A novel scalable, data efficient and correct Markov boundary learning algorithm under faithfulness condition

Sergio Rodrigues de Morais, Alexandre Aussem

A novel scalable, data efficient and correct Markov boundary learning algorithm under faithfulness condition

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

* *INSA-Lyon, LIESP, F-69622 Villeurbanne, France*
sergio.rodrigues-de-morais@insa-lyon.fr

** *Université de Lyon 1, LIESP, F-69622 Villeurbanne, France*
aaussem@univ-lyon1.fr

RÉSUMÉ. *Dans cet article, nous proposons un nouvel algorithme sous contraintes pour l'apprentissage de la couverture de Markov dans de gros volumes de données. Ce dernier combine les avantages de deux algorithmes récents, PCMB et IAMB, en évitant certains écueils. En outre, nous démontrons qu'il est correct sous la condition dite de fidélité. Une évaluation empirique est menée sur plusieurs bases de données synthétiques et réelles, dont la base Thrombin constituée de 139,351 variables, pour évaluer son efficacité.*

ABSTRACT. *In this paper, we discuss a novel scalable, data efficient and correct Markov boundary learning algorithm under faithfulness condition. The latter combines the main advantages of PCMB and IAMB yet avoids some of their drawbacks. An empiric evaluation of our algorithm is provided on synthetic and real sparse databases scaling up to 139,351 variables. Our method is shown to be efficient in terms of both runtime and accuracy.*

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

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

1. Introduction

In this paper, we aim to identify the minimal subset of discrete random variables that is relevant for probabilistic classification in data sets with many variables but few instances (Guyon *et al.*, 2003). A principled solution to this problem is to determine the *Markov boundary* of the class variable T, i.e., the minimal subset of $\mathbf{U}$ (the full set), denoted by $\mathbf{MB}_T$ in the sequel, that renders the rest of $\mathbf{U}$ independent of T (Nilsson *et al.*, 2007).

Following (Peña *et al.*, 2007), we present a novel divide-and-conquer method in order to increase the efficiency of the Markov boundary (MB for short) discovery while still being scalable and correct under the faithfulness condition. The proposed method aims at producing an accurate MB discovery algorithm by combining rough and moderately accurate MB learners based on IAMB. Our algorithm, called MBOR, was designed with a view to keep the conditional test sizes of the tests as small as possible to increase the reliability of the conditional independence tests. MBOR is compared against two recent powerful constraint-based algorithms PCMB (Peña *et al.*, 2007) and Inter-IAMB (Yaramakala *et al.*, 2005). MBOR is proved by extensive empirical simulations on various synthetic and real data bases to be an excellent trade-off between time and quality of reconstruction.

2. Notations and preliminaries

We denote a variable with an upper-case, X, and value of that variable by the same lower-case, x. We denote a set of variables by upper-case bold-face, $\mathbf{Z}$, and we use the corresponding lower-case bold-face, $\mathbf{z}$, to denote an assignment of value to each variable in the set. In this paper, we only deal with discrete random variables. We denote the conditional independence of the variable X and Y given $\mathbf{Z}$, in some distribution P by $X \perp_P Y|\mathbf{Z}$. Similarly, we write $X \perp_{\mathcal{G}} Y|\mathbf{Z}$ if X and Y are d-separated by $\mathbf{Z}$ in the DAG $\mathcal{G}$.

A Markov blanket $\mathbf{M}_T$ of the T is is any set of variables such that T is conditionally independent of all the remaining variables given $\mathbf{M}_T$. A Markov boundary, $\mathbf{MB}_T$, of T is any Markov blanket such that none of its proper subsets is a Markov blanket of T. In general, in a Baysesian network $< \mathcal{G}, P >$, we would want an edge to mean a direct dependency. As we know, the faithfulness entails this :

Definition 1 *Suppose we have a joint probability distribution P of the random variables in some set $\mathbf{U}$ and a DAG $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$. We say that $< \mathcal{G}, P >$ satisfies the faithfulness condition if, based on the Markov condition, G entails all and only conditional independencies in P.*

Theorem 1 *Suppose $< \mathcal{G}, P >$ satisfies the faithfulness condition. Then for each variable X, the set of parents, children of X, and parents of children of X is the unique Markov boundary.*

A proof can be found for instance in (Neapolitan, 2004). A *spouse* of T is a another parent of a T's child node. We denote by $\mathbf{PC}_T$, the unique set of parents and children of T in $\mathcal{G}$ when $< \mathcal{G}, P >$, satisfies the faithfulness condition. Otherwise, $\mathbf{PC}_X^{\mathbf{U}}$ will denote the unique set of the variables that remains dependent on X conditioned on any set $\mathbf{Z} \in \mathbf{U} \setminus \{X, Y\}$.

3. Some problems with constraint-based methods

Constraint-based (CB for short) procedures systematically check the data for independence relationships to infer the structure. The association between two variables X and Y given a conditioning set $\mathbf{Z}$ is a measure of the strength of the dependence with respect to the data base $\mathcal{D}$. It is usually implemented with a statistical measure of association (e.g. χ^2, G^2). CB methods have the advantage of possessing clear stopping criteria and deterministic search procedures. On the other hand, they are prone to several instabilities : namely if a mistake is made early on in the search, it can lead to incorrect edges which may in turn lead to bad decisions in the future, which can lead to even more incorrect edges. This instability has the potential to cascade, creating many errors in the final graph (Dash *et al.*, 2003).

Insufficient data presents a lot of problems when working with statistical inference techniques like the independence test mentioned earlier. This occurs typically when the expected counts in the contingency table are small. The decision of accepting or rejecting the null hypothesis depends implicitly upon the degree of freedom which increases exponentially with the number of variables in the conditional set. So the larger the size of the conditioning test, the less accurate are the estimates of conditional probabilities and hence the less reliable are the independence tests. Another difficulty arises when true- or almost-deterministic relationships (ADR) are observed among the variables. Loosely speaking, a relationship is said to be almost deterministic when the fraction of tuples that violate the deterministic dependency is at most equal to some threshold. True DR are source of unfaithfulness but the existence of ADR among variables doesn't invalidate the faithfulness assumption. Several proposals have been discussed in the literature in order to reduce the cascading effect of early errors that causes many errors to be present in the final graph. The general idea is to keep the size of the conditional sets as small as possible in the curse of the learning process. Another idea is to reduce the degree of freedom of the statistical conditional independence test by some ways. The aim is twofold : to improve the data efficiency and to allow an early detection of ADR. Theses strategies are not discussed here for conciseness, see (Yilmaz *et al.*, 2002; Luo, 2006; Aussem *et al.*, 2007; Rodrigues de Morais *et al.*, 2008) for instance.

4. New method

In this section, we present our MB learning algorithm called MBOR. Like PCMB (Peña *et al.*, 2007) and MMMB (Tsamardinos *et al.*, 2006), MBOR takes a divide-

and-conquer approach that breaks the problem of identifying $\mathbf{MB}_T$ into two subproblems : first, identifying $\mathbf{PC}_T$ and, second, identifying the parents of the children (the spouses) of T. Hence its data-efficiency according to Peña et al. MBOR stands for "Markov Boundary search using the OR condition". This "OR condition" is the **key difference** between MBOR and all the above mentioned correct divide-and-conquer algorithms : two variables *X* and *Y* are considered as neighbors with MBOR if $Y \in PC_X$ OR $X \in PC_Y$, contrary to the other proposals, e.g., MMMB (Tsamardinos *et al.*, 2006), PCMB (Peña *et al.*, 2007), that apply the AND condition to guarantee correctness. Clearly, the OR condition makes it easier for true positive nodes to enter the Markov boundary. Hence the name and the practical efficiency of our algorithm. MBOR is designed to scale up to hundreds of thousands of variables in reasonable time just as PCMB (Peña *et al.*, 2007) does. Moreover, according to Peña et al., this divide-and-conquer approach is supposed to be more data efficient than IAMB (Tsamardinos *et al.*, 2003) and its variants, e.g., Fast-IAMB (Yaramakala, 2004) and Interleaved-IAMB (Yaramakala *et al.*, 2005), because $\mathbf{MB}_T$ can be identified by conditioning on sets much smaller than those used by IAMB. IAMB and its variants seek directly the minimal subset of **U** (the full set) that renders the rest of **U** independent of T, given $\mathbf{MB}_T$.

MBOR (Algorithm 1) works in three steps and it is based on three subroutines called *MBtoPC, PCS* and *MBS* (Algorithms 2-4). Another desirable characteristic of MBOR is to keep the size of the conditional sets to the minimum possible (less than 2) without sacrificing the performance. However, this comes at the expense of the simplicity and legibility of the overall procedure, compared to IAMB for instance. In phase 1, MBOR calls MBS to extract **PCS**, a superset for the parents and children, and **SPS**, a superset for the target spouses (parents of children). Filtering reduces as much as possible the number of variables before proceeding to the MB discovery. In PCS and MBS, the size of the conditioning set in the tests is severely restricted (PCS at lines 3 and 6, MBS at lines 6 and 12). As discussed before, conditioning on larger sets of variables would increase the risk of missing variables that are weakly associated to the target. Phase II finds the parents and children in the restricted set of variables using the OR condition. Therefore, all variables that have T in their vicinity are included in $\mathbf{PC}_T$ (lines 5-7 of MBOR). Phase 3 identifies the target's spouses in **MBS** in exactly the same way PCMB does (Peña *et al.*, 2007).

MBtoPC (Algorithm 2) implements a correct Parents and Children learning procedure. It works in two steps. First, a "weak" MB learner called *CorrectMB* is used at line 1 to output a candidate MB. CorrectMB may be implemented by any correct and fast MB algorithm of the IAMB family. In our implementation, we use Inter-IAMB for its simplicity and performance (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 second step (lines 3-6) of *MBtoPC* removes the spouses of the target. *PCS*(T) aims to output a super set for $\mathbf{MB}_T$ based a scalable and highly data-efficient manner. The correctness of this procedure under the faithfulness condition is guarantied. The proof is provided in the next section.

Algorithm 1 *MBOR*

Require: T : target ; D : data set (**U** is the set of variables)
Ensure: **MB** = Markov boundary of T

Phase I : *Find MB superset (MBS)*
$[\mathbf{PCS}, \mathbf{SS}] = MBS(T, D)$
$\mathbf{MBS} = \mathbf{PCS} \cup \mathbf{SS}$
$\mathcal{D} = \mathcal{D}(\mathbf{MBS})$ *i.e., remove from data set all variables in* ***U/MBS***

Phase II : *Find parents and children of the target*
$\mathbf{PC} = \mathbf{MBtoPC}(T, \mathcal{D})$
for all $X \in \mathbf{PCS} \setminus \mathbf{PC}$ **do**
  if $T \in \mathbf{MBtoPC}(X, \mathcal{D})$ **then**
    $\mathbf{PC} = \mathbf{PC} \cup X$
  end if
end for

Phase III : *Find spouses of the target*
$\mathbf{SP} = \emptyset$
for all $X \in \mathbf{PC}$ **do**
  for all $Y \in \mathbf{MBtoPC}(X, D) \setminus \{\mathbf{PC} \cup T\}$ **do**
    find $\mathbf{Z} \subset \mathbf{MBS} \setminus \{T \cup Y\}$ so that $T \perp Y | \mathbf{Z}$
    if $(T \not\perp Y | \mathbf{Z} \cup X)$ **then**
      $\mathbf{SP} = \mathbf{SP} \cup Y$
    end if
  end for
end for
$\mathbf{MB} = \mathbf{PC} \cup \mathbf{SP}$

Algorithm 2 *MBtoPC*

Require: T : target ; D : data set
Ensure: **PC** : Parents and children of T

$\mathbf{MB} = \mathbf{CorrectMB}(T, D)$
$\mathbf{PC} = \mathbf{MB}$
for all $X \in \mathbf{MB}$ **do**
  if $\exists \mathbf{Z} \subset (\mathbf{MB} \setminus X)$ such that $T \perp X \mid \mathbf{Z}$ **then**
    $\mathbf{PC} = \mathbf{PC} \setminus X$
  end if
end for

Algorithm 3 *PCS*

Require: T : target ; D : data set (**U** is the set of variables)
Ensure: **PCS** : PC superset of T

Phase I : *Remove X if* $T \perp X$
PCS $= \mathbf{U} \setminus T$
for all $X \in$ **PCS do**
  if $(T \perp X)$ **then**
    PCS $=$ **PCS** $\setminus X$
    dSep$(X) = \emptyset$
  end if
end for

Phase II :*Remove X if* $T \perp X|Y$
for all $X \in$ **PCS do**
  for all $Y \in$ **PCS** $\setminus X$ **do**
    if $(T \perp X \mid Y)$ **then**
      PCS $=$ **PCS** $\setminus X$
      dSep$(X) = Y$
    end if
  end for
end for

Algorithm 4 *MBS*

Require: T : target ; D : data set (**U** is the set of variables)
Ensure: [**PCS**, **SPS**], Markov boundary superset of T

Phase I : *Find parents and children superset (PCS)*
PCS $= PCS(T, D)$

Phase II : *Find spouses superset (SPS)*
SPS $= \emptyset$
for all $X \in$ **PCS do**
  SPS$_X = \emptyset$
  for all $Y \in \mathbf{U} \setminus \{T \cup PCS\}$ **do**
    if $(T \not\perp Y|$**dSep**$(Y) \cup X)$ **then**
      SPS$_X =$ **SPS**$_X \cup Y$
    end if
  end for
  for all $Y \in$ **SPS**$_X$ **do**
    for all $Z \in$ **SPS**$_X \setminus Y$ **do**
      if $(T \perp Y|X \cup Z)$ **then**
        SPS$_X =$ **SPS**$_X \setminus Y$
      end if
    end for
  end for
  SPS $=$ **SPS** $\cup$ **SPS**$_X$
end for

5. Proof of correctness under faithfulness condition

Several intermediate theorems are required before we demonstrate MBOR's correctness under faithfulness condition. Indeed, as **MBS** is a subset of **U**, a difficulty arises : a marginal distribution $P^{\mathbf{V}}$ of $\mathbf{V} \subset \mathbf{U}$ may not satisfy the faithfulness condition with any DAG even if $P^{\mathbf{U}}$ does. This is an example of embedded faithfulness, which is defined as follow :

Definition 2 *Let P be a distribution of the variables in $\boldsymbol{V}$ where $\boldsymbol{V} \subset \boldsymbol{U}$ and let $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$ be a DAG. $< \mathcal{G}, P >$ satisfies the embedded faithfulness condition if $\mathcal{G}$ entails all and only the conditional independencies in P, for subsets including only elements of $\boldsymbol{V}$.*

We obtain embedded faithfulness by taking the marginal of a faithful distribution as shown by the next theorem :

Theorem 1 *Let P be a joint probability of the variables in $\boldsymbol{U}$ with $\boldsymbol{V} \subseteq \boldsymbol{U}$ and $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$. If $< \mathcal{G}, P >$ satisfies the faithfulness condition and $P^{\boldsymbol{V}}$ is the marginal distribution of $\boldsymbol{V}$, then $< \mathcal{G}, P^{\boldsymbol{V}} >$ satisfies the* embedded *faithful condition.*

The proof can be found in (Neapolitan, 2004). Note that every distribution doesn't admit an embedded faithful representation. This property is useful to prove the correctness of our MBOR under the faithfulness condition. Let $\mathbf{PC}_X^{\mathbf{U}}$ denote the variables $Y \in \mathbf{U}$ such that there is no set $\mathbf{Z} \in \mathbf{U} \setminus \{X, Y\}$ such that $X \perp_P Y | \mathbf{Z}$. If $< \mathcal{G}, P >$ satisfies the faithfulness condition, $\mathbf{PC}_X^{\mathbf{U}}$ are the parents and children of X in $\mathbf{U}$. Otherwise, $\mathbf{PC}_X^{\mathbf{U}}$ is the unique set of the variables that remains dependent on X conditioned on any set $\mathbf{Z} \in \mathbf{U} \setminus \{X, Y\}$.

Theorem 2 *Let $\boldsymbol{U}$ be a set of random variables and $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$. If $< \mathcal{G}, P >$ satisfies the faithfulness condition, then every target T admits a unique Markov boundary $\boldsymbol{MB}_T^{\boldsymbol{U}}$. Moreover, for all $\boldsymbol{V}$ such that $\boldsymbol{MB}_T^{\boldsymbol{U}} \subseteq \boldsymbol{V} \subseteq \boldsymbol{U}$, T admits a unique Markov boundary over $\boldsymbol{V}$ and $\boldsymbol{MB}_T^{\boldsymbol{V}} = \boldsymbol{MB}_T^{\boldsymbol{U}}$.*

Proof : If $\mathbf{MB}_T^{\mathbf{U}}$ is the Markov boundary of T in $\mathbf{U}$, then T is independent of $\mathbf{V} \setminus \{\mathbf{MB}_T^{\mathbf{U}} \cup T\}$ conditionally on $\mathbf{MB}_T^{\mathbf{U}}$ so $\mathbf{MB}_T^{\mathbf{U}}$ is a Markov blanket in $\mathbf{V}$. Moreover, none of the proper subsets of $\mathbf{MB}_T^{\mathbf{U}}$ is a Markov blanket of T in $\mathbf{V}$, so $\mathbf{MB}_T^{\mathbf{U}}$ is also a Markov boundary of T in $\mathbf{V}$. So if it is not the unique MB for T in $\mathbf{V}$ there exists some other set $\mathbf{S}_T$ not equal to $\mathbf{MB}_T^{\mathbf{U}}$, which is a MB of T in $\mathbf{V}$. Since $\mathbf{MB}_T^{\mathbf{U}} \neq \mathbf{S}_T$ and $\mathbf{MB}_T^{\mathbf{U}}$ cannot be a subset of $\mathbf{S}_T$, there is some $X \in \mathbf{MB}_T^{\mathbf{U}}$ such that $X \notin \mathbf{S}_T$. Since $\mathbf{S}_T$ is a MB for T, we would have $T \perp_P X | \mathbf{S}_T$. If X is a parent or child of T, we would not have $T \perp_{\mathcal{G}} X | \mathbf{S}_T$ which means we would have a conditional independence which is not entailed by d-separation in $\mathcal{G}$ which contradicts the faithfulness condition. If X is a parent of a child of T in $\mathcal{G}$, let Y be their common child in $\mathbf{U}$. If $Y \in \mathbf{S}_T$ we again

would not have $T \perp_{\mathcal{G}} X|\mathbf{S}_T$. If $Y \notin \mathbf{S}_X$ we would have $T \perp_P Y|\mathbf{S}_T$ because $\mathbf{S}_T$ is a MB of T in $\mathbf{V}$ but we do not have $T \perp_{\mathcal{G}} Y|\mathbf{S}_X$ because T is a parent of Y in G. So again we would have a conditional independence which is not a d-separation in $\mathcal{G}$. This proves that there can not be such set $\mathbf{S}_X$. □

Theorem 3 *Let **U** be a set of random variables and T a target variable. Let $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$ be a DAG such that $< \mathcal{G}, P >$ satisfies the faithfulness condition. Let **V** such that $\boldsymbol{MB}_T^{\boldsymbol{U}} \subseteq \boldsymbol{V} \subseteq \boldsymbol{U}$ then, $\boldsymbol{PC}_T^{\boldsymbol{V}} = \boldsymbol{PC}_T^{\boldsymbol{U}}$.*

Proof : Clearly $\mathbf{PC}_T^{\mathbf{U}} \subset \mathbf{PC}_T^{\mathbf{V}}$ as $\mathbf{MB}_T^{\mathbf{U}} \subset \mathbf{V} \subseteq \mathbf{U}$. If $X \in \mathbf{PC}_T^{\mathbf{V}}$ and $X \notin \mathbf{PC}_T^{\mathbf{U}}$, $\exists \mathbf{Z} \subset \mathbf{MB}_T^{\mathbf{U}} \setminus X$ such that $T \perp_P X|\mathbf{Z}$ because all non adjacent nodes may be d-separated in $\mathcal{G}$ by a subset of its Markov boundary. As $\mathbf{MB}_T^{\mathbf{U}} = \mathbf{MB}_T^{\mathbf{V}}$ owing to Theorem 2, so X and T can be d-separated in $\mathbf{V} \setminus \{X, T\}$. Therefore, X cannot be adjacent to T in $\mathbf{V}$. □

Theorem 4 *Let **U** be a set of random variables and T a target variable. Let $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$ be a DAG such that $< \mathcal{G}, P >$ satisfies the faithfulness condition. Let **V** such that $\boldsymbol{MB}_T^{\boldsymbol{U}} \subseteq \boldsymbol{V} \subseteq \boldsymbol{U}$. Under the assumptions that the independence tests are correct and that the learning database is an independent and identically distributed sample from P, MBtoPC(T,**V**) returns $\boldsymbol{PC}_T^{\boldsymbol{U}}$. Moreover, let $X \in \boldsymbol{V} \setminus T$, then T is in the output of MBtoPC$(X, \boldsymbol{V}, \mathcal{D})$ iff $X \in \boldsymbol{PC}_T^{\boldsymbol{U}}$.*

Proof : We prove first that MBtoPC(T,**V**) returns $\mathbf{PC}_T^{\mathbf{U}}$. In the first stage of MBtoPC, InterIAMB(T,**V**) seeks a minimal set $\mathbf{S}_T \in \mathbf{V} \setminus T$ that renders the $\mathbf{V} \setminus \mathbf{S}_T$ independent of T conditionally on $\mathbf{S}_T$. This set is unique owing to Theorem 3, therefore $\mathbf{S}_T = \mathbf{MB}_T^{\mathbf{V}} = \mathbf{MB}_T^{\mathbf{U}}$. In the backward phase, MBtoPC removes the variables $X \in \mathbf{MB}_T^{\mathbf{V}}$ such that $\exists \mathbf{Z} \subset (\mathbf{MB}_T^{\mathbf{V}} \setminus X)$ for which $T \perp X \mid \mathbf{Z}$. These variables are the spouses of T in $\mathcal{G}$, so MBtoPC(T,**V**) returns $\mathbf{PC}_T^{\mathbf{U}}$. Now, if $X \notin \mathbf{PC}_T^{\mathbf{U}}$ then $X \notin \mathbf{PC}_T^{\mathbf{V}}$ owing to Theorem 4. So there is a set $\mathbf{Z} \subset \mathbf{V} \setminus \{X, Y\}$ such that $T \perp X \mid \mathbf{Z}$. Therefore, X cannot be in the output of MBtoPC(T,**V**). □

Theorem 5 *Under the assumptions that the independence tests are correct and that the learning database is an independent and identically distributed sample from a probability distribution P faithful to a DAG $\mathcal{G}$, MBOR(T) returns $\boldsymbol{MB}_T^{\boldsymbol{U}}$.*

Proof : Let **MBS** be the output of MBS$(T, \mathbf{U}, \mathcal{D})$. It is straightforward to show that $\mathbf{MB}_T^{\mathbf{U}} \subset \mathbf{MBS}$. So the Markov boundary of T in **MBS** is that of **U** owing to Theorem 3 so the problem is well defined. In Phase II at line 6, if T is in the output of MBtoPC$(X, \mathbf{V}, \mathcal{D})$ then X should be in the output of MBtoPC$(T, \mathbf{V}, \mathcal{D})$ owing to Theorem 5. So phase II ends up with the $\mathbf{PC}_T^{\mathbf{U}}$. In Phase III, lines 11-18 identify all and only the spouse of T in $\mathcal{G}$ when the faithfulness condition is asssumed as shown in (Peña *et al.*, 2007). When the assumption doesn't hold anymore for $< \mathcal{G}, P^{`}\mathbf{V} >$, we need to show that a fake spouse will note enter the set **SP**. In phase III line 12, it

is easy to see that MBtoPC$(X, \mathbf{V}, \mathcal{D})$ returns a set $\mathbf{PC}_X^{\mathbf{V}}$ that may differ from $\mathbf{PC}_X^{\mathbf{U}}$. Suppose $Y \notin \mathbf{PC}_X^{\mathbf{U}}$ and Y is in the output of MBtoPC$(X, \mathbf{V}, \mathcal{D})$. This means that there exists at least one active path between X and Y in $\mathcal{G}$ that contains a node in $\mathbf{U} \setminus \mathbf{V}$. At lines 13-14, Y is considered as spouse of T if there is a set $\mathbf{Z} \subset \mathbf{MBS}\backslash\{T \cup Y\}$ so that $T \perp Y|\mathbf{Z}$ and $T \not\perp Y|\mathbf{Z} \cup X$. Therefore, this path in $\mathcal{G}$ should necessarily be of the type $T \rightarrow X \leftarrow A \leftrightsquigarrow B \leftrightsquigarrow Y$ where $\leftrightsquigarrow$ denotes an active path otherwise we would not have $T \not\perp Y|\mathbf{Z} \cup X$. As A is a spouse of T, $A \in \mathbf{MB}_T^{\mathbf{U}}$ and so A is in $\mathbf{V}$. Suppose B is not in $\mathbf{V}$, then $A \in$ still d-separates X and Y so Y cannot be in the output of MBtoPC$(X, \mathbf{V}, \mathcal{D})$ since we found a set $\mathbf{Z} \subseteq \mathbf{V}$ such that $X \not\perp_P Y|\mathbf{Z}$. So Y is included in **S**P at line 15 iff Y is a spouse of T in $\mathbf{U}$. □.

6. Experimental validation

In this section, we compare the performance of InterIAMB, PCMB and MBOR through experiments on synthetic and real databases with very few instances compared to the number of variables. They are written in MATLAB and all the experiments are run on a Intel Core 2 Duo T77500 with 2Gb RAM running Windows Vista. To implement the conditional independence test, we calculate the G^2 statistic as in (Spirtes *et al.*, 2000), under the null hypothesis of the conditional independence. The significance level of the test in all compared algorithms is 0.05 except on the high-dimensional THROMBIN data where it is 0.0001. Our implementation breaks ties at random. All three algorithms are correct under the faithfulness condition and are also scalable. We do not consider MMMB and HITON-MB because we are not interested in any algorithm that does not guarantee the correctness under faithfulness assumption. We do not consider GS because IAMB outperforms it (Tsamardinos *et al.*, 2003). Even if PCMB was also shown experimentally in (Peña *et al.*, 2007) to be more accurate than IAMB and its variants, we consider InterIAMB because it is used as a subroutine in MBOR.

6.1. *Synthetic data*

Figure 1 illustrates the results of our experiments on six common BN benchmarks : BREAST-CANCER or ASIA (8 nodes/8 arcs), INSURANCE (27/52), INSULINE (35/52), ALARM (37/46), HAILFINDER (56/66) and CARPO (61/74). These benchmarks are available from the UCI Machine Learning Repository. All three algorithms have been run on each variable for all data sets. Figure 1 (upper part) summarizes graphically the results in terms of missing and extra nodes in the output of the MB averaged over 10 runs for 200, 500 and 1000 i.i.d. samples. The upper part shows the average false positive and and lower part shows the false negative rates. The overall accuracy is very similar for nodes with Markov boundaries with less than 4 variables. For larger MBs, however, the advantages of MBOR against the other two algorithms are far more noticeable. For instance, MBOR consistently outperforms the other algorithms on variable *IPA* in the INSULINE benchmark as may be seen in Table 2.

Figure 2 (lower part) show the performance for nodes with more than 4 variables. Results are averaged over all the above mentioned benchmarks. As observed, MBOR reduces drastically the average number of false negatives compared to PCMB and InterIAMB (up to 40% on INSULINE). This benefit comes at very little expense : the false positive rate is slightly higher. This is not a surprise as PCMB makes it harder for true positives to enter the output.

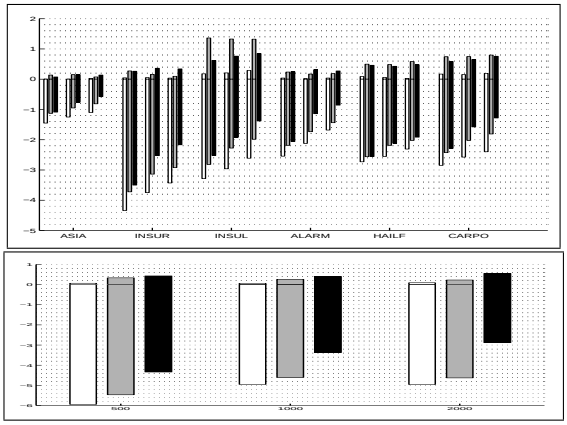

Figure 1. Upper plot : average missing (lower part) and extra (upper part) variables for learning Markov boundaries of all variables of ASIA, INSURANCE, INSULINE, ALARM, HAILFINDER and CARPO networks. The results of PCMB, InterIAMB and MBOR are shaded in white, gray and black respectively. For each benchmark the bars show the results on 200, 500 and 1000 instances respectively. All results are averaged over 10 runs. Lower plot : results are averaged over all benchmarks for nodes that have a MB with more than 4 variables in the MB, for 500, 1000 and 2000 i.i.d. samples.

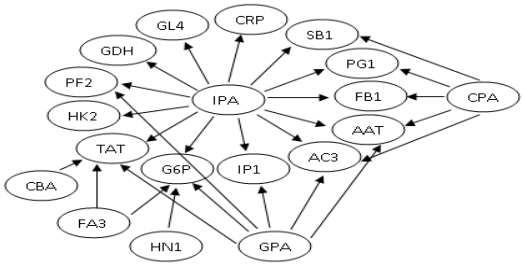

Figure 2. INSULINE benchmark : Markov boundary of the variable *IPA*.

Algorithm	false positive	false negative
PCMB	0.4	11.8
InterIAMB	0	12.6
MBOR	2.1	2.1

Tableau 1. *INSULINE benchmark : number of extra and missing variables for PCMB, Inter-IAMB and MBOR for variable* IPA *run on 1000 instances. Results are averaged over 10 runs.*

6.2. *Real data*

In this section, we assess the performance of the probabilistic classification using the feature subset output by MBOR. To this purpose, we consider several categorical data bases from the UCI Machine Learning Repository in order to evaluate the accuracy of MBOR against InterIAMB and PCMB. The database description and the results of the experiments with the Car Evaluation, Molecular Biology, SPECT heart, Tic-Tac-Toe databases, Wine and Waveform are shown in Table 1. Performance is assessed by hit rate (correct classification rate), relative absolute error (R.A.E.), and Kappa Statistics obtained by 10-fold cross-validation. Kappa can be thought of as the chance-corrected proportional agreement, and possible values range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to -1 (complete disagreement). As may be seen, the classification performance by naive Bayes classifier on the features selected by MBOR has always outperformed that of InterIAMB and PCMB by a noticeable margin, especially on the Molecular Biology database (C).

6.3. *Real data : Thrombin database*

Our last experiments demonstrate the ability of MBOR to solve a real world FSS problem involving thousands of features. We consider the THROMBIN database which was provided by DuPont Pharmaceuticals for KDD Cup 2001. It is exemplary of a real drug design (Cheng *et al.*, 2002). The training set contains 1909 instances characterized by 139,351 binary features. The accuracy of a Naive Bayesian classifier was computed as the average of the accuracy on true binding compounds and the accuracy on true non-binding compounds on the 634 compounds of the test set. As the data is unbalanced, the accuracy is calculated as the average of true positive rate and the true negative rate. A significance level of 0.0001 avoids better than 0.01 the spurious dependencies that may exist in the data due to the large number of features. MBS with $\alpha = 0.0001$ returns a set of 21 variables, SPS select 39 variables and MBOR outputs 5 variables on average in about 3h running time.

Note shown here, the 10 runs return each time a different MB, all of them containing 5 features. They mostly differ by one or two variables. MBOR scores between 36% (really bad) to 66% with an average 53% on average which seems really de-

Data Sets	*Accuracy*	*Algorithms*		
		I.IAMB	PCMB	MBOR
A	Hit Rate	79.11%	79.11%	85.36%
	Kappa	0.5204	0.5204	0.6665
	R.A.E.	56.59%	56.59%	49.88%
B	Hit Rate	76.40%	79.40%	84.27%
	Kappa	0.2738	0	0.4989
	R.A.E.	77.83%	93.92%	71.71%
C	Hit Rate	74.01%	74.01%	95.61%
	Kappa	0.5941	0.5941	0.9290
	R.A.E.	56.94%	56.94%	10.27%
D	Hit Rate	67.95%	67.95%	72.44%
	Kappa	0.1925	0.1951	0.3183
	R.A.E.	85.62%	85.75%	82.53%
E	Hit Rate	76.42%	76.42%	81.32%
	Kappa	0.6462	0.6462	0.7196
	R.A.E.	44.50%	44.50%	29.43%
F	Hit Rate	94.38%	94.38%	98.88%
	Kappa	0.9148	0.9148	0.9830
	R.A.E.	19.08%	19.08%	2.67 %

Tableau 2. *10-fold cross-validation performance by naive Bayes classifier on the selected feautres ontained with Inter-IAMB, PCMB and MBOR. A : Car Evaluation (1728,6). B : SPECT Heart (267,22). C : Splice-junction Gene Sequences (3190,61). D : Tic-Tac-Toe Endgame (958,9). E : Waveform - Version 1 (5000,21). F : Wine (178,13).*

ceiving compared to PCMB and IAMB that achieves respectively 63% and 54% as shown in (Peña *et al.*, 2007). Nonetheless, MBOR is highly variable and was able to identify 3 different MBs that outperform those found by IAMB and 90% of those by PCMB. For instance, the MB which scores 66% contains the two variables 20973, 63855. These two variables, when used conjunctly, score 66,9% which is impressive according to (Cheng *et al.*, 2002; Peña *et al.*, 2007) for such a small feature set. Note that a MB with the four features obtained by the winner of KDD cup 2001 scores 67% accuracy.

The execution time was not reported as it is too dependent on the specific implementation. We were unable to run PCMB on the Thrombin database in reasonable time with *our* MATLAB implementation. On synthetic data, MBOR runs (say) 30% faster than PCMB.

7. Discussion and conclusion

We discussed simple solutions to improve the data efficiency of current constraint-based Markov boundary discovery algorithms. We proposed a novel approach called MBOR. Our experimental results show a clear benefit in several situations : densely connected DAGs, weak associations or approximate functional dependencies among the variables.

8. Bibliographie

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.

Cheng J., Hatzis C., Hayashi H., Krogel M., Morishita S., Page D., Sese J., « KDD Cup 2001 Report », *ACM SIGKDD Explorations*, p. 1-18, 2002.

Dash D., Druzdzel M. J., « Robust Independence Testing for Constraint-Based Learning of Causal Structure. », *UAI*, p. 167-174, 2003.

Guyon I., Elisseeff A., « An Introduction to Variable and Feature Selection. », *Journal of Machine Learning Research*, vol. 3, p. 1157-1182, 2003.

Luo W., « Learning Bayesian Networks in Semi-deterministic Systems », *Canadian Conference on AI*, p. 230-241, 2006.

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

Nilsson R., Peña J., Björkegren J., Tegnér J., « Consistent Feature Selection for Pattern Recognition in Polynomial Time », *Journal of Machine Learning Research*, vol. 8, p. 589-612, 2007.

Peña J., Nilsson R., Björkegren J., Tegnér J., « Towards Scalable and Data Efficient Learning of Markov Boundaries », *International Journal of Approximate Reasoning*, vol. 45, n° 2, p. 211-232, 2007.

Rodrigues de Morais S., Aussem A., Corbex M., « Handling almost-deterministic relationships in constraint-based Bayesian network discovery : Application to cancer risk factor identification », *16th European Symposium on Artificial Neural Networks ESANN'08*, p. 101-106, 2008.

Spirtes P., Glymour C., Scheines R., *Causation, Prediction, and Search*, 2 edn, The MIT Press, 2000.

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

Tsamardinos I., Brown L. E., Aliferis C. F., « The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. », *Machine Learning*, vol. 65, n° 1, p. 31-78, 2006.

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.

Yilmaz Y. K., Alpaydin E., Akin H. L., Bilgiç T., « Handling of Deterministic Relationships in Constraint-based Causal Discovery. », *Probabilistic Graphical Models*, 2002.