Analysis of Nasopharyngeal Carcinoma Data with a Novel Bayesian Network Learning Algorithm

Alexandre Aussem, Sergio Rodrigues de Morais, Marilys Corbex

Analysis of Nasopharyngeal Carcinoma Data with a Novel Bayesian Network Learning Algorithm

Alex Aussem, Sergio Rodrigues de Morais, Marilys Corbex

***Abstract*— Learning the structure of a bayesian network from a data set is NP-hard. In this paper, we discuss a novel heuristic called Polynomial Max-Min Skeleton (PMMS) developped by Tsamardinos et al. in 2005. PMMS was proved by extensive empirical simulations to be an excellent trade-off between time and quality of reconstruction compared to all constraint based algorithms, especially for the smaller sample sizes. Unfortunately, there are two main problems with PMMS : it is unable to deal with missing data nor with datasets containing functional dependencies between variables. In this paper, we propose a way to overcome these problems. The new version of PMMS is first applied on standard benchmarks to recover the original structure from data. The algorithm is then applied on the Nasopharyngeal carcinoma (NPC) made up from only 1289 uncomplete records in order to shed some light into the statistical profile of the population under study.**

***Index Terms*— Basyesian networks, medical decision support systems, epidemiology**

I. Introduction

THE last twenty years have brought considerable advances in the field of computer-based medical systems. These advances have resulted in noticeable improvements in medical care, support for medical diagnosis and computer assisted discovery. Decision support systems based on Bayesian Networks (BN) have proven to be valuable tools that help practitioners in facing challenging medical problems, such as diagnosis [1] by identifying the relevant factors (also called features) involved in the disease, illness or disorders under study from experimental data [2]. In this paper, we discuss efforts to apply a new BN learning method to a real world epidemiological problem, namely the Nasopharyngeal Carcinoma (NPC). The objective is to investigate the role of various environmental factors in the aetiology of NPC. A multi-center case-control study has been undertaken in 2004 by the International Agency for Research on Cancer (IARC) in the Maghreb (Morocco, Algeria and Tunisia), the endemic region of North Africa. The experiments presented in the paper focus on environmental risk factors of NPC in the Maghrebian population. The specific aims are : (1) to provide a statistical profile of the recruited population, (2) to help indentify the important risk factors involved in NPC, and (3) to shed insight on the applicability and limitations of BN methods on small epidemiological data sets obtained from questionnaires.

Bayesian networks, also called belief networks or causal networks [3]–[5], are probabilistic graphical models that offer a coherent and intuitive representation of uncertain domain knowledge. Formally, BN are directed acyclic graphs (DAG) modelling probabilistic dependencies among variables. The graphical part of BN reflects the structure of a problem (usually a graph of causal dependencies in the modeled domain), while local interactions among neighboring variables are quantified by conditional probability distributions. One of the main advantages of BN over other AI schemes for reasoning under uncertainty is that they readily combined expert judgment with knowledge extracted from the data within the probabilistic framework. Often, clinics, hospitals or epidemiologists collect patient data which over time allow for discovering statistical dependencies and, when incorporated into a model, provide a valuable enhancement to the subjective knowledge obtained from expert. Another advantage is that they represent graphically the (possibly causal) independence relationships that may exist in a very parsimonious manner.

Learning a Bayesian network from data requires both identifying the model structure $\mathcal{G}$ and identifying the corresponding set of model parameter values. Given a fixed structure, however, it is straightforward to estimate the parameter values. As a result, research on the problem of learning Bayesian networks from data is focused on methods for identifying one or more "good" DAG structures from data. All independence constraints that hold in the joint distribution represented by any Bayesian network with structure $\mathcal{G}$ can be identified from the structure itself under certain conditions. However, the problem of learning the most probable *a posteriori* Bayesian network (BN) from data is worst-case NP-hard [6].

Constraint-based (CB) causal discovery searches a database for independence relations and constructs graphical structures called "patterns" which represent a class of statistically indistinguishable DAGs. This method contrasts to those based on Bayesian concepts, which typically reduce to a search-and-score procedure on the space of DAGs. Both CB and Bayesian approaches have advantages and disadvantages. Constraint-based approaches are relatively quick, deterministic, and have a well defined stopping criterion; however, they rely on an arbitrary significance level to test for independence, and they can be unstable in the sense that an error early on in the search can have a cascading effect that causes many errors to be present in the final graph [7]. Bayesian approaches have the advantage of being able to flexibly incorporate users' background knowledge in the form of prior probabilities over the structures and they are also capable of dealing with incomplete records in the database (e.g. EM technique). Bayesian approaches are however slow to converge. When data sets are small, the relative benefits of the two approaches are unclear.

Very recently, a new powerful polynomial constraint-based

url: http://comad.univ-lyon1.fr. Email: aaussem@univ-lyon1.fr
Supported by "Ligue contre le Cancer - Comité du Rhône - France"

learning algorithm in $\mathcal{O}(n^4)$ has been proposed by L. Brown et al. [8], [9] particularly well suited to smaller data sets. The algorithm, known as Polynomial Max-Min Skeleton (PMMS), learns the BN skeleton, i.e., the graph of the BN without regard to the direction of the edges, that is, the most difficult task. PMMS employs a smart search strategy for identifying conditional dependencies that exhibits better sample utilization compared to other procedures (e.g. TPDA [10], PC [3]). Although very encouraging results have been reported with PMMS with smaller datasets, it suffer from two difficulties : 1) it is unable to deal with missing data, that is, when some attribute values are reported as unknown, and 2) the method fails to reconstruct the skeleton when some functional dependencies exists among groups of variables. A functional dependency (written $\mathbf{X} \rightarrow Y$) is a constraint between a set of variables, such that, given the value for all $X_j \in \mathbf{X}$, one can functionally (and deterministically) determine the corresponding value of Y. As our NPC dataset is extracted from a questionnaire, it is incomplete and it possibly holds hidden FD.

In this paper, we discuss in detail the way we modify the standard PMMS algorithm to overcome these problems in Section III. In Section IV, we carefully evaluate the strengths and limitations of the new PMMS version on synthetic data with missing attributes generated from the well known Asia and Asia8 benchmarks that include FDs. The new PMMS was implemented in Matlab with the BNT Toolbox [11] and the Toolbox BNTSLP [12]. The method is then applied, in Section V, to the NPC data. The aim is to infer the statistical profile of the recruited population from the questionnaire and to clarify the role of the environment in the aetiology of NPC.

II. Background

For the paper to be accessible to those outside the domain, we recall first the principle of BN. 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. A Bayesian network (BN) is a tuple $< \mathcal{G}, P >$, where $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ is a directed acyclic graph (DAG) with nodes representing the random variables $\mathcal{V}$ and P a joint probability distribution on $\mathcal{V}$. In addition, $\mathcal{G}$ and P must satisfy the Markov condition: every variable, $X \in \mathcal{V}$, is independent of any subset of its non-descendant variables (ND_X) conditioned on the set of its parents $\mathbf{Pa}_i^{\mathcal{G}}$ [5], [13]. 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 $Dep_P(X;Y|\mathbf{Z})$. From the markov condition, (i.e., $\forall X_i, Ind_P(X_i; ND_X|\mathbf{Pa}_i^{\mathcal{G}})$), it is easy to prove that the joint probability distribution P on the variables on $\mathcal{V}$ can be factored as follows :

$$P(\mathcal{V}) = P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\mathbf{Pa}_i^{\mathcal{G}}) \qquad (1)$$

Equation 1 allows a parsimonious decomposition of the joint distribution P. All independence constraints that necessarily hold in the joint distribution represented by any Bayesian network with structure $\mathcal{G}$ can be identified by the d-separation criterion of Pearl (1988) applied to $\mathcal{G}$. In particular, two nodes X and Y are said to be d-separated in a DAG $\mathcal{G}$ given a set of nodes $\mathbf{Z}$ if and only if there is no $\mathbf{Z}$-active path in $\mathcal{G}$ between X and Y; an $\mathbf{Z}$-active path is a simple path for which each node W along the path either (1) has converging arrows (i.e., $A \rightarrow W \leftarrow B$) and W or a descendant of W is in $\mathbf{Z}$ or (2) does not have converging arrows and W is not in $\mathbf{Z}$. By simple, we mean that the path never passes through the same node twice. If two nodes are not d-separated given some set, we say that they are d-connected given that set. We use $Dsep_{\mathcal{G}}(X;Y|\mathbf{Z})$ to denote the assertion that DAG $\mathcal{G}$ imposes the constraint, via d-separation, that for all values z of the set Z, X is independent of Y given $Z = z$.

For illustration purpose, consider the Chest clinic network (also known as the "Asia network") used in [14] and shown in Figure 1. Asia network is small BN for fictitious medical domain, relating whether a patient has tuberculosis, lung cancer or bronchitis, to their X-ray, dyspnea, visit-to-Asia and smoking status. There are for instance two simple chains from A to D : $[A, T, O, D]$ and $[A, T, O, L, S, B, D]$. The first one is open but the second is blocked by O because $T \rightarrow O \leftarrow L$ is convergent. By conditioning upon O ($O \in \mathbf{Z}$), the opposite becomes true. Once the conditional probabilities are set, a doctor may for example be interested in estimating the probability of lung cancer from indications of smoking and X-ray status.

If P is Markov with respect to a DAG $\mathcal{G}$, $Dsep_{\mathcal{G}}(X;Y|\mathbf{Z})$ implies $Ind_P(X_i; ND_X|\mathbf{Pa}_i^{\mathcal{G}})$. Similarly, we say that P is *faithful* with respect to $\mathcal{G}$ if $Ind_P(X_i; ND_X|\mathbf{Pa}_i^{\mathcal{G}})$ implies $Dsep_{\mathcal{G}}(X;Y|\mathbf{Z})$. If P is both Markov and faithful with respect to $\mathcal{G}$, we say that P is perfect with respect to $\mathcal{G}$. We say that P is DAG perfect if there exists a DAG $\mathcal{G}$ such that P is perfect with respect to $\mathcal{G}$. We say that two DAGs $\mathcal{G}$ and $\mathcal{G}'$ are equivalent if the two sets of distributions included by $\mathcal{G}$ and $\mathcal{G}'$ are the same (see for instance [5] for further details).

It may be observed in the Asia network that a functional dependency exists because "Lung Cancer" and "Tuberculosis" functionally determine "Lung Cancer OR Tuberculosis". Therefore, we have $Ind_P(O;X|\{T,L\})$ and $Ind_P(O;D|\{T,L\})$ but neither $Desp_P(O;X|\{T,L\})$ nor $Desp_P(O;D|\{T,L\})$ is true. Also, P is unfaithful or not DAG perfect with respect to $\mathcal{G}$ in Fig. 1. The existence of FDs in the data may arise incidentally in smaller data samples. The less data, the more FDs are expected. It is often the case in questionnaire data owing to hidden redundancies in the questions or misunderstanding. As PMMS fails to work properly in the presence of FD, we will see next how to deal with FDs.

III. Polynomial Max-Min Skeleton revisited

In this section, we recall the principles of PMMS before we discuss the modification made to tackle the two problems : missing data and presence of functional dependencies. The following discussion draws strongly on [8], [9].

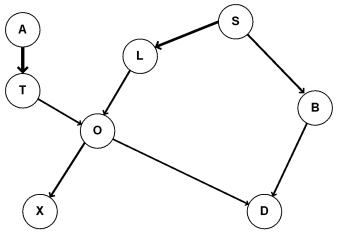

Fig. 1. The Chest-clinic network. A stands for 'visit to Asia', T 'tuberculosis', O 'tuberculosis OR cancer', S 'smoking', B 'bronchitis', D 'dyspnea', X 'X-ray' and L 'lung cancer'.

PMMS assumes the *faithfulness condition*. As all contraint-based algorithm, it exploits the d-separation criterion to construct the BN skeleton. We will denote with $\mathbf{PC}_T^{\mathcal{G}}$ the parents and children of T in the $\mathcal{G}$. This set is unique for all $\mathcal{G}$, such that $< \mathcal{G}, P >$ is faithful and so we will drop the superscript $\mathcal{G}$. We define the minimum association $Assoc(X;Y|\mathbf{Z})$, denoted as $MinAssoc(X;Y|\mathbf{Z})$

$$MinAssoc(X;Y|\mathbf{Z}) = \min_{\mathbf{S}\subseteq\mathbf{Z}} Assoc(X;Y|\mathbf{S}) \qquad (2)$$

i.e., as the minimum association achieved between X and Y over all subsets of $\mathbf{Z}$. $Assoc(X;Y|\mathbf{Z})$, the association between two variables given a conditioning set, can implemented with a number of statistical or information theoretic measures of association, e.g., by the conditional mutual information [10]. In our implementation we prefer a statistically oriented test and we use $1-p$, where p is the p-value returned by the $\chi 2$ test of independence $Ind(X;Y|\mathbf{Z})$ (the smaller the p-value, the higher we consider the association between X and Y) as in [9]. A requirement for $Assoc(X;Y|\mathbf{Z})$ is to return zero when $Ind(X;Y|\mathbf{Z})$.

The Polynomial Max-Min Skeleton algorithm (PMMS) is run on a dataset and returns the skeleton network. PMMS works by calling the Polynomial Max-Min Parents and Children algorithm (PMMPC) for each variable T as the target. PMMPC identifies an approximation of $\mathbf{PC}_T$ given a target node, T, and the data. Once the parents and children set has been discovered for each node, PMMS pieces together the identified edges into the network skeleton. PMMPC(T), the main subroutine of the original PMMS algorithm, discovers the $\mathbf{PC}_T$ using a two-phase scheme. In this paper, we propose to add a third one to handle FD (see Algorithms 1 and 2).

Phase I - The forward phase, variables sequentially enter a candidate parents and children set of T, denoted as CPC, by use of the MAX-MIN heuristic : Select the variable that maximizes the minimum association with T relative to CPC (hence the name of the algorithm). The intuitive justification for the heuristic is to select the variable that remains highly associated with T despite our best efforts to make the variable independent of T. Phase I stops when all remaining variables are independent of the target T given some subset of CPC, i.e., when the maximum minimum association reaches zero. Conditioning on all subsets of CPC to identify the $MinAssoc(X;T|\mathbf{P}C)$ requires an exponential number of calls to $Assoc$. Instead, PMMPC greedily search for the subset $S \in CPC$ that achieves the minimum association between X and T conditioned over the subsets of CPC using the function $GreedyMinAssoc(X;T|\mathbf{CPC}, minval, \mathbf{MinSet})$. Initially $\mathbf{MinSet} = \emptyset$ et $minval = Assoc(X;Y|\emptyset)$. This function augments iteratively MinSet by a single variable. We denote $\min_{x\neq\epsilon}\mathbf{S}$ the smallest x different of ϵ in the set S (if it exists). When data is found to be lacking (i.e. $Assoc = \epsilon$) the growing phase stops and dependence is assumed. However, the value of association is set to the smallest possible value (ϵ). This is discussed further below.

Phase II - The backward phase, PMMPC attempts to remove some of the false positives that may have entered in the first phase. The false positives are removed by testing $Ind_{\mathcal{G}}(X;T|\mathbf{S})$ for all $\mathbf{S} \subseteq \mathbf{Z} \setminus X$. by using again $GreedyMinAssoc$.

Phase III - This phase is new. If a set of variables $\mathbf{X}$ is deterministically related to another variable Y, then each assignment of values $\mathbf{X} = x$ is associated with only one assignment $Y = y$ value. Thus, given an assignment of value to each variable in the set $\mathbf{X}$, one can determine the corresponding value of the values of Y. PPMPC run on target O yields **CPC** = $\{T, L\}$ because X and D are d-separated from O by O's parents $\{T, L\}$. T and L enter first CPC because their associated with O is found to be stronger than $O - X$ and $O - D$. Also, the algorithm fails to include the O's children. To get around this problem, we first test if the association between the current **CPC** and the target T is a FD by calling $IsFuncDep(T;\mathbf{CPC};\mathcal{D})$. In that case, PMMPC is called recursively on the subset $\mathcal{V} \setminus CPC$ ($\mathbf{V}set$ is the forbidden set that is augmented by the current CPC at each recursive call), until no FD is found anymore, and the new CPC will be added to the old one. At the second call to PMMPC, not only T's children but also T's grand-parents will enter CPC. This is not a problem because as PMMS pieces together the identified edges into the network skeleton, the false neighbors will be detected.

It should be noted at this stage that PMMPC and $GreedyMinAssoc$ starts with the smallest conditioning set possible, the empty set, and proceeds with conditioning on larger sets only when the sample allows so. We expect PMMS to better reconstruct the skeleton when the available sample is low relative to the parent and children set sizes compared to other algorithms such as as TPDA [10], PC [3]. To orient the edges of the identified skeleton and further refine the network, various strategies exist. This task is rather easy and will not be discussed here, see [5] for further details.

Algorithm 1 *PMMS*

Require: ***D***: data set
Ensure: ***E*** : skeleton
1: **for all** $X \in \mathbf{V}$ **do**
2: $\quad\mathbf{PC}_X = PMMPC(X, \mathcal{D})$
3: **end for**
4: **for all** $(X, Y) \in \mathbf{V}$ **do**
5: $\quad$**if** $X \in \mathbf{PC}_Y$ & $Y \in \mathbf{PC}_X$ **then**
6: $\quad\quad\mathbf{E} = \mathbf{E} \cup (XY)$
7: $\quad$**end if**
8: **end for**
9: **return E**

Algorithm 2 *PMMPC revisited*

Require: T: target
Vset: variable set s.t. **Vset** $\rightarrow$ T functional dependency. **Vset** $= \emptyset$ at first call of PMMPC
$\mathcal{D}$: data set and ***V*** the variables
Ensure: ***CPC*** : candidate parents and children set of T

Phase I: *Forward*
$CPC = \emptyset$
V=$SetDiff$(***V***,***Vset***);
V=$SetDiff$(***V***,*T*);
repeat
$F = argmax_{X \in V} GMA(X; T; \mathbf{CPC}; Assoc(X; T|\emptyset); \emptyset, \mathcal{D})$
$assoc = max_{X \in V} GMA(X; T; \mathbf{CPC}; Assoc(X; T|\emptyset); \emptyset, \mathcal{D})$
if $assoc \neq 0$ **then**
  CPC $=$ **CPC** $\cup F$
end if
until ***CPC*** *unchanged*

Phase II: *Backward*
for all X $\in$ **CPC do**
  if $GMA(X; T; \mathbf{CPC}; Assoc(X; T|\emptyset); \emptyset, \mathcal{D}) = 0$ **then**
    CPC $=$ **CPC** $\setminus$ **X**
  end if
end for
return CPC

Phase III: *Check Functional Dependency*
if $IsFuncDep(T; \mathbf{CPC}; \mathcal{D})$ **then**
  $\mathbf{CPC}_1 = \mathbf{CPC}$
  $\mathbf{CPC}_2 = PMMPC(T; \mathbf{CPC} \cup \mathbf{Vset}, \mathcal{D})$
  CPC $= \mathbf{CPC}_1 \cup \mathbf{CPC}_2$
end if

GMA = GreedyMinAssoc function
function GMA(X;T;**Z**;minval;**MinSet**;$\mathcal{D}$)
$min = \infty$
$min = min_{S \in Z}\{Assoc(X; T|\mathbf{MinSet} \cup S, \mathcal{D}) \neq \epsilon\}$
$arg = argmin_{S \in Z}\{Assoc(X; T|\mathbf{MinSet} \cup S, \mathcal{D}) \neq \epsilon\}$
if $min < minval$ & $\mathbf{Z} \setminus arg \neq \emptyset$ **then**
  $minval = GMA(X; T; \mathbf{Z} \setminus minarg; min; \mathbf{MinSet} \cup arg; \mathcal{D})$
end if
return *minval*
end funtion

function IsFuncDep(T;**CPC**;$\mathcal{D}$)
if CPC $\rightarrow$ T **then**
  return *TRUE*
else
  return *FALSE*
end if

A. Association measure

The association between two variables given a conditioning set, $Assoc(X; Y|\mathbf{Z})$, can be implemented with a number of statistical or information theoretic measures of association. The only requirement for $Assoc(X; Y|\mathbf{Z})$ is to return zero when $Ind(X; Y|\mathbf{Z})$. In our implementation we prefer a statistically oriented test base on the $\chi 2$ test as in [8], [15]. We recall that, to test the association between X and Y given $\mathbf{Z}$, when the dataset is complete, the $\chi 2$ test calculates a statistic denoted as $\chi^2_{XY|\mathbf{Z}}$. The statistic is compared against a critical value to decide upon of the acceptance or rejection of the null hypothesis of conditional independence. The critical value depend upon the degrees of freedom ν of the $\chi 2$ distribution where

$$\nu = (n_X - 1)(n_Y - 1) \prod_{Z_j \in \mathbf{Z}} n_{Z_j}$$

One can use $1 - p$ instead, where p is the p-value defined as $P(\chi^2(\nu) \in [\chi^2_{XY|\mathbf{Z}}, \infty[)$ (the smaller the p-value, the higher we consider the association between X and T). $1 - p$ is compared against the risk α of the test. The choice of α directly influences the topology of the computed skeleton and will be discussed in the next session.

When n is small compared to ν, the test becomes unreliable. In our experiments, the test is applied whether $n > 10\nu$ (as coded in the BNT-SLP toolbox [12]) , otherwise $Assoc(X; Y|\mathbf{Z})$ returns a constant ϵ which value is chosen such that $0 < \epsilon < 1 - \alpha$. It is important to note that dependence is assumed between the variables whenever the test fails to decide due to lack of sufficient data. With such small positive value, the variables that cannot be d-separated in Phase I of PMMPC because of the lack of data will ultimately enter CPC *after* the variables that were found associated with sufficient data.

Another difficulty emerges as we address the problem of missing values. Simple solutions to handle missing data are either to ignore the cases including unknown entries or to ascribe these entries to an ad hoc dummy state of the respective variables. However, both these solutions are known to introduce potentially dangerous biases in the estimates, see [16] for a discussion. We adopt the *available case analysis*, i.e. $Assoc(X; Y|\mathbf{Z})$ is calculated on the cases having non missing values for X, Y and $\forall j, Z_j \in \mathbf{Z}$. This allows for a better use of the available sample but it relies on the assumption that data are Missing Completely at Random (MCAR), that is, the probability for data to be missing does not depend on $\mathcal{D}$. Unfortunately, there is no way to verify that data are actually MCAR in a particular database and, when this assumption is violated, these estimation methods may suffer of a decrease in accuracy with the consequence of jeopardizing the performance of the skeleton reconstruction algorithm [16]–[18].

Finally, three cases have to be considered : (1) acceptance of null hypothesis and sufficient data ($n > 10\nu$), (2) rejection of null hypothesis and sufficient data ($n > 10\nu$), and (3) rejection because of lack of data $n \leq 10\nu$. It follows :

$$Assoc(X; Y|\mathbf{Z}) = \begin{cases} P(\chi^2(\nu) \in]-\infty, \chi^2_{XY|\mathbf{Z}}])) & \text{case 1} \\ 0 & \text{case 2} \\ \epsilon & \text{case 3} \end{cases} \quad (3)$$

IV. Experiments with Asia

The original version of PMMS was already shown by its authors to outperform other BN (polynomial) learning algorithms on the task of reconstructing the network skeleton by extensive empirical evaluation on various DAG-faithful benchmark networks and small data sets. In this section, we apply the new PMMS version on small incomplete synthetic

data sets generated from given DAGs that have functional dependencies, and compare the computed skeleton against the true skeleton. To compare the quality of the learned skeleton, the number of extra edges and missing edges are presented. The network used in the evaluation are know as Asia (8 binary variables) and Asia8 (64 binary variables) [14]. Asia8 is obtained by tiling eight copies of the smaller Asia network as discussed in [9]. The tiling is performed in a way that maintains the structural and probabilistic properties of the original network in the tiled network. From each network, 10 datasets were randomly sampled in size 1289, the same size as our NPC dataset. 5% of the data was removed completely at random in each data set according to the MCAR hypothesis. The original version of PMMS on the complete dataset is unable to find the edges $O - X$ and $O - D$ due to the FD $TL \rightarrow O$. It yields therefore 25% missing edges at least. Its results are not reported.

The number of missing edges, additional edges, the mean and the standard deviation obtained with the new version of PMMS are reported in Table I and II. The aim is twofold : 1) to construct a skeleton as closed as possible to the true one from the available data, despite the missing values and the very limited size, 2) to evaluate empirically the inevitable errors that will necessarily arise. The choice of the threshold α directly influences the topology of the computed skeleton. As the standard threshold (0.05) used in the $\chi 2$ may be too small with respect to the dataset, we tested two values $\alpha = 0.05$ and $\alpha = 0.1$. Larger values yield poor results.

From Tables 1 and 2, it is observed that more false negatives (missing edges) are observed than false positives (spurious edges). With α set to 0.1 on $Asia8$, 22% false negatives are obtained on average and only 7% false positives. At first sight, this seems counterintuitive because PMMS is admissible in the sense that all variables with an edge to or from a given target T will enter CPC, assuming a perfect association measure. It also apparently contradicts Brown et al's experimental results [8]. There are two reasons : variable A "Visit-to-Asia" is rarely observed, hence the difficulty to infer the edge $A - T$ with less than 5000 complete observations, *a fortiori* with 1289 uncomplete observations, whatever the algorithm used. This is confirmed by the fact that $A - T$ has never been detected. The second reason is due to the limitations of the $\chi 2$ test : when $PMMPC$ is run on the target O, T and L enter first in CPC. In Phase III, the FD $TL \rightarrow O$ is detected, PMMPC is run again without T et L. As X enter CPC, the dependency $O - D$ conditionally to X becomes too weak to be detected because the association $O - X$ is too strong. In other words, as X is observed, the uncertainty about O is drastically reduced and $O - D$ conditionally to X doesn't pass the $\chi 2$ test.

Finally, from these experiments, we conclude that excessively rare or strong (but still random) associations are not detected with very small samples. Assuming dependencies of similar strength in the the NPC data, we expect (say) between 15% and 30% missing edges, but significantly less extra edges.

	Asia			
Risk	*5%*		*10%*	
Edges	false +	false -	false +	false -
Max	1	4	1	3
Min	0	0	0	0
Std. dev.	0.42	1.15	0.42	0.97
Mean	0.2	2.0	0.2	1.5

TABLE I
RESULTS FOR THE ASIA NETWORK : 8 VARIABLES, 1289 OBSERVATIONS AND 5% MISSING DATA.

	Asia8			
Risk	*5%*		*10%*	
Edges	false +	false -	false +	false -
Max	6	20	7	19
Min	3	10	3	10
Std. dev.	1.37	2.86	1.69	2.68
Mean	3.9	16.2	4.8	15.5

TABLE II
RESULTS FOR THE ASIA8 NETWORK : 64 VARIABLES, 1289 OBSERVATIONS AND 5% MISSING DATA.

V. APPLICATION TO NPC DATA

In this section, we apply the new version of PMMS on the NPC epidemiological data made up from 1289 individuals, 61 variables and 5% missing data. The aim of this research was to investigate the role of the environmental factors in the aetiology of NPC, in an endemic region where cancer research has been underdeveloped, and to shed some light into the statistical profile of the Maghrebian population.

In order to clarify the role of some environmental risk factors in the aetiology of NPC, a multi-centre case-control has been undertaken in Morocco, Algeria and Tunisia. A "case" is an individual with a histopathology confirmed diagnosis of NPC ; "controls" are individuals without cancer frequency-matched on sex, age and urban/rural housing which reflects the socio-economic level.

A. NPC Context

Nasopharyngeal carcinoma (NPC) is a malignancy with unusually variable incidence rates across the world. In most parts of the world it is a rare disease but in some regions it occurs in an endemic form. Endemic regions include the southern parts of China, other parts of south-east Asia and the Maghreb. In these countries it is a major public health problem. NPC has severe impact on households and economic situation of the affected families.

It is well established that poor and rural populations are more affected by NPC, almost certainly because they are more exposed to the various environmental risk factors. Epidemiological studies have suggested a large number of environmental risk factors for NPC, including dietary components as well as household and occupational exposures. The common features observed in NPC patients are : 1) a low socio-economic status with poor housing condition characterised by overcrowding and lack of ventilation; and 2) a monotonous diet including the regular consumption of traditionally preserved foods since very early age. There is controversy about the effect of cigarette smoking, alcohol intake and some occupational

exposures. Genetic factors are also thought to contribute to NPC.

B. The data

Patients were interviewed according to a specific questionnaire designed by IARC to collect precise information. This questionnaire includes a "demographic" part examining age, ethnicity, consanguinity, etc and an "environmental" part examining the environmental factors.

The present study attempt to assess the possible impact of 61 potential risk factors suggested by former studies, including : dietary factors, ear, nose and throat infections, housing conditions, tobacco, alcohol, professional exposure to dust, fumes, formaldehyde, traditional preserved food such as pickled vegetables and dried and salted meat or fat, abrupt weaning, and specific traditional medicines, and many others (See legend of Figure 2). The discrete variables have 2 or 3 modalities except age with 11 modalities.

Because the disease is quite uncommon, the sample sizes required for adequate precision and power are typically beyond the scope of any single group. Despite the limited size of our dataset, it is the largest epidemiological study of NPC performed to date. 664 cases and 625 controls were recruited. The data set is made up from 1289 records in total. Also, it has to be noted that the data was not sampled uniformly from the population.

C. Results

To get an idea of the performance of the classifier, we used a 10-fold cross validation technique to : 1) learn the skeleton, to orient the edges in NPC groups of variables, 2) learn the distributions tables using the training set, and finally 3) assess the success rate using the test set that was not used for training. After each training phase, we found the same group A and a unique local structure $1 \rightarrow 30 \leftarrow 31$ called a V-structure (see [5]) among the group A of variables. Inference is readily performed, only the values of the variables 30 and 31 are necessary. The average NPC hit rate is 74% on the test set to be compared to 50% NPC patients, which seems at first sight a very promissing result. Nonetheless, when interpreting arcs as causalities, it seems that NPC is the cause of bad ventilation in the kitchen during childhood (30) and not the opposite ! After discussion with our domain expert, this curious finding reveals a biais well known by the epidemiologists : the sick individuals are inclined to infer themselves the cause of their cancer and have a natural tendency to attribute their disease towhatever reason even if it is not true. In the questionnaire the specific question about the air ventilationduring childhood, is formulated in a way which give too much freedom to self interpretation. So the biais is encoded in the data themseleves and has nothing to do with the learning algorithm.

In conclusion, the skeleton provides interesting information on the habits of the maghrebian population under study but it also reveal a biais in the data owing to psychological side effects. In spite of the significant thematic association found among groups of variables and a 74% hit-rate, our experiments cannot help to untangle the etiological puzzle of this malignant disease.

VI. Conclusion

In this paper, we discussed a new heuristic called PMMS developped by Tsamardinos et al. in 2005. PMMS offers an excellent trade-off between time and quality of reconstruction compared to all algorithms, particularly for the smaller sample sizes. It suffers however from two problems : it is unable to deal with missing data, nor with datasets containing functional dependencies between variables. In this paper, we proposed a way to overcome these problems. The new version of PMMS was first validated on the Asia network benchmark to recover the original structure from data, and then applied on the Nasopharyngeal carcinoma (NPC) data made up from 1289 individuals, 61 variables and 5% missing data. Although the skeleton obtained shed some light into the statistical profile of the population under study, the strongest risk factor to NPC (bad kitchen ventilation) is due to a biais in the data.

References

[1] H. Wasyluk, A. Onisko, and M. J. Druzdzel, "Support of diagnosis of liver disorder based on causal bayesian network model," *Medical Science Monitor*, no. 7, pp. 327–332, 2001.

[2] I. Tsamardinos, C. F. Aliferis, and A. R. Statnikov, "Algorithms for large scale markov blanket discovery." in *FLAIRS Conference*, 2003, pp. 376–381.

[3] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. The MIT Press, 2000.

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

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

[6] D. M. Chickering and C. Meek, "Monotone dag faithfulness : A bad assumption." *Technical Report MSR-TR-2003-16, Microsoft Research, Redmond WA*, 2003.

[7] D. Dash and M. J. Druzdzel, "A hybrid anytime algorithm for the construction of causal models from sparse data." in *UAI*, 1999, pp. 142–149.

[8] L. E. Brown, I. Tsamardinos, and C. F. Aliferis, "A comparison of novel and state-of-the-art polynomial bayesian network learning algorithms." in *In the Proceedings of the Twentieth National Conference on Artificial Intelligence AAAI*, 2005, pp. 739–745.

[9] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm." *Machine Learning*, vol. 65, no. 1, pp. 31–78, 2006.

[10] J. Cheng, R. Greiner, J. Kelly, D. A. Bell, and W. Liu, "Learning bayesian networks from data: An information-theory based approach." *Artif. Intell.*, vol. 137, no. 1-2, pp. 43–90, 2002.

[11] K. Murphy, "The bayesnet toolbox for matlab," in *Computing Science and Statistics: Proceedings of Interface*, vol. 33, 2001. [Online]. Available: http://www.ai.mit.edu/ murphyk/Software/BNT/bnt.html

[12] P. Leray and O. Francois, "BNT structure learning package: Documentation and experiments," Laboratoire PSI, Tech. Rep., 2004. [Online]. Available: $http : //bnt.insa - rouen.fr/programmes/BNT\_StructureLearning\_v1.3.pdf$

[13] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press, 2000.

[14] S. Lauritzen and D. Speigelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Royal statistical Society B*, vol. 50, pp. 157–224, 1988.

[15] A. Aussem, Z. Kebaili, M. Corbex, and F. D. Marchi, "Apprentissage de la structure des réseaux bayésiens à partir des motifs fréquents corrélés : application à l'identification des facteurs environnementaux du cancer du nasopharynx." in *Actes des journées Extraction et Gestion des Connaissances, EGC'06*. RNTI-E-6, Cépaduès-Éditions, 2006, pp. 651–662.

[16] M. Ramoni and P. Sebastiani, "Robust learning with missing data." *Machine Learning*, vol. 45, no. 2, pp. 147–170, 2001.

[17] N. Friedman, "The bayesian structural EM algorithm." in *UAI*, 1998, pp. 129–138.

[18] D. Dash and M. J. Druzdzel, "Robust independence testing for constraint-based learning of causal structure." in *UAI*, 2003, pp. 167–174.

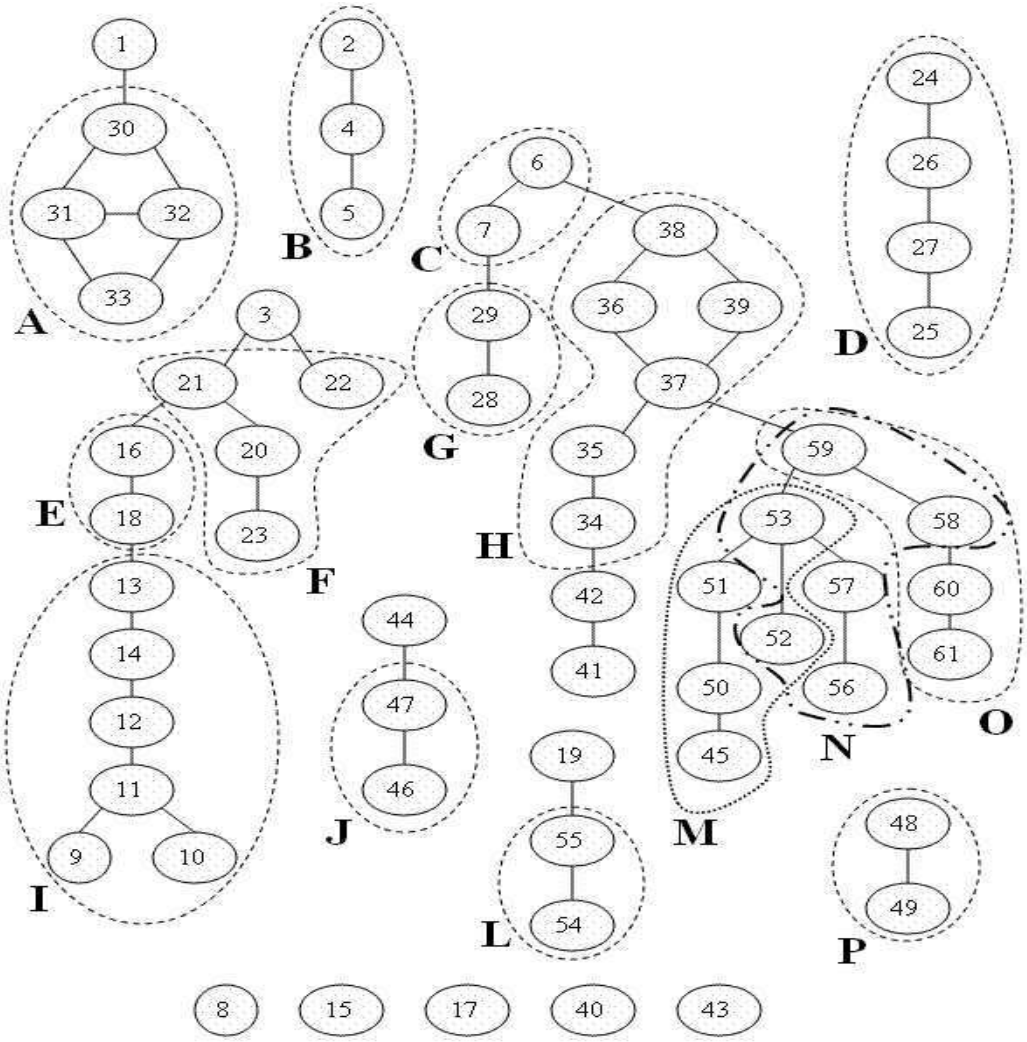

Fig. 2. The skeleton obtained with the new PMMS version. In dotted line: the groups of thematic variables. Lexical : **NPC** 1, age of interview for control individuals and age at cancer for cases 2, sex 3, instruction 4, professional category 5, lodging ch. and ad. 6 7, parents consanguinity 8, otitis 9, pharyngitis 10, cold 11, asthma 12, eczema 13, allergy 14, chemical manure and pesticide 15, chemical products 16, smoke 17, dust 18, formaldehyde 19, alcohol 20, tabac 21, neffa 22, cannabis 23, housing type ch. and ad. 24 25, separated beds ch. and ad. 26 27, animal in the house ch. et ad. 28 29, kitchen ventilation ch. et ad. 30 31, house ventilation ch. and ad. 32 33, incense ch. and ad. 34 35, kanoun and tabouna ch. et ad. 36 37, wood fire ch. et ad. 38 39, brest feeding and age of weaning and way of weaning 40 41 42, contact with adult saliva 43, traditional childhood treatments 44, hot pepper 45, smen and fat ch. and ad. 46 47, vegetables and fruits ch. and ad. 48 49, house made harrissa ch and ad. 50 51, industrial harrissa ch. ad. 52 53, house made proteins ch. and ad. 54 55, industrial proteins ch. and ad. 56 57, industrial canned vegetables ch. and ad. 58 59, house made canned vegetables ch. and ad. 60 61. ch.=childhood and ad=adult.