# Réseaux Bayésiens à Densités Conditionelles

**Santiago Cortijo, Marvin Lasserre, Christophe Gonzales**

*Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6.*
*F-75005 Paris, France*
*prenom.nom@lip6.fr*

RÉSUMÉ. *Nous proposons dans cet article un modèle graphique probabiliste fondé sur les réseaux bayésiens permettant de raisonner avec des variables aléatoires continues et discrètes. Il s'agit d'une variante du modèle ctdBN (Cortijo, Gonzales, 2018). Nous montrons qu'il préserve le bon compromis entre précision et complexité de l'inférence atteint par les ctdBNs, tout en résolvant des problèmes soulevés par l'apprentissage de structure des ctdBNs. Dans cet article, nous introduisons notre nouveau modèle et présentons un algorithme d'apprentissage de sa structure.*

ABSTRACT. *We propose in this paper a Bayesian network-based probabilistic graphical model for reasoning over both continuous and discrete random variables. Our model is a variant of the ctdBN model (Cortijo, Gonzales, 2018). We show that it preserves the good trade-off between faithfulness and inference complexity reached by ctdBNs, while addressing some issues raised by ctdBN structure learning. In the paper, we introduce our new model and present an algorithm for its structure learning.*

MOTS-CLÉS : *Réseaux bayésiens, Variables aléatoires continues, inférence, apprentissage*

KEYWORDS: *Bayesian network, Continuous random variable, inference, learning*

DOI:10.3166/RIA.0.1-9 

## 1. Introduction

Depuis leur introduction dans les années 80, les réseaux bayésiens (RB) sont devenus l'un des modèles les plus populaires pour manipuler les incertitudes (Pearl, 1988). De par leur définition, ils souffrent cependant d'une limitation importante: ils ne peuvent gérer que des variables discrètes. Malheureusement, dans la réalité, il n'est pas rare que certaines variables soient de nature continue. Aussi, d'autres modèles ont été proposés dans la littérature, tels que les modèles gaussiens conditionnels (CG) et leurs extensions aux variables discrètes (Lauritzen, 1992 ; Lerner *et al*., 2001), les mixtures d'exponentielles tronquées (MTE) (Moral *et al*., 2001 ; Cobb *et al*., 2006 ; Rumí, Salmerón, 2007), les mixtures de fonctions de bases tronquées (MTBF) (Langseth *et al*., 2012), les mixtures de polynômes (MOP) (Shenoy, West, 2011) et les réseaux bayésiens à densités conditionnelles tronquées (ctdBN) (Cortijo, Gonzales, 2018).

Modéliser des distributions de probabilité mixtes nécessite un compromis entre pouvoir expressif du modèle et complexité des algorithmes d'apprentissage et d'inférence. Ainsi, les modèles CG ne sont pas très *expressifs* dans la mesure où les dépendances entre variables qu'ils encodent sont uniquement linéaires. En revanche, ils sont très efficaces en termes d'inférence et d'apprentissage. À l'inverse, les MTEs, MTBFs et MOPs peuvent être très précis mais au prix d'une inférence complexe. Entre ces deux extrêmes, les ctdBNs, qui combinent un RB classique avec des densités conditionnelles tronquées, présentent un bon compromis entre expressivité et complexité d'inférence. En effet, ils sont presque aussi expressifs qu'un MTE mais leur inférence est aussi rapide que celle des modèles CG. Leur principal défaut réside dans l'apprentissage conjoint de leur structure et des discrétisations nécessaires aux densités tronquées, qui engendre des effets de bord indésirables. Afin de pallier cela, nous introduisons ici une variante des ctdBNs, composée d'un RB classique et de densités de probabilité conditionnelles non tronquées, nommée "*réseau bayésien à densités conditionnelles*" (cdBN). Outre leurs bonnes propriétés en apprentissage, les cdBNs sont plus expressifs que les ctdBNs, tout en offrant des inférences aussi rapides.

L'article est organisé de la manière suivante : dans la prochaine section, nous rappelons le modèle des ctdBNs et expliquons pourquoi, dans le cadre de l'apprentissage conjoint de structure et de discrétisation, l'extension des scores classiques tels que K2, BDeu, BIC, *etc.*, aux distributions mixtes est nécessairement vouée à l'échec. Pour éviter cet écueil, dans la troisième section, nous introduisons les cdBNs et proposons un algorithme d'apprentissage de structure ainsi qu'un algorithme d'inférence et sa complexité. Le lien entre ctdBNs et cdBNs est également discuté. Enfin, une conclusion et des perspectives sont proposées.

## 2. Modèle ctdBN et verrous d'apprentissage

Dans la suite de l'article, les lettres capitales (possiblement indicées) représentent des variables aléatoires et les symboles en caractères gras des ensembles. Afin de discriminer variables continues et discrètes, nous notons $\mathring{X}_i$ une variable continue et $X_i$ une variable discrète. Sans perte de généralité, pour n'importe quelle variable $\mathring{X}_i$, $X_i$

représente son équivalent discrétisé. Nous notons respectivement $\mathbf{X_D} = \{X_1, \ldots, X_d\}$, $\mathring{\mathbf{X}}_\mathbf{C} = \{\mathring{X}_{d+1}, \ldots, \mathring{X}_n\}$ et $\mathbf{X_C} = \{X_{d+1}, \ldots, X_n\}$ l'ensemble des variables discrètes ne provenant pas de discrétisation, l'ensemble des variables continues et l'ensemble des variables discrétisées. Enfin, quels que soient la variable $X$ ou l'ensemble de variables $\mathbf{Y}$ ou $\mathring{\mathbf{Y}}$, $\Omega_X$ (resp. $\Omega_\mathbf{Y}$ ou $\Omega_{\mathring{\mathbf{Y}}}$) correspond à son domaine.

Dans les ctdBNs, les variables aléatoires continues sont d'abord discrétisées :

DÉFINITION 1 (Discrétisation). — *Une discrétisation d'une variable $\mathring{X}_i$ est une fonction $\Delta_{\mathring{X}_i} : \Omega_{\mathring{X}_i} \rightarrow \{0, \ldots, g_i\}$ définie par une suite croissante de $g_i$ cutpoints $\{t_1, t_2, ..., t_{g_i}\} \subset \Omega_{\mathring{X}_i}$ telle que :*

$$\Delta_{\mathring{X}_i}(\mathring{x}_i) = \begin{cases} 0 & si\ \ \mathring{x}_i < t_1, \\ k & si\ \ t_k \leq \mathring{x}_i < t_{k+1},\ pour\ tout\ k \in \{1, ..., g_i - 1\} \\ g_i & si\ \ \mathring{x}_i \geq t_{g_i} \end{cases}$$

Ainsi, la variable discrétisée $X_i$ correspondant à $\mathring{X}_i$ possède un domaine fini $\{0, \ldots, g_i\}$ et, après discrétisation de toutes les variables continues, l'incertitude sur l'ensemble des variables discrètes et discrétisées peut être représentée par un RB classique. L'idée clef des ctdBNs est d'augmenter la précision du modèle en incorporant dans ce RB des densités conditionnelles tronquées capturant l'information contenue dans les variables aléatoires $\mathring{X}_i$ qui est perdue dans leur version discrétisée $X_i$ :

DÉFINITION 2 (Densité conditionnelle tronquée). — *Soit $\mathring{X}_i$ une variable aléatoire continue. Soit $\Delta_{\mathring{X}_i}$ une discrétisation de $\mathring{X}_i$ avec l'ensemble de cutpoints $\{t_1, t_2, ..., t_{g_i}\}$. Enfin, soit $X_i$ une variable aléatoire discrète avec pour domaine $\Omega_{X_i} = \{0, \ldots, g_i\}$. Une densité conditionnelle tronquée est une fonction $f(\mathring{X}_i|X_i) : \Omega_{\mathring{X}_i} \times \Omega_{X_i} \mapsto \mathbb{R}_0^+$ satisfaisant les propriétés suivantes :*

*1. $f(\mathring{x}_i|x_i) = 0$ pour tout $x_i \in \Omega_{X_i}$ et $\mathring{x}_i \notin [t_{x_i}, t_{x_i+1})$ avec, par abus de notation $t_0 = \inf \Omega_{\mathring{X}_i}$ et $t_{g_i+1} = \sup \Omega_{\mathring{X}_i}$;*

*2. l'équation suivante est vérifiée : $\int_{t_{x_i}}^{t_{x_i+1}} f(\mathring{x}_i|x_i)\, d\mathring{x}_i = 1$, pour tout $x_i \in \Omega_{X_i}$.*

La définition d'un ctdBN en découle directement :

DÉFINITION 3 (ctdBN). — *Un ctdBN est un couple $(\mathcal{G}, \boldsymbol{\theta})$ où :*

*1. $\mathcal{G} = (\mathbf{X}, \boldsymbol{\mathcal{A}})$ est un graphe dirigé acyclique,*

*2. $\mathbf{X} = \mathbf{X_D} \cup \mathbf{X_C} \cup \mathring{\mathbf{X}}_\mathbf{C}$,*

*3. $\boldsymbol{\mathcal{A}}$ est un ensemble d'arcs tel que chaque nœud $\mathring{X}_i \in \mathring{\mathbf{X}}_\mathbf{C}$* n'a aucun enfant et exactement un parent correspondant à $X_i$.

*4. Enfin, $\boldsymbol{\theta} = \boldsymbol{\theta}_\mathbf{D} \cup \boldsymbol{\theta}_\mathbf{C} \cup \mathring{\boldsymbol{\theta}}_\mathbf{C}$, où $\boldsymbol{\theta}_\mathbf{D} = \{P(X_i|\mathbf{Pa}(X_i))\}_{i=1}^d$ et $\boldsymbol{\theta}_\mathbf{C} = \{P(X_i|\mathbf{Pa}(X_i))\}_{i=d+1}^n$ sont l'ensemble des tables de probabilités conditionnelles des variables discrètes et discrétisées $X_i$ sachant leurs parents $\mathbf{Pa}(X_i)$ dans $\mathcal{G}$, et $\mathring{\boldsymbol{\theta}}_\mathbf{C} = \{f(\mathring{X}_i|X_i)\}_{i=d+1}^n$ est l'ensemble des densités conditionnelles tronquées des variables aléatoires continues de $\mathring{\mathbf{X}}_\mathbf{C}$ sachant leur version discrétisée.*

*Le ctdBN encode la distribution de probabilité mixte sur* $\mathbf{X}$ *comme le produit des éléments dans* $\boldsymbol{\theta}$.

Dans (Cortijo, Gonzales, 2018), il est démontré que les ctdBNs sont fidèles dans le sens où n'importe quelle distribution de probabilité mixte peut être approchée à $\epsilon$ près par un ctdBN, et ce quel que soit $\epsilon > 0$. De plus, la Propriété 3 ci-dessus garantit que la complexité de l'inférence probabiliste dans les ctdBNs est identique à celle des RBs définis sur l'ensemble $\mathbf{X_D} \cup \mathbf{X_C}$ des variables discrètes et discrétisées, assurant ainsi des calculs rapides.

Malheureusement, apprendre conjointement par maximum de vraisemblance la structure de modèles basés sur des densités conditionnelles tronquées ainsi que la discrétisation des variables aléatoires continues est, comme nous le montrons ci-dessous, voué à l'échec. Or, c'est ce que font les articles étendant les scores classiques (BDeu, *etc.*) afin de tenir compte des variables continues (Monti, Cooper, 1998 ; Friedman, Goldszmidt, 1996). Rappelons que ceux-ci représentent des vraisemblances de structures graphiques $p(\mathcal{G}|\mathcal{D})$, où $\mathcal{D}$ est la base d'apprentissage (discrète). L'apprentissage conjoint de structure graphique et de discrétisation consiste donc à déterminer $(\mathcal{G}^*, \boldsymbol{\Delta}^*) = \text{Argmax}_{\mathcal{G},\boldsymbol{\Delta}} p(\mathcal{G}, \boldsymbol{\Delta}|\mathring{\mathcal{D}})$, où $\mathring{\mathcal{D}}$ représente la base d'apprentissage non discrétisée et $\boldsymbol{\Delta}$ est l'ensemble des discrétisations des variables continues. En supposant que l'échantillon $\mathring{\mathcal{D}}$ est i.i.d., $p(\mathcal{G}, \boldsymbol{\Delta}|\mathring{\mathcal{D}})$ s'exprime via la formule de Bayes sous la forme $p(\mathring{\mathcal{D}}|\mathcal{G}, \boldsymbol{\Delta}) = \int_{\boldsymbol{\theta}} \prod_{j=1}^{N} p(\mathring{\mathbf{x}}^{(j)}|\mathcal{G}, \boldsymbol{\theta}, \boldsymbol{\Delta})\pi(\boldsymbol{\theta}|\mathcal{G}, \boldsymbol{\Delta})d\boldsymbol{\theta}$, où $N$ est le nombre d'enregistrements dans $\mathring{\mathcal{D}}$, $\mathring{\mathbf{x}}^{(j)}$ est le $j$-ème enregistrement et $\pi(\boldsymbol{\theta}|\mathcal{G}, \boldsymbol{\Delta})$ est un *a priori* sur les paramètres $\boldsymbol{\theta}$ du modèle graphique. En exploitant la factorisation de $p$ selon $\mathcal{G}$, le score d'une structure $\mathcal{G}$ d'un ctdBN est donc :

$$p(\mathring{\mathcal{D}}|\mathcal{G}, \boldsymbol{\Delta}) = \int_{\boldsymbol{\theta}} \prod_{j=1}^{N} \prod_{i=1}^{n} P(x_i^{(j)}|\mathbf{Pa}(x_i^{(j)}), \boldsymbol{\theta}, \boldsymbol{\Delta}) \prod_{i=d+1}^{n} f(\mathring{x}_i^{(j)}|x_i^{(j)}, \boldsymbol{\theta}, \boldsymbol{\Delta})\pi(\boldsymbol{\theta}|\mathcal{G}, \boldsymbol{\Delta})\, d\boldsymbol{\theta}.$$

Or, afin de maximiser l'équation ci-dessus, il suffit de choisir n'importe quelle variable continue, disons $\mathring{X}_{i_0}$, de la discrétiser de manière arbitraire, excepté pour un intervalle de discrétisation $[t_k, t_{k+1})$ rendu arbitrairement petit mais contenant au moins un élément $\mathring{x}_{i_0}^{(j)}$ de la base de donnée $\mathring{\mathcal{D}}$. Par définition, $f$ est une densité tronquée, donc $\int_{t_k}^{t_{k+1}} f(\mathring{x}_{i_0}|x_{i_0}^{(j)}, \boldsymbol{\theta}, \boldsymbol{\Delta})d\mathring{x}_{i_0} = 1$. L'intervalle $[t_k, t_{k+1})$ étant arbitrairement petit, l'intégrale ne peut être égale à un que si la densité $f(\mathring{x}_{i_0}|x_{i_0}^{(j)}, \boldsymbol{\theta}, \boldsymbol{\Delta})$ est arbitrairement grande. Par conséquent, peu importe la structure $\mathcal{G}$, la discrétisation ci-dessus de $\mathring{X}_{i_0}$ implique que $p(\mathring{\mathcal{D}}|\mathcal{G}, \boldsymbol{\Delta})$ tend vers $+\infty$. Il est donc impossible d'apprendre conjointement des structures et des discrétisations raisonnables par maximum de vraisemblance. Ce problème se rencontre lorsqu'on utilise le score proposé dans (Monti, Cooper, 1998). Dans (Friedman, Goldszmidt, 1996), les auteurs contournent ce problème en ajoutant à leur score MDL un terme fondé sur l'entropie appelé $DL_\Lambda(\Lambda)$ qui, lorsqu'on l'examine attentivement, se révèle ne pas être de nature MDL. En effet, l'information qu'il encode peut être significativement compressée en utilisant d'autres schémas d'encodage. Mais ce terme contrebalance l'impact de $f(\mathring{x}_{i_0}^{(j)}|x_{i_0}^{(j)}, \boldsymbol{\theta}, \boldsymbol{\Delta})$ et fait tendre les discrétisations vers des discrétisations uniformes.

## 3. Réseaux bayésiens à Densités Conditionnelles

Dans cette section, nous proposons une variante du modèle ctdBN qui évite l'inconvénient mentionné dans la section précédente. L'idée clé est simple: substituer les densités *tronquées* par des densités non-tronquées. Cela conduit au modèle suivant:

DÉFINITION 4 (Réseaux bayésiens à Densités Conditionnelles (cdBN)). — *La définition d'un cdBN est exactement la même que celle d'un ctdBN sauf que, dans la Propriété 4,* $\mathring{\boldsymbol{\theta}}_{\mathbf{C}} = \{f(\mathring{X}_i|X_i)\}_{i=d+1}^{n}$ *est l'ensemble des densités conditionnelles* non tronquées *des variables aléatoires continues de* $\mathring{\mathbf{X}}_{\mathbf{C}}$ *sachant leur contrepartie discrétisée. Autrement dit, pour n'importe quel* $x_i \in \Omega_{X_i}$, $f(\mathring{X}_i|x_i) : \Omega_{\mathring{X}_i} \mapsto \mathbb{R}_0^+$ *est tel que* $\int_{\mathring{x}_i \in \Omega_{\mathring{X}_i}} f(\mathring{x}_i|x_i)d\mathring{x}_i = 1$.

La Figure 1 montre un exemple de cdBN : les cercles en pointillés représentent les variables continues. Remarquons qu'elles sont seulement liées à leur contrepartie « discrétisée ». Le point clé dans les cdBNs réside dans le fait que, contrairement aux ctdBNs, chaque fonction de densité conditionnelle est définie sur l'ensemble du domaine de définition de $\mathring{X}_i$. Ainsi, il n'y a plus de discrétisation, le modèle s'apparente plus à une mixture de fonctions de densité. C'est cette caractéristique qui garantit que les fonctions de score peuvent bien être définies pour les cdBNs.

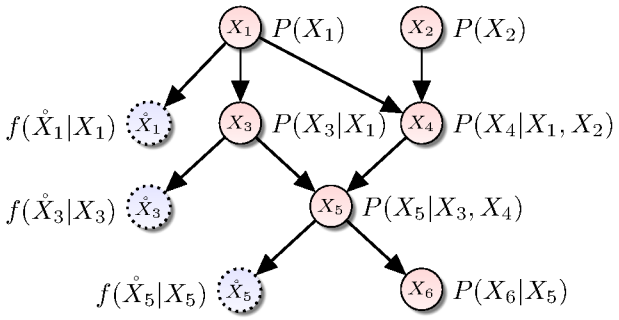


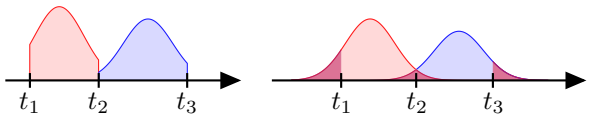

FIGURE *1. Un cdBN.*

PROPOSITION 5. — *Un cdBN est une représentation compacte d'une distribution de probabilité mixte.*

PROPOSITION 6. — *Soit* $\mathbf{\Omega_D} = \prod_{i=1}^{d} \Omega_{X_i}$, $\mathbf{\Omega_C} = \prod_{i=d+1}^{n} \Omega_{X_i}$ *et* $\mathring{\mathbf{\Omega}}_{\mathbf{C}} = \prod_{i=d+1}^{n} \mathring{\Omega}_{\mathring{X}_i}$ *les domaines joints des variables discrètes, discrétisées et continues. Soit* $p : \mathbf{\Omega_D} \times \mathring{\mathbf{\Omega}}_{\mathbf{C}} \mapsto \mathbb{R}$ *une distribution de probabilité mixte, Lipschitz par rapport aux variables continues de* $\mathring{\mathbf{X}}_{\mathbf{C}}$*, c'est-à-dire qu'il existe une constante* $M > 0$ *telle que, pour tout couple* $(\mathring{x}, \mathring{y})$ *d'éléments de* $\mathbf{\Omega_D} \times \mathring{\mathbf{\Omega}}_{\mathbf{C}}$ *tel que* $x_i = y_i$ *pour tout* $i \in \{1, \ldots, d\}$*, on a* $|p(\mathring{x}) - p(\mathring{y})| \leq M||\mathring{x} - \mathring{y}||$*, où* $||\mathring{x} - \mathring{y}||$ *représente la norme L2 du vecteur* $(\mathring{x} - \mathring{y})$*. Alors, pour tout* $\epsilon \in ]0, 1[$*, il existe un cdBN* $\mathcal{B} = (\mathcal{G}, \boldsymbol{\theta})$ *dont l'ensemble des nœuds est* $\mathbf{X} = \mathbf{X_D} \cup \mathbf{X_C} \cup \mathring{\mathbf{X}}_{\mathbf{C}}$ *et qui approche* $p$ *à* $\epsilon$ *près, c'est-à-dire que* $\mathcal{B}$ *représente une distribution de probabilité mixte* $q : \mathbf{\Omega_D} \times \mathbf{\Omega_C} \times \mathring{\mathbf{\Omega}}_{\mathbf{C}} \mapsto \mathbb{R}$

*telle que, pour tout* $(y, \mathring{x}) \in \mathbf{\Omega_D} \times \mathring{\mathbf{\Omega}}_\mathbf{C}$*, on a* $|q(y, x, \mathring{x}) - p(y, \mathring{x})| \leq \epsilon$*, où* $x$ *est la contrepartie discrétisée de* $\mathring{x}$*.*

La figure 2 illustre la différence entre les fonctions de densité utilisées dans les ctdBNs et celles utilisées dans les cdBNs : dans un ctdBN, la fonction de densité affectée à l'intervalle de discrétisation $[t_k, t_{k+1})$ ne prend en compte que la distribution des valeurs de la variable continue définie sur cet intervalle; elle ne tient jamais compte des valeurs qui sont en dehors de $[t_k, t_{k+1})$ même si elles sont très proches de cet intervalle. Dans un cdBN, la fonction de densité prend en compte les valeurs dans l'intervalle $[t_k, t_{k+1})$ mais également celles situées à l'extérieur (les parties violettes de la figure). Ces dernières, étant sur les queues de la distribution, auront des valeurs de fonction de densité inférieures à celles dans l'intervalle $[t_k, t_{k+1})$. Ainsi, le modèle cdBN peut être interprété comme une version robuste des ctdBNs par rapport aux discrétisations. En effet, lorsque les ctdBNs et les cdBNs sont exploités pour la prise de décision, la distribution d'intérêt n'est jamais $p(\mathbf{X})$ mais plutôt $p(\mathbf{X_D}, \mathring{\mathbf{X}}_\mathbf{C}) = \sum_{\mathbf{X_C}} p(\mathbf{X})$, c'est-à-dire que les variables discrétisées non-observées sont marginalisées. Or, après cette sommation, les fonctions de densité deviennent des mixtures de densités. Par exemple, supposons que $p(\mathbf{X}) = P(X_1) f(\mathring{X}_1 | X_1)$ et que $\Omega_{X_1} = \{0, \ldots, g_1\}$. En définissant $\pi_i = P(X_1 = i)$, nous avons $p(\mathring{X}_1) = \sum_{i=0}^{g_1} \pi_i f(\mathring{X}_1 | X_1 = i)$. Comme illustré dans la figure 2.(b), ces mixtures sont généralement assez lisses, de sorte que de petites variations sur les cutpoints $t_k$ dans la discrétisation ont un faible impact sur les distributions $p(\mathbf{X_D}, \mathring{\mathbf{X}}_\mathbf{C})$ pour le cdBN. À l'inverse, dans les ctdBNs, de petites variations sur les points de discrétisation peuvent avoir un impact beaucoup plus important sur $p(\mathbf{X_D}, \mathring{\mathbf{X}}_\mathbf{C})$ car les mixtures de densités présentent des discontinuités (cf. la figure 2.(a)).




(a) ctdBN (b) cdBN

FIGURE 2. *Densités pour les ctdBNs v.s. densités pour les cdBNs.*

### *3.1. L'Inférence dans les cdBNs*

L'inférence dans les cdBNs est exactement la même que dans les ctdBNs: comme les structures dans les deux modèles sont identiques, l'inférence peut être réalisée en utilisant le même algorithme à base d'arbres de jonction (cf. (Cortijo, Gonzales, 2018) pour plus de détails). La complexité de cet algorithme est la suivante:

PROPOSITION 7. — *Soit* $w$ *la largeur d'arbre (treewidth) de l'arbre de jonction utilisé pour l'inférence et soit* $k$ *la taille de domaine maximale des variables aléatoires discrètes et discrétisées. Soit* $n$ *le nombre de variables aléatoires dans le cdBN et soit* $\overline{I}$ *la complexité moyenne du calcul d'une intégrale d'une fonction de densité conditionnelle* $f(\mathring{X}_i | x_i)$*. Soit* $\overline{J}$ *la complexité moyenne du calcul d'une mixture de densités*

*sur une variable $\mathring{X}_i$. La complexité du calcul des distributions marginales a posteriori de toutes les variables aléatoires dans un cdBN est alors en $O(nk(k^w + \overline{I} + \overline{J}))$.*

### *3.2. Apprentissage de cdBN*

L'avantage principal des cdBNs par rapport aux ctdBNs réside dans leur apprentissage structurel à partir des données. Comme nous l'avons déjà vu, les densités tronquées utilisées par les ctdBNs ne permettent pas d'adapter les scores classiques utilisés dans l'apprentissage de la structure des RBs comme K2, BDeu, BIC, *etc.* Heureusement, avec les cdBNs, ces scores peuvent être adaptés ainsi que l'algorithme d'apprentissage. Pour le comprendre, considérons le cdBN de la figure 1 appris à partir d'une base de données $\mathring{\mathcal{D}}$. Dans cette base de données, seules les variables $\mathring{X}_1, X_2, \mathring{X}_3, X_4, \mathring{X}_5, X_6$ sont observées, les variables $X_1, X_3, X_5$ sont cachées. Donc, l'apprentissage de cdBNs consiste à apprendre un modèle avec des variables cachées. Cela suggère simplement d'utiliser l'EM structurel (SEM) proposé dans (Friedman, 1998 ; Peña *et al.*, 2000). Ceci est résumé dans l'algorithme 1.

**Input :** Une base de données $\mathring{\mathcal{D}}$
**Output :** un cdBN $\mathcal{B} = (\mathcal{G}, \mathbf{\Theta})$
1 $t \leftarrow 0$
2 $\mathcal{G}_0 \leftarrow$ une structure initiale de cdBN
3 $\mathbf{\Theta}_0 \leftarrow$ Paramètres MAP de $\mathcal{G}_0$ étant donné $\mathring{\mathcal{D}}$
4 **repeat**
  // **E step**
5  **foreach** $\mathcal{G}$ *dans le voisinage* $\mathbf{N}(\mathcal{G}_t)$ *de* $\mathcal{G}_t$ **do**
6  Affecter score $Sc(\mathcal{G}) = \log p(\mathcal{G}|\mathring{\mathcal{D}})$ à $\mathcal{G}$
  // **M step**
7  $\mathcal{G}_{t+1} \leftarrow \text{Argmax}_{\mathcal{G} \in \mathbf{N}(\mathcal{G}_t)} Sc(\mathcal{G})$
8  $\mathbf{\Theta}_{t+1} \leftarrow$ Paramètres MAP de $\mathcal{G}_{t+1}$ étant donné $\mathring{\mathcal{D}}$
9  $t \leftarrow t + 1$
10 **until** $\mathcal{G}_{t+1} = \mathcal{G}_t$;
11 **return** *cdBN* $\mathcal{B} = (\mathcal{G}_t, \mathbf{\Theta}_t)$

**Algorithme 1 :** Apprentissage de cdBNs avec SEM.

Certains détails doivent être précisés à propos de cet algorithme. Pour cela, supposons qu'une base de données $\mathring{\mathcal{D}}$ contienne $N$ enregistrements, chacun étant défini sur un ensemble de variables aléatoires discrètes $\mathbf{X_D} = \{X_1, \ldots, X_d\}$ et un ensemble de variables aléatoires continues $\mathring{\mathbf{X}}_\mathbf{C} = \{\mathring{X}_{d+1}, \ldots, \mathring{X}_n\}$. Supposons que la base de données est complète, c'est-à-dire que, pour chaque enregistrement, les valeurs de toutes les variables aléatoires sont observées. Enfin, soit $\mathbf{X_C} = \{X_{d+1}, \ldots, X_n\}$ et $\mathbf{X} = \mathbf{X_D} \cup \mathbf{X_C} \cup \mathring{\mathbf{X}}_\mathbf{C}$ l'ensemble des variables latentes et l'ensemble de toutes les variables de notre modèle. Sur la ligne 2 de l'algorithme, un graphe initial est fourni. Habituellement, dans le contexte des RBs, il s'agit d'un graphe sans arc. Ici, l'équivalent

pour les cdBNs est un graphe $\mathcal{G}_0 = (\mathbf{X}, \boldsymbol{\mathcal{A}}_0)$, où $\boldsymbol{\mathcal{A}}_0 = \{(X_i, \mathring{X}_i) : i = d+1, \ldots, n\}$. Sur la ligne 3, les paramètres optimaux des densités conditionnelles ainsi que des probabilités marginales des $X_i$ doivent être déterminés. Tous les $X_i$, $i \leq d$, sont indépendants des autres variables et leur détermination par MAP (maximum *a posteriori*) est donc bien connue, notamment lorsque l'*a priori* sur les paramètres de leur CPT est une distribution de Dirichlet (Heckerman *et al*., 1995). Tous les couples $(X_i, \mathring{X}_i)$, $i = d+1, \ldots, n$ sont mutuellement indépendants, donc les paramètres de leurs distributions peuvent être déterminés indépendamment. Pour les variables $\mathring{X}_i$, nous devons d'abord supposer que les densités conditionnelles appartiennent à une famille de distributions donnée. Supposons que cette famille soit celle des distributions normales $\mathcal{N}(\mu, \tau^{-1})$, où $\tau = 1/\sigma^2$ est la précision. Il est bien connu que leur distribution conjuguée est la distribution Gamma-Normale $N\Gamma(\mu_0, \lambda_0, \alpha_0, \beta_0)$, qui peut être utilisée comme *a priori* sur les paramètres des densités conditionnelles. Maintenant, notons que $\sum_{X_i} P(X_i) f(\mathring{X}_i | X_i)$ est une mixture de distributions normales, dont les paramètres optimaux pour $f$ sont déterminés par MLE (estimation de Maximum de vraisemblance) comme ceux d'une mixture de Gaussiennes et, par MAP, comme une mixture de fonctions Gamma-Normale. Cela peut être effectué efficacement par un algorithme EM classique.

Sur la Ligne 6, $Sc(\mathcal{G}) = \log p(\mathcal{G}|\mathring{\mathcal{D}})$ doit être calculé. Par le theorème de Bayes, nous avons que $Sc(\mathcal{G}) = \log p(\mathring{\mathcal{D}}|\mathcal{G}) + \log(p(\mathcal{G})/p(\mathcal{D}))$. En supposant un *a priori* uniforme sur toutes les structures graphiques, le deuxième terme est constant et n'a pas besoin d'être pris en compte. Malheureusement, il n'existe pas de forme analytique pour calculer $\log p(\mathring{\mathcal{D}}|\mathcal{G})$ car certaines variables sont cachées. Dans l'algorithme SEM, ce terme est approché de la manière suivante: Soit $\mathbf{X}_\mathbf{D}^{(m)}, \mathring{\mathbf{X}}_\mathbf{C}^{(m)}$ (resp. $\mathbf{x}_\mathbf{C}^{(m)}$) les variables observées (resp. non observées) dans le $m$-ème enregistrement de la base de données, et soit $\mathbf{X}_\mathbf{C}^{(\mathcal{D})} = \{\mathbf{X}_\mathbf{C}^{(m)}\}_{m=1}^N$, $\mathbf{X}_\mathbf{D}^{(\mathcal{D})} = \{\mathbf{X}_\mathbf{D}^{(m)}\}_{m=1}^N$ et $\mathring{X}_\mathbf{C}^{(\mathcal{D})} = \{\mathring{X}_\mathbf{C}^{(m)}\}_{m=1}^N$, l'union des variables sur l'ensemble de *tous* les enregistrements. Alors, le score SEM affecté à $\mathcal{G}$ est donné par:

$$\begin{aligned} Sc(\mathcal{G}) &\approx \sum_{\mathbf{x}_\mathbf{C}^{(\mathcal{D})}} p(\mathbf{x}_\mathbf{C}^{(\mathcal{D})} | \mathbf{x}_\mathbf{D}^{(\mathcal{D})}, \mathring{x}_\mathbf{C}^{(\mathcal{D})}, \mathcal{G}_t, \boldsymbol{\Theta}_t) \times \log p(\mathbf{x}_\mathbf{D}^{(\mathcal{D})}, \mathbf{x}_\mathbf{C}^{(\mathcal{D})}, \mathring{x}_\mathbf{C}^{(\mathcal{D})} | \mathcal{G}). \\ &\approx \sum_{\mathbf{x}_\mathbf{C}^{(\mathcal{D})}} P(\mathbf{x}_\mathbf{C}^{(\mathcal{D})} | \mathbf{x}_\mathbf{D}^{(\mathcal{D})}, \mathring{x}_\mathbf{C}^{(\mathcal{D})}, \mathcal{G}_t, \boldsymbol{\Theta}_t) \times \\ &\qquad \left[ \sum_{i=1}^{n} Sc(X_i | \mathbf{Pa}(X_i)) + \sum_{i=d+1}^{n} Sc(\mathring{X}_i | X_i) \right], \end{aligned} \tag{1}$$

où le dernier terme correspond au score de chaque nœud du cdBN conditionnellement à ses parents. La partie qui correspond à $Sc(X_i|\mathbf{Pa}(X_i))$, $i = 1, \ldots, n$, ne concerne que les variables discrètes et peut par conséquent être calculée exactement comme dans (Friedman, 1998). La partie correspondant à $Sc(\mathring{X}_i|X_i)$ est exactement la même pour n'importe quelle structure candidate $\mathcal{G}$ et peut donc être ignorée. L'apprentissage des cdBNs peut alors être effectué en utilisant SEM.

## 4. Conclusion

Un nouveau modèle graphique appelé *cdBN* a été introduit dans cet article pour représenter de manière compacte des distributions de probabilité mixtes. L'avantage de ce modèle par rapport à d'autres modèles similaires réside dans son mécanisme d'inférence rapide mais aussi dans la possibilité d'un apprentissage efficace de sa structure. Une étude plus quantitative des cdBNs, notamment au niveau de l'algorithme d'apprentissage, est prévue pour de prochains travaux.

## Bibliographie

Cobb B., Shenoy P., Rumí R. (2006). Approximating probability density functions in hybrid Bayesian networks with mixtures of truncated exponentials. *Statistics and Computing*, vol. 16, n° 3, p. 293–308.

Cortijo S., Gonzales C. (2018). On conditional truncated densities Bayesian networks. *International Journal of Approximate Reasoning*, vol. 92, p. 155-174.

Friedman N. (1998). The Bayesian structural EM algorithm. In *Proc. of uai*, p. 129–138.

Friedman N., Goldszmidt M. (1996). Discretizing continuous attributes while learning Bayesian networks. In *proc. of icml*, p. 157–165.

Heckerman D., Geiger D., Chickering D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, vol. 20, n° 3, p. 197-243.

Langseth H., Nielsen T., Rumí R., Salmerón A. (2012). Mixtures of truncated basis functions. *International Journal of Approximate Reasoning*, vol. 53, n° 2, p. 212–227.

Lauritzen S. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, vol. 87, p. 1098–1108.

Lerner U., Segal E., Koller D. (2001). Exact inference in networks with discrete children of continuous parents. In *Proc. of uai*, p. 319–328.

Monti S., Cooper G. (1998). A multivariate discretization method for learning Bayesian networks from mixed data. In *Proc. of uai*, p. 404–413.

Moral S., Rumí R., Salmerón A. (2001). Mixtures of truncated exponentials in hybrid Bayesian networks. In *Proc. of ecsqaru*, vol. 2143, p. 156–167.

Pearl J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufman.

Peña J., Lozano J., Larrañaga P. (2000). An improved Bayesian structural EM algorithm for learning Bayesian networks for clustering. *Pattern Recognition Letters*, vol. 21, n° 8, p. 779–786.

Rumí R., Salmerón A. (2007). Approximate probability propagation with mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, vol. 45, n° 2, p. 191–210.

Shenoy P., West J. (2011). Inference in hybrid Bayesian networks using mixtures of polynomials. *International Journal of Approximate Reasoning*, vol. 52, n° 5, p. 641–657.