Apprentissage de structure de réseaux bayésiens à partir de réseaux markoviens

A Markov Network-based Approach for Learning the Structure of Bayes Nets

C. Gonzales N. Jouve

LIP6 – université Paris VI
8, rue du capitaine Scott, 75015 Paris
{Christophe.Gonzales,Nicolas.Jouve}@lip6.fr

Résumé

Sous l'hypothèse que les données ont été générées selon une loi de probabilité isomorphe à un graphe orienté sans circuit et que les données sont en nombre suffisant, nous proposons une méthode d'apprentissage de structure des réseaux bayésiens exploitant les propriétés de trois représentations graphiques de l'indépendance probabiliste : les réseaux bayésiens (RB), les réseaux markoviens (RM) et les graphes essentiels (GE). La méthode se déroule comme suit : i) apprendre un RB $\mathcal{B}$ dans un espace contraint par un ordre topologique heuristique ; ii) le moraliser en un RM $\mathcal{G}$; iii) optimiser $\mathcal{G}$ en $\mathcal{G}^*$, le RM optimal, par des tests d'indépendance ; iv) orienter $\mathcal{G}^*$ en un RB $\mathcal{B}'$ de même graphe moral que $\mathcal{B}^*$, le RB optimal ; v) raffiner $\mathcal{B}'$ par recherche locale dans les GE jusqu'à atteindre un équivalent de $\mathcal{B}^*$. Nous fournissons des garanties d'optimalité et de complexité.

Mots-Clefs

Réseaux bayésiens, apprentissage de structure, raisonnement probabiliste.

Abstract

Assuming that the generative probabilistic distribution of the data is DAG-isomorph and that the database is sufficiently large, we derive a method for learning the structure of Bayesian networks that exploits the properties of three graphical models of probabilistic independence : Bayesian networks (BN), Markov networks (MN) and essential graphs (EG). It consists in i) learning a BN, say $\mathcal{B}$, in a space constrained by a heuristic topological order ; ii) moralizing it, hence resulting in a MN $\mathcal{G}$; iii) using independence tests to optimize $\mathcal{G}$ into the optimal MN $\mathcal{G}^*$; iv) adding orientations to $\mathcal{G}^*$ to get a new BN $\mathcal{B}'$ sharing its moral graph with the optimal BN $\mathcal{B}^*$; v) refining $\mathcal{B}'$ into a BN equivalent to $\mathcal{B}^*$ using local search in the EG space.Optimality and complexity results are provided.

Keywords

Bayesian networks, graphical structure learning, probabilistic reasoning.

1 Introduction

Le raisonnement en situation d'incertitude est une préoccupation importante en IA. Par la compacité de leur représentation en machine et la rapidité de leurs calculs, les réseaux bayésiens (RB) permettent une mise en œuvre particulièrement efficace du raisonnement probabiliste et ont ainsi donné lieu à de très nombreuses applications (systèmes de diagnostic [11, 20], filtres de spams [17], modélisation d'utilisateurs [6], détection d'intrusions [8] etc.). Un RB est un graphe orienté sans circuit représentant des indépendances entre variables aléatoires. Sa construction requiert donc d'identifier ces indépendances, ce qui constitue l'un des obstacles principaux à l'utilisation de ce modèle dans des problèmes complexes. En pratique, cette tâche peut être confiée à un ou plusieurs expert. Toutefois un tel travail, cognitivement délicat, a de grandes chances d'être entaché d'incohérences et ne repose de toute façon que sur l'estimation subjective des experts. En outre, cette procédure est lourde, ce qui est d'autant plus pénalisant que l'on peut être amené, dans certains cas, à remettre en cause assez souvent la structure du réseau pour tenir compte de nouvelles informations – dans les systèmes de détection d'intrusion par exemple. C'est pourquoi, depuis quelques années, on note un intérêt croissant pour les procédures permettant un apprentissage automatique du réseau à partir de données. Malheureusement, apprendre un RB optimal par rapport aux données est un problème NP-complet [1].

Pour tenter d'y répondre, trois familles d'algorithmes ont vu le jour. La première (IC [14], PC [21] etc.) consiste à déterminer, dans un premier temps, un graphe non orienté tel que toute variable est indépendante des variables qui ne lui sont pas adjacentes conditionnellement à un certain sous-ensemble X du reste des variables, puis, dans un second temps, à orienter ce graphe pour obtenir un RB. La détermination des ensembles X étant exponentielle, ces algorithmes sont peu efficaces sur les problèmes de grande taille. La deuxième famille (K2 [3], SEM [4] etc.) procède par recherche locale dans l'espace des RB, après y avoir défini un voisinage et une fonction de score. Toutefois, la présence, dans le voisinage d'un RB, de beaucoup

d'autres RB équivalents (au sens où ils encodent les mêmes indépendances) conduit à de nombreuses évaluations de la même classe d'équivalence et génère de multiples optima locaux. Pour pallier cela, la troisième famille (GES [2], EQ [12] etc.) déplace la recherche dans l'espace des graphes essentiels (GE), c'est-à-dire l'espace des classes d'équivalence des RB. Cet espace peut être parcouru sans rencontrer d'optima locaux non globaux mais présente l'inconvénient de nécessiter des opérateurs de voisinage nombreux et/ou complexes.

Il existe un troisième espace dans lequel on peut apprendre les indépendances de la loi de probabilité : c'est l'espace des réseaux markoviens (RM) [5]. Les RM sont des graphes non orientés, tels que toute variable est indépendante conditionnellement à ses voisines du reste des variables. Cet espace est assez proche de celui utilisé par des algorithmes tels que IC ou PC, à ceci près que l'ensemble X est ici défini localement autour de chaque variable aléatoire, réduisant ainsi drastiquement la recherche de cet ensemble. Si, comme celui des GE, cet espace peut être parcouru sans risque de tomber sur un optimum local non global,son exploration ne nécessite en revanche, selon la méthode employée, qu'un ou deux opérateurs, d'évaluations simples. Une recherche dans cet espace a donc davantage de chances de converger rapidement. Cependant, les RM sont peu utilisés pour l'apprentissage des RB car les deux types de réseaux ne caractérisent pas les mêmes distributions.

Dans cet article, nous proposons un nouvel algorithme d'apprentissage permettant de mettre à profit les avantages de l'espace des RM pour obtenir rapidement un réseau bayésien optimal. Plus précisément, après avoir introduit formellement les différents types de réseaux dans la section 2 et présenté un exemple motivant l'utilisation des phases successives de l'algorithme dans la section 3, nous détaillons ces différentes phases dans les trois sections suivantes.

2 Notations et généralités

Dans cette section, nous présentons les notions et notations indispendables pour la suite.

On se donne un ensemble $\mathcal{V}$ de n variables aléatoires à valeurs discrètes. Nous notons en capitales grasses ($\mathbf{X}, \mathbf{Y}, \ldots$) des sous-ensembles de $\mathcal{V}$ et en capitales maigres ($A, T, X_i, \ldots$) des éléments de $\mathcal{V}$ (la lettre P désigne toutefois une probabilité). Les valeurs possibles d'une variable s'appellent des *modalités* et nous notons r_Y (resp. r_i) le nombre de modalités de Y (resp.X_i). Une *valuation* de $\mathcal{V}$ est un n-uplet $(x_i)_{1\leq i\leq n}$, où $\forall i, x_i$ est la valeur de X_i. Les *données* $\mathcal{D}$ sont une collection de m valuations de $\mathcal{V}$ que l'on suppose générée conformément à une loi de probabilité $P(\mathcal{V})$. On dit que P est *strictement positive* (noté $P > 0$) si $P(v) > 0$ pour toute valution v de $\mathcal{V}$. Dans la suite, nous allons être amenés à considérer des graphes dont les nœuds sont les variables de $\mathcal{V}$. Nous parlerons alors indifféremment de nœud ou de variable. En général, nous notons $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ un graphe non orienté (GNO) et $\mathcal{B} = (\mathcal{V}, \mathcal{A})$ un graphe orienté sans circuit (GOSC). Pour un GOSC, **Pa**(X) désigne les parents de X. La paire $(X, Y) \in \mathcal{E}$ désigne une arête entre X et Y et le couple $(X, Y) \in \mathcal{A}$ l'arc entre X et Y pointant vers Y. Rappelons enfin qu'une *clique* est un sous-graphe complet maximal au sens de l'inclusion et que, lorsque $P(\mathbf{Y}) \neq 0$, la probabilité de $\mathbf{X}$ *conditionnellement* à $\mathbf{Y}$ (on dit aussi *sachant* $\mathbf{Y}$) se note $P(\mathbf{X} \mid \mathbf{Y})$ et vaut par définition $\frac{P(\mathbf{X},\mathbf{Y})}{P(\mathbf{Y})}$.

Définition 1 *$\mathbf{X}$ et $\mathbf{Y}$ sont dits indépendants conditionnellement à $\mathbf{Z}$, ce que l'on note $\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$, si*

$$P(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})P(\mathbf{Y} \mid \mathbf{Z}).$$

L'ensemble $\mathbf{Z}$ peut être vide et l'on parle alors d'*indépendance marginale* entre $\mathbf{X}$ et $\mathbf{Y}$. La loi P induit donc une relation ternaire $(\cdot) \perp\!\!\!\perp_P (\cdot) \mid (\cdot)$ qui se caractérise par l'axiomatique des semi-graphoïdes [13]. Une définition équivalente de l'indépendance nous est fournie par le théorème 1 [9] :

Théorème 1

$$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \Longleftrightarrow \exists f, g : P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = f(\mathbf{X}, \mathbf{Z})g(\mathbf{Y}, \mathbf{Z}).$$

Dans les deux cas, il s'agit d'une factorisation de la loi jointe de $\mathbf{X}$, $\mathbf{Y}$ et $\mathbf{Z}$ faisant intervenir $\mathbf{X}$ et $\mathbf{Y}$ dans des facteurs séparés. Dans la première décomposition, les facteurs sont des probabilités conditionnelles, tandis que dans la seconde, ce ne sont que des fonctions quelconques.

Réseaux markoviens (RM) et bayésiens (RB) sont des représentations graphiques de la relation d'indépendance : la nature des chaînes ou des chemins liant deux variables dans le graphe rend compte d'éventuelles indépendances entre elles. Ces indépendances encodées par le réseau expriment, par l'intermédiaire des caractérisations ci-dessus, une factorisation la loi de probabilité jointe des variables. C'est cette factorisation qui rend possible l'exploitation calculatoire d'une loi jointe multidimensionnelle, de taille combinatoire, en la décomposant en une série d'éléments de petites dimensions. Nous allons voir que les RM sont des GNO s'appuyant sur la seconde caractérisation, et les RB des GOSC s'appuyant sur la première.

2.1 Réseaux markoviens et bayésiens

La définition suivante fournit le critère graphique de représentation de l'indépendance pour les RM.

Définition 2 *Soit $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ un GNO. On note $\mathbf{X} \perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}$ et on dit que $\mathbf{X}$ et $\mathbf{Y}$ sont séparés par $\mathbf{Z}$ si toute chaîne entre $\mathbf{X}$ et $\mathbf{Y}$ possède un nœud dans $\mathbf{Z}$.*

Définition 3 *Une loi P est isomorphe à un GNO s'il existe un GNO $\mathcal{G}$ tel que*

$$\mathbf{X} \perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z} \Longleftrightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}.$$

On dit alors que $\mathcal{G}$ est une représentation parfaite de P.

Malheureusement, beaucoup de lois simples n'ont pas de représentation non orientée parfaite. Par exemple, la loi P_1 définie sur $\{A, B, C\}$ et ayant pour unique indépendance $A \perp\!\!\!\perp B$ n'est pas isomorphe à un GNO. On va donc s'intéresser à une représentation moins exigeante.

Définition 4 *Un GNO $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ est un réseau markovien pour P si*

$$\mathbf{X} \perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z} \Longrightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}.$$

Cette fois, toute loi possède au moins un RM : le graphe complet. Un RM est donc, à travers la relation $\perp\!\!\!\perp_P$, une représentation graphique non orientée de P. Plus précisément, le théorème 2 dû à Hammersley et Clifford (voir par exemple [10]) caractérise la décomposition de la loi de probabilité relative à un RM.

Théorème 2 *Soient $\mathcal{G}$ un GNO et $\mathcal{C}$ l'ensemble de ses cliques. Soit $P > 0$. $\mathcal{G}$ est un RM pour P si et seulement si*

$$P(\mathcal{V}) = \prod_{\mathbf{C} \in \mathcal{C}} \psi_{\mathbf{C}}(\mathcal{V}),$$

où $\psi_{\mathbf{C}}$ ne dépend que des $X \in \mathcal{V}$ tels que $X \in \mathbf{C}$.

Notons que les fonctions ψ sont nécessairement strictement positives. Les décompositions les plus abouties étant bien sûr les plus intéressantes, on préfèrera les réseaux encodant le plus possible d'indépendances. Un RM pour P est dit *minimal* (RMmin) si aucun de ses graphes partiels (c'est-à-dire dont on a enlevé au moins une arête) n'est un RM pour P. Pearl [13] a montré que si $P > 0$, il existe un unique RMmin pour P, que nous notons alors $\mathcal{G}^*$.
Le critère graphique d'indépendance pour les RB est un peu plus complexe. Si, dans un GOSC, les nœuds R, S et T forment une chaîne alors S est dit *convergent* sur cette chaîne si les deux arcs pointent vers lui, et *non convergent* sinon.

Définition 5 *Soit $\mathcal{B} = (\mathcal{V}, \mathcal{A})$ un GOSC. On note $\mathbf{X} \perp_{\mathcal{B}} \mathbf{Y} \mid \mathbf{Z}$ et on dit que $\mathbf{X}$ et $\mathbf{Y}$ sont d-séparés par $\mathbf{Z}$ si pour toute chaîne entre $\mathbf{X}$ et $\mathbf{Y}$ il existe un nœud S de la chaîne tel que*
- *si S est convergent sur la chaîne, ni S ni aucun de ses descendants n'appartiennent à $\mathbf{Z}$,*
- *sinon, S appartient à $\mathbf{Z}$.*

Définition 6 *Une loi P est isomorphe à un GOSC s'il existe un GOSC $\mathcal{B}$ tel que*

$$\mathbf{X} \perp_{\mathcal{B}} \mathbf{Y} \mid \mathbf{Z} \Longleftrightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}.$$

On dit alors que $\mathcal{B}$ est une représentation parfaite de P.

Là encore, il existe des lois ne satisfaisant pas à cette propriété. Par exemple, la loi P_2 définie sur $\{A, B, C, D\}$ et ayant pour indépendances $A \perp\!\!\!\perp D \mid \{B, C\}$ et $B \perp\!\!\!\perp C \mid \{A, D\}$ n'est pas isomorphe à un GOSC.

Définition 7 *Un GOSC $\mathcal{B} = (\mathcal{V}, \mathcal{A})$ est un réseau bayésien pour P si*

$$\mathbf{X} \perp_{\mathcal{B}} \mathbf{Y} \mid \mathbf{Z} \Longrightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}.$$

Ainsi, le graphe complet est un RB pour toute loi. A l'instar des RM, un RB est donc une représentation graphique, orientée cette fois, encodant les indépendances de P. Le théorème 3, dû à Verma [22], explicite la décomposition induite sur P par un RB.

Théorème 3 *$\mathcal{B}$ est un RB pour P si et seulement si*

$$P(\mathcal{V}) = \prod_{i=1}^{n} P(X_i | \mathbf{Pa}(X_i)). \tag{1}$$

Un RB pour P est dit *minimal* (RBmin) si aucun de ses graphes partiels (c'est-à-dire dont on a enlevé au moins un arc) n'est un RB pour P.
RM et RB fournissent deux représentations graphiques de l'indépendance probabiliste différentes malgré leurs similarités. Il existe en effet des ensembles d'indépendances représentables par un modèle mais pas par l'autre, et réciproquement. Par exemple, la loi P_1, qui n'avait pas de représentation non orientée parfaite, est isomorphe au RB de la figure 1.b. Réciproquement, la loi P_2, qui n'avait pas de représentation orientée parfaite, est isomorphe au RM de la figure 1.a. L'espace des RB est malgré tout plus expressif que celui des RM. Comme, en plus, la décomposition en probabilités conditionnelles se manipule et s'interprète beaucoup plus aisément que celle en simples facteurs positifs, on utilise plutôt les RB que les RM lorsqu'il s'agit de modéliser un problème.

(a) diamant (b) V-structure

FIG. 1 – Différence sémantique entre RB et RM

Passer d'un RB à un RM sans perdre trop d'informations est relativement aisé et nécessite la moralisation du RB.

Définition 8 *Soit $\mathcal{B} = (\mathcal{V}, \mathcal{A})$ un GOSC. Une V-structure est un triplet $(A, C, B) \in \mathcal{V}^3$ tel que (i) $(A, C) \in \mathcal{A}$ et $(B, C) \in \mathcal{A}$ et (ii) A et B ne sont pas adjacents.*

Le graphe de la figure 1.b est une V-structure.

Définition 9 *Le graphe moral d'un GOSC $\mathcal{B}$ est un GNO $\mathcal{G}$ qui s'obtient à partir de $\mathcal{B}$ (i) en ajoutant pour chaque V-structure (A, C, B) une arête (A, B), appelée arête de moralisation, puis (ii) en supprimant l'orientation des arcs.*

Le graphe moral du GOSC de la figure 1(b) est le GNO complet à trois nœuds. On montre facilement que le graphe

moral d'un RB pour P est un RM pour P. En revanche, passer d'un RM à un RB est une tâche plus ardue, pour laquelle les sections 5 et 6 proposent une solution.
Enfin, il convient de noter que Pearl [13], contrairement à beaucoup d'auteurs qui lui sont postérieurs, réserve les appellations « réseau markovien » et « réseau bayésien » aux seuls réseaux minimaux. De plus, alors qu'il identifie, comme nous venons de le présenter, un RB à un graphe, les RB sont désormais le plus souvent définis par le couple composé par la *structure – i.e.* le graphe – et les tables de probabilités conditionnelles qu'elle induit – *i.e.* les facteurs de la décomposition du théorème 3.

2.2 Optimalité

Cherchons maintenant à caractériser un « bon » réseau. Pour les RM, on note $\mathcal{G}' \leq \mathcal{G}$ si toute loi P décomposable selon $\mathcal{G}'$, au sens du théorème 2, l'est aussi selon $\mathcal{G}$. Autrement dit, $\mathcal{G}' \leq \mathcal{G}$ si toute indépendance encodée par $\mathcal{G}$ l'est aussi par $\mathcal{G}'$. En ce sens, $\mathcal{G}$ est un modèle « contenant » $\mathcal{G}'$, c'est-à-dire plus général. Deux RM $\mathcal{G}$ et $\mathcal{G}'$ sont dits *équivalents*, ce qu'on note $\mathcal{G} \simeq \mathcal{G}'$, si $\mathcal{G}' \leq \mathcal{G}$ et $\mathcal{G} \leq \mathcal{G}'$. On note de plus $\mathcal{G}' < \mathcal{G}$ si $\mathcal{G}' \leq \mathcal{G}$ et $\mathcal{G} \not\simeq \mathcal{G}'$. Ces relations sont définies de façon similaire sur l'ensemble des RB, en s'appuyant sur la décomposition du théorème 3. Les RB équivalents sont caractérisés par le théorème 4 [15].

Définition 10 *Le squelette d'un GOSC est le GNO obtenu par la suppression de l'orientation de ses arcs.*

Théorème 4 *Deux RB sont équivalents si et seulement si ils ont le même squelette et le même ensemble de V-structures.*

Un RM $\mathcal{G}$ (resp. RB $\mathcal{B}$) pour P est dit *optimal au sens de l'inclusion* s'il n'existe pas de RM $\mathcal{G}'$ (resp. RB $\mathcal{B}'$) pour P tel que $\mathcal{G}' < \mathcal{G}$. Notons que si pour les RM, lorsque $P > 0$, minimalité et optimalité pour l'inclusion se confondent, ce n'est pas le cas pour les RB, comme le montre l'exemple suivant.

Exemple 1 *Soit P_3 la loi ayant pour unique indépendance $A \perp\!\!\!\perp B$. Considérons les DAG de la figure 2. P_3 est isomorphe à $\mathcal{B}_1$, qui est un RB pour P_3 à la fois minimal et optimal. $\mathcal{B}_2$ est un RB pour P_3 minimal mais pas optimal puisque $\mathcal{B}_1$ est un RB pour P_3 tel que $\mathcal{B}_1 < \mathcal{B}_2$. Enfin, $\mathcal{B}_3$ est un RB pour P_3 qui n'est ni optimal, pour la même raison que $\mathcal{B}_2$, ni minimal puisque $\mathcal{B}_1$, qui est un RB pour P_3, est un graphe partiel de $\mathcal{B}_3$.*

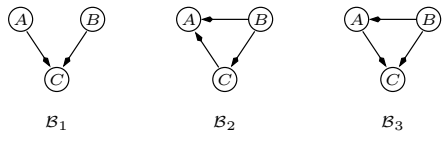


FIG. 2 – Différence entre minimalité et optimalité

Si P est isomorphe à un GOSC, on est assuré de l'existence d'un RB optimal, que nous noterons $\mathcal{B}^*$, et s'il en existe plusieurs, tous sont équivalents à $\mathcal{B}^*$. De plus, sous cette hypothèse, on peut facilement montrer à partir des axiomatiques développées par Pearl [13] que $P > 0$, ce qui autorise à parler de $\mathcal{G}^*$. $\mathcal{G}^*$ est alors l'unique RM optimal pour P et c'est de plus le graphe moral de $\mathcal{B}^*$. Dans cet article, *nous faisons l'hypothèse que P est isomorphe à un GOSC* pour nous intéresser à la recherche d'un RB appartenant à la classe d'équivalence de $\mathcal{B}^*$. Pour ce faire, nous recherchons d'abord $\mathcal{G}^*$.

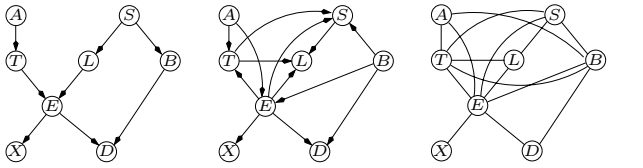
3 Présentation synthétique de notre méthode d'apprentissage

Cette section motive et présente la méthode d'apprentissage que nous proposons.
Apprendre la structure du RB optimal $\mathcal{B}^*$ représentant P revient à rechercher des indépendances conditionnelles de cette loi par l'intermédiaire de la base de données, supposée être un échantillon de P. L'espace des structures possibles est super-exponentiel en le nombre de variables (le nombre a_k de GOSC à k nœuds est $\sum_{i=1}^{k}(-1)^{i+1}\binom{k}{i}2^{i(k-1)}a_{k-i} = k^{2^{\mathcal{O}(k)}}$ [16]) et [1] a montré que trouver $\mathcal{B}^*$ est un problème NP-complet. Aussi peut-on être tenté d'effectuer une recherche gloutonne ou une recherche locale heuristique dans l'espace des RB, de manière à obtenir, sinon un réseau optimal, du moins un « bon » réseau, au sens d'un certain score. Dans ce cadre, un RB est considéré voisin d'un autre s'ils ne diffèrent que d'un arc. Malheureusement la structure de ce voisinage est telle qu'il y a dans le voisinage d'un RB beaucoup d'autres RB qui lui sont équivalents et cela pose principalement deux problèmes. D'une part, comme il peut y avoir un très grand nombre de RB au sein d'une classe d'équivalence, l'algorithme de recherche peut passer beaucoup de temps à évaluer des structures qui sont en fait toutes équivalentes entre elles. D'autre part, la présence de voisins de même score conduit à de nombreux optima locaux de la fonction de score, que les algorithmes gloutons ou les métaheuristiques ont de grandes difficultés à éviter.
Pour remédier à cela, il a été proposé [12, 2] d'utiliser l'espace des classes d'équivalences des RB. Une telle classe peut être caractérisée graphiquement par son graphe essentiel (GE), qui est un graphe partiellement orienté (GPO).

Définition 11 *Le graphe essentiel d'un RB $\mathcal{B}$ est un GPO de même squelette que $\mathcal{B}$ et dont les arcs sont exactement ceux qui sont orientés identiquement dans tous les RB $\mathcal{B}'$ tels que $\mathcal{B}' \simeq \mathcal{B}$.*

Intuitivement, les arcs du GE correspondent aux V-structures – communes à tous les RB de la classe d'équivalence (cf. théorème 4) – et aux arcs qui ne pourraient être orientés différemment sans entraîner de circuit ou de nouvelle V-structure. Par exemple, le graphe de la figure 4.c est le graphe essentiel de la figure 4.b et le graphe de la figure 4.d est celui du graphe de la figure 4.e. On dit d'un GPO qu'il est *instanciable* s'il existe un RB dont il est le GE

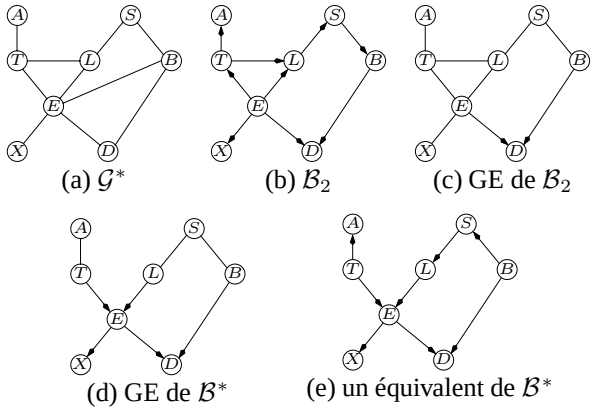
et on appelle *instanciation* d'un GE tout RB de la classe d'équivalence qu'il caractérise.
Dans l'espace des GE, on dispose donc d'un représentant unique par classe d'équivalence, évitant ainsi un certain nombre de problèmes rencontrés dans l'espace des GOSC. L'espace des GE possède même la propriété remarquable que l'on peut s'y déplacer sans jamais tomber sur un optimum local non global, ce qui assure la convergence vers $\mathcal{B}^*$. Un tel parcours nécessite toutefois, à chaque mouvement, l'exploration d'un voisinage exponentiel en n [2].
Nous proposons une méthode, assurant également la convergence vers $\mathcal{B}^*$, mais menant une partie de la recherche dans l'espace des RM, où le voisinage est de taille polynomiale en n. Comme de plus le nombre d'états traversés est borné polynomialement, nous obtenons finalement une complexité globale meilleure que celle des méthodes existantes. Plus précisément, la méthode que nous proposons se compose de trois phases. La première est l'obtention du RM optimal $\mathcal{G}^*$ par recherche dans l'espace des RM. Cette recherche est initialisée grâce à un premier RB, obtenu par une recherche gloutonne dans un espace de RB contraints par un ordre arbitraire des variables. Cet espace a la même structure que celui des RM. La deuxième phase oriente $\mathcal{G}^*$ en un RB $\mathcal{B}$, et nous assure d'une propriété fondamentale : $\mathcal{G}^*$, graphe moral de $\mathcal{B}^*$, est aussi celui de $\mathcal{B}$. La troisième phase, enfin, mène une recherche dans l'espace des GE à partir du GE de $\mathcal{B}$, où l'évaluation du voisinage, grâce à la propriété ci-dessus, est plus légère que dans les méthodes existantes.



(a) $\mathcal{B}^*$ (b) $\mathcal{B}_1$ (c) graphe moral de $\mathcal{B}_1$

FIG. 3 – Construction du RM initial

Voici, sur un exemple, le cheminement général de notre méthode, dont le détail sera exposé dans les sections suivantes. Supposons la loi P isomorphe au GOSC de la figure 3.a. On construit d'abord un RB contraint par un ordre arbitraire des variables, afin de pouvoir initialiser correctement la recherche de $\mathcal{G}^*$. Si par exemple l'ordre choisi est A, B, E, D, T, S, L, X, on peut obtenir le GOSC $\mathcal{B}_1$ (figure 3.b). $\mathcal{B}_1$ est bien un RB pour P – on peut vérifier que $\mathcal{B}^* \leq \mathcal{B}_1$ – et son ordre topologique respecte l'ordre choisi. Son graphe moral (figure 3.c) est un RM pour P et on va en ôter de manière gloutonne toutes les arêtes superflues, grâce à des tests d'indépendances. On obtient ainsi $\mathcal{G}^*$ (figure 4.a). Ensuite, on oriente $\mathcal{G}^*$ en en enlevant éventuellement quelques arêtes, selon la procédure décrite dans la section 5, pour aboutir à un GOSC $\mathcal{B}_2$ (figure 4.b). $\mathcal{B}_2$ est un RB pour P et son graphe moral est $\mathcal{G}^*$. Enfin, on effectue une recherche dans l'espace des GE, décrite section 6, menant du GE de $\mathcal{B}_2$ (figure 4.c) au GE de $\mathcal{B}^*$ (figure 4.d). On instancie alors ce GE en un RB équivalent à $\mathcal{B}^*$ (figure 4.e).



(a) $\mathcal{G}^*$ (b) $\mathcal{B}_2$ (c) GE de $\mathcal{B}_2$
(d) GE de $\mathcal{B}^*$ (e) un équivalent de $\mathcal{B}^*$

FIG. 4 – Recherche du GE optimal

4 Obtention de $\mathcal{G}^*$

La première étape de notre apprentissage consiste donc à appliquer une procédure gloutonne de type K2 [3] afin d'obtenir un premier RB nous permettant de déduire un RM. Nous utilisons l'algorithme K2_BIC qui reprend le principe de K2 mais utilise le score BIC [18], qui possède de meilleures propriétés que celui utilisé dans [3].

Algorithme K2_BIC($\mathcal{V}$,$\mathcal{D}$)
Entrée : $\mathcal{V} = \{X_1, \ldots, X_n\}$: ensemble ordonné de variables aléatoires, $\mathcal{D}$: base de données sur $\mathcal{V}$
Sortie : un RB $\mathcal{B}$
1. $\mathcal{A} \leftarrow \emptyset$ et $S_{new} \leftarrow S_{BIC}((\mathcal{V}, \mathcal{A}), \mathcal{D})$
2. **pour** $i = 1, \ldots, n$ **faire**
3. **répéter**
4. $S_{old} \leftarrow S_{new}$
5. $X_j \leftarrow \text{argmax}_{X_k : k<i \text{ et } (X_k, X_i) \notin \mathcal{A}} \{s_{BIC}(X_i, \mathbf{Pa}^{\mathcal{A}}(X_i) \cup X_k) - s_{BIC}(X_i, \mathbf{Pa}^{\mathcal{A}}(X_i))\}$
6. $S_{new} \leftarrow S_{BIC}((\mathcal{V}, \mathcal{A} \cup \{(X_j, X_i)\}), \mathcal{D})$
7. **Si** $S_{new} > S_{old}$ **alors** $\mathcal{A} \leftarrow \mathcal{A} \cup \{(X_j, X_i)\}$
8. **jusqu'à** $S_{new} \leq S_{old}$
9. **fait**
10. **retourner** $\mathcal{B} = (\mathcal{V}, \mathcal{A})$

L'étape élémentaire d'un tel algorithme consiste à comparer deux RB différant par un arc, au moyen d'une fonction de score $S(\mathcal{B}, \mathcal{D})$, pour décider de conserver ou non cet arc. Cette fonction peut posséder différentes propriétés. D'abord, $S(\mathcal{B}, \mathcal{D})$ est dite *décomposable* si elle s'écrit

$$S(\mathcal{B}, \mathcal{D}) = \sum_{i=1}^{n} s^{\mathcal{B}}(X_i, \mathbf{Pa}^{\mathcal{B}}(X_i)). \qquad (2)$$

La différence de score entre deux RB $\mathcal{B}$ et $\mathcal{B}'$ différant seulement par un arc (X, Y) peut alors être calculée localement par $s^{\mathcal{B}'}(Y, \mathbf{Pa}^{\mathcal{B}'}(Y)) - s^{\mathcal{B}}(Y, \mathbf{Pa}^{\mathcal{B}}(Y))$. Ensuite, on dit que $S(\mathcal{B}, \mathcal{D})$ *préserve l'équivalence* si, lorsque deux

RB $\mathcal{B}$ et $\mathcal{B}'$ sont équivalents, alors $S(\mathcal{B},\mathcal{D}) = S(\mathcal{B}',\mathcal{D})$. Enfin, $S(\mathcal{B},\mathcal{D})$ est dite *localement cohérente* si, lorsque deux RB $\mathcal{B}$ et $\mathcal{B}'$ diffèrent uniquement par le fait que $\mathcal{B}'$ contient l'arc (X_i, X_j) et pas $\mathcal{B}$, alors :

$$\text{si } X_j \not\perp\!\!\!\perp_P X_i|\mathbf{Pa}^{\mathcal{B}}(X_j), \text{ alors } S(\mathcal{B}',\mathcal{D}) > S(\mathcal{B},\mathcal{D}) \quad (3)$$

$$\text{si } X_j \perp\!\!\!\perp_P X_i|\mathbf{Pa}^{\mathcal{B}}(X_j), \text{ alors } S(\mathcal{B}',\mathcal{D}) < S(\mathcal{B},\mathcal{D}). \quad (4)$$

Le score BIC utilisé est défini par :

$$S_{BIC}(\mathcal{B},\mathcal{D}) = \log P(\mathcal{D}|\mathcal{B},\hat{\theta}) - \frac{1}{2} Dim(\mathcal{B}) \log m,$$

où $\hat{\theta}$ représente les valeurs des probabilités du RB obtenues par maximum de vraisemblance. Le second terme pénalise les modèles trop complexes en tenant compte de la dimension, définie par $Dim(\mathcal{B}) = \sum_{i=1}^{n} Dim(X_i,\mathcal{B})$, où

$$Dim(X_i,\mathcal{B}) = (r_i - 1)q_i \quad \text{et} \quad q_i = \prod_{X_j \in \mathbf{Pa}(X_i)} r_j.$$

Le premier terme, lui, rend compte de l'adéquation du modèle $\mathcal{B}$ aux données $\mathcal{D}$ et se calcule par

$$\log P(\mathcal{D}|\mathcal{B},\hat{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}}, \quad (5)$$

où N_{ijk} représente le nombre d'enregistrements de $\mathcal{D}$ dans lesquels la valeur de X_i est sa $k$ème modalité et la valeur de l'ensemble des parents de X_i est la $j$ème modalité possible, et $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. On montre que le score BIC est décomposable, localement cohérent et préserve l'équivalence. Toutefois, la cohérence locale ne peut être établie qu'à la limite, quand la taille m de la base de données croît. Aussi, dans toute la suite, on fera *l'hypothèse que les données sont en nombre suffisant*.

K2_BIC part du graphe vide pour y ajouter des arcs. Il nécessite un ordre – *a priori* arbitraire – sur les nœuds et, parcourant les nœuds dans ledit ordre, insère de manière gloutonne un arc pointant de A vers B, où A précède B dans l'ordre, quand cela améliore le score de la structure.

Lemme 5 *Si les données sont en nombre suffisant, quel que soit l'ordre sur les variables, l'algorithme* K2_BIC($\mathcal{V}$,$\mathcal{D}$) *produit un RB pour P.*

Esquisse de preuve : Supposons que ce ne soit pas le cas, alors il existe une indépendance dans le GOSC $\mathcal{B}$ produit par K2_BIC non contenue dans P. Autrement dit, il existe X_i et Y tels que $X_i \not\perp\!\!\!\perp Y|\mathbf{Pa}(X_i)$ selon la loi P tandis que $X_i \perp\!\!\!\perp Y|\mathbf{Pa}(X_i)$ selon $\mathcal{B}$. On montre alors que rajouter un arc entre X_i et Y augmente le score sans créer de circuit. Donc K2_BIC n'aurait pas produit $\mathcal{B}$. ♦

Notons que la construction du RB est plus ou moins efficace selon l'ordre choisi et une bonne heuristique consiste, lorsque l'on dispose d'une telle information, à ordonner les variables causes avant leurs conséquences. Il est également possible de lancer plusieurs fois l'algorithme avec des ordres différents afin d'en trouver un intéressant, c'est-à-dire aboutissant à un RB sans trop d'arcs. K2_BIC a une complexité en $O(mn^4r)$, où $r = max_{\mathcal{V}} r_X$.

On s'intéresse maintenant au graphe moral du RB obtenu avec cet algorithme. Bien évidemment, ce graphe est un RM mais n'a pas de raison d'être $\mathcal{G}^*$. Or, dans le cadre où nous nous plaçons, $\mathcal{G}^*$ est unique et tout autre RM s'en déduit par ajout d'arêtes. Aussi, pour atteindre $\mathcal{G}^*$, il suffit de supprimer les arêtes (X, Y) inutiles, c'est-à-dire celles dont la suppression ne remet pas en cause le fait que l'on a bien un RM. Dans ce cas, on sait que $X \perp\!\!\!\perp Y|\mathbf{Z}$ quel que soit le sous-ensemble $\mathbf{Z}$ de $\mathcal{V}$ tel que toute chaîne du RM entre X et Y, à l'exception de celle ne contenant que l'arête (X, Y), passe par un nœud de $\mathbf{Z}$. En particulier, on peut choisir pour $\mathbf{Z}$ l'ensemble des voisins de X privé de Y. Pour savoir si on peut se passer de l'arête (X, Y), il suffit donc de déterminer si on a bien $X \perp\!\!\!\perp Y|\mathbf{Z}$. Pour cela, on effectue un test statistique indiquant dans quelle mesure on dévie de l'indépendance conditionnelle :

Définition 12 *Soient trois groupes de variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ et une base de données $\mathcal{D}$. Soient r_X, r_Y, r_Z le nombre de modalités de ces groupes. La déviance par rapport à l'indépendance conditionnelle $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$ est égale à :*

$$dev(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}) = 2\sum_{i=1}^{r_X}\sum_{j=1}^{r_Y}\sum_{k=1}^{r_Z} N_{ijk} \log \frac{N_{ijk}N_k}{N_{ik}N_{jk}}.$$

$dev(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$ suit une loi du χ^2 à $(r_X - 1)(r_Y - 1)r_Z$ degrés de liberté.

Sachant la loi du χ^2 suivie par la déviance, on peut faire un test d'hypothèse statistique classique pour déterminer s'il y a effectivement indépendance de X et Y conditionnellement à $\mathbf{Z}$. Remarquons que cela revient à effectuer un test du G^2. On peut maintenant définir une fonction supprimant une arête donnée d'un RM si celle-ci est jugée inutile :

Algorithme SUPPR_MARKOV($(\mathcal{V},\mathcal{E})$,$(X,Y)$,$\mathcal{D}$)
Entrée : $(\mathcal{V},\mathcal{E})$: réseau markovien, (X, Y) : arête à tester, $\mathcal{D}$: base de données
Sortie : le RM sans l'arête si celle-ci est superflue
1. déterminer un $\mathbf{Z} \subset \mathcal{V}$ tel que toute chaîne de $(\mathcal{V},\mathcal{E})$ entre X et Y, à l'exception de celle ne contenant que l'arête (X, Y), passe par un nœud de $\mathbf{Z}$
2. $D \leftarrow dev(X \perp\!\!\!\perp Y|\mathbf{Z})$
3. effectuer test d'hypothèse sur D
4. **Si** test positif **alors** $\mathcal{E} \leftarrow \mathcal{E}\backslash\{(X, Y)\}$
5. **retourner** $(\mathcal{V},\mathcal{E})$

Cet algorithme peut être appelé de manière systématique pour obtenir $\mathcal{G}^*$ à partir de n'importe quel RM :

Algorithme MIN_MARKOV($\mathcal{G} = (\mathcal{V},\mathcal{E})$,$\mathcal{D}$)
Entrée : $\mathcal{G}$: un RM, $\mathcal{D}$: base de données
Sortie : le RM minimal $\mathcal{G}^*$
1. **pour** toute arête (X, Y) de $\mathcal{E}$ **faire**
2. $\quad \mathcal{G} \leftarrow$ SUPPR_MARKOV($\mathcal{G}$,(X, Y),$\mathcal{D}$)
3. **fait**
4. **retourner** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

Lemme 6 *Si la loi P est isomorphe à un GOSC et les données en nombre suffisant alors, quels que soient le RM passé en premier argument et l'ordre dans lequel sont examinées les arêtes (X, Y) à la ligne 1, l'algorithme* MIN_MARKOV *renvoie le RM optimal $\mathcal{G}^*$.*

Esquisse de preuve : La fonction SUPPR_MARKOV teste l'indépendance de X et de Y conditionnellement à des ensembles séparants **Z**. Si un tel ensemble existe alors, en particulier, $X \perp\!\!\!\perp Y|(\mathcal{V}\backslash\{X, Y\})$. Dans ce cas, il est impossible de se voir interdire l'élimination de l'arête (X, Y) à la $k$ième passe de la boucle 1–3 si on peut le faire à la $k + 1$ème passe. Les arêtes n'ont donc besoin d'être examinées qu'une seule fois, et ce dans n'importe quel ordre. Le graphe résultant de MIN_MARKOV est donc un RMmin. Comme $P > 0$, il est de plus optimal. ♦

Le test statistique pouvant être réalisé en $O(mn)$, MIN_MARKOV a une complexité en $O(mn^3)$. Cet algorithme permet théoriquement de retrouver $\mathcal{G}^*$, pour P isomorphe à un GOSC, à partir de n'importe quel RM et donc en particulier du graphe complet. Toutefois, le test statistique effectué par la fonction SUPPR_MARKOV s'accomode très mal d'ensembles conditionnants de cardinal trop élevé. Il est donc préférable de passer en argument à MIN_MARKOV un RM déjà relativement épuré, comme celui obtenu par la moralisation du RB fourni par K2_BIC.

5 Orientation de $\mathcal{G}^*$

Dans cette section, nous décrivons ORIENTE_MARKOV, l'algorithme qui transforme le RM $\mathcal{G}^*$, obtenu à l'issue de la section précédente, en un graphe orienté $\mathcal{B}$ qui est un RB pour P et a le même graphe moral que $\mathcal{B}^*$, à savoir $\mathcal{G}^*$.

5.1 Description de l'algorithme

Cet algorithme s'applique à un RM $\mathcal{G}$ et procède par récurrence sur les nœuds de $\mathcal{G}$ (lg 2) en en éliminant un à chaque étape. L'élimination d'un nœud (lg 8) ne peut être réalisée que s'il existe un nœud du $\mathcal{G}$ courant (sous-graphe du $\mathcal{G}$ initial) qui n'appartient qu'à une seule clique (lg 3). La proposition 8 assure que, si l'on exécute ORIENTE_MARKOV sur $\mathcal{G}^*$, il existera à chaque pas de la boucle 2–12 un X à choisir sur la ligne 3.

Algorithme ORIENTE_MARKOV($\mathcal{G} = (\mathcal{V}, \mathcal{E})$)
Entrée : réseau markovien $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
Sortie : graphe orienté $\mathcal{B} = (\mathcal{V}, \mathcal{A})$
1. $\mathcal{A} \leftarrow \emptyset$; $\mathcal{V}' \leftarrow \mathcal{V}$
2. **tant que** $|\mathcal{V}| > 1$ **faire**
3. $\quad$ Choisir un $X \in \mathcal{V}$ appartenant à une unique clique de $\mathcal{G}$
4. $\quad \mathbf{E} \leftarrow \{Y \in \mathcal{V} : (Y, X) \in \mathcal{E}\}$
5. $\quad$ **pour tout** $Y \in \mathbf{E}$ **faire**
6. $\quad\quad \mathcal{E} \leftarrow \mathcal{E}\backslash\{(Y, X)\}$ et $\mathcal{A} \leftarrow \mathcal{A} \cup \{(Y, X)\}$
7. $\quad$ **fait**
8. $\quad \mathcal{V} \leftarrow \mathcal{V}\backslash\{X\}$
9. $\quad$ **pour toute** paire (Y, Z) d'éléments de **E** **faire**
10. $\quad\quad \mathcal{G} \leftarrow$ SUPPR_MARKOV$(\mathcal{G}, (Y, Z), \mathcal{D})$
11. $\quad$ **fait**
12. **fait**
13. **retourner** $\mathcal{B} = (\mathcal{V}', \mathcal{A})$

Lemme 7 *Soit $\mathcal{B}$ un RB et $\mathcal{G}$ son graphe moral. Soit X une variable appartenant à une seule clique de $\mathcal{G}$ alors il existe un RB $\mathcal{B}'$ tel que : i) $\mathcal{B} \leq \mathcal{B}'$; ii) $\mathcal{B}'$ a pour graphe moral $\mathcal{G}$; et iii) X n'a pas d'enfant dans $\mathcal{B}'$.*

Esquisse de preuve : On montre que si X appartient à une seule clique, ses voisins ne peuvent avoir de parent dans $\mathcal{B}$ en dehors de la clique de X. La suite de la démonstration est assez similaire à celle du théorème 4 de [2] : on montre que l'on peut effectuer une séquence de retournements et d'ajouts d'arcs modifiant $\mathcal{B}$ en $\mathcal{B}'$ et assurant que chaque retournement/ajout laisse le graphe moral invariant. ♦

Proposition 8 *Si P est isomorphe à un GOSC, si l'on applique* ORIENTE_MARKOV *à $\mathcal{G}^*$ alors, à chaque pas de la boucle 2–12, il existe au moins un nœud X se trouvant dans une seule clique.*

Esquisse de preuve : $\mathcal{G}^*$ est le graphe moral de $\mathcal{B}^*$. D'après le lemme 7, on peut transformer $\mathcal{B}^*$ en un nouveau RB $\mathcal{B}'$ tel que $\mathcal{G}^*$ est le graphe moral de $\mathcal{B}'$ et X n'a pas d'enfant dans $\mathcal{B}'$. Les lignes 6, 8 et 9–11 produisent le graphe moral de $\mathcal{B}'$. Chaque étape de la boucle 2–12 itère sur ce processus. Donc on peut toujours trouver un X appartenant à une seule clique. ♦

Revenons à l'algorithme : une fois le nœud X à éliminer choisi (lg 3), on le supprime du $\mathcal{G}$ courant (lg 8), ainsi que ses arêtes adjacentes (lg 5–7), et on ajoute à $\mathcal{B}$ un arc pointant de chaque voisin de X vers X (lg 5–7). L'idée est donc de construire un RB $\mathcal{B}'$ se différenciant de $\mathcal{B}^*$ par l'ordre topologique des variables. Les lignes 9–11 assurent qu'à chaque étape de la boucle 2–12, le graphe $\mathcal{G}$ est toujours le graphe moral d'un RB.

Proposition 9 *Si P est isomorphe à un GOSC et si les données sont en nombre suffisant, l'application de* ORIENTE_MARKOV *à $\mathcal{G}^*$ renvoie un RB pour P, de graphe moral $\mathcal{G}^*$.*

Esquisse de preuve : Le lemme 7 nous fournit une séquence de retournements et d'ajouts d'arcs permettant de passer de $\mathcal{B}^*$ à $\mathcal{B}$. La séquence inverse, associée aux modifications de $\mathcal{G}^*$ des lignes 9–11, permet de retrouver $\mathcal{B}^*$ à partir du graphe produit par ORIENTE_MARKOV. Par conséquent, ORIENTE_MARKOV renvoie un RB. De plus, par construction, son graphe moral est évidemment le RM passé en argument de la fonction, c'est-à-dire $\mathcal{G}^*$. ♦

ORIENTE_MARKOV a une complexité en $O(mn^4)$. Le fait de déjà disposer, à ce stade de la méthode d'apprentissage, d'un RB $\mathcal{B}$ possèdant le même graphe moral que $\mathcal{B}^*$ assure que les algorithmes d'inférence probabiliste comme ceux de Jensen [7] ou de Shafer-Shenoy [19] effectueront les mêmes calculs dans $\mathcal{B}$ et $\mathcal{B}^*$. Autrement dit, pour les applications où l'intelligibilité du modèle n'est pas requise, c'est-à-dire pour lesquelles le RB n'est qu'un élément algorithmique, $\mathcal{B}^*$ n'est pas plus intéressant que le RB $\mathcal{B}$, que nous obtenons en temps polynomial.
Notons pour finir que le choix (lg 3) du nœud à éliminer influe considérablement sur la qualité du RB fourni par ORIENTE_MARKOV. Une heuristique intéressante pour ce choix consiste à classer les nœuds candidats en fonction de la taille de l'unique clique à laquelle ils appartiennent et d'en choisir un pour lequel cette taille est la plus petite. On peut montrer qu'orienter en priorité les nœuds appartenant à une clique de cardinal 2 ne perd aucune indépendance.

5.2 Exemple

Prenons un exemple, illustré par la figure 5. Supposons la loi P isomorphe à $\mathcal{B}^*$. À l'issue de l'apprentissage décrit dans la section 4, on obtient le graphe moral $\mathcal{G}^*$ de $\mathcal{B}^*$ que l'on passe en argument à l'algorithme ORIENTE_MARKOV. La proposition 8 nous assure que l'on trouvera toujours un nœud à éliminer et qu'il y aura donc $n-1$ passages dans la boucle 2–12 (ici $n=7$). Pour $1 \leq i \leq 6$, les graphes $\mathcal{G}_i$ et $\mathcal{B}_i$ représentent respectivement l'état des variables $\mathcal{G}$ et $\mathcal{B}$ après le $i^{\text{ème}}$ tour de boucle. Les nœuds de $\mathcal{G}^*$ n'appartenant qu'à une seule clique sont F et R. À la première itération, ce sont donc eux les candidats à l'élimination et c'est F qui est choisi (lg 3). On supprime donc le nœud F (lg 8) et les arêtes (F,C) et (J,F) (lg 6) du graphe $\mathcal{G}$ courant, à savoir $\mathcal{G}^*$, pour obtenir $\mathcal{G}_1$. Comme il n'y a dans $\mathcal{D}$ aucune indépendance[1] entre C et J, l'appel à SUPPR_MARKOV (lg 10) ne permet pas de supprimer l'arête (C,J) que l'on retrouve donc bien dans $\mathcal{G}_1$. Enfin, on rajoute dans $\mathcal{B}$ les arcs (C,F) et (J,F) (lg 6) pour obtenir $\mathcal{B}_1$. À la deuxième itération, seul R est candidat à l'élimination car c'est le seul nœud de $\mathcal{G}_1$ appartenant à une seule clique. On supprime donc R et ses arêtes adjacentes dans $\mathcal{G}_1$. De plus, il existe cette fois une indépendance entre K et L conditionnellement à $\{C,H,J\}$ que l'on détecte par l'appel à SUPPR_MARKOV. L'arête (K,L) est donc supprimée pour obtenir $\mathcal{G}_2$. Notons que cette arête, que l'on supprime de $\mathcal{G}_1$ avant de l'avoir ajoutée dans $\mathcal{B}_1$, était une arête de moralisation de $\mathcal{B}^*$. Enfin, on ajoute les arcs (K,R) et (L,R) à $\mathcal{B}_1$ pour obtenir $\mathcal{B}_2$. À la troisième itération, K et L sont candidats à l'élimination et c'est L qui est choisi. L n'ayant qu'un seul voisin, il suffit de le retirer de $\mathcal{G}_2$, avec son arête adjacente, pour obtenir $\mathcal{G}_3$ et d'ajouter à $\mathcal{B}_2$ l'arc (J,L) pour obtenir $\mathcal{B}_3$. Les itérations suivantes sont similaires et l'algorithme nous fournit finalement le graphe $\mathcal{B}_6 = \mathcal{B}$.
Il s'agit d'un GOSC et l'on peut vérifier que c'est bien un RB pour $\mathcal{D}$ en s'assurant, d'une part, que toute indépendance obtenue par d-séparation dans $\mathcal{B}$ apparaît également dans $\mathcal{B}^*$ et, d'autre part, que le retrait d'un arc quelconque entraînerait la perte de cette propriété. Enfin, son graphe moral est le même que celui de $\mathcal{B}^*$, à savoir $\mathcal{G}^*$. Comparons maintenant notre RB $\mathcal{B}$ et l'optimum $\mathcal{B}^*$, dont on souhaite se rapprocher. On peut se rendre compte que le fait d'avoir mal orienté un arc comme (C,H) ne crée pas de dépendance importune dans $\mathcal{B}$ car (C,H) n'est qu'une arête, et non arc, dans le GE de $\mathcal{B}$. En revanche, il existe dans P une indépendance marginale entre C et F que l'on ne retrouve pas dans $\mathcal{B}$. Cela résulte du fait qu'à la première itération, on a éliminé le nœud F avant d'avoir eu l'occasion de détecter par un appel à SUPPR_MARKOV que l'une de ses arêtes adjacentes, en l'occurrence (C,F), correspondait à une arête de moralisation de $\mathcal{B}^*$. L'arête (C,F) a donc été orientée avant d'avoir pu être supprimée, engendrant ainsi dans $\mathcal{B}$ une dépendance (potentielle) absente de $\mathcal{B}^*$. Ce mauvais ordonnancement des opérations aurait pu être évité si, à la première itération, on avait choisi d'éliminer R plutôt que F. Plus généralement, on peut montrer que si l'on choisit d'éliminer, parmi l'ensemble des candidats, le nœud qui possède le rang le plus élevé dans l'ordre topologique de $\mathcal{B}^*$, l'algorithme ORIENTE_MARKOV renvoie exactement $\mathcal{B}^*$. Malheureusement, on ne connaît pas cet ordre et le travail de l'algorithme décrit dans la section 6, est précisément de modifier l'ordre topologique de $\mathcal{B}$, en retournant certains de ses arcs, pour converger vers l'ordre de $\mathcal{B}^*$ et pouvoir ainsi lancer les appels à SUPPR_MARKOV qu'ORIENTE_MARKOV n'a pas pu effectuer.

6 Recherche de $\mathcal{B}^*$ dans l'espace des graphes essentiels

La dernière étape de notre apprentissage consiste à « raffiner » le RB $\mathcal{B}$ trouvé précédemment de manière à obtenir le RB optimal $\mathcal{B}^*$ ou un de ses équivalents. Comme nous venons de le voir, la différence fondamentale entre $\mathcal{B}$ et $\mathcal{B}^*$ tient à l'ordre topologique des nœuds. Modifier cet ordre tout en essayant de préserver au maximum les indépendances encodées dans le graphe suggère d'effectuer ces modifications dans l'espace des GE. En effet, les RB qui partagent un même GE représentent les mêmes ensembles d'indépendances tout en ayant des ordres topologiques différents.
On sait que $\mathcal{B}$ est un RB pour P et, par conséquent, que

[1] Pour cet exemple, on peut le voir en observant $\mathcal{B}^*$, à partir duquel les données sont supposées avoir été générées. Toutefois, dans un problème réel, $\mathcal{B}^*$ n'est évidemment pas connu et seul l'appel à SUPPR_MARKOV permet de déterminer une éventuelle indépendance.

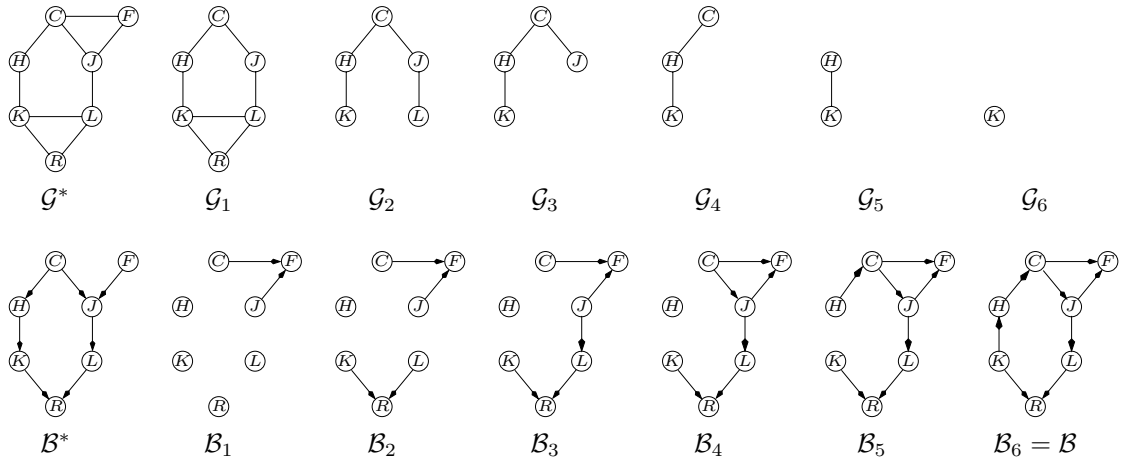


FIG. 5 – Graphes de l'exemple de la sous-section 5.2

toute instanciation du GE de $\mathcal{B}$ est également un RB pour P. D'après Chickering [2], pour obtenir $\mathcal{B}^*$, il suffit d'appliquer la deuxième phase de son algorithme GES, autrement dit : tant que l'on peut trouver un couple de nœuds (X, Y) ainsi qu'un ensemble $\mathbf{H}$ pour lesquels l'application de la fonction DELETE décrite ci-après produit un nouveau GPO instanciable de meilleur score que le score courant, on applique DELETE(X,Y,$\mathbf{H}$). Pour un arc ou une arête (X, Y), et pour tout sous-ensemble $\mathbf{H}$ des voisins de Y (les nœuds reliés à Y par une arête) qui sont aussi adjacents à X (reliés à X par une arête ou un arc), DELETE(X,Y,$\mathbf{H}$) modifie le GE courant en supprimant l'arête ou l'arc (X, Y) puis, pour chaque nœud $H \in \mathbf{H}$: 1) oriente les arêtes (Y, H) de Y vers H ; et 2) oriente les arêtes (X, H) de X vers H. Enfin, DELETE propage les contraintes imposées par ces opérations pour obtenir un nouveau GE. Pour un couple (X, Y), notons $\mathbf{NA}_{Y,X}$ l'ensemble des voisins de Y qui sont aussi adjacents à X. [2] montre qu'il suffit d'examiner les couples (X, Y) et les sous-ensembles $\mathbf{H}$ de $\mathbf{NA}_{Y,X}$ tels que $\mathbf{NA}_{Y,X} \backslash \mathbf{H}$ forment un sous-graphe complet car seuls ces sous-ensembles permettent d'obtenir des GPO instanciables. La complexité de DELETE pour chaque couple (X, Y) de nœuds adjacents est donc en $O(2^{|\mathbf{NA}_{Y,X}|})$ tests, soit, globalement, en $O(mn^3 2^s)$, où s est le nombre maximal de voisins communs aux deux extrémités d'une arête ou d'un arc du GE de $\mathcal{B}$. Comme il y a au plus $\frac{n(n-1)}{2}$ mouvements, la complexité globale de la seconde phase de GES est en $O(mn^5 2^s)$.
Notre algorithme est similaire à celui de Chickering à ceci près que, tandis que ce dernier doit systématiquement reparcourir pour chaque appel à DELETE l'ensemble des couples (X, Y) de nœuds adjacents, nous pouvons nous restreindre à inspecter chaque couple une et une seule fois, comme le montre la proposition suivante.

Proposition 10 *Supposons P est isomorphe à un GOSC et soit $\mathcal{B}$ le RB obtenu à l'issue de la section précédente. Si l'on applique la fonction* DELETE *à un couple (X, Y) de nœuds adjacents du GE de $\mathcal{B}$ et si elle nous indique que (X, Y) ne doit pas être supprimée, alors (X, Y) n'aura pas à être supprimée ultérieurement.*

Esquisse de preuve : Supposons qu'à l'étape k de l'algorithme le score BIC indique qu'un arc ou une arête (X, Y) ne doit pas être supprimé du GE actuel. Supposons qu'à l'étape $k + 1$, on supprime un arc ou une arête (Z, T) et que, après cette suppression, la fonction de score indique qu'il faut supprimer (X, Y). Comme P est isomorphe à un GOSC et que, à chaque étape, les instanciations des GE ont tous le même graphe moral $\mathcal{G}^*$, (Z, T) est une arête de moralisation de $\mathcal{B}^*$. Si l'on ne peut pas supprimer (X, Y) à l'étape k, il y a une chaîne d-séparante entre X et Y dans le GE de l'étape k passant par (Z, T) et bloquée soit par Z, soit par T. La suppression de (Z, T) fait apparaître une V-structure (Z, U, T). Si, à l'étape $k + 1$, il existe une chaîne non d-séparante entre X et Y, celle-ci passe soit par l'arc (Z, U) soit par l'arc (T, U). On montre alors que, dans tous les cas, cela aurait impliqué l'existence d'une chaîne non d-séparante à l'étape k. ♦

Corollaire 11 *Soit $\mathcal{B}$ le RB obtenu à l'issue de la section précédente. Si l'on applique successivement la fonction* DELETE *à tous les arcs et arêtes du GE de $\mathcal{B}$, et ce dans n'importe quel ordre, on obtient $\mathcal{B}^*$.*

Esquisse de preuve : D'après la proposition 10, il suffit de visiter les arcs/arêtes une et une seule fois. De plus, d'après la cohérence locale de la fonction de score, tout arc/arête que l'on peut enlever augmente le score. On peut donc exécuter DELETE pour n'importe quel ordre des couples, on finira par aboutir à $\mathcal{B}^*$. ♦

La complexité de cette phase est donc en $O(mn^3 2^s)$. Finalement, comme par ailleurs les deux premières phases de notre méthode sont d'une complexité polynomiale alors que la première phase de GES, qui leur correspond, est exponentielle, la complexité de notre méthode est globalement meilleure que celle de GES, seul algorithme à notre connaissance à posséder la même garantie d'optimalité.

Remarquons enfin que cette phase de l'algorithme possède une propriété *anytime*. À la fin de la section précédente, nous avons déjà obtenu un « bon » RB dans le sens où celui-ci possède le même graphe moral que le RB optimal $\mathcal{G}^*$. À chaque étape de la phase actuelle, on transforme le RB courant $\mathcal{B}$ en un RB $\mathcal{B}'$, toujours de même graphe moral, et de meilleure qualité au sens où l'on a $\mathcal{B}' < \mathcal{B}$. De plus, il est possible de paramétrer l'algorithme de façon à ne considérer que les sous-ensembles de $\mathbf{NA}_{Y,X}$ de cardinal inférieur à une certaine valeur $t < s$. Il est ainsi possible de réduire cette dernière phase à une complexité polynomiale en contrepartie de l'éventuelle non-suppression de certaines arêtes de moralisation.

7 Conclusion et perspectives

Dans cet article, nous avons montré, lorsque la loi est isomorphe à un GOSC et les données en nombre suffisant, comment obtenir efficacement d'abord un RB de même graphe moral que l'optimum, puis un des équivalents de l'optimum. Nous proposons pour cela une méthode en plusieurs phases utilisant l'espace des RB, l'espace des RM et l'espace des GE. Ces résultats sont toutefois préliminaires et nécessitent d'effectuer des expérimentations, en particulier pour se comparer à GES en pratique. Par ailleurs, dans ce papier, nos démonstrations se fondent sur le fait que la loi ayant engendré la base de données est isomorphe à un GOSC. Il serait intéressant d'abandonner cette hypothèse. Le problème devient alors plus difficile à traiter car il peut alors exister plusieurs RB optimaux pour l'inclusion non équivalents entre eux.

Références

[1] D. Chickering. Learning Bayesian networks is NP-complete. In D. Fisher and H. Lenz, editors, *Learning from Data : Artificial Intelligence and Statistics V*, pages 121–130. Springer-Verlag, 1996.

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

[3] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4) :309–347, 1992.

[4] N. Friedman. The Bayesian structural EM algorithm. In *Proceedings of UAI*, 1998.

[5] Y. Huang and Y. Xiang. Learning Bayesian networks by learning decomposable Markov networks first. In *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering*, 1999.

[6] A. Jameson. Numerical uncertainty management in user and student modeling : An overview of systems and issues. *User Modeling and User-Adapted Interaction*, 5 :193–251, 1996.

[7] F. Jensen. *An Introduction to Bayesian Networks*. Springer-Verlag, North America, 1996.

[8] C. Kruegel, D. Mutz, F. Robertson, and F. Valeur. Bayesian event classification for intrusion detection. In *proceedings of ACSAC*, 2003.

[9] S. L. Lauritzen. *Lectures on Contingency Tables*. University of Aalborg Press, 2nd edition, 1982.

[10] S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[11] P. Leray and O. Francois. Réseaux Bayésiens pour la classification - méthodologie et illustration dans le cadre du diagnostic médical. *Revue d'Intelligence Artificielle*, 18, 2004.

[12] P. Munteanu and M. Bendou. The EQ framework for learning equivalence classes of Bayesian networks. In *Proceedings of the IEEE International Conference on Data Mining*, pages 417–424, 2001.

[13] J. Pearl. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufman, 1988.

[14] J. Pearl. *Causality : models, reasoning and inference*. Cambridge University Press, 2000.

[15] J. Pearl, D. Geiger, and T. Verma. The logic of influence diagrams. In R. Oliver and J. Smith, editors, *Influence Diagrams, Belief Networks and Decision Analysis*. John Wiley and Sons, 1989.

[16] R. Robinson. Counting labeled acyclic digraphs. In F. Harary, editor, *New directions in Graph Theory*, pages 239–273. Academic Press, 1973.

[17] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk e-mail. In *proceedings of AAAI, Workshop on Learning for Text Categorization*, 1998.

[18] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464, 1978.

[19] G. Shafer. Probabilistic expert systems. The Society of Industrial and Applied Mathematics, 1996.

[20] C. Skaanning, F. Jensen, U. Kjærulff, P. Pelletier, and L. Rostrup-Jensen. Printing system diagnosis - a Bayesian network application. In *9th International Workshop on Principles of Diagnosis*, 1998.

[21] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer Verlag, 1993.

[22] T. S. Verma. Causal networks : Semantics and expressiveness. Technical Report R-65, Cognitive Systems Laboratory, U. of California, Los Angeles, 1986.