A maximum entropy approach to learn Bayesian networks from incomplete data

Giorgio Corani and Cassio P. de Campos

Abstract This paper addresses the problem of estimating the parameters of a Bayesian network from incomplete data. This is a hard problem, which for computational reasons cannot be effectively tackled by a full Bayesian approach. The workaround is to search for the estimate with maximum posterior probability. This is usually done by selecting the highest posterior probability estimate among those found by multiple runs of Expectation-Maximization with distinct starting points. However, many local maxima characterize the posterior probability function, and several of them have similar high probability. We argue that high probability is necessary but not sufficient in order to obtain good estimates. We present an approach based on maximum entropy to address this problem and describe a simple and effective way to implement it. Experiments show that our approach produces significantly better estimates than the most commonly used method.

1 Introduction

Bayesian networks (BN) are well-established probabilistic graphical models that can represent joint probability distributions over a large number of random variables in a compact and efficient manner by exploiting their conditional independences, encoded through a directed acyclic graph. Inferring BNs from data sets with missing values is a very challenging problem even if the graph is given [10]. This paper focuses on inferring the parameters of a BN with *known graph* from incomplete data samples, under the assumption that missingness satisfies MAR (missing-at-random). The missing data make the log-likelihood function non-concave and multimodal; the most common approach to estimate the parameters is based on the

Giorgio Corani
Dalle Molle Institute for Artificial Intelligence, Manno, Switzerland, e-mail: giorgio@idsia.ch

Cassio P. de Campos
Dalle Molle Institute for Artificial Intelligence, Manno, Switzerland, e-mail: cassio@idsia.ch

Expectation-Maximization (EM) algorithm [14, 8]. In this case, EM can be used to search for estimates that maximize the posterior probability of the data (rather than the likelihood [17, Sec. 1.6], as it was originally designed [8]). Maximizing the posterior probability rather than the likelihood is recommended, as it generates BN parameter estimates which are less prone to overfitting [14, 13].

With abuse of notation, in the following we refer to the posterior probability of the data as the MAP score. Maximizing the posterior probability of the data is by far the most used idea to infer BN parameters, even if it does not offer the same advantages of a full Bayesian estimation. For instance, because it does not integrate over the posterior, it cannot average over the uncertainty about the parameter estimates. On the other hand, estimation can be performed by fast algorithms, such as EM, while the computational cost of the full Bayesian approach to infer BN parameters is simply prohibitive, especially in domains with many variables. EM almost always converges to a local maximum of the MAP score, so multiple starts from different initialization points are adopted with the aim of avoiding bad local maxima, and eventually the estimate corresponding to the highest MAP score is selected.

One could expect an improvement in the estimation of the parameters by using an algorithm that always obtains the *global* maximum solution of the MAP score instead of a local one, something that cannot be guaranteed with EM. To check this conjecture, we implement an optimization framework which is ensured to find, at least in small-sized problems, the global maximum score. For large domains, such task is computationally intractable, as the problem is known to be NP-hard. However, we show empirically that the global solver produces worse parameter estimates than EM itself does, despite finding estimates with higher MAP scores. The global maximum of the MAP score seems thus to be subject to some type of overfitting, highlighting severe limitations in the correlation between MAP score and the quality of the parameter estimates. In turn, this opens a question about whether selecting the estimate with highest MAP score is the best approach. Different EM runs typically achieve very close values of the MAP score, and yet return largely different parameter estimates [13, Chap. 19]. Selecting the parameter estimate which maximizes the MAP score is not a robust choice, since the difference in score among competing estimates can be very thin. We note that approaches such as the Bayesian Information Criterion (BIC) do not constitute a solution to this problem: since all the competing estimates refer to the same graph, the BIC (and other similar approaches) would simply select the estimate with highest MAP score.

In view of such considerations, we propose the following idea to estimate BN parameters: One should select the *least informative estimate, namely the maximum entropy one, among those which have a high MAP score*. The maximum entropy criterion can be stated as: "when we make inferences on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information which we do have" [12]. Thus, our criterion is applied in two steps: i) computation of the highest MAP score; ii) selection of the maximum entropy estimate, among those with high MAP score. We implement our criterion on top of both our new global solver and on top of a multi-start EM procedure. The idea of using entropy to estimate parameters of BNs from incomplete samples has

been previously advocated [4, 5, 7, 11, 22], yet our approach and those considerably differ: either they work with continuous variables and very few parameters, or they employ other inference approaches, such as the imprecise Dirichlet model [21], to later use entropy as criterion. We deal with discrete variables with BNs of a great numbers of parameters, and interpret the entropy criterion in a softer manner (as explained later on). Entropy methods have also been applied before for dealing with the uncertainty about the missingness mechanism, where the nature of the censoring data is unknown [2, 19] (we instead assume MAR).

This paper is divided as follows. Section 2 presents the estimation problem and the methods to tackle it. Expectation-Maximization (EM) (Section 2.1) and a non-linear formulation (Section 2.2) are described, which are then compared in Section 2.3. The entropy-based idea is presented in Section 2.4. Section 3 presents experiments comparing the methods. Finally, Section 4 presents our concluding remarks.

2 Methods

We adopt Bayesian networks as framework for our study. Therefore, we assume that the reader is familiar with their basic concepts [13]. A Bayesian network (BN) is a triple $(\mathscr{G}, \mathscr{X}, \mathscr{P})$, where $\mathscr{G}$ is a directed acyclic graph with nodes associated to random variables $\mathscr{X} = \{X_1, \ldots, X_n\}$ over discrete domains $\{\Omega_{X_1}, \ldots, \Omega_{X_n}\}$ and $\mathscr{P}$ is a collection of probability values $p(x_j|\pi_j)$ with $\sum_{x_j \in \Omega_{X_j}} p(x_j|\pi_j) = 1$, where $x_j \in \Omega_{X_j}$ is a category or state of X_j and $\pi_j \in \times_{X \in \Pi_j} \Omega_X$ a (joint) state for the parents Π_j of X_j in $\mathscr{G}$. In a BN, every variable is conditionally independent of its non-descendants given its parents, according to $\mathscr{G}$. Given its independence assumptions, the joint probability distribution represented by a BN is obtained by $p(\mathbf{x}) = \prod_j p(x_j|\pi_j)$, where $\mathbf{x} \in \Omega_{\mathscr{X}}$ and all x_j, π_j (for every j) agree with $\mathbf{x}$. Nodes of the graph and their associated random variables are used interchanged. The graph $\mathscr{G}$ and the variables $\mathscr{X}$ (and their domains) are assumed to be known; $\theta_{\mathbf{v}|\mathbf{w}}$ is used to denote an estimate for $p(\mathbf{v}|\mathbf{w})$ (with $\mathbf{v} \in \Omega_{\mathbf{V}}$, $\mathbf{w} \in \Omega_{\mathbf{W}}$, $\mathbf{V}, \mathbf{W} \subseteq \mathscr{X}$).

We denote as $\mathbf{y}^i$ the i-th *incomplete* instance and by $\mathbf{Y}^i \subseteq \mathscr{X}$ the set of observed variables of the i.i.d. sampled instance i. Given the *incomplete* training data $\mathbf{y} = (\mathbf{y}^1, \ldots, \mathbf{y}^N)$ with N instances such that each $\mathbf{y}^i \in \Omega_{\mathbf{Y}^i}$, we denote by $N_{\mathbf{u}}$ the number of instances of $\mathbf{y}$ that are consistent with the state configuration $\mathbf{u} \in \Omega_{\mathbf{U}}$, where $\mathbf{U} \subseteq \mathscr{X}$. Parameters are estimated by maximizing the posterior probability given $\mathbf{y}$:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} S_\theta(\mathbf{y}) = \operatorname*{argmax}_{\theta} \left(\sum_{i=1}^{N} \log \theta_{\mathbf{y}^i} + \alpha(\theta) \right), \tag{1}$$

where α represents the prior:

$$\alpha(\theta) = \log \prod_{j=1}^{n} \prod_{x_j} \prod_{\pi_j} \theta_{x_j|\pi_j}^{\alpha_{x_j,\pi_j}}, \text{ and } \alpha_{x_j,\pi_j} = \frac{\text{ESS}}{|\Omega_{X_j}| \cdot |\Omega_{\Pi_j}|},$$

where ESS stands for *equivalent sample size*, which we set to one, as usually done in the literature [13]. The argument $\mathbf{y}$ of S_θ is omitted from now on (S means *score*).

In the experiments, in order to evaluate the quality of estimates, we measure the Kullback-Leibler (KL) divergence between the joint distribution represented by the true BN and the estimated BN (which we name *joint metric*); moreover, we also use the joint marginal distribution of all leaf nodes (named *reasoning metric*). The latter measures how close a reasoning about those leaf variables with the estimated model is to that of the true model:

$$\mathrm{KL}_{\mathscr{P}(\mathbf{Z})}(\theta) = \sum_{\mathbf{z}\in\Omega_{\mathbf{Z}}} p(\mathbf{z}) \log\left(\frac{p(\mathbf{z})}{\theta_{\mathbf{z}}}\right),$$

where $\mathbf{Z}$ are the leaves and $p(\mathbf{z}) = \sum_{\mathbf{x}\in\Omega_{\mathscr{X}\setminus\mathbf{Z}}} p(\mathbf{x},\mathbf{z})$ (and respectively for $\theta_{\mathbf{z}}$). This metric requires marginalizing out all non-leaf variables, so it involves all variables in the computation. Because of that, local errors in the estimates can compensate each other, and tend to smooth the differences among methods.

2.1 Expectation-maximization

For a complete data set (that is, $\mathbf{Y}^i = \mathscr{X}$ for all i), we have a concave function on θ:

$$S_\theta = \sum_{j=1}^{n} \sum_{x_j} \sum_{\pi_j} N'_{x_j,\pi_j} \log \theta_{x_j|\pi_j},$$

where $N'_{x_j,\pi_j} = N_{x_j,\pi_j} + \alpha_{x_j,\pi_j}$, and the estimate $\hat{\theta}_{x_j|\pi_j} = N'_{x_j,\pi_j} / (\sum_{x_j} N'_{x_j\pi_j})$ achieves highest MAP score. In the case of incomplete data, we have

$$S_\theta = \sum_{i=1}^{N} \log \sum_{\mathbf{z}^i} \prod_{j=1}^{n} \theta_{x_j^i|\pi_j^i} + \alpha(\theta), \tag{2}$$

where $\mathbf{x} = (\mathbf{y}^i, \mathbf{z}^i) = (x_1^i, \ldots, x_n^i)$ represents a joint state for all the variables in instance i. No closed-form solution is known, and one has to directly optimize: $\max_\theta S_\theta$, subject to

$$\forall_j \forall_{\pi_j} : \sum_{x_j} \theta_{x_j|\pi_j} = 1, \quad \forall_j \forall_{x_j} \forall_{\pi_j} : \theta_{x_j|\pi_j} \geq 0. \tag{3}$$

The most common approach to optimize this function is to use the EM method, which completes the data with the expected counts for each missing variable given the observed variables, that is, variables Z_j^i are completed by "weights" $\hat{\theta}^k_{Z_j|\mathbf{y}^i}$ for each i, j of a missing value, where $\hat{\theta}^k$ represents the current estimate at iteration k. This idea is equivalent to weighting the chance of having $Z_j^i = z_j$ by the (current) distribution of Z_j given $\mathbf{y}^i$ (this is known as the *E-step*, and requires computations

over the BN instantiated with $\mathscr{P} = \hat{\theta}^k$ to obtain the estimated probability of missing values). Using these weights together with the actual counts from the data, the sufficient statistics values $N^k_{x_j,\pi_j}$ are computed for every x_j,π_j, and the next (updated) estimate $\hat{\theta}^{k+1}$ is obtained as if the data were complete: $\hat{\theta}^{k+1}_{x_j|\pi_j} = N'^k_{x_j,\pi_j}/(\sum_{x_j} N'^k_{x_j\pi_j})$, where $N'^k_{x_j,\pi_j} = N^k_{x_j,\pi_j} + \alpha_{x_j,\pi_j}$ (this is the *M-step*). Because in the first step there is no current estimate $\hat{\theta}^0$, an initial guess has to be used. Using the score itself to test convergence, this procedure achieves a saddle point of Eq. (1), which is usually a local optimum of the problem, and may vary according to the initial guess $\hat{\theta}^0$. Hence, it is common to execute multiple runs of EM with distinct initial guesses and then to take the estimate with highest score among them.

2.2 Non-linear solver

In order to understand whether the good/bad quality of estimates is not simply a product of EM pitfalls to properly optimize Eq.(1), we build a systematic way to translate the parameter estimation into a compact non-linear optimization problem, which is later (globally) solved with an optimization suite. The idea of directly optimizing the score function is not new (see e.g. [17]). Nevertheless, we are not aware of a method that translates the original score function into a simple formulation using symbolic variable elimination. The main issue regards the internal summations

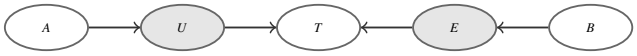

Fig. 1 Network BN_1, used in the description of the algorithm and later in the experiments; nodes affected by the missingness process have a grey background.

of Equation (2), because there is an exponential number of terms. We process them using a symbolic version of a variable elimination procedure as in [3], but the elimination method is run with target $\theta_{\mathbf{y}^i}$. Instead of numerical computations, it generates the polynomial constraints that precisely describe $\theta_{\mathbf{y}^i}$ in terms of (the still unknown) local conditional probability values of the specification of the BN. Because these values are to be found, they become the variables to be optimized in the polynomials. To clarify the method, we take the Example of Figure 1, where E,U might be missing, while the others are always observed. In this example, we need to write the constraints that describe θ_{a^i,b^i,t^i}, because these are instances in the data with missing u^i and e^i. The score function is: $\hat{\theta} = \mathrm{argmax}_\theta \max_s s$, subject to Eqs. (3) and

$$s \leq \alpha(\theta) + \sum_{i \in N^M} \log \theta_{a^i,b^i,t^i} + \sum_{i \in N^{\neg M}} \log \theta_{a^i,b^i,t^i,e^i,u^i}\,, \tag{4}$$

where N^M, $N^{\neg M}$ are index sets of the instances with and without missing values, respectively. Note that an extra optimization variable s was introduced to make the

objective function become a constraint (for ease of expose). All $\theta_{\mathbf{v}|\mathbf{w}}$ (for each possible argument $\mathbf{v},\mathbf{w}$) and s are *unknowns* to be optimized by the solver. The summation in Equation (4) can be shortened by grouping together elements related to the same states, and variables with no missing value can still be factorized out, obtaining:

$$\begin{aligned} s \leq \alpha(\theta) + \sum_a N_a \log \theta_a + \sum_b N_b \log \theta_b + \sum_{i \in N^M} \log \theta_{t^i|a^i,b^i} \\ + \sum_{a,u} N^{\neg M}_{a,u} \log \theta_{u|a} + \sum_{b,e} N^{\neg M}_{b,e} \log \theta_{e|b} + \sum_{e,t,u} N^{\neg M}_{e,t,u} \log \theta_{t|e,u}. \end{aligned} \tag{5}$$

This equation is automatically built by the symbolic variable elimination procedure. As one can see, in this particular example the marginal distributions $p(A)$ and $p(B)$ can be estimated by the standard closed-form solution, as they are roots of the network and (in this example) their corresponding data are always complete. However, this is not true for every term in the equation. For instance, the summation $\sum_{i \in N^M} \log \theta_{t^i|a^i,b^i}$, where the sum runs over the categories a^i, b^i, t^i, comes in Eq.(5) and involves elements that are not direct part of the network specification. It is exactly the job of the symbolic variable elimination to obtain the extra constraints:

$$\theta_{t^i|a^i,e} = \sum_u \theta_{u|a^i} \cdot \theta_{t^i|u,e}\,, \qquad \theta_{t^i|a^i,b^i} = \sum_e \theta_{e|b^i} \cdot \theta_{t^i|a^i,e}. \tag{6}$$

These equations tie together the *auxiliary* optimization unknowns (such as $\theta_{t^i|a^i,e}$ and $\theta_{t^i|a^i,b^i}$) and the actual parameter estimates of interest, which are part of the specification of the network (such as $\theta_{u|a^i}$, $\theta_{t^i|u,e}$, $\theta_{e|b^i}$). We emphasize that these derivations are not done by hand (with the user interaction), but instead they are *automatically* processed by the symbolic variable elimination procedure. The left-hand side of Equation (6) comes from the symbolic (variable) elimination of u, while the right-hand side comes from the symbolic elimination of e. Together, they create a mathematical correspondence between $\theta_{t^i|a^i,b^i}$ and actual network parameters. After the symbolic preprocessing, it is up to the polynomial programming solver to optimize the non-linear problem. We have implemented an adapted version of the reformulation-linearization technique [20], which is a global solver for it.

To make a parallel, the EM algorithm would have to compute $p(E,U|a^i,b^i,t^i)$ (with $\mathscr{P} = \hat{\theta}^k$) for each instance i in the data set, in order to obtain the sufficient statistics of iteration k. Each such computation is in fact a procedure of similar complexity to the one we just did. The main difference between the methods is that we do not work with numbers but with a symbolic version of the computation.

2.3 Global solver vs EM-MAP

We define as EM-MAP the approach which performs multiple runs of EM using different initialization points, eventually selecting the estimate corresponding to the highest MAP score. As explained in Section 2.2, we implemented a global solver

based on non-linear programming. In our experiments, the global solver achieved slightly higher MAP scores than EM-MAP, yielding however worse parameter estimates than EM-MAP. This phenomenon can be seen from Figure 2, where points above the diagonal indicate better estimate for EM-MAP (using the *joint metric* as criterion) and points below the diagonal indicate better estimate for the global solver. Similar results were found in many experiments (not shown) comparing the global solver vs. EM-MAP, which suggests that selecting the estimate with the highest MAP score has drawbacks, being for instance subject to overfitting.

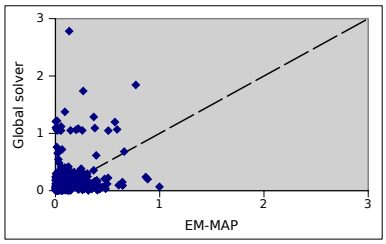

Fig. 2 Scatter plot of KL-divergences in the joint metric over data sets produced by BN_1(detailed in the experiments section). Points *above* the diagonal show a worse estimate for the global solver, compared to EM-MAP.

2.4 Discriminating high-score estimates by entropy

In order to overcome the drawback just described, we propose the following criterion: to pick the parameter estimate with *maximum entropy, among those which have a high MAP score*. To identify the estimates with high MAP score we adopt a criterion similar to the Bayes factor. When discriminating among two competing models m_1 and m_2 on the basis of the data $\mathbf{y}$, the evidence in favor of m_1 can be considered substantial only if the Bayes factor $P(\mathbf{y}|m_1)/P(\mathbf{y}|m_2)$ is at least some threshold, for instance 2 or 3 [9], where $P(\mathbf{y}|m)$ represents the *marginal* likelihood given m: $P(\mathbf{y}|m) = \int P(\mathbf{y}|m,\theta)p(\theta)d\theta$. Because of the challenges that come with the missing data, we adopt a ratio of MAP scores (a full Bayesian approach would integrate over the parameters, but such computation would be intractable). We assume that if the ratio of the MAP scores among two competing parameter estimates is less than 2, there is no substantial evidence for preferring one over the other. To choose among the competing estimates with high MAP score (whose MAP score is at least a half of the maximum MAP score known for the data set under consideration), we use the maximum entropy, thus choosing the least informative estimate given the available information [12]. This approach differs from standard maximum entropy inferences previously reported [11, 22] since we first check for high score estimates, and then maximize entropy among them. It can be formally written as:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \sum_{j=1}^{n} \sum_{\pi_j} \sum_{x_j} \theta_{x_j|\pi_j} \log \theta_{x_j|\pi_j} \quad \text{subject to} \quad S_\theta + c \geq s^* , \tag{7}$$

where the function to be maximized is the *local entropy* of a Bayesian network [16], $s^* = \max_\theta S_\theta$ is the highest MAP score for the problem (which we compute before running this optimization), and c is the logarithm of the ratio of the MAP scores.

The optimization of Eq. (7) maximizes entropy, being however constrained to ensure that the MAP score S_θ is high. In fact, we found our result to be robust on the threshold c when letting it vary between 2 and 3.

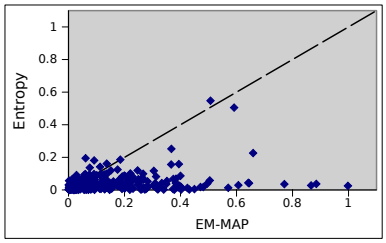

Fig. 3 Scatter plot of KL-divergences in the joint metric over data sets produced by BN_1(detailed in the experiments section). Points *below* the diagonal show a better estimate for the global solver coupled with the entropy criterion compared to EM-MAP.

If it is to use the previously mentioned global solver, the optimization is tackled in two steps: first by globally optimizing the MAP score as already described, and then by solving Eqs. (7). The quality of the estimates obtained in this case dramatically improves, as can be seen from Figure 3. As the problem is NP-hard, we cannot expect this solver to obtain a global optimal solution in all problem instances. Because of that, we adapted our idea to work also within EM. In this case, we select, among the various estimates generated by the multi-start EM, the maximum entropy estimate among those which have a high MAP score. The high MAP score is checked by computing the ratio of the MAP score with respect to the highest MAP score obtained in the different EM runs (thus, the computation of the highest score is only done in an approximate fashion). We call the resulting approach *EM-entropy*. This differs from the maximum entropy approach described before, because we focus only on the many estimates generated by the EM runs. The great benefit is that the implementation becomes straightforward: only a few changes on top of an already running EM suffice. The drawback of EM-entropy is that the true maximum entropy estimate might well be a non-optimum estimate in terms of score. Because of that, even if we increase the number of EM runs, the empirical entropy is still confined to saddle points of the score function, and the resulting estimate may differ. Nevertheless, the experiments will later show that the EM-entropy also produces significantly better estimates than MAP. An insight of the reason for which EM-entropy outperforms EM-MAP is given by Figure 4, which shows an experiment where higher MAP scores do not necessarily imply a better estimate; instead, when comparing estimates that already have high MAP score (the right-most points

in Figure 4), entropy is more discriminative than the MAP score itself and has also a stronger correlation with the Kullback-Leibler (KL) divergence.

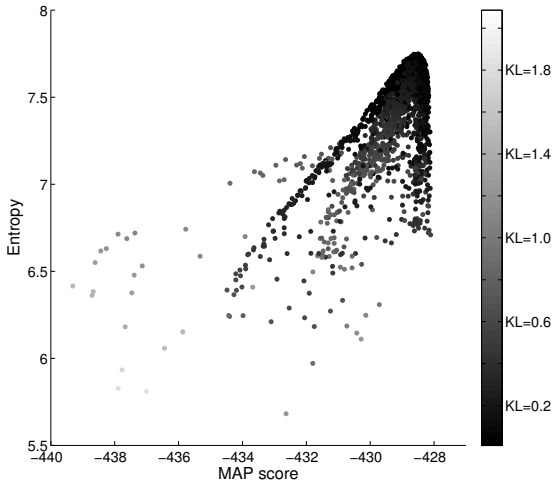

Fig. 4 Relation between KL divergence, entropy and score; darker points represent lower KL divergence between true and estimated joint distributions. The figure refers to one thousand EM runs performed on an incomplete training set of 200 samples.

In any BN with more than a couple of variables, the number of parameters to estimate becomes quickly large and there is only a very small (or no) region of the parameter space with estimates that achieve the very same global maximum value. However, a feasibility region defined by a small percentage away from the maximum score is enough to produce a whole region of estimates, indicating that the region of high score estimates is almost (but not exactly) flat. This is expected in a high-dimensional parameter space of BNs.

3 Experiments

We perform a empirical study using different BN graphs, sample sizes N and missingness process mp. The experiments were run using the open-source Bayesian Network Toolbox (BNT) [18] for Matlab.

A triple $\langle \text{BN graph}, N, mp \rangle$ identifies a *setting*; for each setting, we perform 300 *experiments*, each defined as follows: a) instantiation of the *reference BN*; b) sampling of N complete instances from the reference BN; c) application of the missingness process; d) execution of EM from 30 different initializations; e) execution of

our solvers and estimation procedure using the different methods. We evaluate the quality of the estimates through the *joint metric* and the *reasoning metric* already introduced. We analyze the significance of the differences through the non-parametric Friedman test with significance level of 1%. By a post-hoc procedure applied on the statistic of the test, we generate a rank of methods for each setting and each metric.

The first set of experiments regards the BN graph of Figure 1 (named BN_1), which has been used in previous sections to illustrate the methods. Variables A (binary) and B (ternary) have uniform distributions and are always observed; variables U, E and T are binary (assuming states true and false); the value of T is defined by the logical relation $T = E \wedge U$. Variable T is always observed, while U and E are affected by the missingness process: in particular, both U and E are observed if and only if T is true. Therefore, E and U are either both observed and positive, or non-observed. The missingness process is MAR [13, Sec. 19.1.2] because given T (always observed) the probability of U and E to be missing does not depend on their actual values; E and U are missing in about 85% of the sampled instances. We assume the conditional probabilities of T to be known, thus focusing on the difficulty of estimating the probabilities related to variables U and E. For both $\langle \mathrm{BN}_1, 100, \mathrm{MAR} \rangle$ and $\langle \mathrm{BN}_1, 200, \mathrm{MAR} \rangle$ and for both the joint and the reasoning metric, the Friedman test returned the following rank: 1st) entropy; 2nd) EM-entropy; 3rd) EM-MAP. The boxplots in the first row of Figure 3 show that the entropy-based methods largely improve over EM-MAP; interestingly, the simple EM-entropy already delivers much of the gain achieved by the more sophisticated entropy method which relies on globally optimal solvers.

In a second set of experiments we use the graph $A \rightarrow B \rightarrow C$, which we call BN_2. We consider two different configurations of number of states for each node: 5-3-5 (meaning A,C with 5 states and B with 3) and 8-4-8 (A,C with 8 states and B with 4). In both cases, we make B randomly missing in 85% of the instances. Each experiment now includes an additional step, namely the generation of random parameters of the reference BN. From the viewpoint of how realistic is this experiment, one may see BN_2 as a subnetwork (possibly repeated many times) within a much larger BN. For instance, if we see A as the joint parent set of B, and C as the joint children of B, this experiment regards the very same challenges of estimating a node's parameters (in this case B) with missing values in a BN of irrespective number of variables. This graph also captures the BN that could be used for clustering with EM [6]. Despite the simple graph of this BN, the estimation task requires to estimate from incomplete samples a non-negligible number of parameters, referring to nodes B and C: respectively $2 \cdot 5 + 4 \cdot 3 = 22$ and $8 \cdot 3 + 7 \cdot 4 = 52$, for each used configuration. To these numbers, one should add the marginals of A, which are however inferred from complete samples and whose estimate is thus identical for all methods. We adopted N=300 for the 5-3-5 configuration and N=500 for the 8-4-8 configuration. In both settings and the two metrics, we obtained the same rank: 1st) entropy; 2nd) EM-entropy; 3rd) EM-MAP. It is worth noting again that the simple EM-entropy improves over EM-MAP. The boxplots are shown in the second row of Figure 3.

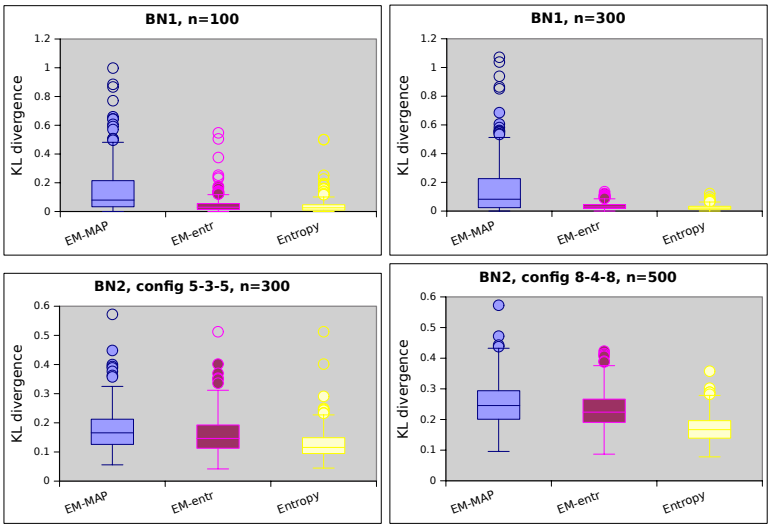

Fig. 5 Boxplot of KL-divergences for the joint metric over 300 runs of the experiment with BN_1 and BN_2(the scale changes between top and bottom graphs.)

		n=100		*n*=200	
Net	q=	30%	60%	30%	60%
Asia	*joint*	0.96	0.90	0.96	0.91
Asia	*reasoning*	0.92	0.86	0.99	0.89
Alarm	*joint*	0.93	0.88	0.93	0.89
Alarm	*reasoning*	0.95	0.94	0.97	0.96
Random20	*joint*	0.94	0.89	0.92	0.89
Random20	*reasoning*	0.92	0.88	0.97	0.92

Table 1 Relative medians of KL divergence, i.e., medians of entropy are presented (experiment-wise) divided by the median of MAP. Smaller numbers indicate better performance; in particular, values smaller than 1 indicate a smaller median than MAP.

To further compare the behavior of EM-entropy and EM-MAP, we run experiments using well-known BNs: i) the Asia network (8 binary variables, 2 leaves) [15], ii) the Alarm network (37 variables with 2 to 4 states each, and 8 leaves) [1] and iii) BNs with randomly generated graphs with 20 variables. In each experiment, we randomly re-generated the parameters of the reference networks. In the case of randomly generated BN graphs, the experimental procedure also includes the generation of the random graph, which is accomplished before drawing the parameters. Given two variables X_i and X_j, an arc from X_i to X_j is randomly included with probability 1/3 if $i < j$ (no arc is included if $j \geq i$, which ensures that the graph is acyclic and has no loops). Furthermore, the maximum number of parents of each variable is set to 4 and the number of states per variable is randomly selected from 2 to 4. After

that, the experiments follow the same workflow as before. We consider a MCAR[1] process, which makes each single value missing with probability q; we use q equals to 30% and 60%; we moreover consider sample sizes N of 100 and 200.

In all these experiments, EM-entropy performs significantly better than EM-MAP, with respect to both the joint and the reasoning metric. The quantitative difference of performance can be seen in Table 1, which reports the *relative medians* of metrics, namely the medians of EM-entropy in a certain task, divided by the median of EM-MAP in the same task. The improvement of the median over EM-MAP ranges from 1% to 14%; most importantly, it is consistent, occurring in all settings. As a final remark, the difference in performance increases when the estimation task is more challenging, typically when the percentage of missing data increases.

4 Conclusions

The most common approach to estimate the parameters of a Bayesian network in presence of incomplete data is to search for estimates with maximum posterior probability (MAP). MAP estimation is no harder than maximum likelihood estimation, over which it should be preferred because it yields estimates that are more resilient to overfitting. MAP estimation is much faster than full Bayesian estimation, but does not offer the same advantages of the latter. Many local maxima are usually present and several of them present high posterior probability. Selecting the one which maximizes it is not robust, since the difference among these competing estimates is generally very thin.

We presented an approach to select the least informative estimate, namely the maximum entropy one, among those which have a high posterior probability; our empirical analyses indicate that this approach consistently improves the quality of results. The approach has been implemented with a global solver developed by us and within EM, obtaining in both cases a significant improvement when compared to MAP. In particular, the EM-entropy method for inferring Bayesian networks can be promptly implemented on top of any existing EM implementation for that task. As a future work, we plan to apply these ideas in more general settings of parameter estimation problems from incomplete samples, not only restricted to Bayesian networks.

Acknowledgements The research in this paper has been partially supported by the Swiss NSF grant no. 200021_146606/1.

[1] MCAR (or *missing completely at random*) indicates that the probability of each value being missing does not depend on the value itself, neither on the value of other variables.

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 2nd European Conference on Artificial Intelligence in Medicine*, volume 38, pages 247–256, 1989.
2. R. G. Cowell. Parameter learning from incomplete data for Bayesian networks. In *Proc. of the 7th International Workshop on Artificial Intelligence and Statistics*. Morgan Kaufmann, 1999.
3. C. P. de Campos and F. G. Cozman. Inference in credal networks using multilinear programming. In *Proc. of the 2nd Starting AI Researcher Symposium*, pages 50–61, Valencia, 2004. IOS Press.
4. C. P. de Campos and Q. Ji. Improving Bayesian network parameter learning using constraints. In *Proc. of the 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.
5. C. P. de Campos and Q. Ji. Bayesian networks and the imprecise Dirichlet model applied to recognition problems. In Weiru Liu, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 6717 of *Lecture Notes in Computer Science*, pages 158–169. Springer Berlin / Heidelberg, 2011.
6. C. P. de Campos, P. M. V. Rancoita, I. Kwee, E. Zucca, M. Zaffalon, and F. Bertoni. Discovering subgroups of patients from DNA copy number data using NMF on compacted matrices. *PLoS ONE*, 8(11):e79720, 2013.
7. C. P. de Campos, L. Zhang, Y. Tong, and Q. Ji. Semi-qualitative probabilistic networks in computer vision problems. *Journal of Statistical Theory and Practice*, 3(1):197–210, 2009.
8. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.
9. I. J. Good. Studies in the History of Probability and Statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika*, 66:393–396, 1979.
10. D. Heckerman. A tutorial on learning with Bayesian networks. *Learning in Graphical Models*, 89:301–354, 1998.
11. B. Huang and A. Salleb-Aouissi. Maximum entropy density estimation with incomplete presence-only data. In *Proc. of the Twelfth International Conference on Artificial Intelligence and Statistics: JMLR W&CP 5*, pages 240–247, 2009.
12. E. T. Jaynes. On the rationale of maximum-entropy methods. *Proc. of the IEEE*, 70(9):939–952, 1982.
13. D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT press, 2009.
14. S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19(2):191–201, 1995.
15. S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B*, 50(2):157–224, 1988.
16. T. Lukasiewicz. Credal Networks under Maximum Entropy. In *Proc. of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 363–370. Morgan Kaufmann Publishers Inc., 2000.
17. G. M. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.
18. K. P. Murphy. The Bayes Net Toolbox for MATLAB. In *Computing Science and Statistics*, volume 33, 2001.
19. M. Ramoni and P. Sebastiani. Robust learning with missing data. *Machine Learning*, 45(2):147–170, 2001.
20. H. D. Sherali and C. H. Tuncbilek. A global optimization algorithm for polynomial programming problems using a reformulation-linearization technique. *Journal of Global Optimization*, 2:101–112, 1992.
21. P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.
22. S. Wang, D. Schuurmans, F. Peng, and Y. Zhao. Combining statistical language models via the latent maximum entropy principle. *Machine Learning*, 60(1-3):229–250, 2005.