Application des réseaux bayésiens à l'analyse des facteurs impliqués dans le cancer du Nasopharynx

Alexandre Aussem*, Sergio Rodrigues de Morais*, Marilys Corbex**

*Université de Lyon 1,
EA 2058 PRISMa, F-69622 Villeurbanne
aaussem@univ-lyon1.fr,
**Unité d'épidémiologie génétique,
Centre International de Recherche sur le Cancer (CIRC),
150 cours Albert Thomas - 69280 Lyon Cedex 08
corbex@iarc.fr

Résumé. L'apprentissage de la structure des réseaux bayésien à partir de données est un problème NP-difficile. Une nouvelle heuristique de complexité polynômiale, intitulée Polynomial Max-Min Skeleton (PMMS), a été proposée en 2005 par Tsamardinos et al. et validée avec succès sur de nombreux bancs d'essai. PMMS présente, en outre, l'avantage d'être performant avec des jeux de données réduits. Néanmoins, comme tous les algorithmes sous contraintes, celui-ci échoue lorsque des dépendances fonctionnelles (déterministes) existent entre des groupes de variables. Il ne s'applique, par ailleurs, qu'aux données complètes. Aussi, dans cet article, nous apportons quelques modifications pour remédier à ces deux problèmes. Après validation sur le banc d'essai $Asia$, nous l'appliquons aux données d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC) de 1289 observations, 61 variables et 5% de données manquantes issues d'un questionnaire. L'objectif est de dresser un profil statistique type de la population étudiée et d'apporter un éclairage utile sur les différents facteurs impliqués dans le NPC.

1 Introduction

L'apprentissage de la *structure* des réseaux bayésiens (RB) à partir de données est un problème ardu ; la taille de l'espace des graphes orientés sans circuits (*DAG* en anglais) est super-exponentielle en fonction du nombre de variables et le problème combinatoire associé est NP-difficile (Chickering et al., 2004). Deux grandes familles de méthodes existent pour l'apprentissage de la structure des RB : celles fondées sur la satisfaction de contraintes d'indépendance conditionnelle entre variables et celles à base de score fondées sur la maximisation d'un score (BIC, MDL, BDe, etc.). Les deux méthodes ont leurs avantages et leurs inconvénients. Les méthodes sous contraintes sont déterministes, relativement rapides et bénéficient des critères d'arrêt clairement définis. Les contraintes imposées à la structure du graphe proviennent des informations statistiques sur les dépendances et indépendances conditionnelles observées dans les données. Elles reposent cependant sur un niveau de signification arbitraire

du test d'indépendance employé. En outre, les erreurs commises au début peuvent se répercuter en cascade dans la suite de l'execution de l'algorithme, et conduire à un graphe erroné. Les méthodes à base de score ont, quant à elles, l'avantage d'incorporer des probabilités a priori sur la structure du graphe et de traiter plus facilement les données manquantes. En revanche, elles sont facilement piégées dans les nombreux minima locaux et le graphe final obtenu dépend fortement des conditions initiales.

Plusieurs méthodes ont été proposées durant ces quinze dernières années mais quelques avancées prometteuses ont été réalisées très récemment. Dans un article paru en 2006, Tsamardinos et al. montrent par des simulations exhaustives sur une vingtaine de bancs d'essais (Child, Insurance, Alarm, Hailfinder etc.) l'avantage significatif d'un algorithme sous contraintes, dénommé Min-Max Hill-Climbing (MMHC), au regard des algorithmes majeurs (score et contraintes) en fonction de plusieurs métriques de performance (Tsamardinos et al., 2006). Son inconvénient, toutefois, est sa complexité au pire cas en $\mathcal{O}(n2^n)$ où n désigne le nombre de nœuds. En réalité, aucun algorithme exact n'échappe à une complexité exponentielle car les méthodes exactes reposent toutes sur une recherche exhaustive des indépendances entre deux variables X et Y conditionnellement à un ensemble $\mathbf{Z}$. Cette recherche nécessite $\mathcal{O}(2^{|Z|})$ opérations au pire cas. C'est pourquoi toutes les approches polynômiales reposent sur une heuristique particulière pour parcourir les ensembles $\mathbf{Z}$.

Brown et al. ont donc proposé une version polynomiale en $\mathcal{O}(n^4)$ de MMHC dénommée Polynomial Min-Max Skeleton (PMMS) (Brown et al., 2005) en adoptant une heuristique ingénieuse. L'algorithme séduit par ses nombreux attraits : outre sa grande simplicité, les auteurs ont montré empiriquement l'excellent compromis entre faible complexité et qualité de reconstruction comparé aux autres algorithmes, surtout en présence de faibles jeux de données. Cet avantage est décisif à nos yeux car nous disposons d'une base de données d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC) de seulement 1289 observations. L'idée est donc d'utiliser PMMS afin de construire par apprentissage la structure du RB associé aux données. Néanmoins, comme tous les algorithmes sous contraintes, PMMS échoue lorsque des dépendances fonctionnelles (DF) déterministes existent entre des groupes de variables. Une DF, notée $\mathbf{X} \rightarrow Y$, est une contrainte entre un ensemble de variables, telle que tout ensemble de valeurs prises par les $X_j \in \mathbf{X}$ determine la valeur de Y de façon univoque. Or rien n'exclue cette éventualité compte tenu du faible nombre d'observations. En outre, PPMS ne s'applique qu'aux données complètes, ce qui n'est pas notre cas. Aussi, dans cet article, nous apportons quelques modifications algorithmique à PMMS pour remédier à ces deux problèmes : le traitement des DF et son adaptation aux données manquantes.

Après un rappel indispensable de la problématique et des principes de l'algorithme, ces modification sont présentées en détail et validées sur deux bancs d'essai Asia et Asia8 avec une DF entre trois variables, 1289 données et 5% de données manquantes pour nous ramener au cas du NPC. La nouvelle version a été développée sous Matlab à l'aide de la Toolbox BNT de (Murphy, 2001) et de la Toolbox BNT-SLP de (Leray et Francois, 2004). Ensuite, nous appliquons la méthode aux données du NPC de 1289 observations. A la différence des travaux préliminaires menés dans (Aussem et al., 2006) avec seulement 10 variables binaires synthétiques et sans données manquantes, une base plus vaste de 61 variables qualitatives ordinales et 5% de données manquantes est analysée. L'objectif est de dresser un profil statistique type de la population étudiée et d'apporter un éclairage utile sur les différents facteurs impliqués dans le NPC.

2 Préliminaires

Notons l'indépendance conditionnelle entre X et Y sachant l'ensemble $\mathbf{Z}$ dans une loi de probabilité P par $Ind_P(X;Y|\mathbf{Z})$ et la dépendance par $Dep_P(X;Y|\mathbf{Z})$. Les lettres majuscules en gras, $\mathbf{Z}$, désignent des ensembles de variables aléatoires, les autres majuscules, X, désignent des variables uniques, les minuscules ($\mathbf{X} = \mathbf{x}$, $X = x$) désignent les attributs ou modalités des variables. Soit P, une loi de probabilité conjointe sur un ensemble de variables aléatoires $\mathcal{V}$, et $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ un graphe orienté sans circuit (DAG en anglais). On dira que le tuple $< \mathcal{G}, P >$ est un réseau bayésien si $< \mathcal{G}, P >$ vérifie la condition dite de Markov : chaque variable, $X \in \mathcal{V}$, doit être indépendante de ses non descendantes (ND_X) dans $\mathcal{G}$ conditionnellement à ses parents (Neapolitan, 2004; Pearl, 2000). Cette condition se note $Ind_P(X; ND_X|\mathbf{Pa}_i^{\mathcal{G}})$ où $\mathbf{Pa}_i^{\mathcal{G}}$ désigne l'ensemble des parents de X_i dans $\mathcal{G}$. La condition de Markov implique la factorisation de la loi jointe :

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

Cette propriété importante montre qu'il suffit de stocker les valeurs de $P(X_i = x_i|\mathbf{Pa}_i^{\mathcal{G}} = \mathbf{pa}_i)$ pour toutes les valeurs de x_i et les possibles instantiations conjointes de $\mathbf{pa}_i$ dans une table de probabilités. A partir de $P(\mathcal{V})$, toute requête portant sur une ou plusieurs variables d'intérêt conditionnellement à d'autres (les observations partielles) peut être obtenue par *inférence* (voir par ex. Naim et al. (2004); Neapolitan (2004)). La propriété de Markov impose en revanche une condition forte aux lois de probabilité P qui peuvent être représentées par le même graphe $\mathcal{G}$.

Contraintes du graphe - La *d-séparation* est un critère important qui permet de caractériser *graphiquement* toutes les contraintes d'indépendance des lois P qui peuvent être représentées par un même DAG. Pour éclaircir son rôle, il faut introduire la notion de chaîne d'information bruité. Par chaîne, on entend une succession d'arcs orientés entre X et Y mais vus comme des arêtes non orientés (Neapolitan, 2004). Pour comprendre la *d-séparation*, il faut symboliser les noeuds sur ces chaînes par des vannes d'information, ouvertes ou fermées selon le cas. Un chemin est dit *ouvert* si toutes les vannes sont ouvertes auquel cas il laisse passer l'information. A l'inverse, si l'une des vanne est bloquée, la chaîne est dite *bloquée*. En extrapolant, l'information qu'apporte X sur Y peut se voir comme la somme des flots sur tous les chaînes ouvertes reliant X à Y. Il reste à spécifier le mécanisme d'ouverture et de fermeture des vannes. Il existe 3 types de connexions : les connexions en série $A \rightarrow B \rightarrow C$ ou $A \leftarrow B \leftarrow C$, la connexion divergente $A \leftarrow B \rightarrow C$ et la connexion convergente $A \rightarrow B \leftarrow C$. Formellement, une chaîne entre X et Y est bloquée par un ensemble de noeuds $\mathbf{Z}$ s'il existe un noeud sur cette chaîne vérifiant l'une des conditions : (1) W n'est pas un noeud convergent et $W \in \mathbf{Z}$, (2) W est un noeud convergent et, ni W, ni aucun de ses descendants ne sont dans $\mathbf{Z}$. Deux noeuds X et Y sont dits *d-séparés* par $\mathbf{Z}$ dans le graphe $\mathcal{G}$, noté $Dsep_{\mathcal{G}}(X;Y|\mathbf{Z})$, si tous les chaînes (simples) entre X et Y sont bloqués par $\mathbf{Z}$. La *d-séparation* dresse un parallèle élégant entre l'algorithmique des graphes et le calcul des indépendances conditionnelles dans une distribution. Prenons l'exemple du réseau de la dyspnée *Asia* (Lauritzen et Speigelhalter, 1988) de la Figure 1. Il existe 2 chaînes simples entre A et D : $[A, T, O, D]$ et $[A, T, O, L, S, B, D]$. La première est ouverte mais la seconde bloquée en O car $T \rightarrow O \leftarrow L$ est convergente. En conditionnant par rapport à O ($O \in \mathbf{Z}$), on trouve l'inverse.

Réseau bayésien appliqué aux données du cancer

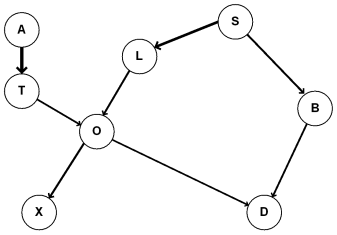

FIG. 1 – *Réseau Asia original. A désigne 'visite en Asie', T 'tuberculose', O OU logique entre T et L, S 'fumeur', B 'bronchite', D 'dyspnée' (difficulté respiratoire), X 'rayons X' et L 'cancer de la langue'.*

Condition de fidélité - $\mathcal{G}$ et P sont dits fidèles (*faithful*) l'un à l'autre ssi toutes les indépendances conditionnelles sont strictement identifiées par les d-*séparations*, i.e., $Dsep_{\mathcal{G}}(X;Y|\mathbf{Z})$ $\Leftrightarrow Ind(X;Y|\mathbf{Z})$. On parle alors de réseau bayésien $< \mathcal{G}, P >$ fidèle. PMMS repose sur l'hypothèse de fidélité ; l'algorithme construit un DAG sensé être fidèle à la loi de probabilité P sous-jacente aux données. Cela pose un problème car toutes les distributions ne sont pas fidèles à un DAG, c'est le cas notamment du réseau $Asia$ en raison de la variable O. O est un OU logique entre T et L, du coup $Ind_P(O;X|\{T,L\})$ est vérifié sans pour autant avoir $Dep_P(O;X|\{T,L\})$. L'existence de dépendances fonctionnelles (DF), accidentelles ou non, entre des variables, est hélas fréquent. C'est souvent le cas dans les données sont issues de questionnaires pour de multiples raisons (e.g. questions redondantes ou mal comprises, réponses groupées etc.).

3 Polynomial Min-Max Skeleton revisité

Dans ce paragraphe, nous rappelons le principe de PMMS avant de présenter les modifications que nous avons opérées pour traiter l'existence des DF et les données incomplètes. PMMS exploite ingénieusement le critère de d-*séparation*. Il construit itérativement le voisinage, parents et enfants, de chaque variable cible T, en observant que ce sont les seuls noeuds qui ne peuvent être d-séparés de T. Notons par $\mathbf{PC}_T$ les nœuds parents et enfants du nœud T dans $\mathcal{G}$. $\mathbf{PC}_T$ est unique à tous les DAG tels que $< \mathcal{G}, P >$ soient fidèles, il ne dépend donc pas de $\mathcal{G}$. PMMS emploie une mesure d'association probabiliste conditionnelle notée $Assoc(X;Y|\mathbf{Z})$ pour calculer $MinAssoc(X;Y|\mathbf{Z})$:

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

C'est-à-dire les plus petites associations entre X et Y pour tous les sous-ensembles $\mathbf{S}$ de $\mathbf{Z}$. PMMS appelle successivement la procédure Polynômial Min-Max Parents and Children PMMPC pour chaque variable de $\mathcal{G}$ (voir algorithme 1). PMMPC identifie $\mathbf{PC}_T$ étant donné une variable cible T. Ainsi, connaissant le voisinage direct de chaque variable cible, il suffit de connecter les noeuds pour obtenir le squelette du graphe (non connecté). La version originale de PMMPC opère en 2 phases. Nous y avons adjoint une troisième phase pour traiter les DF (voir algorithme 2).

Phase I - Les variables entrent séquentiellement dans un ensemble de candidats noté **CPC** à l'aide d'une heuristique Max-Min. L'idée est de sélectionner itérativement les variables qui ne peuvent pas être d-séparées par l'ensemble **CPC** courant. Celle qui rentre est celle qui présente l'association résiduelle la plus forte avec la cible T malgré notre effort pour bloquer toutes les chaînes les reliant. Le calcul de $MinAssoc(X;Y|\mathbf{C}PC)$ requiert normalement un nombre d'appels exponentiel à la fonction $Assoc$. Dans PMMS, ce calcul est réduit à l'aide de l'heuristique gloutonne $GreedyMinAssoc(X;T|\mathbf{CPC},minval,\mathbf{MinSet})$; $minval$ est l'estimation courante du minimum d'association et **MinSet** est l'estimation courante de $\mathbf{S} \subseteq \mathbf{Z}$ qui réalise ce minimum. Initialement $\mathbf{MinSet} = \emptyset$ et $minval = Assoc(X;Y|\emptyset)$; **MinSet** croît itérativement par adjonction d'une variable de **CPC** après l'autre jusqu'à ne plus pouvoir décroître l'association résiduelle. La notation $\min\{x \neq \epsilon\}$ désigne le plus petit x différent de ϵ s'il existe. Si les données sont en nombre insuffisant, alors $Assoc = \epsilon$, la phase de croissance est stoppée et la dépendance est supposée. Néanmoins la valeur de l'association est fixée à la plus petite valeur possible, ϵ, comme discuté au chapitre 3.1.

Phase II - Dans cette phase *backward*, les faux-positifs entrés par erreur dans la phase I (voir Tsamardinos et al. (2006)) sont itérativement éliminés de **CPC**. Pour ce faire, on teste pour chaque X de **CPC** si $Dsep_{\mathcal{G}}(X;T|\mathbf{S})$ pour tout $\mathbf{S} \subseteq \mathbf{Z} \setminus X$ à vérifiant si $MinAssoc$ $(T;S|\mathbf{S}) = 0$. L'opération est encore réalisée avec l'heuristique $GreedyMinAssoc$.

Phase III - Cette étape est nouvelle. Si **CPC** $\rightarrow T$ est une DF, le graphe G n'est pas fidèle avec P, auquel cas T sera d-séparé de tous ses enfants par **CPC** alors que c'est faux. Dans le cas d'$Asia$ par exemple, PPMPC lancé sur la cible O conduit à **CPC** = $\{T, L\}$ car l'association de T et L avec O est la plus forte, et par suite, ni X ni D ne pourront plus entrer dans **CPC**. Aussi, pour y ajouter les enfants de la cible T, nous testons si la relation **CPC** $\rightarrow T$ est un DF par un appel à $IsFuncDep(T;\mathbf{CPC};\mathcal{D})$ (simple parcours de l'hypercube de contingences). Si la DF est observée, un nouvel appel récursif à PMMPC est fait en retirant **CPC** à l'ensemble des variables V jusqu'aucune DF ne subsiste. On récupère au final un ensemble **CPC** qui contient non seulement les enfants de T mais aussi ses grands-parents. Dans Asia, le CPC de O sera au final l'ensemble $\{T, L, X, D, A, S\}$ mais PMMS ne connectera pas O avec ses grands-parents A et S car réciproquement O ne sera ni dans le CPC de A, ni dans celui de S. Ainsi, la phase 3 permet de détecter les DF susceptibles de tromper PMMPC (et uniquement celles-ci) pour ensuite trouver tous les parents et enfants de la variable cible.

Au final, PMMS construit un graphe non orienté (le squelette) qui est sensé, une fois les arêtes dirigées, représenter la structure du réseau bayésien. Comment diriger les arêtes du graphe du squelette ? Il faut garder à l'esprit que plusieurs DAG peuvent encoder la même loi de probabilité conjointe, seule importe la position des V-structures (i.e., $X \leftarrow Y \leftarrow Z$ tel que X et Z ne soit pas connectés). Cet article ne porte que sur l'apprentissage du squelette, la partie de loin la plus difficile. Pour la recherche des V-structures et la direction des arcs, le lecteur est invité à consulter (Naim et al., 2004; Neapolitan, 2004) pour de plus amples informations.

3.1 Mesure d'association

La mesure d'association $Assoc(X;Y|\mathbf{Z})$ choisie repose sur la statistique du χ^2 comme celle utilisée dans (Aussem et al., 2006; Brown et al., 2005). Le test du χ^2 est couramment utilisé dans les tests d'adéquation de loi et les tests d'indépendance, mais son usage est ici "dévoyé" pour en faire une mesure d'association. Sans données manquantes, l'association vaut simplement $1-p$, où p désigne la *p-value* définie par $P(\chi^2(\nu) \in [\chi^2_{XY|\mathbf{Z}}, \infty[)$ si p est inférieur

Algorithme 1 *PMMS revisité*

ENTRÉES: ***D*** : matrice des données
SORTIES: ***E*** : squelette du réseau bayésien

pour tout $X \in$ **V faire**
  $\mathbf{PC}_X = PMMPC(X, \mathcal{D})$
fin pour
pour tout $(X, Y) \in$ **V faire**
  si $X \in \mathbf{PC}_Y$ & $Y \in \mathbf{PC}_X$ **alors**
    $\mathbf{E} = \mathbf{E} \cup (XY)$
  finsi
fin pour
return E

au risque α du test, et zéro sinon. Le $\chi^2_{XY|\mathbf{Z}}$ suit une loi χ^2 à $\nu = (n_X - 1)(n_Y - 1)c$ degrés de liberté où $c = \prod_{Z_j \in \mathbf{Z}} n_{Z_j}$ est le produit du nombre de modalités de chaque variable dans **Z**. Intuitivement, plus la valeur de p est petite, plus l'association entre X et Y sachant **Z** est forte. Aussi, une valeur de p supérieure au seuil α indiquera une association nulle. En pratique, le test n'est utilisé que si n est suffisamment grande devant ν. Dans le cas contraire, il faudrait idéalement procéder à des regroupements de modalités voisines. Des méthodes heuristiques existent pour évaluer empiriquement le nombre effectif de degrés de liberté (voir Tsamardinos et al. (2006)). Dans notre cas, le test est appliqué dès lors que $n > 10\nu$ comme dans l'heuristique PC Spirtes et al. (2000), sinon $Assoc(X; Y|\mathbf{Z})$ retourne une constante ϵ inférieur au risque du test $0 < \epsilon < 1 - \alpha$ et l'on suppose la dépendance, faute de pouvoir statuer.

La présence de données manquantes dans les données du cancer pose une difficulté supplémentaire. L'apprentissage de RB en présence de données manquantes est en soi un domaine de recherche actif dans lequel plusieurs solutions ont été proposées, en majorité pour les algorithmes bayésiens (à base de score) François (2006). Nous optons pour une solution simple connue sous le nom "available case analysis" : le $\chi^2_{XY|\mathbf{Z}}$ est calculé uniquement sur les observations pour lesquelles les n variables X, Y et $Z_j \in \mathbf{Z}$ sont présentes. Les autres sont ignorées. Ce faisant, le risque est toutefois d'introduire un biais dans les estimateurs (Ramoni et Sebastiani, 2001; Friedman, 1998; Dash et Druzdzel, 2003) en particulier si l'hypothèse *Missing Completely at Random* (MCAR) n'est pas vérifiée, i.e., le processus de perte est indépendant de la valeur des données complétées $\mathcal{D}$ (observées et manquantes).

On distingue donc trois cas : 1) le cas où l'indépendance est supposée, 2) le cas du manque de données $n \leq 10\nu$ et enfin 3) le cas l'hypothèse d'indépendance est rejetée :

$$Assoc(X; Y|\mathbf{Z}) = \begin{cases} P(\chi^2(\nu) \in]-\infty, \chi^2_{XY|\mathbf{Z}}])) & \text{si} \chi^2_{XY|\mathbf{Z}} > \chi^2_{1-\alpha}(\nu) \text{ et } n > 10\nu \\ 0 & \text{si} \chi^2_{XY|\mathbf{Z}} \leq \chi^2_{1-\alpha}(\nu) \text{ et } n > 10\nu \\ \epsilon & \text{si } n \leq 10\nu \end{cases} \quad (3)$$

4 Validation sur Asia

Pour illustrer et valider ces modifications dans un contexte le plus proche possible de nos données du cancer, nous avons utilisé le réseau de la dyspnée *Asia* (Lauritzen et Speigelhalter,

Algorithme 2 *PMMPC*

ENTRÉES: T : variable cible
Vset : ensemble des variables telles que **Vset** $\rightarrow$ T soit une dépendance fonctionnelle. **Vset** $= \emptyset$ au premier appel de PMMPC
$\mathcal{D}$: matrice des données avec ***V*** variables
SORTIES: ***CPC*** : ensemble de parents et enfants de T
Phase I : *Forward*
$CPC = \emptyset$
V=$SetDiff$(***V***,***Vset***) ;
V=$SetDiff$(***V***,*T*) ;
repeat
$F = argmax_{X \in V} GreedyMinAssoc(X; T; \textbf{CPC}; Assoc(X; T|\emptyset); \emptyset, \mathcal{D})$
$assoc = max_{X \in V} GreedyMinAssoc(X; T; \textbf{CPC}; Assoc(X; T|\emptyset); \emptyset, \mathcal{D})$
si $assoc \neq 0$ **alors**
  CPC $=$ **CPC** $\cup F$
finsi
until ***CPC*** *unchanged*
Phase II : *Backward*
pour tout X $\in$ **CPC faire**
  si $GreedyMinAssoc(X; T; \textbf{CPC}; Assoc(X; T|\emptyset); \emptyset, \mathcal{D}) = 0$ **alors**
    CPC $=$ **CPC** $\setminus$ **X**
  finsi
fin pour
return CPC
Phase III : *Check Functional Dependence*
si $IsFuncDep(T; \textbf{CPC}; \mathcal{D})$ **alors**
  $\textbf{CPC}_1 = \textbf{CPC}$
  $\textbf{CPC}_2 = PMMPC(T; \textbf{CPC} \cup \textbf{Vset}, \mathcal{D})$
  CPC $= \textbf{CPC}_1 \cup \textbf{CPC}_2$
finsi

fonction GreedyMinAssoc(X ;T ;**Z** ;minval ;**MinSet** ;$\mathcal{D}$)
$min = \infty$
$min = min_{S \in Z}\{Assoc(X; T|\textbf{MinSet} \cup S, \mathcal{D}) \neq \epsilon\}$
$arg = argmin_{S \in Z}\{Assoc(X; T|\textbf{MinSet} \cup S, \mathcal{D}) \neq \epsilon\}$
si $min < minval$ & **Z** $\setminus arg \neq \emptyset$ **alors**
  $minval = GreedyMinAssoc(X; T; \textbf{Z} \setminus minarg; min; \textbf{MinSet} \cup arg; \mathcal{D})$
finsi
return *minval*
end function

fonction IsFuncDep(T ;**CPC** ;$\mathcal{D}$)
si CPC $\rightarrow$ T **alors**
  return *TRUE*
sinon
  return *FALSE*
finsi

1988) comme banc d'essai, dont la structure est déjà connue, pour engendrer par tirage aléatoire 10 bases de 1289 observations et 8 caractères. Dans chaque base, 5% des données ont été effacées aléatoirement selon le protocole *MCAR*. Ensuite nous avons fait de même avec le réseau Asia8 (i.e. la réplication de Asia huit fois à l'identique) pour obtenir un nombre de variables proche de celui du cancer (61 variables pour NPC, 64 pour Asia8). La version de PMMS classique avec une base *complète* de 1289 données ne trouve jamais les liens (O, X) et (O, D) à cause de la dépendance fonctionnelle $TL \rightarrow O$, soit au minimum 25% de faux négatifs, la méthode est donc d'emblée écartée.

Les résultats obtenus avec la nouvelle version de PMMS sont consignés dans la Table 1 en fonction du risque α du test. L'objectif est de montrer la qualité de la reconstruction du squelette (Figure 1) malgré le faible nombre de données, partiellement manquantes, et de choisir empiriquement le meilleur seuil α. Des tests avec $\alpha > 0.1$ ne sont pas affichés car ils se sont avérés décevants. Dix jeux de données ont été synthétisés à partir du RB original. La Table 1 affiche le nombre d'arêtes en trop et en moins (faux positifs et faux négatifs), le maximum et le minimum obtenus avec PMMS. Ces valeurs doivent être comparées au nombre d'arêtes du RB original : *Asia* comporte 8 arêtes et *Asia8* 64. D'une manière générale, on observe plus d'arêtes négatives que positives et une plus grande variabilité pour les faux négatifs. Avec $\alpha = 0.1$ sur $Asia8$, on tourne à environ 22% de faux négatifs et 7% de faux positifs. Le nombre important de faux négatifs est contre-intuitif et va à l'encontre des expériences menées dans (Brown et al., 2005). Il y a deux raisons à cela. La première raison est liée au fait que l'événement "visite en Asie" est très rare, d'où la difficulté de déceler le lien (A, T) avec moins de 5000 observations complètes, a fortiori avec 1289 incomplètes, et ce, quelque soit l'algorithme utilisé. D'ailleurs, (A, T) n'a jamais été décelée sur les 10 simulations. La seconde raison est plus complexe : Lorsque $PMMPC$ est lancé avec O comme argument, T et L vont entrer les premiers dans CPC. Dans la Phase III, la dépendance fonctionnelle $TL \rightarrow O$ sera détectée et PMMPC sera relancé sans T et L. X rentrera dans CPC en premier. Le problème est que, X et O étant très forte, l'association entre O et D conditionnellement à X devient trop faible pour être détectée. Il semblerait donc qu'avec si peu de données, les associations trop fortes ou trop rares entre variables engendrent des erreurs de reconstruction. Cela n'est pas à imputer à PMMS mais à la mesure d'association. Au final, si l'on suppose une structure de corrélation du NPC similaire à celle d'Asia avec le même nombre de données, on s'attend à avoir (disons) entre 15% et 30% d'arêtes manquantes, mais beaucoup moins d'arêtes en trop.

5 Application au cancer

5.1 Les données

Nous appliquons la nouvelle version de PMMS aux données d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC) dans la lignée des travaux de recherche récents (Aussem et al., 2006; Antal et al., 2004; Getoor et al., 2004). Pour clarifier le rôle de l'environnement dans l'étiologie du NPC, l'Unité d'épidémiologie génétique du CIRC a mené en 2004 une étude multicentrique de cas-témoins dans la région endémique du Maghreb. 650 cas (personnes atteintes du cancer) et 639 témoins ont été recrutés. Les données extraites du questionnaires comportent 1289 enregistrements et 61 caractères pertinents ont été extraits d'un questionnaire de 450 questions. Il faut savoir que les données ne sont pas issues d'un tirage

	Asia				*Asia8*			
Risque	*5%*		*10%*		*5%*		*10%*	
Arêtes	faux +	faux -	faux +	faux -	faux +	faux -	faux +	faux -
Max	1	4	1	3	6	20	7	19
Min	0	0	0	0	3	10	3	10
Ecart type	0.42	1.15	0.42	0.97	1.37	2.86	1.69	2.68
Moyenne	0.2	2.0	0.2	1.5	3.9	16.2	4.8	15.5

TAB. 1 – *Bancs d'essai Asia (8 arcs) et Asia8 (64 arcs) avec 1289 observations et 5% de données manquantes : reconstruction du squelette avec PMMS (revisité) en fonction du risque α du test du χ^2. 10 simulations ont été menées pour chaque réseau. Les résultats comptent le nombre d'arêtes en trop et en moins (faux positifs et faux négatifs)*

i.i.d. dans la population puisque, comme toute étude épidémiologique classique, les individus témoins ont été choisis en fonction des cas (51% de cas et 49% de témoins). Le NPC présente une incidence très variable selon les régions du monde. C'est un cancer relativement rare sauf en Chine, en Asie du sud-est et au Maghreb, où les taux d'incidence sont élevés. Dans ces régions, le NPC est un problème majeur de santé publique. Les études ont suggéré l'existence d'un grand nombre de facteurs de risques environnementaux incluant habitudes alimentaires et environnement domestique et professionnel. L'idée est d'apporter un éclairage utile et pertinent sur les différents facteurs causes impliquées dans le NPC et dresser un profil statistique type de la population étudiée, que ne permet pas nécessairement la régression logistique, modèle couramment employée en épidémiologie, et si possible d'émettre des hypothèse sur les facteurs impliqués dans le NPC.

5.2 Résultats

La base est contient 5% de données manquantes mais on ignore si l'hypothèse $MCAR$ est valide. Le nombre de modalité varie de 2 (binaire) à 11 pour la classe d'âge. La majorité des variables sont codées en trois classe "pas du tout", "un peu" et "beaucoup". La variable à expliquer est "cas NPC", les autres sont les variables explicatives, voir le lexique de la Figure 2. Elles portent sur les conditions socio-professionnelles, l'habitat, l'exposition aux produits toxiques, la nourriture industrielle ou faite maison, les maladies, allergies, les drogues locales etc. Le squelette du RB obtenu avec la nouvelle version de PMMS est représenté sur la Figure 2 avec $\alpha = 0.1$. Il s'agit d'une visualisation graphique des interactions entres les variables qui dresse le profil statistique de la population considérée. En observant uniquement la structure du graphe, des groupes thématiques cohérents de variables de A à P ont été exhibés avec notre expert. Leur homogénéité est frappante. A est la le seul groupe lié au NPC (variable 1), il est lié à l'aération de l'habitat (variables 30, 31 et 32 mais pas 33 car indépendamment de l'orientation des arcs, 1 peut être d-séparée de 33 par un sous-ensemble de 30, 31 et 32) ; B, conditions socio-professionnelles liées à l'âge. C, lieu d'habitat ; D, catégorie de logement ; E, produits toxiques et fumées ; F, drogues ; G, animaux domestiques ; H, encens, parfums et feu de bois ; I, maladies ; J, graisse rance ; L, protéines maison ; M, piment et harrissa ; N, nourriture industrielle ; O, conserves ; P légumes et fruits. On retrouve donc des résultats

de bon sens : les hommes (3) sont plus enclins à fumer, à consommer des drogues (F) et être exposés à des produits toxiques au travail (E), que l'exposition aux fumées (encens, parfums, feu de bois etc.) (H) est plus fréquente dans les gourbis que les appartements en villes (C) ; que la nourriture industrielle (N) se consomme en conserve (O) ; que d'une manière générale, les habitudes acquises à l'enfance se conservent à l'âge adulte ; que les produits toxiques (E) provoquent des maladies (I) etc. Nous avons de plus testé les qualités du clasifieur selon le principe du *10-fold cross validation*. Après chaque apprentissage, on retrouve toujours la V-structure $1 \rightarrow 30 \leftarrow 31$, l'inférence est alors immédiate et seules 30 et 31 doivent être renseignées. Le taux de réussite moyen est de 74% sur les données en test à comparer aux 51% d'individus atteints par le NPC, ce qui semble a priori un bon résultat. Toutefois, le NPC semble être la cause de la mauvaise aération cuisine à l'enfance et non l'inverse d'après l'orientation des arcs ! Après discussion avec l'expert, ces résultats curieux confirment surtout un biais dit de classement : les individus atteints du NPC (cancer des voies respiratoires) sont plus enclins à chercher les causes de leur maladie dans la mauvaise aération de leur habitat. Souffrant de difficultés respiratoires, ils ont une tendance à imputer à tort la cause de leur cancer à l'aération. Ce résultat curieux n'est donc pas à imputer à PMMS mais aux données elles-mêmes.

En conclusion, ce graphe nous renseigne, certes, sur le mode de vie des sujets maghrébins mais révèle par ailleurs des biais propres aux comportements psychologique des individus, à la façon dont ils comprennent (ou non) les questions. Malgré la pertinence des groupes de variables obtenus et leurs associations, il ne dit rien en revanche sur les "causes potentielles" du NPC, en gardant à l'esprit que, même en supposant l'hypothèse de *suffisance causale* et la fiabilité de la mesure d'association, il est impossible de déceler les relations causales à partir de données sans mener des expérimentations supplémentaires.

6 Conclusion

Nous avons présenté une nouvelle heuristique polynômiale, intitulée Polynomial Max-Min Skeleton (PMMS), proposée en 2006 et validée avec succès sur de nombreux bancs d'essai par Tsamardinos et al. Deux modifications ont été apportées à PMMS pour traiter les dépendances fonctionnelles et les données incomplètes. Après validation sur le banc d'essai $Asia$, nous l'avons appliquée aux données d'une étude épidémiologique du cancer du nasopharynx.

Références

Antal, P., G. Fannes, D. Timmerman, Y. Moreau, et B. D. Moor (2004). Using literature and data to learn bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine 30*(3), 257–281.

Aussem, A., Z. Kebaili, M. Corbex, et F. D. Marchi (2006). 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*, pp. 651–662. RNTI-E-6, Cépaduès-Éditions.

Brown, L. E., I. Tsamardinos, et C. F. Aliferis (2005). 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*, pp. 739–745.

Chickering, D. M., D. Heckerman, et C. Meek (2004). Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research 5*, 1287–1330.

Dash, D. et M. J. Druzdzel (2003). Robust independence testing for constraint-based learning of causal structure. In *UAI*, pp. 167–174.

François, O. (2006). *De l'indentification de la structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplètes*. Thèse de doctorat de l'INSA Rouen.

Friedman, N. (1998). The bayesian structural EM algorithm. In *UAI*, pp. 129–138.

Getoor, L., J. T. Rhee, D. Koller, et P. Small (2004). Understanding tuberculosis epidemiology using structured statistical models. *Artificial Intelligence in Medicine 30*(3), 233–256.

Lauritzen, S. et D. Speigelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Royal statistical Society B 50*, 157–224.

Leray, P. et O. Francois (2004). BNT structure learning package : Documentation and experiments. Technical report, Laboratoire PSI.

Murphy, K. (2001). The bayesnet toolbox for matlab. In *Computing Science and Statistics : Proceedings of Interface*, Volume 33.

Naim, P., P. Wuillemin, P. Leray, O. Pourret, et A. Becker (2004). *Réseaux bayesiens*. Eyrolles.

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

Pearl, J. (2000). *Causality : Models, Reasoning, and Inference*. Cambridge, England : Cambridge University Press.

Ramoni, M. et P. Sebastiani (2001). Robust learning with missing data. *Machine Learning 45*(2), 147–170.

Spirtes, P., C. Glymour, et R. Scheines (2000). *Causation, Prediction, and Search* (2 ed.). The MIT Press.

Tsamardinos, I., L. E. Brown, et C. F. Aliferis (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning 65*(1), 31–78.

Summary

Learning the structure of a bayesian network from a data set is NP-hard. Several methods have been proposed. In this paper, we discuss a new heuristic called PMMS developed 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 algorithms, particularly 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 in groups of variables. In this paper, we propose a way to overcome these problems. The new version of PMMS is first successfully applied on the Asia network benchmark to recover the original structure from data. The algorithm is then applied on the nasopharyngeal cancer (NPC) data in order to shed some light into the statistical profile of the population under study.

Réseau bayésien appliqué aux données du cancer

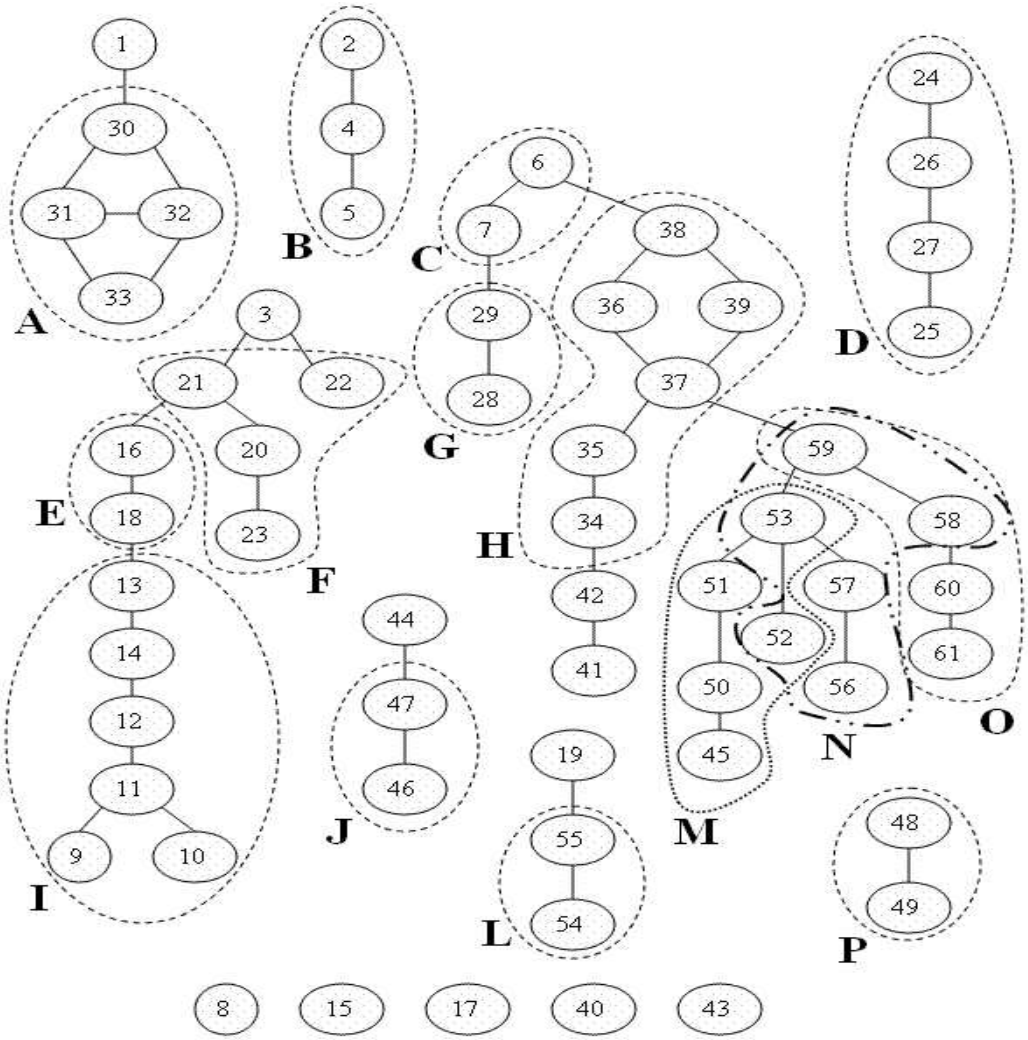

FIG. 2 – *Squelette du RB obtenu avec PMMS revisité. En pointillé : les groupes de variables thématiques. Lexique :* **NPC** *1, age à l'interview pour les témoins et au cancer pour les cas 2, sexe 3, niveau d'instruction 4, catégorie professionnelle 5, habitat dans l'enf. et ad. 6 7, parents consanguins 8, fréquentes otites 9, fréquentes angines 10, fréquents rhume 11, asthme 12, eczéma 13, allergie 14, engrais chimiques et pesticides 15, produits chimiques 16, exposition aux fumées 17, exposition aux poussières 18, exposition aux formaldéhydes 19, consommation d'alcool 20, consommation de tabac 21, consommation de neffa 22, consommation de cannabis 23, type de logement enf. et ad. 24 25,lits séparés enf. et ad. 26 27, animaux dans la maison enf. et ad. 28 29, aération cuisine enf. et ad. 30 31, aération maison enf. et ad. 32 33, exposition aux fumées d'encens enf. et ad. 34 35, exposition aux fumées de kanoun et tabouna enf. et ad. 36 37, exposition aux fumées de feu de bois enf. et ad. 38 39, allaitement et age au sevrage et modalité de sevrage 40 41 42, contact avec la salive adulte par le sol ou les aliments 43, traitements traditionnels enfance 44, piment 45, smen et graisse enf. et ad. 46 47, legumes fruits agrumes enf. et ad. 48 49, harrissa maison enf. ad. 50 51, harrissa industrielle enf. ad. 52 53, protéines maison enf. ad. 54 55, protéines industrielles enf. ad. 56 57, conserves légumes industrielles enf. ad. 58 59, conserves légumes maison enf. ad. 60 61. enf=enfance et ad=adulte.*