A Conservative Feature Selection Algorithm with Informatively Missing Data

Alexandre Aussem, Sergio Rodrigues de Morais

A Conservative Feature Selection Algorithm with Informatively Missing Data

Alexandre Aussem* — **Sergio Rodrigues de Morais****

* *Université de Lyon 1, LIESP, F-69622 Villeurbanne, France*
aaussem@univ-lyon1.fr
** *INSA-Lyon, LIESP, F-69622 Villeurbanne, France*
sergio.rodrigues-de-morais@insa-lyon.fr

RÉSUMÉ. *Cet article présente un nouvel algorithme conservateur de sélection de variables avec données manquantes. Il est conservateur au sens où il fait une hypothèse au pire cas sur le processus de perte des données. Il s'applique en particulier aux données manquantes vérifiant l'hypothèse IM ("informatively missing"), i.e., quand les données manquantes ne peuvent être inférées à partir des seules données disponibles. L'algorithme est fondé sur la recherche de la couverture de Markov de la variable cible. Une évaluation empirique est menée sur plusieurs bases de données synthétiques et réelles pour évaluer son efficacité.*

ABSTRACT. *This papers introduces a novel conservative feature subset selection method with informatively missing data, i.e., when data is not missing at random but due to an unknown censoring mechanism. This is achieved in the context of determining the Markov blanket (MB) of the target variable in a Bayesian network. The method is conservative in the sense that it constructs the MB that reflects the worst-case assumption about the missing data mechanism, when the missing values cannot be inferred from the available data only. An application of the method on synthetic and real-world incomplete data is carried out to illustrate its practical relevance.*

MOTS-CLÉS : *Réseaux Bayésiens, données manquantes, couverture de Markov, classification probabiliste, selection de variables.*

KEYWORDS: *Bayesian networks, missing data, Markov boundary, probabilistic classification, feature subset selection.*

1. Introduction

A principled solution to the feature subset selection problem is to determine the Markov boundary (MB) of a target variable T (Koller *et al.*, 1996; Margaritis *et al.*, 1999; Pearl, 1988). The Markov boundary of T, denoted by $\mathbf{MB}_T$, is defined as any minimal subset of $\mathbf{V}$ (the full set) that renders the rest of $\mathbf{V}$ independent of T. In recent years, they have been a growing interest in inducing the MB automatically from data. Very powerful correct, scalable and data-efficient constraint-based (CB) algorithms have been proposed recently, e.g., PCMB (Peña *et al.*, 2005), IAMB (Tsamardinos *et al.*, 2003) or its variants : Fast-IAMB (Yaramakala, 2004) and Inter-IAMB (Yaramakala *et al.*, 2005). These methods yields compact MB by heeding independencies in the data. They systematically check the data for independence relations and use those relationships to infer necessary features in the MB. When no entry is missing in the database, the MB can be estimated efficiently with these methods. An excellent trade-off between time and quality of reconstruction is systematically obtained. Unfortunately, when the database is incomplete, i.e., some entries are reported as unknown, the simplicity and efficiency of these methods are lost.

Scoring approaches are capable of dealing with incomplete records in the database based on the expectation maximization (EM) principle (Dempster *et al.*, 1977). The EM algorithm (Friedman, 1997; Friedman, 1998; Francois *et al.*, 2006; Francois *et al.*, 2007) and Gibbs sampling (Geman *et al.*, 1984) are notorious solutions to handle incomplete data sets, but both methods assume implicitly that data are missing at random. Under this assumption, the missing values can be inferred from the available data. However, this assumption does not hold and it is hard, if not impossible, to test in practice. The decrease in accuracy my be severe with EM-based methods when the assumption is violated. More recently, Robust Bayesian Estimator (RBE) (Ramoni *et al.*, 2001) methods were proposed to learn conditional probability distributions from incomplete data sets without making any assumption about the missing data mechanism. The major feature of the RBE is to produce probability estimates that are robust with respect to different types of missing data. This robustness is achieved by providing probability intervals containing the estimates that can be learned from all completed data sets.

In this paper, we transpose Ramoni's ideas (Ramoni *et al.*, 2001) to the problem of finding a *conservative* Markov blanket of a variable with information missing entries. By conservative, we mean the MB that renders the rest of the variables independent of the target without making any assumption about the unknown censoring mechanism. The problem is addressed in this paper by maximizing the conditional dependence measure over all possible ways to fill the missing data. The idea is exemplified with the G^2 statistic in this paper but its generalization is fairly easy. The intuition behind the method is similar to that of Ramoni et al. (Ramoni *et al.*, 2001). The idea is that, when no information about the pattern of missing data is available, an incomplete database contains the set of all possible estimates and this paper provides a characterization of these constraints. This conservative test addresses the main shortcoming of CB methods with missing data : the difficulty of performing a classical (non bayesian)

independence test when some entries are missing without making any assumption about the missing data mechanism.

The remainder of this paper describes our approach. Sections 2 and 3 establish some notation and reviews the background and motivation of the research. Section 4 describes the theoretical framework of the method, while Section 5 applies the method to synthetic and real incomplete data sets.

2. Constraint-based Markov blanket discovery

We denote the conditional independence of the variable X and Y given $\mathbf{Z}$, in some distribution P with $Ind_P(X;Y|\mathbf{Z})$, dependence as $Dep_P(X;Y|\mathbf{Z})$. Constraint-Based (CB for short) learning methods systematically check the data for independence relations and use those relationships to infer necessary features to be included in the MB. They rely on a probabilistic association measure between X and Y conditionally on $\mathbf{Z}$ denoted by $Assoc(X;Y|\mathbf{Z})$. In our implementation, we use a well-known CB method called IAMB (Tsamardinos *et al.*, 2003) along with a statistically oriented conditional independence test based on the G-test :

$$G = 2\sum_{i=1}^{m}\sum_{j=1}^{p}\sum_{k=1}^{q} n(i,j,k)\ln\frac{n(i,j,k)n(\cdot,\cdot,k)}{n(i,\cdot,k)n(\cdot,j,k)}. \quad [1]$$

where $n(i,j,k)$ is the number of times simultaneously $X = x_i$, $Y = x_j$ and $\mathbf{Z} = \mathbf{z_k}$ in the sample, that is, the value of the cell (i,j,k) in the contingency table. A distribution P is said *faithful* with respect to $\mathcal{G}$ if the d-separations in the DAG identify all and only the conditional independencies in P. Suppose $< \mathcal{G}, P >$ satisfies the faithfulness condition, it may be shown that : (1) the set of parents, children and parents of children of X is the *unique* Markov boundary of X, and (2) X and Y are not adjacent in $\mathcal{G}$ iff $\exists \mathbf{Z} \in \mathcal{V} \setminus \{X \cup Y\}$ such that $Ind_P(X;Y|\mathbf{Z})$ (see (Neapolitan, 2004) for instance). Tsamardinos et al. prove in [19] that IAMB is correct under the faithfulness assumption.

2.1. *Dealing with missing entries*

When the database is complete, that is, all entries are known, the computation of $Assoc(X;Y|\mathbf{Z})$ is straightforward. Unfortunately, the expression cannot be computed when entries are reported as unknown. Suppose for simplicity that $\mathbf{Z}$ is empty and that we wish to estimate the maximum value of $Assoc(X,Y)$ from an incomplete database $\mathcal{D} = \{X_j, Y_j\}_{j=1,\ldots,n}$ in which some entries of the variable X_j and Y_j are unknown. These unknown entries give rise to multiple types of incomplete cases that are relevant to the estimation of $n(i,j)$. For instance, Table I shows a database with several entries missing denoted by " ?". $n(?,j)$ will for instance denote the number

Case	1	2	3	4	5	6	7	8
X	1	2	1	?	2	?	1	1
Y	2	1	?	1	2	?	?	2

Tableau 1. *Cases with some items missing.*

	$Y=1$	$Y=2$	$Y=?$
$X=1$	0	2	2
$X=2$	1	1	0
$X=?$	1	0	1

Tableau 2. *Extended contingency table.*

of cases where X is missing and Y take value y_j. $n(?, j) = 1$ in this example. The extended contingency table is shown in Table 2. The task is to estimate the $n(i, j)$ values. The issues involved in estimating theses values from an incomplete data set $\mathcal{D}$ are better explained if we regard $\mathcal{D}$ as the result of a deletion process applied to a complete but unknown database $\mathcal{D}_c$. We define a consistent completion of $\mathcal{D}$ to be any complete database $\mathcal{D}_c$ from which we can obtain $\mathcal{D}$ using some deletion process. The set of consistent completions $\mathcal{D}_c$ is given by all databases in which the unknown entries are replaced by one of the possible values of the unobserved variables.

2.2. *Deletion process*

According to (Dempster *et al.*, 1977), the assumptions about the missing data mechanisms may be classified into three categories :

– missing completely at random (MCAR) : the probability that an entry is missing is independent of both observed and unobserved values in the data set ;

– missing at random (MAR) : the probability that an entry is missing is a function of the observed values in the data set ;

– informatively missing (IM) : the probability that an entry is missing depends on both observed and unobserved values in the data set.

In order to specify the deletion processes, a dummy binary variable R_i may be associated with each variable X_i. When R_i takes value '1', the entry $X_i = x_i$ is not observed and vice-versa. When the probability distribution of each R_i is independent of $X_1, \ldots, X_n$, the data may be seen as MCAR. When this probability distribution is a function of the observed values in the data set, data are MAR. Now, when this probability distribution is a function of the observed and unobserved entries, data are IM. The expectation maximization (EM) algorithm (Dempster *et al.*, 1977) and Gibbs

	$Y=0$	$Y=1$
$X=1$	0	4
$X=2$	3	1

Tableau 3. *Completed contingency table maximizing G.*

sampling (Geman *et al.*, 1984) are well known solutions to handle incomplete data sets but they rely on the assumption that data are MAR. The problem is that MCAR and MAR assumptions are hard, if not impossible, to test. On the other hand, one cannot simply infer the missing entries from the observed ones anymore when the data is IM. Hence the need for a general approach dealing with the IM worst case censoring mechanism.

3. A conservative Markov blanket

The solution we propose is based on the idea that, even with no information on the missing data mechanism, an incomplete data set $\mathcal{D}$ constrains the set of estimates that can be induced from its consistent completions. Following this principle, we introduce the conservative statistical test with no assumptions about the missing data mechanism. It is conservative in the sense that makes always the worst case assumption by assuming dependency when independency cannot be guaranteed in all the distributions associated with the consistent data completions. Let $Assoc(X;Y|\mathbf{Z};\mathcal{D}_c)$ be the value of $Assoc(X;Y|\mathbf{Z})$ evaluated on complete set $\mathcal{D}_c$.

Definition 1 *Let* $Assoc(X;Y|\mathbf{Z})$ *be an conditional association measure.* $ConsAssoc(X;Y|\mathbf{Z})$ *is called a* conservative *association measure with respect to* $Assoc(X;Y|\mathbf{Z})$ *if, for all incomplete database* $\mathcal{D}$*,* $ConsAssoc(X;Y|\mathbf{Z};\mathcal{D}) \geq Assoc(X;Y|\mathbf{Z};\mathcal{D}_c)$ *for every consistent data completion* $\mathcal{D}_c$ *obtained from* $\mathcal{D}$*.*

Using $ConsAssoc(X;Y|\mathbf{Z};\mathcal{D})$ yields a conservative test in the sense that it always take the worst-case assumption about the missing data mechanism to decide whether X and Y are conditionally independent. It is implicitly assumed in the definition that $ConsAssoc(X;Y|\mathbf{Z};\mathcal{D}_c) = Assoc(X;Y|\mathbf{Z};\mathcal{D}_c)$ for any completion $\mathcal{D}_c$ obtained from $\mathcal{D}$. In general, whatever the way the missing database is completed, we would want an edge to mean a direct dependency when the CB algorithm is run on these data. As we know, the faithfulness entails this. The following theorem shows that a conservative Markov blanket can be obtained using IAMB$(T,\mathcal{D},ConsAssoc)$ (i.e., IAMB run with the conservative test) as shown next :

Theorem 1 *Suppose the independence tests are correct and that the learning database* $\mathcal{D}_c$ *is an independent and identically distributed sample from a probability distribution* P *faithful to a DAG* $\mathcal{G}$*. Suppose given an incomplete database* $\mathcal{D}$ *that was*

obtained from $\mathcal{D}_c$ by some missing data mechanism. Then, IAMB$(T, \mathcal{D}, ConsAssoc)$ returns a conservative Markov blanket of X.

Proof : If $X \in \mathbf{PC}_T$ (where $\mathbf{PC}_T$ denote the set of parents and children in $\mathcal{G}$) then X remains dependent on T conditioned on any set $\mathbf{Z} \in \mathbf{U} \setminus \{X, Y\}$. Therefore, from definition 1, we have $\alpha < Assoc(X; Y|\mathbf{Z}; \mathcal{D}_c) \leq ConsAssoc(X; Y|\mathbf{Z}; \mathcal{D})$ for all $\mathbf{Z} \in \mathbf{U} \setminus \{X, Y\}$. So X is necessarily in the output of IAMB$(T, \mathcal{D})$ run with the conservative test. Now, if $X \in \mathbf{MB}_T \setminus \mathbf{PC}_T$ (where $\mathbf{MB}_T$ denote the MB of T), then X is spouse of T owing to the faithfulness assumption. So there is a variable $Y \in \mathbf{PC}_T$ such that $T \not\perp X|\mathbf{Z} \cup Y$ for all $\mathbf{Z} \in \mathbf{U} \setminus \{T, X\}$. Recall that IAMB works in two steps. In the first stage, candidate nodes are added sequentially to the current candidate MB set when they are not found independent on T conditioned on the current MB. The extra nodes are removed from MB in the second step. As Y will enter this set during the first stage, X will also enter but it will never leave this set in Phase II because $T \not\perp X|\{\mathbf{Z} \cup Y\}$ for all $\mathbf{Z} \in \mathbf{U} \setminus \{T, X\}$. □.

4. A conservative independence test

This section shows how to design practically a conservative test based on the G statistic. For sake of clarity, $\mathbf{Z}$ is supposed empty. The case $\mathbf{Z}$ non empty will be discussed in the sequel. As a shorthand, we note $\sum_{j=1}^{p} n(i,j) = n(i,\cdot)$ and $\sum_{i=1}^{m} n(i,j) = n(\cdot,j)$. Let $n^0(i,j)$ be the number of non missing cases in cell (i,j) and let x_{ij} (resp. y_{ij} and z_{ij}) be the number of additional cases that are affected to cell (i,j) owing to $n(?,j)$ (resp. $n(?,j)$ and $n(?,?)$). The value that would be computed from the complete data set (if known) is

$$n(i,j) = n^0(i,j) + x_{ij} + y_{ij} + z_{ij}, \ \forall i,j. \qquad [2]$$

The information conveyed by the incomplete cases impose several constraints on the variables x_{ij}, y_{ij} and z_{ij}. In order to identify the maximum estimate of G, we

have to consider the problem of maximizing G subject to equality and inequality constraints :

$$(\mathcal{P}) \begin{cases} \text{Max G} & = & 2\sum_{i=1}^{p}\sum_{j=1}^{m} n(i,j) \ln \dfrac{n(i,j)}{n(i,\cdot)n(\cdot,j)/n}, \\ \text{subject to} & & \\ \sum_{j=1}^{p} x_{ij} & = & n(i,?),\ \forall i, \\ \sum_{i=1}^{m} y_{ij} & = & n(?,j),\ \forall j, \\ \sum_{i=1}^{m}\sum_{j=1}^{p} z_{ij} & = & n(?,?), \\ n(i,j) & = & n^0(i,j) + x_{ij} + y_{ij} + z_{ij},\ \forall i,j, \\ x_{ij} & \geq & 0,\ \forall i,j, \\ y_{ij} & \geq & 0,\ \forall i,j. \\ z_{ij} & \geq & 0,\ \forall i,j. \end{cases} \quad [3]$$

$(\mathcal{P})$ is combinatorial problem with $3mp$ variables. If we relax the integrity constraints the problem may be solved using nonlinear programming techniques. From *Karush-Kuhn-Tucker* (KKT) Theorem, the problem of maximizing G subject to equality and inequality constraints is obtained by optimizing the Lagrange function $\mathcal{L}$ with respect to $3mp + m + p + 1$ parameters subject to an extended condition set. Let $\mathbf{x} = (x_{11}, \ldots, x_{mp}) \in \Re^{mp}$, $\mathbf{y} = (y_{11}, \ldots, y_{mp}) \in \Re^{mp}$ and $\mathbf{z} = (z_{11}, \ldots, z_{mp}) \in \Re^{mp}$ be a candidate solution in C. From the KKT conditions, if $\mathbf{x}$ is a local extrema of the $Minf(x)$ subject to $h_i(x) = 0$ for $i = 1, \ldots, n$ and $g_j(x) \geq 0$ $j = 1, \ldots, p$, there exist $\lambda = (\lambda_1, \ldots, \lambda_n) \in \Re^n$ and a positive vector $\mu = (\mu_1, \ldots, \mu_m) \in \Re^m$ such that $\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x) + \sum_{i=1}^{p} \mu_i \nabla g_i(x) = 0$ and $\mu_i g_i(x) = 0, \forall i$.

We need to express ∇G by taking the derivative of G with respect to x_{ij} (y_{ij} and z_{ij} play a symmetric role),

$$\frac{1}{2}\frac{\partial G}{\partial x_{ij}} = \frac{1}{2}\frac{\partial G}{\partial y_{ij}} = \frac{1}{2}\frac{\partial G}{\partial z_{ij}} = \ln \frac{n(i,j)}{n(i,\cdot)n(\cdot,j)/n} - 1 \quad [4]$$

Define λ_i^x and μ_i^x (resp. λ_i^y and μ_i^y) and λ^z and μ_i^z, the Lagrange multipliers associated to the constraints. From the KKT conditions, the maximum is obtained when

$$n(i,j) = e^{\beta_{ij}} \cdot n(i,\cdot)n(\cdot,j)/n,\ \forall i,j. \quad [5]$$

where the parameters β_{ij} are solution of

$$\left\{ \begin{array}{rcl} \beta_{ij} & = & \ln n - (\lambda_i^x + \mu_{ij}^x), \\ \beta_{ij} & = & \ln n - (\lambda_j^y + \mu_{ij}^y), \\ \beta_{ij} & = & \ln n - (\lambda^z + \mu_{ij}^z). \end{array} \right. \quad [6]$$

with μ_{ij}^x, μ_{ij}^y and μ_{ij}^z solution of

$$\left\{ \begin{array}{l} \mu_{ij}^x x_{ij} = 0,\ \mu_{ij}^y y_{ij} = 0,\ \mu_{ij}^z z_{ij} = 0, \\ \mu_{ij}^x \geq 0,\ \mu_{ij}^y \geq 0,\ \mu_{ij}^z \geq 0. \end{array} \right. \quad [7]$$

This is a nonlinear system that does not seem to admit a closed-form solution. $\beta_{ij} \neq 0$ is interpreted as a local deviation of cell (i, j) to independence. $\beta_{ij} > 0$ when $n(i, j)$ is over-represented and vice-versa. The next theorem will help characterize G's maximum.

Theorem 2 *If* $(\mathbf{x}, \mathbf{y})$ *is a solution of* $(\mathcal{P})$*, then the following assertions hold :*

– For all i*, there is unique* $j^\star$ *such that* $x_{ij^\star} \neq 0$ *and* $\forall j \neq j^\star$, $\beta_{ij} \leq \beta_{ij^\star}$,

– For all j*, there is unique* $i^\star$ *such that* $y_{i^\star j} \neq 0$*, and* $\forall i \neq i^\star$, $\beta_{ij} \leq \beta_{i^\star j}$,

– There is unique pair $(i^\star, j^\star)$ *such that* $z_{i^\star j^\star} \neq 0$*,and* $\forall (i, j) \neq (i^\star, j^\star)$, $\beta_{ij} \leq \beta_{i^\star j^\star}$,

Proof : Consider $G_{ijk}(x) = G(n_{11}, \ldots, n_{ij} - x, \ldots, n_{ik} + x, \ldots, n_{pm})$ for all (i, j, k) over the domain $0 \leq x \leq \min(n_{ij}, n_{i,?})$. $G'(x) = 0$ iff $n(i, j)/n(\cdot, j) = n(i, k)/n(\cdot, k)$. This is the minimum of G. From KKT conditions, the maximum is reached only when an inequality is active (at $x = 0$ or at $x = \min(n_{ij}, n_{i,?})$). The same results hold for $G_{ijk}(x) = G(n_{11}, \ldots, n_{ij} - y, \ldots, n_{kj} + y, \ldots, n_{pm})$. The maximum of G requires all $G_{ijk}(x)$ to be locally maximized, otherwise G can be increased by permuting the missing cases across lines or columns. The maximum for G can be characterized further. Let i and j be such that $x_{ij} \neq 0$ and $y_{ij} \neq 0$. $\beta_{ij} = \ln n - \lambda_i^x = \ln n - \lambda_i^y$, since $\mu_{ij}^x = 0$ and $\mu_{ij}^y = 0$. Therefore, $\lambda_i^x = \lambda_i^y$. For all $k \neq i$, $\beta_{kj} \leq \beta_{ij}$ and for all $k \neq j$, $\beta_{ik} \leq \beta_{ij}$. This completes the proof.

Owing to Theorem 2, to maximize G, a single cell $(i, j^\star)$ along each line i (resp. column j) of the contingency table should receive all the missing entries $n(i, ?)$ (resp. $n(?, j)$). In addition, a single cell $(i^\star, j^\star)$ increases the most the G value. $(i^\star, j^\star)$ is such that $\beta_{ij} \leq \beta_{i^\star j^\star}$ for all (i, j). In order to avoid the burdensome numerical resolution of the optimization problem $\mathcal{P}$, we propose in the next section a simple heuristic that greedily completes the missing entries for the test to be conservative.

4.1. *Extension to conditional G-tests*

While the previous section only discussed the case where **Z** is empty, Theorem 2 can easily be generalized to conditional tests $ConsAssoc(X;Y|\mathbf{Z};\mathcal{D})$ although it is not discussed here for sake of conciseness. With the above result in mind, we devised a greedy heuristic termed *GreedyGmax* in order to approximate the maximum in order $\mathcal{O}(n)$ when n stands for the number of cells in the contingency table. The idea is to select sequentially the triple (i,j,k) that increases most G. From Theorem 2, we know the cell that will increase G the most is the one for which $\beta_{(i,j,k)}$ is the largest. Since the $\beta_{(i,j,k)}$ depend on the missing entries, they are unknown. Our idea is to estimate $\beta_{(i,j,k)}$ by $\beta^0_{(i,j,k)}$ using the available data only in Equation 5 and fill all the possible missing entries in the cell with the highest $\beta^0_{(i,j,k)}$ value. This procedure is then repeated until the database is completed. The Greedymax heuristic is depicted in Algorithm 2. $\beta^0_{(i,j,k)}$ are computed at line 5 and they are sorted in decreasing order at line 7.

Algorithm 1 *GreedyGmax*

Require: *X*, *Y* : testing variables ; **Z** : conditioning set ; *D* : an incomplete data set ;
Ensure: *Gmax* : an upper bound for the G-statistic ;
1: **for all** *i*, *j*, *k* **do**
2: **Compute** $n^o(i,j,k)$
3: **end for**
4: **for all** *i*, *j*, *k* **do**
5: $\beta^0_{(i,j,k)} = n^0(i,j,k)n^0(\cdot,\cdot,k)/n^0(i,\cdot,k)n^0(\cdot,j,k)$
6: **end for**
7: **Cell=Sort** $\{\beta^0_{(i,j,k)}\}$
8: $idx = 1$
9: **repeat**
10: Fill **Cell**(idx) with as much as possible missing entries
11: $idx = idx + 1$
12: **until** *D* is complete
13: *Gmax* = G-statistic on completed data

It should be noted that the method is not yet conservative strictly speaking, as the greedy maximization heuristic is not guaranteed to find the global maximum. We shall argue, however, that the probability of a completion mechanism to yield $G > GreedyGmax$ is sufficiently small that it can be considered as zero in practice. For sake of illustration, Table 3 shows the completion obtained when *GreedyGmax* is applied to data shown in Table 1. In this simple example, the heuristic yields the maximum for G.

5. Experimental evaluation

This section reports the results of an experiment based on real-world data with missing entries. The aim of these experiments is to show that the MB returned by the

conservative method can reveal interesting dependencies that may have been missed by standard approaches. This raises an interesting question : is a conservative MB learning algorithm useful in the context of feature selection ? In other word, can we design more efficient probabilistic classifiers by use of a conservative Markov blanket ? What are the missing rates and/or the missing mechanisms for which the methods works best ?

In this section, we consider the Interleaved Incremental Association Markov Boundary (Inter-IAMB) (Tsamardinos *et al.*, 2003; Aliferis *et al.*, 2003) as our reference Markov boundary discovery algorithm. Inter-IAMB is variant of IAMB that has been proposed to improve its data efficiency while still being correct under faithfulness assumptions. Inter-IAMB seeks directly the minimal subset of **V** (the full set) that renders the rest of **V** independent of T, then $\mathbf{MB}_T$ as IAMB does (Tsamardinos *et al.*, 2003). The key difference between IAMB and Inter-IAMB is that the shrinking phase is interleaved into the growing phase in Inter-IAMB.

The original G-test is applied using the *available case analysis* technique, i.e., using for the estimation of $Ind(X, Y|Z)$ only the instances where X, Y and Z are non missing. The significance level for the independence tests is 0.01. We compare the accuracy of Inter-IAMB with the standard G-test versus Inter-IAMB with the *conservative* G-test based on $GreedyGmax(X, Y|\mathbf{Z})$. In our implementation, Inter-IAMB considers both tests to be reliable when the number of instances in D is at least ten times the number of degrees of freedom and skips it otherwise. Skipping the test means the variables are assumed to be independent without actually performing the test.

5.1. *Limits of the conservative test*

Before we assess the benefits of this approach, let us first gauge its limits on a simple example. Consider two discrete variables X and Y with 5 modalities each. 200 databases were generated, each containing 2000 independent and identically distributed samples. Figure 1 plots the average and standard deviation of the p-values obtained from the conservative Gmax test between 2 independent variables as a function of the ratio of missing data. The missing mechanism is MCAR. As may be seen, both variables are considered dependent for a significance level of 0.01 as soon as the ratio of missing data is superior to (say) 3%. So the present approach will surely be inefficient for larger missing rates.

5.2. *Synthetic data*

This section reports the results of an experiment on a toy problem. The aim is to evaluate the effectiveness of the Gmax approach when the data is informatively missing. The toy Bayesian network applied for this experiment is composed of four random variables as depicted in Figure 2. All variables were subject to a deletion process. We associated several pairs (i, j) of variable with a dummy variable R_{ij}

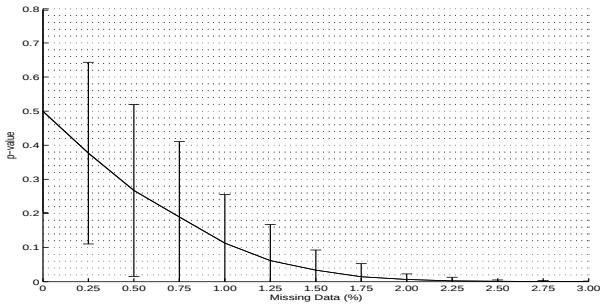
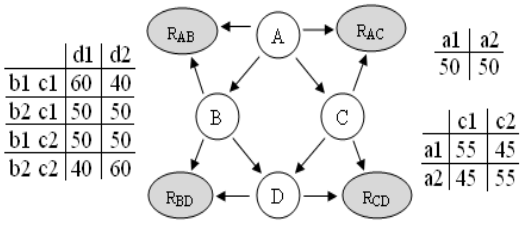


Figure 1. p-values obtained from the conservative Gmax test between 2 independent variables as a function of the ratio of missing data.

that took on one of the two values 0 and 1 with some probability. As seen in Fig. 2, the original graph was augmented by the 4 variables $R_{AB}, R_{AD}, R_{BC}, R_{CD}$. The dummy variables have all the same probability table. They were chosen such that : when the parents of the dummy variable have different values, no data is removed. Otherwise, the entry for both parents have MR% chances to be removed. Since the distribution of the variables R_{ij} depends of the unobserved values in the data set, all values removed with this process is IM. This deletion process was repeated with ten times with different missing rates. As may be seen, the conservative approach returns lesser and lesser false negatives as the missing rate increases. The price to pay in terms of false positives nodes seems affordable for the missing rates that we are considering.



	d1	d2
b1 c1	60	40
b2 c1	50	50
b1 c2	50	50
b2 c2	40	60

a1	a2
50	50

	c1	c2
a1	55	45
a2	45	55

Figure 2. Benchmark with extra dummy variables. The probability tables are shown. Values represent percentages.

5.3. *Real data*

For the second experiment we used data on Congressional Voting Records, available from the Machine Learning Repository at the University of California, Irvine (Blake, Keogh, Merz, 1998). The data set describes votes for each of the 435 member

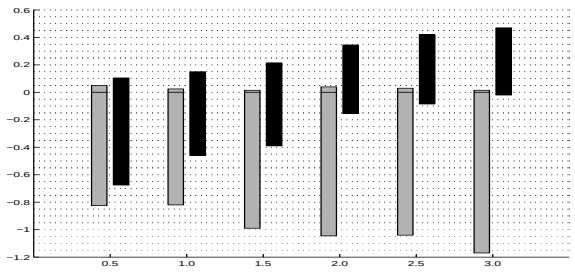


Figure 3. Average missing and extra variables for learning Markov boundaries of all variables of the toy problem in Fig.2. In black : the conservative InterIAMB using the GreedyGmax heuristic. In grey : the standard InterIAMB. All results are averaged over 10 runs.

of the US House of Representative on the 16 key issues during the 1984. Hence, the data set consists of 435 cases on 16 binary attributes and two classes that represent the party affiliation. There are 289 values reported as unknown. Although these missing entries amount to 4% of the data set, the number of incomplete cases is 203, more than 45% of the total. An important feature of this data is that unknown entries, and hence what member of the US House of Representative did not vote on, can be predictive. We have applied InterIAMB in this data set as a solution to the feature subset selection for probabilistic classification. The table below shows the results of classification using the variables selected by InterIAMB by applying the G statistic and the GreedyGmax heuristic. One can see the same algorithm provides a better feature selection for probabilistic classification when applying the GreedGmax heuristic. Standard InterIAMB outputs 4 variables : *adoption-of-the-budget-resolution*, *physician-fee-freeze*, *anti-satellite-test-ban* and *synfuels-corporation-cutback*. InterIAMB with Gmax outputs 5 variables : the same 4 except *anti-satellite-test-ban* plus 2 others : *synfuels-corporation-cutback* and *education-spending*. It is interesting to note here that the first MB is not included in the second as we would expect from Theorem 1. This comes from an unwanted side-effect in practice : the cascading effect of early test errors causes errors to be present in the output MB. Therefore, it is not surprising. As may be seen, the Gmax technique slightly outperforms in terms of classification accuracy by about 2% on average.

5.4. *Real data : Nasopharyngeal Carcinoma epidemiological data*

In this section, we apply the method on Nasopharyngeal Carcinoma (NPC) epidemiological data that was made available to us by the International Agency for Research on Cancer (IARC) (Feng *et al.*, 2007; Aussem *et al.*, 2007). This database is not yet available in the public domain. The original data is made up from 1289 instances and 321 discrete features to evaluate the risk of factors of NPC. The variables have 2 or 3 modalities. Among the 321 variables, we selected only the ones that have less than

	Gtest	Gmax
Correct classification rate	92.41%	94.25%
True positive rate	0.94	0.95
True negative rate	0.88	0.92
False positive rate	0.077	0.113
False negative rate	0.045	0.052
Kappa statistic	0.8391	0.8786

Tableau 4. *Congressional Voting Records : 10-fold cross-validation results with a naive Bayes classifier using the features selected by Inter-IAMB.*

	Gtest	*Gmax*
Correct classification rate	0.6059	0.6160
True positive rate	0.6340	0.6717
True negative rate	0.5760	0.5568
False positive rate	0.4240	0.4432
False negative rate	0.3660	0.3283
Kappa statistic	0.2102	0.2292

Tableau 5. *NPC data : 10-fold cross-validation results with a naive Bayes classifier using the features selected by Inter-IAMB.*

3% missing entries. This reduces the problem to 145 variables. The data was collected during a multi-center case-control study that has been undertaken in 2004 by the IARC) in the Maghreb (Morocco, Algeria and Tunisia), the endemic region of North Africa. Patients were interviewed according to a specific questionnaire. As all case control studies in epidemiology, half of population is comprised of individuals that are disease positive (the cases), and the other half (the control group) come from the same population that gave rise to the cases. As observed in Table 5, the Gmax technique yields here again a slight improvement in terms of overall accuracy. The false negative rate is important here as it represents a failure in detecting the disease. The latter was reduced by 4%.

6. Conclusion

In this paper, we discussed a conservative constraint-based Markov blanket learning method from incomplete data. The method is conservative in the sense that it constructs a Markov blanket that reflects the worst-case assumption about the missing data mechanism. An application of the method on synthetic and real-world incomplete data was carried out to illustrate its practical relevance. The method was shown to yield a benefit for databases with no more than 5% missing data. Future substantiation

through more experiments with other statistical association measures (e.g., Mutual Information, G^2 test) are currently being undertaken and comparisons with other data analysis techniques will be reported in due course.

7. Bibliographie

Aliferis C., Tsamardinos I., Statnikov A., « HITON : a novel Markov Blanket algorithm for optimal variable selection », *AMIA Annu Symp Proc*, p. 21-26, 2003.

Aussem A., Rodrigues de Morais S., Corbex M., « Nasopharyngeal Carcinoma Data Analysis with a Novel Bayesian Network Skeleton Learning », *11th Conference on Artificial Intelligence in Medicine AIME 07*, p. 326-330, 2007.

Dempster A. P., Laird N. M., Rubin D. B., « Maximum likelihood from incomplete data via the EM algorithm », *J. Roy. Statist. Soc. Ser. B*, vol. 39, n° 1, p. 1-38, 1977.

Feng B., al., « Dietary risk factors for nasopharyngeal carcinoma in Maghrebian countries. », *International Journal of Cancer*, vol. 121, n° 7, p. 1550-1555, 2007.

Francois O., Leray P., « Learning the Tree Augmented Naive Bayes Classifier from incomplete datasets », *The third European Workshop on Probabilistic Graphical Models PGM'06*, Prague, Czech Republic, p. 91-98, 2006.

Francois O., Leray P., « Generation of Incomplete Test-Data using Bayesian Networks », *Proceedings of IEEE IJCNN, International Joint Conference on Neural Networks*, Orlando, USA, p. 1-6, 2007.

Friedman N., « Learning Belief Networks in the Presence of Missing Values and Hidden Variables. », *ICML*, p. 125-133, 1997.

Friedman N., « The Bayesian Structural EM Algorithm. », *UAI*, p. 129-138, 1998.

Geman S., Geman D., « Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images », *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, n° 6, p. 721-741, Nov., 1984.

Koller D., Sahami M., « Toward Optimal Feature Selection », *ICML*, p. 284-292, 1996.

Margaritis D., Thrun S., « Bayesian Network Induction via Local Neighborhoods. », *NIPS*, p. 505-511, 1999.

Neapolitan R. E., *Learning Bayesian Networks*, Prentice Hall, 2004.

Peña J., Björkegren J., Tegnér J., « Scalable, Efficient and Correct Learning of Markov Boundaries under the Faithfulness Assumption », *8th European Conference on Symbolic and Quantitative Approaches to Reasoning under Uncertainty (ECSQARU 2005)*, vol. 21, Lecture Notes in Artificial Intelligence 3571, p. 136-147, 2005.

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

Ramoni M., Sebastiani P., « Robust Learning with Missing Data. », *Machine Learning*, vol. 45, n° 2, p. 147-170, 2001.

Tsamardinos I., Aliferis C. F., Statnikov A. R., « Algorithms for Large Scale Markov Blanket Discovery. », *FLAIRS Conference*, p. 376-381, 2003.

Yaramakala S., « Fast Markov Blanket Discovery. », *MS-Thesis, Iowa State University*, 2004.

Yaramakala S., Margaritis D., « Speculative Markov Blanket Discovery for Optimal Feature Selection. », *ICDM*, p. 809-812, 2005.