Efficient Bayesian Network Learning Using EM or Pairwise Deletion

Olivier C. H. François

University of Reading, Sustainable Urban Environments Research Division,
URS building, 3n38, Whiteknights, PO Box 219, RG6 6AW, Reading, United Kingdom.
http://ofrancois.tuxfamily.org

Abstract

In previous work, we have seen how to learn a TAN classifier from incomplete dataset using the Expectation Maximisation algorithm (François and Leray, 2006). In this paper, we study differences for Bayesian network structure learning between estimating probabilities using the EM algorithm or using Pairwise Deletion. We have implemented these two estimation techniques with greedy search learning methods in several spaces: Trees, Directed Acyclic Graphs, Completed Partially Directed Acyclic Graphs or Tree Augmented Naive Bayes structures. An experimental study shows strengths and weaknesses of using the EM algorithm or Pairwise Deletion on classification tasks.

1 Introduction

Bayesian networks introduced by (Kim and Pearl, 1987) are a formalism of probabilistic reasoning used increasingly in decision aid, diagnosis and complex systems control (Jensen, 1996; Pearl, 1998; Naïm et al., 2004).

Let $\mathbb{X} = \{X_1, \ldots, X_n\}$ be a set of discrete random variables. A *Discrete Bayesian network* $\mathcal{B} =< \mathcal{G}, \Theta >$ is defined by a directed acyclic graph (DAG) $\mathcal{G} =< \mathbb{N}, \mathbb{U} >$ where $\mathbb{N}$ represents the set of nodes (one node for each variable) and $\mathbb{U}$ the set of edges,AND parameters $\Theta = \{\theta_{ijk}\}_{1\leqslant i\leqslant n, 1\leqslant j\leqslant q_i, 1\leqslant k\leqslant r_i}$the set of conditional probability tables of each node X_i knowing its parents' state $Pa(X_i)$ (with r_i and q_i as respective cardinalities of X_i and $Pa(X_i)$).

Determination of Θ (when $\mathcal{G}$ is given) is often based on expert knowledge, but several learning methods based on data have appeared. However, most of these methods only deal with complete data cases.

We will therefore first recall the issues relating to structural learning and review the various ways of dealing with incomplete data for structure determination.

Because of the super-exponential size of the search space (Gillispie and Lemieux, 2001), exhaustive search for the best Bayesian network structure is impossible. Many heuristic methods have been proposed to determine the structure of a Bayesian network. Some of them rely on human expert knowledge, others use real data which are -most of the time- completely observed.

We are here more specifically interested in score-based methods, primarily MWST proposed by (Chow and Liu, 1968) and applied to Bayesian networks in (Heckerman et al., 1995), then GS algorithm (Chickering et al., 1995), and finally GES algorithm proposed by (Chickering and Meek, 2002). GS is a greedy search carried out in DAG spaces where the interest of each structure located near the current structure is assessed by means of a Bayesian score like BDe (Heckerman et al., 1994) or a BIC/MDL type measurement (equation 1). As (Friedman, 1997), we consider that the BIC/MDL score is a function of the graph $\mathcal{G}$ and the parameters Θ, generalising the classical definition of the BIC score which is defined with our notation by $BIC(\mathcal{G}, \Theta^*)$ where Θ^* is obtained by maximising the likelihood or $BIC(\mathcal{G}, \Theta)$ score for a given $\mathcal{G}$ which is given by

$$BIC(\mathcal{G}, \Theta) = \log P(\mathcal{D}|\mathcal{G}, \Theta) - \frac{\log N}{2}\text{Dim}(\mathcal{G}) \quad (1)$$

where $\mathrm{Dim}(\mathcal{G})$ is the number of parameters used for the Bayesian network representation and N is the size of the dataset $\mathcal{D}$.

The BIC score is decomposable. It can be written as the sum of local score computed for each node of the graph:

$$BIC(\mathcal{G},\Theta) = \sum_i bic(X_i, P_i, \Theta_{Xi|P_i}) \tag{2}$$

where $bic(X_i, P_i, \Theta_{Xi|P_i}) =$

$$\sum_{X_i=x_k} \sum_{P_i=pa_j} N_{ijk} \log \theta_{ijk} - \frac{\log N}{2} \mathrm{Dim}(\Theta_{Xi|P_i})$$

with N_{ijk} the occurrence number of $\{X_i = x_k$ and $P_i = pa_j\}$ in $\mathcal{D}$.

An improvement of the greedy search in DAG space over CPDAG space have also been proposed by (Chickering, 2002b). The MWST principle is rather different. This algorithm determines the best tree that links all the variables, using a mutual information measurement (Chow and Liu, 1968) or the BIC score variation when two variables become linked (Heckerman et al., 1994).

The aim is to compare improvement of learning algorithms in these different spaces when the dataset is incomplete and when using two different methods for estimating probability: pairwise deletion (ACA, for available cases analysis) and the Expectation-Maximisation algorithm.

2 Dealing with incomplete data

2.1 Nature of missing data.

Let $\mathcal{D} = \{X_i^l\}_{1\leqslant i\leqslant n, 1\leqslant l\leqslant N}$ our dataset, with $\mathcal{D}_o$ the observed part of $\mathcal{D}$, $\mathcal{D}_m$ the missing part and $\mathcal{D}_{co}$ the set of completely observed cases in $\mathcal{D}_o$. Let also $\mathcal{M} = \{M_{il}\}$ with $M_{il} = 1$ if X_i^l is missing, 0 if it is not.

$$\mathcal{D}_m = \{X_i^l / M_{il} = 1\}_{\,1\leqslant i\leqslant n, 1\leqslant l\leqslant N}$$
$$\mathcal{D}_o = \{X_i^l / M_{il} = 0\}_{\,1\leqslant i\leqslant n, 1\leqslant l\leqslant N}$$
$$\mathcal{D}_{co} = \{[X_1^l \dots X_n^l]/[M_{1l} \dots M_{nl}] = [0 \dots 0]\}_{\,1\leqslant l\leqslant N}$$

Dealing with missing data depends on their nature. (Rubin, 1976) identified several types of missing data:

- MCAR (Missing Completely At Random): $P(\mathcal{M}|\mathcal{D}) = P(\mathcal{M})$, the probability for data to be missing does not depend on $\mathcal{D}$,
- MAR (Missing At Random): $P(\mathcal{M}|\mathcal{D}) = P(\mathcal{M}|\mathcal{D}_o)$, the probability to be missing depends on observed data,
- NMAR (Not Missing At Random): the probability for data to be missing depends on both observed and missing data.

MCAR and MAR cases are the easiest to solve as observed data include all necessary information to estimate missing data distribution. The case of NMAR is trickier as outside information has to be used to model missing data distribution. Many methods try to rely more on all the observed data. Among them are *sequential updating* (Spiegelhalter and Lauritzen, 1990), *Gibbs sampling* (Geman and Geman, 1984), and the EM algorithm. More recently, *bound and collapse* (Ramoni and Sebastiani, 1998) and *robust Bayesian estimator* (Ramoni and Sebastiani, 2000) try to resolve this task whatever the nature of missing data.

EM has been first proposed by (Dempster et al., 1977) and adapted by (Lauritzen, 1995) to the learning of the parameters of a Bayesian network whose structure is known. Let $\log P(\mathcal{D}|\Theta) = \log P(\mathcal{D}_o, \mathcal{D}_m|\Theta)$ be the data log-likelihood. $\mathcal{D}_m$ being an unmeasured random variable, this log-likelihood is also a random variable function of $\mathcal{D}_m$. By establishing a reference model Θ^*, it is possible to estimate the probability density of the missing data $P(\mathcal{D}_m|\Theta^*)$ and therefore to calculate $Q(\Theta : \Theta^*)$ expectation of the previous log-likelihood:

$$Q(\Theta : \Theta^*) = E_{\Theta^*}\left[\log P(\mathcal{D}_o, \mathcal{D}_m|\Theta)\right] \tag{3}$$

So $Q(\Theta : \Theta^*)$ is the likelihood expectation of any set of parameters Θ calculated using a distribution of the missing data $P(\mathcal{D}_m|\Theta^*)$. Equation 3 can be re-written as follows:

$$Q(\Theta : \Theta^*) = \sum_{i=1}^{n} \sum_{X_i=x_k} \sum_{P_i=pa_j} N_{ijk}^* \log \theta_{ijk} \tag{4}$$

where $N_{ijk}^* = E_{\Theta^*}[N_{ijk}] = N \times P(X_i{=}x_k, P_i{=}pa_j|\Theta^*)$ is obtained by inference in the network $< \mathcal{G}, \Theta^* >$ if the $\{X_i, P_i\}$ are not completely measured, or else only by mere counting.

We are also interested in pairwise deletion which is a method that uses all available data. Cases are removed when they have missing data on the variables involved in that particular computation. This method is very efficient computationally but it assumes that the data are missing completely at random (MCAR). If not, it introduces a bias.

We are interested in studying the results quality that could be expected using these two methods named EM and ACA.

2.2 Determining structure Θ when data are incomplete.

The main methods for structural learning with incomplete data use the EM principle : *Alternative Model Selection EM* (AMS-EM) (Friedman, 1997) and *Bayesian Structural EM* (BS-EM) (Friedman, 1998). We can also cite the *Hybrid Independence Test* proposed in (Dash and Druzdzel, 2003) that can use EM to estimate the essential sufficient statistics that are then used for an independence test in a constraint-based method. (Myers et al., 1999) proposes a structural learning method based on genetic algorithm and MCMC.

2.3 General principle and scoring metric

In practice, to perform a maximisation in the joint space $\{\mathcal{G},\Theta\}$, we must distinguish these two steps[1] :

$$\mathcal{G}^i = \operatorname*{argmax}_{\mathcal{G}} \; Q(\mathcal{G},\bullet:\mathcal{G}^{i-1},\Theta^{i-1}) \qquad (5)$$

$$\Theta^i = \operatorname*{argmax}_{\Theta} \; Q(\mathcal{G}^i,\Theta:\mathcal{G}^{i-1},\Theta^{i-1}) \qquad (6)$$

where $Q(\mathcal{G},\Theta:\mathcal{G}^*,\Theta^*)$ is the expectation of the likelihood of any Bayesian network $<\mathcal{G},\Theta>$ calculated using a distribution of the missing data $P(\mathcal{D}_m|\mathcal{G}^*,\Theta^*)$.

Note that the first search (equation 5) in the space of possible graphs takes us back to the initial problem, i.e. the search for the best structure in a super-exponential space. However, with *Generalised EM* it is possible to look for a better solution to function Q, rather than the best possible one, without affecting the algorithm convergence properties. This search for a better solution can then be done in a limited space, like for example $\mathcal{V}_{\mathcal{G}}$, the set of the neighbours of graph $\mathcal{G}$ that have been generated by removal, addition or inversion of an arc interpreted either in the DAG space or in the CPDAG space.

We now have to choose the function Q that will be used for structural learning. The likelihood used for parameter learning is not a good indicator to determine the best graph since it gives more importance to strongly connected structures. Moreover, it is impossible to calculate marginal likelihood when data are incomplete, so that it is necessary to rely on an efficient approximation like those reviewed by (Chickering and Heckerman, 1996). In complete data cases, the most frequently used measurements are the BIC/MDL score and the Bayesian BDe score. These two scoring metrics are locally consistent but the BIC/MDL score includes a penalty term and we have chosen it for this study.

$$Q^{BIC}(\mathcal{G},\Theta:\mathcal{G}^*,\Theta^*) = E_{\mathcal{G}^*,\Theta^*}\left[\log P(\mathcal{D}_o,\mathcal{D}_m|\mathcal{G},\Theta)\right] - \tfrac{1}{2}\mathrm{Dim}(\mathcal{G})\log N \qquad (7)$$

As the BIC score is decomposable, so is Q^{BIC}:

$$Q^{BIC}(\mathcal{G},\Theta:\mathcal{G}^*,\Theta^*) = \sum_i Q^{bic}(X_i,P_i,\Theta_{Xi|P_i}:\mathcal{G}^*,\Theta^*) \qquad (8)$$

where $Q^{bic}(X_i,P_i,\Theta_{Xi|P_i}:\mathcal{G}^*,\Theta^*) =$

$$\sum_{X_i=x_k}\sum_{P_i=pa_j} N^*_{ijk}\log\theta_{ijk} - \frac{\log N}{2}\mathrm{Dim}(\Theta_{Xi|P_i}) \qquad (9)$$

3 Tested structural learning algorithms

3.1 Greedy Search

3.1.1 SEM

Friedman has proposed two versions of his Bayesian network greedy search algorithm based on suggestions of (Heckerman et al., 1995) but adapted for incomplete datasets : AMS-EM (Friedman, 1997) and BS-EM (Friedman, 1998). We have chosen to use the method AMS-EM that we simply recall SEM as many people do as it could be used with the BIC criterion as explain above.

[1] the notation $Q(\mathcal{G},\bullet:\dots)$ used in equation 5 stands for $E_\Theta[Q(\mathcal{G},\Theta:\dots)]$ for Bayesian scores or $Q(\mathcal{G},\Theta^o:\dots)$ where Θ^o is obtained by likelihood maximisation

3.1.2 GS-ACA

A new implementation a the greedy search inspired by (Cooper and Hersovits, 1992; Heckerman et al., 1994) using pairwise deletion to deal with incomplete dataset is also tested.

3.2 Greedy Equivalent Search

3.2.1 GES-EM

Recent work by (Chickering, 2002a; Castelo and Kocka, 2002; Auvray and Wehenkel, 2002) show that we could take advantage of using the CPDAGs space. Such a space has less equal scoring models than the DAGs space, as many DAGs have the same representation in a unique CPDAG.

A search algorithm in the Markov equivalent space named GES for *Greedy Equivalent Search* has been proposed by (Meek, 1997). It consists in two iterative steps. First one builds iteratively a graph by adding dependence links to the current essential graph (*i.e.* CPDAG) while the second step consists in removing iteratively arcs that are no more needed in the model. The optimality of this method (*conjecture of Meek*) has been proved by (Kočka et al., 2001; Chickering, 2002b).

The GES-EM algorithm keep the principles of the SEM method using the Markov equivalent space to build neighbourhood of the current graph at each steps.

3.2.2 GES-ACA

A version of this method using the available cases analysis (*i.e.* pairwise deletion) is also implemented using the BIC criterion.

3.3 Maximum Weight Spanning Tree

3.3.1 General principle of MWST-EM

(François and Leray, 2004) have shown that, in complete data cases, the MWST algorithm was able to find a simple structure very rapidly (the best tree connecting all the nodes in the space), which could be used as judicious initialisation by the GS algorithm. (Heckerman et al., 1995) suggests using the variation of any decomposable score instead of the mutual information originally used in MWST. Using this remark, we could therefore implement the MWST algorithm using the EM algorithm to manage incomplete datasets.

MWST-EM deals with the choice of the initial structure. The choice of an oriented chain graph linking all the variables proposed by (Friedman, 1997) seems judicious here, since this chain graph also belongs to the tree space. The MWST algorithm used a similarity function between two nodes which is based on the BIC score variation whether X_j is linked to X_i or not. This function can be summed up in the following symmetrical matrix :

$$\begin{aligned}\left[M_{ij}^{Q}\right] = \; & \Big[Q^{bic}(X_i, P_i = X_j, \Theta_{Xi|X_j} : \mathcal{T}^*, \Theta^*) \\ & -Q^{bic}(X_i, P_i = \emptyset, \Theta_{Xi} : \mathcal{T}^*, \Theta^*)\Big] \end{aligned} \quad (10)$$

Running maximum (weight) spanning algorithms like Kruskal's or Prim's on matrix M enables us to obtain the best tree that maximises the sum of the local scores on all the nodes, i.e. function Q^{BIC} of equation 8.

This method looks for the best tree-DAG among the neighbours of the current graph. With MWST-EM, we can directly get the best tree that maximises function Q at each step and then this method converge in few steps.

3.3.2 Trees as an initialisation of DAGs greedy search : SEM+T and GS+T-ACA

MWST-EM will serve as initialisation of the SEM algorithms proposed by Friedman. This variant of the structural EM algorithm will be called SEM+T (François, 2006). MWST-ACA will serve as initialisation of the GS-ACA, the resulting method is called GS+T-ACA.

3.3.3 MWST-ACA

We could adapt this algorithm using pairwise deletion. For evaluating the probability $p_{ijk} = \mathbb{P}(X_i = x_{i,k}\, and\, Pa(X_i) = pa_{i,j})$, we no longer need an iterative method as the EM algorithm. To deduce $N_{ijk} = p_{ijk} \times N$, we need to evaluate p_{ijk} on a new sub-dataset that contain only the complete cases of the variables $X_i \cup Pa(X_i)$.

This method is the only direct method to learn a Bayesian network from incomplete dataset in our knowledge.

3.3.4 Extension to classification problems : TAN-EM and TAN-ACA

For classification tasks (where data are incomplete), many studies like those of (Keogh and Pazzani, 1999; Leray and François, 2004b) use a structure based on an augmented naive Bayesian network, where observations (i.e. all the variables except class) are linked to the very best tree (*TAN, Tree Augmented Naive Bayes*). (Geiger, 1992) showed it was the tree obtained by running MWST on the observations. It is therefore possible to extend this specific structure to classification problems when data are incomplete by running a specific version MWST-EM where the class node is considered as a parent of each other nodes. This algorithm will be called TAN-EM.

A version of this method using pairwise deletion named TAN-ACA is also tested.

3.4 Experimental tests

3.4.1 Datasets and evaluation techniques

The experiment stage aims at evaluating all these structure learning methods on incomplete datasets: `Hepatitis`, `Horse`, `House`, `Mushrooms` and `Thyroid` (Blake and Merz, 1998).

We indicate classification rates obtained by the best run on three of the different methods as well as the likelihood and the learning time of the best model on these 3 runs. We also give an 95%-confidence interval based on equation 11 for each classification rate based on (Bennani and Bossaert, 1996).

$$I(\alpha, N) = \frac{T + \frac{Z_\alpha^2}{2N} \pm Z_\alpha \sqrt{\frac{T(1-T)}{N} + \frac{Z_\alpha^2}{4N^2}}}{1 + \frac{Z_\alpha^2}{N}} \qquad (11)$$

where N is the sample size, T is the classifier good classification percentage and $Z_\alpha = 1.96$ for $\alpha = 95\%$.

All implementation were done with the *Structure Learning Package* (Leray and François, 2004a) for the *Bayes Net Toolbox* (Murphy, 2001).

3.4.2 Results and interpretations

The results are summed up in table 1. First, we could see that even if the Naive Bayes classifier often gives good results, the other tested methods allow to obtain better classification rates. Whilst all runs of Naive Bayes classifier and ACA methods give same results, EM methods do not always give same results because of the first parameters estimation random initialisation. We have also noticed (not reported here) that TAN methods seem the stabler methods concerning the evaluated classification rate while MWST methods seem to be the less ones.

The method GS-EM could obtain very good structures. Then, initialising it with the results of MWST-EM gives stabler results (see (Leray and François, 2005) for a more specific study of this point).

In our tests, except for `Hepatitis` dataset that have only 90 learning samples, TAN methods always obtain structures that lead to better classification rates in comparison with the other structure learning methods.

Remark that MWST methods could occasionally give good classification rates even if the class node is connected to a maximum of two other attributes. In that case, it could be a good hint of most relevant attributes to the class node.

Regarding the log-likelihood reported in table 1, we see that GS-ACA give best results while TAN methods finds structures that can also lead to a good approximation of the underlying probability distribution of the data, even with a strong constraint on the graph structure.

In these experiments, we could confirm that ACA methods could outperform EM methods on classification for GS and GES learning methods but not systematically. Results are similar for MWST and TAN methods for classification but ACA leads to better log-likelihoods. Classification rates are different but ACA methods could beat EM methods as often as EM methods could beat ACA methods for these two algorithms.

Finally, the table 1 illustrates that TAN and MWST methods have about the same complexity (regarding the computational time) and are a good compromise between Naive Bayes classifiers and Greedy Searches either in DAGs or CPDAG spaces.

Method	HEPATITIS	HORSE	HOUSE	MUSHROOMS	THYROID
sizes	20; 90;65; 8%	28; 300;300; 88%	17; 290;145; 46%	23; 5416;2708; 31%	22; 2800;972; 30%
NB	**73.8%** [62.0;83.0]	73.5% [62.0;82.6]	89.7% [83.6;93.6]	94.4% [93.5;95.2]	**96.0%** [94.6;97.1]
	-1122 (0s)	-1540 (0s)	-1404 (0s)	-41147 (0s)	-15728 (0s)
MWST-ACA	58.5% [46.3;69.6]	**82.4%** [71.6;89.6]	90.3% [84.4;94.2]	75.0% [73.3;76.6]	77.4% [74.6;79.9]
	-847 (2s)	-1240 (**16s**)	-1282 (5s)	-31447 (178s)	**-15359** (96s)
MWST-EM	**75.4%** [63.7;84.2]	**82.4%** [71.6;89.6]	82.1% [75.0;87.5]	60.3% [58.5;62.2]	93.8% [92.1;95.2]
	-1114 (**45s**)	-1306 (299s)	-1462 (67s)	-39773 (1389s)	-16912 (2254s)
TAN-ACA	64.6% [52.5;75.1]	73.5% [62.0;82.6]	**93.1%** [87.8;96.2]	**98.4%** [97.8;98.8]	**95.9%** [94.4;97.0]
	-1123 (2s)	-1319 (15s)	-1284 (**4s**)	**-20453** (**183s**)	-15894 (**86s**)
TAN-EM	64.6% [52.5;75.1]	77.9% [66.7;86.2]	91.7% [86.1;95.2]	**98.4%** [97.8;98.8]	**97.0%** [95.7;97.9]
	-1186 (71s)	-1546 (307s)	-1339 (185s)	-33885 (2345s)	-16292 (1936s)
GS-ACA	67.7% [55.6;77.8]	**80.9%** [70.0;88.5]	91.7% [86.1;95.2]	76.7% [75.0;78.2]	77.4% [74.6;79.9]
	-865 (55s)	**-1052** (774s)	-1289 (71s)	-25256 (9086s)	-15394 (2537s)
SEM	64.6% [52.5;75.1]	51.5% [39.8;62.9]	67.6% [59.6;74.7]	74.9% [73.2;76.5]	93.8% [92.1;95.2]
	-1091 (156s)	-1442 (977s)	-1483 (982s)	-50969 (22562s)	-16197 (963s)
GS+T-ACA	58.5% [46.3;69.6]	77.9% [66.7;86.2]	**93.1%** [87.8;96.2]	77.1% [75.5;78.6]	77.4% [74.6;79.9]
	-826 (16s)	**-1052** (603s)	**-1233** (52s)	**-20469** (5050s)	-15391 (856s)
SEM+T	64.6% [52.5;75.1]	51.5% [39.8;62.9]	**93.1%** [87.8;96.2]	74.9% [73.2;76.5]	93.8% [92.1;95.2]
	-1112 (341s)	-1447 (2190s)	-1485 (1094s)	-50969 (30417s)	-15729 (5492s)
GES-ACA	64.6% [52.5;75.1]	**82.4%** [71.6;89.6]	**93.8%** [88.6;96.7]	77.1% [75.5;78.6]	**96.1%** [94.7;97.1]
	-866 (76s)	-1160 (536s)	-1293 (123s)	-23462 (6350s)	-15535 (515s)
GES-EM	64.6% [52.5;75.1]	51.5% [39.8;62.9]	68.3% [60.3;75.3]	74.9% [73.2;76.5]	93.8% [92.1;95.2]
	-1101 (240s)	-1446 (1120s)	-1522 (1062s)	-38947 (54748s)	-16197 (1545s)

Table 1: Two first lines *: dataset names; number of attributes; dataset length; test dataset length; percentage of incomplete entries.* Following lines *: method names; best good classification percentage on three runs; 95%-confidence interval; selected model likelihood; learning time in seconds on a laptop 2.4GHz with Matlab®R2006a.*

4 Conclusions and prospects

Bayesian networks are a tool of choice for reasoning in uncertainty. However, most of the time, Bayesian network structural learning only deal with complete data.

Usually EM principle is used for structure learning as it has been proved to be optimal when the dataset is MAR as ACA is known to introduce a bias when the dataset is MAR and not only MCAR. In those experiments, we have supposed that learning datasets are NMAR as they are real life issued datasets and as it is difficult to know the kind of the dataset if you have not artificially built it. There is no more reasons to use EM rather than ACA during the learning process as both methods are biased.

In this study, ACA (available cases analysis or pairwise deletion) has empirically been compared to EM for Bayesian structure learning. First results show that this method is quite efficient and not very complex. By using it, it is possible to find structures which have a good likelihood and lead to good classification rates, and to do really less time than using the EM algorithm. This first conclusive experiment stage is not final. We are now planning to test and evaluate these algorithms on a wider range of problems.

Moreover, we know the limitation of the BIC criterion and we need to try other criterions: some specific to classification problems as an adaptation of the classification likelihood LCL/LL_c or of ICL to structure learning or more general ones as AIC, AICc, BDe, BDeu, BDγ, MDL/IMDL to study which one perform well with ACA or EM as we have noticed big graphical differences in learnt structures depending on the used method (not reported here because of space limitation).

Acknowledgements

We would like to thank Philippe Leray for the useful discussions we have had.

References

[Auvray and Wehenkel2002] V. Auvray and L. Wehenkel. 2002. On the construction of the inclusion boundary neighbourhood for markov equivalence classes of bayesian network structures. In Adnan Darwiche and Nir Friedman, editors, *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02)*, pages 26–35, S.F., Cal. Morgan Kaufmann Publishers.

[Bennani and Bossaert1996] Y. Bennani and F. Bossaert. 1996. Predictive neural networks for traffic disturbance detection in the telephone network. In *Proceedings of IMACS-CESA'96*, Lille, France.

[Blake and Merz1998] C.L. Blake and C.J. Merz. 1998. UCI repository of machine learning databases.

[Castelo and Kocka2002] R. Castelo and T. Kocka. 2002. Towards an inclusion driven learning of bayesian networks. Technical Report UU-CS-2002-05, Institute of information and computing sciences, University of Utrecht.

[Chickering and Heckerman1996] D. Chickering and D. Heckerman. 1996. Efficient Approximation for the Marginal Likelihood of Incomplete Data given a Bayesian Network. In *UAI'96*, pages 158–168. Morgan Kaufmann.

[Chickering and Meek2002] D. Chickering and C. Meek. 2002. Finding optimal bayesian networks. In Adnan Darwiche and Nir Friedman, editors, *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02)*, pages 94–102, S.F., Cal. Morgan Kaufmann Publishers.

[Chickering et al.1995] D. Chickering, D. Geiger, and D. Heckerman. 1995. Learning bayesian networks: Search methods and experimental results. In *In Proceedings of Fifth Conference on Artificial Intelligence and Statistics*, pages 112–128.

[Chickering2002a] D.M. Chickering. 2002a. Learning equivalence classes of bayesian-network structures. *Journal of machine learning research*, 2:445–498.

[Chickering2002b] D.M. Chickering. 2002b. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, November.

[Chow and Liu1968] C.K. Chow and C.N. Liu. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467.

[Cooper and Hersovits1992] G. Cooper and E. Hersovits. 1992. A bayesian method for the induction of probabilistic networks from data. *Maching Learning*, 9:309–347.

[Dash and Druzdzel2003] D. Dash and M.J. Druzdzel. 2003. Robust independence testing for constraint-based learning of causal structure. Proceedings of The Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI03), pp 167-174.

[Dempster et al.1977] A. Dempster, N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38.

[François and Leray2004] O. François and P. Leray. 2004. Evaluation d'algorithmes d'apprentissage de structure pour les réseaux bayésiens. In *14ieme Congrès francophone de Reconnaissance des formes et d'Intelligence artificielle*, pages 1453–1460.

[François and Leray2006] O.C.H. François and P. Leray. 2006. Learning the tree augmented naive bayes classifier from incomplete datasets. In *Proceedings of the Third European Workshop on Probabilistic Graphical Models (PGM'06)*, pages 91–98, Prague, Czech Republic, september.

[François2006] Olivier François. 2006. *De l'identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplètes*. Ph.D. thesis, Institut National des Sciences Appliquées de Rouen (INSA), http://ofrancois.tuxfamily.org/these.html.

[Friedman1997] N. Friedman. 1997. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the 14th International Conference on Machine Learning*, pages 125–133. Morgan Kaufmann.

[Friedman1998] N. Friedman. 1998. The bayesian structural EM algorithm. In Gregory F. Cooper and Serafín Moral, editors, *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 129–138, San Francisco, July. Morgan Kaufmann.

[Geiger1992] D. Geiger. 1992. An entropy-based learning algorithm of bayesian conditional trees. In *Uncertainty in Artificial Intelligence: Proceedings of the Eighth Conference (UAI-1992)*, pages 92–97, San Mateo, CA. Morgan Kaufmann Publishers.

[Geman and Geman1984] S. Geman and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November.

[Gillispie and Lemieux2001] S.B. Gillispie and C. Lemieux. 2001. Enumerating markov equivalence classes of acyclic digraph models. In *Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference (UAI-2001)*, pages 171–177, San Francisco, CA. Morgan Kaufmann Publishers.

[Heckerman et al.1994] D. Heckerman, D. Geiger, and M. Chickering. 1994. Learning Bayesian networks: The combination of knowledge and statistical data. In Ramon Lopez de Mantaras and David Poole, editors, *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 293–301, San Francisco, CA, USA, July. Morgan Kaufmann Publishers.

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

[Jensen1996] F.V. Jensen. 1996. *An introduction to Bayesian Networks*. Taylor and Francis, London, United Kingdom.

[Keogh and Pazzani1999] E. Keogh and M. Pazzani. 1999. Learning augmented bayesian classifiers: A comparison of distribution-based and classification-based approaches. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, pages 225–230.

[Kim and Pearl1987] J.H. Kim and J. Pearl. 1987. Convice; a conversational inference consolidation engine. *IEEE Trans. on Systems, Man and Cybernetics*, 17:120–132.

[Kočka et al.2001] T. Kočka, R.R. Bouckaert, and M. Studený. 2001. On characterization inclusion of bayesian networks. In J Breese and D. Koller, editors, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence.*, pages 261–268. Morgan Kaufmann.

[Lauritzen1995] S. Lauritzen. 1995. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19:191–201.

[Leray and François2004a] P. Leray and O. François. 2004a. BNT structure learning package: Documentation and experiments. Technical Report 2004/PhLOF, Laboratoire PSI, INSA de Rouen.

[Leray and François2004b] P. Leray and O. François. 2004b. Réseaux bayésiens pour la classification - méthodologie et illustration dans le cadre du diagnostic médical. *Revue d'Intelligence Artificielle*, 18(2/2004):169–193.

[Leray and François2005] P. Leray and O. François. 2005. Bayesian Network Structural Learning and Incomplete Data. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 2005), Espoo, Finland*, pages 33–40.

[Meek1997] C. Meek. 1997. *Graphical Models: Selecting causal and statistical models*. Ph.D. thesis, Carnegie Mellon University.

[Murphy2001] K. Murphy. 2001. The BayesNet Toolbox for Matlab. Computing Science and Statistics: Proceedings of Interface, 33.

[Myers et al.1999] J.W. Myers, K.B. Laskey, and T.S. Lewitt. 1999. Learning bayesian network from incomplete data with stochastic search algorithms. In *the Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI99)*.

[Naïm et al.2004] P. Naïm, P.-H. Wuillemin, P. Leray, O. Pourret, and A. Becker. 2004. *Réseaux bayésiens*. Eyrolles, ISBN : 2-212-11137-1.

[Pearl1998] J. Pearl. 1998. Graphical models for probabilistic and causal reasoning. In Dov M. Gabbay and Philippe Smets, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Volume 1: Quantified Representation of Uncertainty and Imprecision*, pages 367–389. Kluwer Academic Publishers, Dordrecht.

[Ramoni and Sebastiani1998] M. Ramoni and P. Sebastiani. 1998. Parameter estimation in Bayesian networks from incomplete databases. *Intelligent Data Analysis*, 2:139–160.

[Ramoni and Sebastiani2000] M. Ramoni and P. Sebastiani. 2000. Robust learning with missing data. *Machine Learning*, 45:147–170.

[Rubin1976] D.B. Rubin. 1976. Inference and missing data. *Biometrika*, 63:581–592.

[Spiegelhalter and Lauritzen1990] D. J. Spiegelhalter and S. L. Lauritzen. 1990. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605.