldsymbol{\Pi}^j \setminus \{X^i\} \not\perp\!\!\!\perp X^i | X^j$, i.e. X^i depends on the parents of a child of X^i *given* the child.[2] This induced dependence is because of the converging connection, $X^i \rightarrow X^j \leftarrow X^k$ with $X^j \in \boldsymbol{\Lambda}^i, X^k \in \boldsymbol{\Pi}^j$. This means that if a child of X^i is not chosen to be part of the predictive variables, then in fact we know that the parents of that child are irrelevant as well; we need not consider any of them. Therefore a parent X^k of a child X^j is only considered *in conjunction* with X^j as a predictor of X^i. Hence, the set of potential predictors for X^i is:

$$\boldsymbol{\Pi}^i \cup \boldsymbol{\Lambda}^i \cup \{(X^j, X^k) | X^j \in \boldsymbol{\Pi}^k \setminus \{X^i\}, X^k \in \boldsymbol{\Lambda}^i\},$$

[2] For ease of exposition we here assume for $X^j \in \boldsymbol{\Lambda}^i$ that $\boldsymbol{\Pi}^j$ and $\boldsymbol{\Pi}^i$ are disjoint. If they are not, then the induced dependence is simply a direct dependence (see figure 3).

from which the n best predictors have to be chosen as the parent set of X^i in m'. Figure 3 illustrates the different predictors. We thus have to test each predictor, and select the n best.

If we want to check how well a predictor, say X^j, is able to predict X^i, we compute the following probability:

$$\Pr(\mathcal{D}^i|\mathcal{D}^j, \mathbf{\Pi}^i = \{X^j\}),$$

where $\mathcal{D}^i$ denotes the i'th column of the data sample referring to X^i. Hence, given that X^j is a parent of X^i (thus a predictor), and given the values for the predictor from the data, how well can we predict the values of X^i from the data sample?

If we rewrite the above probability, we get:

$$\Pr(\mathcal{D}^i|\mathcal{D}^j, \mathbf{\Pi}^i = \{X^j\}) = \frac{\Pr(\mathcal{D}^i, \mathcal{D}^j|\mathbf{\Pi}^i = \{X^j\})}{\Pr(\mathcal{D}^j|\mathbf{\Pi}^i = \{X^j\})},$$

the fraction of two marginal likelihoods; in the numerator the data sample $(\mathcal{D}^i, \mathcal{D}^j)$ and in the denominator data sample $\mathcal{D}^j$, conditional on a model with X^j as a single parent of X^i.

Under the Dirichlet assumption, the marginal likelihood has a known functional form, and the above ratio is (where $\mathbf{\Pi}^i$ is a set with a predictor consisting of one or two variables):

$$\Pr(\mathcal{D}^i|\mathcal{D}^{\mathbf{\Pi}^i}, \mathbf{\Pi}^i) = \prod_{\boldsymbol{\pi}^i} \frac{\frac{\Gamma\left(\alpha(\boldsymbol{\pi}^i)\right)}{\prod_{x^i} \Gamma\left(\alpha(x^i, \boldsymbol{\pi}^i)\right)}}{\frac{\Gamma\left(s(\boldsymbol{\pi}^i) + \alpha(\boldsymbol{\pi}^i)\right)}{\prod_{x^i} \Gamma\left(s(x^i, \boldsymbol{\pi}^i) + \alpha(x^i, \boldsymbol{\pi}^i)\right)}}.$$

This ratio is only defined when the prior counts $\alpha(x^i, \boldsymbol{\pi}^i) > 0$. We let $\alpha(x^i, \boldsymbol{\pi}^i) = 1$, indicating that one instance of the configuration $(x^i, \boldsymbol{\pi}^i)$ has been observed prior to observing the data.

As mentioned previously, the functional form of the marginal likelihood as given above, only holds for complete data. In order to use the proposed score to determine the predictive quality, we only use the records where we have observations on X^i as well as on the predictor(s). We note that under the general MAR assumption (not MCAR), records where the predictor is missing but X^i is observed also should be taken into account when collecting the sufficient statistics. We however assume that omitting those records does not introduce any severe bias. We don't actually need an entirely correct probability; we merely rank the predictors according to this score, and assume that the bias introduced does not influence the ranking in a disastrous way.

Another point which we need to mention is that although we check each predictor on an individual basis, there is no real guarantee that the n best predictors *jointly* are good predictors. The general idea is however that n predictors that score well individually, when taken together produce even better predictions.

In order to calculate the predictive quality, we need to collect the sufficient statistics from the data sample. This can be done by one traversal through the sample. The statistics thus obtained are not the same as the ones required in the previous section; the n best predictors need to be determined before we can collect $s^*(\cdot)$.

10 Summary

The theory presented thus far and the assumptions made during the derivations can be summarised as follows:

1. The predictive distribution for case l can be written as a product $\prod_{k=1}^{r(l)} \Pr(U_l^k|\boldsymbol{o}_l, m)$. We restrict this dependence to include only the Markov blanket $\mathbf{\Phi}^i$, such that the predictive distribution becomes $\prod_{k=1}^{r(l)} \Pr(U_l^k|\boldsymbol{\phi}_l^k, m)$.
2. m is transformed into m' as depicted in figures 1 and 2.
3. If the cardinality of the Markov blanket is large in m, we select the n best predictors from the blanket as shown in figure 3 for m', using the conditional marginal likelihood score as a measure of predictive quality.
4. We estimate all $\theta_{x^i|\boldsymbol{\pi}^i}$ for $\Pr(X^i|\mathbf{\Phi}^i, m')$ by considering observed cases using the match-function. The univariate predictive distribution for a missing value in a case with (partly) missing predictors, is obtained by "summing out" the sufficient statistics $s^*(\cdot)$ for the variable(s) in question.

In step 3 and 4 the data sample is traversed once. Actually using the univariate predictive distributions is part of the "counting-mechanism" employed by the complete data method. When a record is consulted with missing data, instead of counting, the predictive distribution is called, which returns the fraction of a count corresponding to the prediction of the missing value(s) in that particular record.

11 Experiments

In this section we evaluate our method, henceforth called MBP (Markov Blanket Predictor), for generating approximate predictive distributions. MBP was implemented in C++ using STL, and was run on a 2 GHz machine under Windows 2000.

	0–10%		10–20%		20–30%		30–40%		
	EM	MBP	EM	MBP	EM	MBP	EM	MBP	AC
$\Pr(A)$	0.475	0.475	0.467	0.466	0.484	0.484	0.482	0.482	0.477
$\Pr(B\|C)$	0.130	0.130	0.132	0.132	0.131	0.131	0.133	0.134	0.122
$\Pr(B\|\bar{C})$	0.290	0.290	0.289	0.289	0.284	0.285	0.282	0.284	0.312
$\Pr(C\|A,E)$	0.510	0.510	0.488	0.489	0.509	0.504	0.477	0.482	0.501
$\Pr(C\|\bar{A},\bar{E})$	0.481	0.481	0.469	0.472	0.491	0.490	0.504	0.507	0.545
$\Pr(C\|A,\bar{E})$	0.615	0.613	0.617	0.598	0.598	0.580	0.626	0.588	0.569
$\Pr(C\|\bar{A},E)$	0.368	0.369	0.407	0.420	0.392	0.413	0.383	0.417	0.440
$\Pr(D\|E)$	0.470	0.470	0.469	0.469	0.480	0.480	0.472	0.473	0.501
$\Pr(D\|\bar{E})$	0.604	0.604	0.603	0.603	0.610	0.610	0.606	0.607	0.577
$\Pr(E\|A)$	0.478	0.477	0.473	0.465	0.467	0.460	0.448	0.444	0.458
$\Pr(E\|\bar{A})$	0.622	0.621	0.628	0.620	0.620	0.613	0.585	0.582	0.600
$\Pr(F\|B)$	0.163	0.162	0.175	0.174	0.163	0.163	0.154	0.154	0.158
$\Pr(F\|\bar{B})$	0.874	0.875	0.870	0.870	0.871	0.871	0.873	0.874	0.867

Table 1: Parameter estimates for the model in figure 4 for different fractions of missing data according to the mechanism in figure 5. A bar indicates negation of the binary variable.

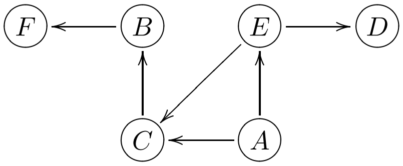

Figure 4: Model to be fitted.

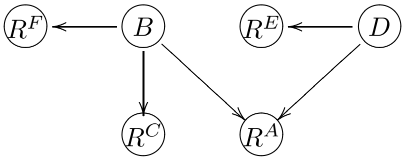

Figure 5: Missing data mechanism.

11.1 Parameter estimation First we compare MBP with standard EM for fitting a BN about probable risk factors of coronary heart disease. The data set [8] consists of 1841 records with 6 binary variables. The model is given in figure 4. Incomplete sets were generated by applying the missingness mechanism in figure 5 on the complete sample. This graph defines how response R^i of variable i depends on observed variables. Since R^i only depends on fully observed variables, it is clear that the missing data mechanism fullfills the MAR condition, but not the MCAR condition.

Four incomplete sets were generated with 0–10%, 10–20%, 20–30% and 30–40% missing values. The probability of non-response of variable i conditional on a parent configuration of R^i was selected from the specified interval.

The expected sufficient statistics obtained via MBP were used for estimating the parameters of the model. EM was run until a convergence threshold of 0.001. In table 1 the MBP and the EM estimates are shown. MBP requires 3 passes through the sample to estimate the parameters, whereas EM requires several passes, every time performing expensive inference in the BN on a per record basis. As the table suggests, the differences between estimates obtained with EM and MBP are small. At 30–40% missing data, only $\Pr(C|A,\bar{E})$ and $\Pr(C|\bar{A},E)$ suffer slightly.

For 30–40% missing data, we included for comparison, the parameter estimates obtained with available case analysis (column AC). AC relies heavily on the MCAR assumption. We see that MBP on the other hand can cope with MAR missing data as in this example. MBP is more robust against departures from the MCAR assumption compared to available cases analysis.

Estimating the parameters using MBP is almost instantaneous no matter how large the fraction of missing data is. For EM the number of passes through the data depends on the fraction of missing items; more passes were required for larger fractions of missing items.[3]

11.2 Model learning Next we evaluate MBP in a model selection context. We compare MBP to the SEM implementation of Friedman which can be downloaded

[3] The convergence rate of EM actually depends on the so-called *fraction of missing information* which does not necessarily change dramatically for different fractions of missing items.

at http://www.cs.huji.ac.il/labs/compbio/LibB. For MBP we implemented a greedy search hill-climber as the one described in [13]. The neighbourhood consists of all single additions, removals and reversals of arcs. In contrast to most hill-climbers that traverse the space of DAG modes, our hill-climber traverses the essential graph space by way of simulation. This is achieved through repeated covered arc reversals moving between equivalent DAG models. We refer to [5] and [11] for a thorough treatment of these matters.

We use the marginal likelihood as the scoring metric, with a BDeu [4] prior equivalent sample size of 1 (equivalent models receive the same marginal likelihood score). For MBP we select the 5 best predictors. The SEM implementation was run with default parameters, except for the equivalent sample size which we set to 1 (the default is 5).

We considered two benchmark BNs for the experiments: The ALARM network with 37 vertices and 46 arcs [1] about intensive care patient monitoring, and the Insurance network with 27 vertices and 52 arcs [2] classifying car insurance applications.

For the ALARM network 1000 and 5000 records were sampled. Incomplete sets were generated by applying a missingness mechanism where 18 variables (selected at random) could be missing, half of these according to an MCAR mechanism and the other half according to a MAR mechanism. Three incomplete sets were generated with 0–10%, 10–20% and 20–30% missing values on the 18 variables. For the Insurance network 1000 and 2500 records were sampled, and incomplete sets were generated where 14 variables could be missing, half of them according to a MAR mechanism and the other half according to an MCAR mechanism.

The models learned using SEM and MBP were compared using the marginal likelihood score from 10,000 records sampled from the BN learned (DAG model plus parameters) from the smaller complete data sample (1000 and 5000 for ALARM, 1000 and 2500 for Insurance). Hence, the larger this score, the better a learned model is able to predict the 10,000 records that represent the real underlying distribution. We do not consider the original networks as the golden standards because the relatively small learning samples do not necessarily support the original data generating networks anymore. It is more reasonable to compare to the models learned from these smaller complete data samples; after all we can't expect to do better than what actually *can* be learned from these smaller complete samples.

In table 2 the results for the ALARM network are shown. At no time does MBP produce worse models than SEM from a prediction point of view; in fact the score is better for both 1000 and 5000 records. The same conclusion holds for the Insurance network in table 3.

	1000		5000	
	SEM	MBP	SEM	MBP
0–10%	-99973	-98906	-110105	-107173
10–20%	-99824	-98792	-110669	-107012
20–30%	-99849	-98800	-111044	-107617

Table 2: Log marginal likelihood score for the ALARM network given 1000 and 5000 records. The score is based on 10,000 records sampled from the BNs learned from the 1000 and 5000 complete records.

	1000		2500	
	SEM	MBP	SEM	MBP
0–10%	-160251	-156628	-144174	-143189
10–20%	-161129	-156610	-149236	-145711
20–30%	-159087	-158944	-144979	-144887

Table 3: Log marginal likelihood score for the Insurance network given 1000 and 2500 records. The score is based on 10,000 record sampled from the BNs learned from the 1000 and 2500 complete records.

That MBP scores better we partly attribute to the fact that MBP does not get stuck in local optima while traversing DAG space due to covered arc reversals. Also, in contrast to EM (used by SEM), MBP will not get trapped in local optima due to an entirely different approach to solving the missing data problem compared to EM. We observed that the models learned using SEM were more complex compared to models learned with MBP.

As in parameter learning, MBP is significantly faster than SEM. At the same time, MBP is considerably simpler to implement than SEM; no BN inference engine is required.

12 Conclusion

We have developed an efficient method for generating approximate predictive distributions for missing data when learning Bayesian networks. It integrates easily with existing learning approaches for complete data, both for model learning and parameter learning.

The experiments show that the method works well for both MCAR and the more general MAR missing data mechanisms. It is a very fast method, yet the experiments indicate that the results obtained are comparable to those of EM and SEM. The gain in efficiency is especially noticeable for large data sets where EM and SEM are relatively slow.

References

[1] I.A. Beinlich, H.J. Suermondt, R.M. Chavez, and G.F. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proc. of the European Conf. on AI in Medicine*, 1989.

[2] J. Binder, D. Koller, S. J. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.

[3] J. P. L. Brand. *Development, Implementation and Evaluation of Muliple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets*. PhD thesis, Erasmus University Rotterdam, 1999.

[4] W. Buntine. Theory refinemement on Bayesian networks. In B. D'Ambrosio, P. Smets, and P. Bonissone, editors, *Proc. of the Conf. on Uncertainty in AI*, 1991.

[5] R. Castelo and T. Kocka. On inclusion-driven learning of Bayesian networks. *J. of Machine Learning Research*, 4:527–574, 2003.

[6] R. G. Cowell, A. P. Dawid, and P. Sebastiani. A comparison of sequential learning methods for incomplete data. *Bayesian Statistics*, 5:533–541, 1995.

[7] 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.

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

[9] N. Friedman. Learning Bayesian networks in the presence of missing values and hidden variables. In *Intl. Conf. on Machine Learning*, pages 125–133, 1997.

[10] 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.

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

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

[13] 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.

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

[15] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. Wiley and Sons, 1987.

[16] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, 1997.

[17] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[18] 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.

[19] M. Ramoni and P. Sebastiani. Parameter Estimation in Bayesian networks from incomplete databases. *Intelligent Data Analysis Journal*, 2(1), 1998.

[20] C. Riggelsen. MCMC learning of Bayesian network models by Markov blanket decomposition. In J. Gama, R. Camacho, P. Bazdil, A. Jorge, and L. Torgo, editors, *European Conf. on Machine Learning*, pages 329–340, 2005.

[21] C. Riggelsen. Learning parameters of Bayesian networks from incomplete data via importance sampling. *Intl. J. of Approximate Reasoning*, 2006. To appear.

[22] C. Riggelsen and A. Feelders. Learning Bayesian network models from incomplete data using importance sampling. In R. G. Cowell and Z. Ghahramani, editors, *Proc. of Artificial Intelligence and Statistics*, pages 301–308, 2005.

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