Apprentissage de modèles continus à grandes dimensions en utilisant l'information mutuelle et les réseaux bayésiens de copules

Marvin Lasserre, Régis Lebrun, Pierre-Henri Wuillemin

Apprentissage de modèles continus à grandes dimensions en utilisant l'information mutuelle et les réseaux bayésiens de copules

Marvin Lasserre,[1] **Régis Lebrun,** [2] **Pierre-Henri Wuillemin** [1]

[1] Laboratoire d'Informatique de Paris 6, 4 place Jussieu, 75005 Paris, France
[2] Airbus Central Research & Technology, 22 rue du Gouverneur Général Eboué, 92130 Issy les Moulineaux, France
marvin.lasserre@lip6.fr, regis.lebrun@airbus.com, pierre-henri.wuillemin@lip6.fr

Résumé

Nous proposons un nouveau cadre pour l'apprentissage de modèles graphiques non-paramétriques à partir de données d'observation continues. Notre méthode se base sur les informations mutuelles conditionnelle et multivariée afin de découvrir des indépendances et la trace de causalité entre des variables aléatoires (comme Verny et al. (2017) pour les modèles discrets). Pour estimer ces quantités à partir des données, nous proposons des estimateurs non-paramétriques reposant sur la copule de Bernstein empirique et qui sont construits en exploitant la relation entre l'information mutuelle et l'entropie de la copule (Ma and Sun 2011; Belalia et al. 2017). À notre connaissance, cette relation n'est documentée que pour le cas bivarié et, pour les besoins de nos algorithmes, elle est ici étendue à l'information mutuelle conditionnelle et l'information mutuelle multivariée. Ce cadre conduit à un nouvel algorithme pour l'apprentissage de réseaux bayésiens continus non-paramétriques. De plus, nous utilisons ces estimateurs pour accélérer l'algorithme BIC proposé dans Elidan (2010) en tirant profit de la décomposition de la fonction de vraisemblance en une somme d'informations mutuelles (Koller and Friedman 2009). Enfin, les performances et la complexité temporelle de notre méthode pour l'apprentissage de la structure d'un réseau bayésien sont comparées avec d'autres techniques de l'état de l'art. Notre nouvel algorithme obtient des résultats supérieurs aux autres et en particulier, il nécessite moins de données pour retrouver la structure de référence et se généralise mieux sur des données qui ne sont pas échantillonnées à partir de distributions gaussiennes.

1 Introduction

La modélisation de distributions continues multivariées est une tâche d'un intérêt central en statistique et en apprentissage automatique avec de nombreuses applications en sciences et en ingénierie. En général, les distributions de grande dimension sont difficiles à manipuler et peuvent conduire à des calculs complexes. Les réseaux bayésiens (BN pour *Bayesian Network*) exploitent les indépendances conditionnelles entre variables aléatoires pour réduire la complexité de la distribution de probabilité jointe en l'exprimant comme un ensemble de distributions de probabilité conditionnelles (CPD pour *Conditional Probability Distribution*) de dimension inférieure. Ces indépendances sont encodées au sein d'un graphe orienté sans circuit (DAG pour *Directed Acyclic Graph*) (Pearl 2014; Koller and Friedman 2009) dont chaque nœud est associé à une CPD. La représentation des CPDs est complexe et a donné lieu à de nombreuses solutions différentes : discrétisation, représentation paramétrique (par exemple à l'aide de l'hypothèse gaussienne), approximation à l'aide de fonctions de base tronquées (Shenoy and West 2011; Langseth et al. 2012), etc.

D'un autre côté, la fonction copule permet de modéliser la structure de dépendance entre variables continues, en s'abstrayant du comportement marginal de chaque variable. D'un point de vue constructif, cela permet de dissocier le choix des marginales de celui de la structure de dépendance. En pratique, cependant, les copules sont limitées à quelques variables et il est difficile de construire ou de manipuler des copules de grande dimension.

Le modèle des réseaux bayésiens de copules (CBN pour *Copula Bayesian Network*) (Elidan 2010) tire parti à la fois de la théorie des copules et des BNs pour modéliser des distributions continues multivariées de grande dimension. Plusieurs autres tentatives de fusionner les deux modèles ont été proposées, telles que les *pair-copulas* (Czado 2010), le modèle *Vine* (Bedford, Cooke et al. 2002) ou encore les *cumulative distribution networks* (Huang 2009), mais le modèle des CBNs reste le plus attrayant puisqu'il utilise le même langage graphique qu'un BN classique.

Cet article se concentre sur l'apprentissage de modèles graphiques continus à partir de données d'observation. Bien qu'il existe de nombreux algorithmes d'apprentissage dans le cas discret (Neapolitan et al. 2004), ils peuvent difficilement être étendus aux modèles continus (Romero, Rumí, and Salmerón 2006). Ceci est principalement dû au trop grand nombre de paramètres de ces modèles, ce qui rend le calcul de scores ou de statistiques de test compliqués. Elles peuvent être appliquées à des modèles plus simples tels que le modèle gaussien linéaire (Lauritzen and Wermuth 1989), mais le modèle lui-même manque d'expressivité. Pour la raison évoquée dans le dernier paragraphe, le modèle des CBNs donne accès à des techniques d'apprentissage similaires à celles utilisées pour les BNs discrets. Des méthodes basées sur une fonction de score (Elidan 2010) et sur les contraintes (Lasserre, Lebrun, and Wuillemin 2020)

ont été proposées pour apprendre la structure d'un CBN à partir d'un ensemble de données. La dernière méthode, appelée CPC, repose sur un algorithme PC et a démontré de meilleures performances que la première, qui s'appuie sur un score BIC et une optimisation par recherche locale. Toutefois, il est bien connu (Colombo and Maathuis 2014) que de telles méthodes basées sur des contraintes souffrent de la nécessité d'un ordre sur les variables et peuvent conduire à des résultats significativement différents en fonction de celui-ci. Dans le cas discret, l'algorithme MIIC (Verny et al. 2017) évite le choix arbitraire d'un ordre en en déterminant un à partir de l'information mutuelle.

Les contributions de cet article sont les suivantes. Premièrement, nous étendons le lien entre l'entropie de la copule et l'information mutuelle démontré dans Ma and Sun (2011), à l'information mutuelle conditionnelle et à l'information mutuelle à trois points. Nous décrivons ensuite les estimateurs non-paramétriques de ces quantités basés sur la copule empirique de Bernstein. Suite à cela, nous utilisons ces estimateurs (i) pour accélérer l'algorithme BIC présenté dans Elidan (2010) en utilisant la décomposition de la fonction de vraisemblance en une somme d'informations mutuelles et (ii) pour proposer un nouvel algorithme d'apprentissage pour les CBNs non-paramétriques. Enfin, un benchmark comparant ces méthodes et l'algorithme CPC est réalisé sur des données synthétiques.

Cet article s'organise de la manière suivante. La section 2 passe en revue les concepts nécessaires sur la théorie des copules et présente le modèle des CBNs. La section 3 développe le lien entre l'information mutuelle et l'entropie de la copule, puis présente les estimateurs utilisés par notre algorithme et la version améliorée de l'algorithme BIC présentés dans la section 4. Ces méthodes sont comparées à l'algorithme CPC dans la section 5 et, pour finir, la section 6 résume et donne plusieurs perspectives à nos travaux.

Cet article traduit et apporte quelques améliorations à Lasserre, Lebrun, and Wuillemin (2021b).

2 Réseaux bayésiens de copules

Considérons un vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_D)$ dont les composantes X_i prennent leurs valeurs x_i dans les domaines Ω_i. Une structure de BN $\mathcal{G}$ est un DAG dont les nœuds $\mathbf{X} = \{X_1, \ldots, X_D\}$ représentent un ensemble de variables aléatoires. Soient $\mathbf{Pa}_i$ et $\mathbf{ND}_i$ respectivement les parents et les non-descendants de X_i dans $\mathcal{G}$. On dit d'une distribution de probabilité multivariée P définie sur un ensemble de variables $\mathbf{X}$, qu'elle se factorise selon $\mathcal{G}$, si elle peut être écrite sous la forme

$$P(X_1, \ldots, X_D) = \prod_{i=1}^{D} P(X_i|\mathbf{Pa}_i). \tag{1}$$

Ainsi, $\mathcal{G}$ encode l'ensemble des indépendances :

$$\mathcal{I}(\mathcal{G}) = \{(X_i \perp \mathbf{ND}_i|\mathbf{Pa}_i)\}. \tag{2}$$

Un BN est une paire $\mathcal{B} = (\mathcal{G}, P)$ où $\mathcal{G}$ est une structure de BN et P est une distribution de probabilité jointe se factorisant sur $\mathcal{G}$. À chaque nœud X_i de la structure du BN est associée sa distribution de probabilité conditionnelle (DPC) $P(X_i|\mathbf{Pa}_i)$ apparaissant dans la factorisation de P. Les CPDs sont généralement représentées par des tables de probabilité conditionnelle dans le cas discret, alors qu'il n'existe pas de modèle général pour le cas continu. Le modèle gaussien linéaire (Lauritzen and Wermuth 1989) $f(x_i|\mathbf{pa}_i) = \mathcal{N}(\beta_{i0} + \beta_i^T \mathbf{pa}_i; \sigma_i^2)$ permet d'effectuer des calculs probabilistes et des estimations rapides mais manque d'expressivité tandis que les modèles basés sur des mélanges de fonctions (Langseth et al. 2012) sont expressifs mais difficiles à apprendre.

Le modèle des CBNs, introduit dans (Elidan 2010), paramètre les CPDs avec des fonctions copules dont nous donnons à présent la définition :

Definition (Fonction copule)**.** *Soit $\mathbf{U} = \{U_1, \ldots, U_D\}$ un vecteur aléatoire dont les composantes U_i sont uniformément distribuées sur $\mathbb{I} = [0, 1]$. Une fonction copule à D dimensions est une fonction de répartition sur $\mathbb{I}^D$:*

$$C(u_1, \ldots, u_D) = \mathbb{P}(U_1 \leq u_1, \ldots, U_D \leq u_D)$$

En tant que fonction de répartition sur $\mathbb{I}^D$ avec des marginales uniformes, la copule respecte les propriétés suivantes :

1. *$C(u_1, \ldots, u_D) = 0$ s'il existe i tel que $u_i = 0$,*
2. *$C(1, \ldots, 1) = 1$,*
3. *$C(1, \ldots, u_i, \ldots, 1) = u_i$.*

La relation qui existe entre la distribution jointe et ses marginales univariées est un résultat central de la théorie des copules (Sklar 1959) :

Theorem (Sklar 1959)**.** *Soit F une fonction de répartition multivariée, Soit $\mathbf{X}$ un vecteur aléatoire, F sa fonction de répartition jointe et F_i ses fonctions de répartitions marginales à 1 dimension. Il existe une fonction copule C telle que*

$$F(x_1, \ldots, x_D) = C(F_1(x_1), \ldots, F_D(x_D)). \tag{3}$$

De plus, si chaque F_i est continu alors C est unique.

Le théorème de Sklar peut être utilisé pour construire de nouvelles copules à partir de distributions multivariées connues. Pour cela, il suffit d'inverser la relation (3) :

$$C(u_1, \ldots, u_D) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_D))$$

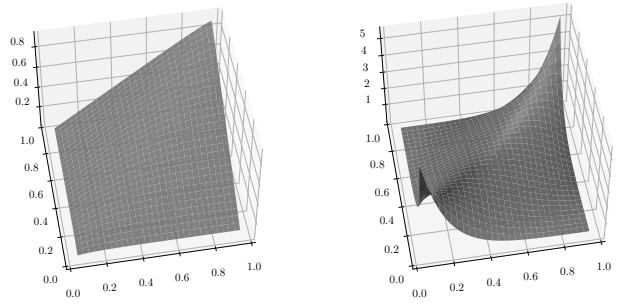
où $u_i = F_i(x_i)$. Par exemple, en prenant la distribution normale multivariée $F = \Phi_\rho$ paramétrée par une matrice de corrélation ρ, on obtient la copule gaussienne (Nelsen 2007) (voir figure 1) :

$$C_G(u_1, \ldots, u_D) = \Phi_\rho(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_D))$$

où ϕ est la distribution normale unidimensionnelle. La fonction densité de la copule est obtenue par dérivation de C : $c(u_1, \ldots, u_D) = \frac{\partial^D C(u_1, \ldots, u_D)}{\partial u_1 \ldots \partial u_D}$. De même, la dérivation de l'équation (3) conduit au corollaire suivant :

Corollary. *Soit $\boldsymbol{x}$ un vecteur aléatoire, F sa fonction de répartition jointe et F_i ses marginales, f sa fonction densité jointe et f_i ses marginales. La densité de la copule c relie la densité jointe à ses marginales unidimensionnelles f_i :*

$$f(x_1, \ldots, x_D) = c(F_1(x_1), \ldots, F_D(x_D)) \prod_{i=1}^{D} f_i(x_i). \tag{4}$$



(a) Fonction de répartition (b) Fonction densité

FIGURE 1 – Copule gaussienne à deux dimensions avec un paramètre de corrélation $\rho_{12} = 0,8$ (les figures ont été obtenues par intégration numérique).

Cette formule généralise le cas où les variables sont indépendantes et où la distribution jointe peut être écrite comme le produit de ses marginales : $f(\boldsymbol{x}) = \prod_{i=1}^{D} f(x_i)$. De la même manière que les marginales encodent le comportement individuel de chaque variable, la fonction copule et sa densité encodent la dépendance entre les variables aléatoires. Ceci est intéressant d'un point de vue constructif puisque le choix des marginales peut être séparé de celui de la structure de dépendance. Cela conduit à la définition d'un CBN donnée par (Elidan 2010) :

Definition (Réseau bayésien de copule)**.** *Un réseau bayésien de copules est un triplet $\mathcal{C} = (\mathcal{G}, \Theta_C, \Theta_f)$ qui encode la densité jointe $f(\mathbf{x})$. Θ_C est un ensemble de fonctions densité de copules locales $c_i\left(F(x_i), \{F(pa_{ik_i})\}\right)$, où $k_i = |\mathbf{pa}_i|$, et Θ_f est un ensemble de densités marginales f_i. À chaque nœud de $\mathcal{G}$, une copule et une fonction marginale sont associées. $f(\mathbf{x})$ se factorise alors comme suit* [1]*,*

$$f(\mathbf{x}) = \prod_{i=1}^{D} R_{c_i}(F(x_i), \{F(pa_{ik_i})\}) f(x_i), \tag{5}$$

où $R_{c_i}(F(x_i), \{F(pa_{ik_i})\}) = \frac{c_i(F(x_i), \{F(pa_{ik_i})\})}{c_i(\{F(pa_{ik_i})\})}$.

Un exemple de CBN est donné sur la figure 2.

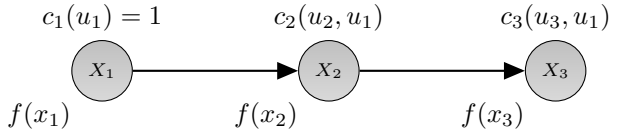


FIGURE 2 – Un CBN à trois variables. Pour tout i, $f(x_i)$ est une densité marginale et $c_i(.)$ est une copule. La densité jointe est :
$f(x_1, x_2, x_3) = f(x_1)f(x_2)f(x_3)c_2(F(x_2), F(x_1))c_3(F(x_3), F(x_2))$

1. Lorsque le contexte l'indique clairement, l'indice i sera supprimé afin d'alléger les notations.

3 Théorie des copules et théorie de l'information

Il a été démontré que l'information mutuelle continue est l'entropie négative de la copule (Ma and Sun 2011). Nous généralisons cette relation pour l'information mutuelle multivariée et l'information mutuelle conditionnelle et l'utilisons pour définir des estimateurs qui seront utilisés dans la section suivante pour mettre en œuvre un algorithme MIIC continu.

Avant d'introduire l'information mutuelle, nous rappelons les définitions de l'entropie différentielle et relative.

Definition (Entropie différentielle)**.** *L'entropie différentielle h sur un ensemble $\boldsymbol{S} \subseteq \boldsymbol{X}$ de variables est donnée par :*

$$h(\boldsymbol{S}) = -\int_{\Omega_{\boldsymbol{S}}} f(\boldsymbol{s}) \log f(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}.$$

Definition (Entropie relative)**.** *L'entropie relative $D(f||g)$ entre deux densités f et g est définie par*

$$D(f||g) = \int_{\Omega_{\mathbf{x}}} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} \mathrm{d}\mathbf{x}. \tag{6}$$

L'information mutuelle est définie comme l'entropie relative entre la densité jointe et ses marginales.

Definition (Information mutuelle)**.** *L'information mutuelle entre deux variables aléatoires X_i et X_j est donnée par :*

$$\begin{aligned} I(X_i; X_j) &= D(f(x_i, x_j)||f(x_i)f(x_j)) \\ &= \iint_{\Omega_i \times \Omega_j} f(x_i, x_j) \log \frac{f(x_i, x_j)}{f(x_i)f(x_j)} \, \mathrm{d}x_i \mathrm{d}x_j. \end{aligned} \tag{7}$$

Par conséquent, puisque $D(f||g) \geq 0$ (Cover and Thomas 2012), l'information mutuelle est également positive. De plus, elle est nulle si et seulement si les variables sont indépendantes, ce qui en fait une bonne mesure de dépendance. L'information mutuelle quantifie la dépendance entre deux variables et est nulle si et seulement si les variables sont indépendantes. Diverses mesures de dépendance, telles que le rho de Spearman ou le tau de Kendall, sont des fonctionnelles de la densité de la copule (Genest and Favre 2007). L'information mutuelle ne fait pas exception puisqu'il s'agit de l'entropie négative de la copule (Ma and Sun 2011) :

Definition (Entropie de la copule)**.** *L'entropie de la copule h_c pour un vecteur aléatoire $\boldsymbol{U}$ est donnée par :*

$$h_c(\boldsymbol{U}) = -\int_{[0,1]^{|\boldsymbol{U}|}} c(\boldsymbol{U}) \log c(\boldsymbol{U}) \mathrm{d}\boldsymbol{u} \tag{8}$$

Theorem. *L'information mutuelle est l'entropie négative de la copule*

$$I(X_i, X_j) = -h_c(X_i, X_j). \tag{9}$$

Nous étendons maintenant cette relation à l'information mutuelle conditionnelle dont la définition est donnée par :

Definition (Information mutuelle conditionnelle)**.** *L'information mutuelle conditionnelle entre X_i et X_j conditionnée par un ensemble de variables $\boldsymbol{U} \subseteq \boldsymbol{X}$. est définie par :*

$$I(X_i; X_j|\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{U}}[D(f(x_i, x_j|\boldsymbol{u})||f(x_i|\boldsymbol{u})f(x_j|\boldsymbol{u})]$$
$$= \iiint_{\Omega_i \times \Omega_j \times \Omega_{\boldsymbol{U}}} f(x_i, x_j, \boldsymbol{u}) \times \log\left(\frac{f(x_i, x_j, \boldsymbol{u})f(\boldsymbol{u})}{f(x_i, \boldsymbol{u})f(x_j, \boldsymbol{u}))}\right) \mathrm{d}x_i \mathrm{d}x_j \mathrm{d}\boldsymbol{u}.$$

De par sa définition, l'information mutuelle conditionnelle est positive en utilisant la positivité de l'entropie relative. Il est facile de montrer à partir de sa définition que

$$I(X_i; X_j|\boldsymbol{U}) = h(X_i, \boldsymbol{U}) + h(X_j, \boldsymbol{U}) - h(X_i, X_j, \boldsymbol{U}) - h(\boldsymbol{U}). \quad (10)$$

En utilisant le lemme suivant démontré dans (Ma and Sun 2011),

Lemma. *L'entropie différentielle peut être écrite comme la somme de l'entropie de chaque variable et de l'entropie de la copule :*

$$h(X_1, \ldots, X_D) = \sum_{i=1}^{D} h(X_i) + h_c(X_1, \ldots, X_D) \quad (11)$$

la relation entre l'information conditionnelle et l'entropie de la copule est facilement obtenue.

Theorem. *L'information mutuelle conditionnelle est liée à l'entropie de la copule par :*

$$I(X_i; X_j|\boldsymbol{U}) = h_c(X_i, \boldsymbol{U}) + h_c(X_j, \boldsymbol{U}) - h_c(X_i, X_j, \boldsymbol{U}) - h_c(\boldsymbol{U}) \quad (12)$$

La définition de l'information mutuelle a été étendue à un ensemble de variables par (McGill 1954) :

$$I(X_1; \ldots; X_D) = \sum_{\boldsymbol{T} \subseteq \boldsymbol{X}} (-1)^{|\boldsymbol{T}|+1} h(\boldsymbol{T}).$$

En particulier, le cas $n = 3$ appelé information à trois points et qui nous intéresse pour la section suivante, est donné par :

$$I(X_i; X_j; X_k) = h(X_i) + h(X_j) + h(X_k) - h(X_i, X_j) - h(X_i, X_k) - h(X_j, X_k) + h(X_i, X_j, X_k). \quad (13)$$

Une remarque importante est que l'information mutuelle multivariée n'est plus nécessairement positive et peut prend ses valeurs dans $\mathbb{R}$. Il s'avère que la négativité de l'information mutuelle entre trois variables est la signature d'une v-structure dans le graphe associé. Comme pour l'information mutuelle conditionnelle, nous voulons relier l'information à trois points à l'entropie de la copule et plus précisément à l'information conditionnelle à trois points.

En regardant de plus près 7, 10 et 13, on peut prouver que l'information à trois points vérifie la relation

$$I(X_i; X_j; X_k) = I(X_i; X_j) - I(X_i; X_j|X_k).$$

Ceci est parfois considéré comme une définition de l'information à trois points et est utilisé ici pour définir l'information conditionnelle à trois points :

Definition (Information conditionnelle à trois points)**.** *L'information conditionnelle à trois points est définie comme suit :*

$$I(X_i; X_j; X_k|\mathbf{U}) = I(X_i; X_j|\mathbf{U}) - I(X_i; X_j|X_k, \mathbf{U})$$

En remplaçant $I(X_i; X_j|\mathbf{U})$ par (12) et en utilisant la même relation avec $\{X_k, \boldsymbol{U}\}$ au lieu de de $\mathbf{U}$, le résultat recherché peut être dérivé pour l'information conditionnelle à trois points :

Theorem. *L'information conditionnelle à trois points est liée à l'entropie de la copule par :*

$$I(X_i; X_j; X_k|\boldsymbol{U}) = h_c(X_i, \boldsymbol{U}) + h_c(X_j, \boldsymbol{U}) + h_c(X_k, \mathbf{U}) - h_c(X_i, X_j, \boldsymbol{U}) - h_c(X_i, X_k, \boldsymbol{U}) - h_c(X_j, X_k, \boldsymbol{U}) + h_c(X_i, X_j, X_k, \boldsymbol{U}) - h_c(\boldsymbol{U}). \quad (14)$$

Dans le cas où $\boldsymbol{U} = \emptyset$, cela se simplifie en

$$I(X_i; X_j; X_k) = h_c(X_i, X_j, X_k) - h_c(X_i, X_j) - h_c(X_j, X_k) - h_c(X_i, X_k) \quad (15)$$

Enfin, toutes les quantités précédentes peuvent être estimées à partir d'un ensemble de données de taille M en utilisant l'estimateur suivant de l'entropie de la copule :

$$\hat{h}_c(\mathbf{X}) = -\sum_{m=1}^{M} \hat{c}(\mathbf{x}[m]) \log(\hat{c}(\mathbf{x}[m])), \quad (16)$$

où $\hat{c}$ est n'importe quel modèle de copule estimé à partir des données d'observation. Afin d'obtenir un estimateur non paramétrique, la copule empirique de Bernstein (Sancetta and Satchell 2004) sera utilisée dans notre algorithme mais une version alternative utilisant une copule gaussienne sera utilisée à titre de comparaison.

4 Apprentissage des réseaux bayésiens de copules

Les CBNs possèdent la même interprétation graphique des indépendances que les BNs classiques (c'est-à-dire la d-séparation), ce qui permet d'utiliser des techniques similaires pour apprendre leurs structures à partir d'un ensemble de données. Ces algorithmes peuvent être grossièrement divisés en deux catégories : les méthodes basées sur une fonction de score et les méthodes basées sur les contraintes. Les méthodes basées sur une fonction de score considèrent la tâche d'apprentissage comme une sélection de modèle et sont guidées par une utilise la fonction de score pour mesurer l'adéquation du modèle avec les données d'observation. Cependant, l'ensemble des DAG étant super-exponentiel relativement au nombre de nœuds, des méthodes de recherche locale sont nécessaires pour maximiser le score. Les méthodes basées sur les contraintes, quant à elles, considèrent le graphe comme un ensemble d'indépendances conditionnelles (2) et utilisent des tests d'indépendance conditionnelle (CIT pour *Conditional Independence Test*) pour obtenir des informations sur la structure sous-jacente. L'algorithme CMIIC présenté dans cette section est une

méthode hybride qui suit un schéma basé sur les contraintes utilisant également une fonction de score basée sur la théorie de l'information. Ce score permet d'éviter l'ordre arbitraire sur les variables qui est souvent nécessaire pour les algorithmes basés sur les contraintes, et qui peut conduire à des résultats très différents.

Amélioration de l'algorithme BIC continu (CBIC)

Dans (Elidan 2010), une méthode utilisant une fonction de score est utilisée pour apprendre la structure d'un CBN. Le score proposé est le critère d'information bayésien (BIC) (Schwarz 1978). Son expression pour une structure de CBN $\mathcal{G}$ est donnée par :

$$\mathcal{S}_{BIC}(\mathcal{G}:\mathcal{D}) = \ell(\mathcal{D}:\hat{\theta},\mathcal{G}) - \frac{1}{2}\log(M)|\Theta_{\mathcal{G}}|,$$

où ℓ est la log-vraisemblance, $\hat{\theta}$ sont les estimateurs du maximum de vraisemblance (MLE) pour les paramètres et $|\Theta_{\mathcal{G}}|$ est le nombre de paramètres libres associé à la structure du graphe. En utilisant la factorisation de la densité jointe (5), nous avons :

$$\ell(\mathcal{D}:\mathcal{G}) = \sum_{m=1}^{M}\sum_{i=1}^{D}\log R_{c_i}\left(u_i[m],\pi_{i1}[m],\ldots,\pi_{ik_i}[m]\right)$$

où $u_i = F(x_i)$ et $\pi_{ij} = F(\text{pa}_{ij})$. Les R_{c_i} sont calculés en utilisant une copule gaussienne paramétrée par une matrice de corrélation Σ. Il peut être difficile de trouver directement les MLEs pour Σ en grande dimension, c'est pourquoi un proxy est utilisé. Celui-ci s'appuie sur la relation $\Sigma_{ij} = \sin(\frac{\pi}{2}\tau_{ij})$ entre le tau de Kendall τ_{ij} et la matrice de corrélation Σ_{ij} qui se vérifie pour toute distribution elliptique (Lindskog, McNeil, and Schmock 2003). Enfin, le score BIC est maximisé à l'aide de l' algorithme TABU list (Glover and Laguna 1998) et de plusieurs itérations de l'algorithme avec un graphe initial aléatoire. L'inconvénient de cette technique est que le score doit être calculé sur l'ensemble du graphe à chaque fois qu'une modification locale est effectuée. Comme amélioration de cet algorithme, nous proposons ici de remplacer le facteur R_i par son expression dans la fonction de vraisemblance, ce qui donne :

$$\ell(\mathcal{D}:\hat{\theta},\mathcal{G}) = M\sum_{i=1}^{D}\hat{I}(X_i;\mathbf{Pa}_i),$$

où $\hat{I}(X_i;\mathbf{Pa}_i) = \hat{h}_c(X_i,\mathbf{Pa}_i) - \hat{h}_c(\mathbf{Pa}_i)$. Cette dernière équation permet de calculer la variation du score pour chaque opération effectuée lors de la recherche locale dans l'espace du graphe, évitant ainsi de le calculer sur le graphe entier (voir p.818 de (Koller and Friedman 2009) pour plus de détails).

Algorithme PC continu (CPC)

L'algorithme PC introduit par (Spirtes et al. 2000) peut être divisé en trois étapes principales : l'apprentissage du squelette, la recherche des v-structures et la propagation des contraintes. La recherche du squelette consiste à supprimer les liens du graphe complet non orienté sur $\mathbf{X}$. Pour ce faire, un CIT est utilisé pour les paires de variables connectées (X_i, X_j) étant donné un sous-ensemble $\boldsymbol{S}$ de leurs voisins communs $\boldsymbol{Adj}(X_i, X_j)$. S'il s'avère que $X_i \perp X_j|\boldsymbol{S}$, $\boldsymbol{S}$ est alors appelé l'ensemble séparateur de X_i et X_j, noté **Sepset**(X_i, X_j), et le lien est retiré du graphe. Les tests sont réalisés en augmentant itérativement la taille $l = |\boldsymbol{S}|$ de l'ensemble conditionnant jusqu'à ce que tous les ensembles d'adjacence dans le graphe soient plus petits que l. Une fois cette première étape terminée, les triplets $X_i - X_k - X_j$ tels que X_i et X_j ne sont pas voisins et que X_k n'est pas dans **Sepset**(X_i, X_j), sont orientés comme des v-structures : $X_i \rightarrow X_k \leftarrow X_j$. Enfin, les liens restants sont orientées sous la contrainte qu'aucune nouvelle v-structure ne soit ajoutée au graphe à moins que cela n'implique l'ajout d'un cycle orienté. L'ordre dans lequel les paires de variables et leurs ensembles d'adjacence sont traités n'est pas unique. Pourtant, il a un effet direct sur la recherche du squelette et les ensembles de séparation. En effet, le squelette est mis à jour après chaque suppression de lien et les ensembles d'adjacence des variables qui sont traités ensuite peuvent changer. Changer l'ordre peut alors conduire à des CIT différents. Pour plus d'informations sur l'algorithme PC, voir la page 84 de Spirtes et al. (2000) et Colombo and Maathuis (2014). Une version continue reposant sur un CIT non-paramétrique, appelée CPC, a été proposée dans Lasserre, Lebrun, and Wuillemin (2020, 2021a) pour apprendre la structure d'un CBN.

Un nouvel algorithme d'apprentissage pour les CBN : MIIC continu (CMIIC)

L'algorithme MIIC comprend les trois mêmes étapes principales que l'algorithme CPC : apprentissage du squelette, orientation des v-structures et propagation des contraintes. Cependant, il utilise l'information mutuelle afin d'ordonner les nœuds et de surmonter le problème de l'ordre discuté dans le cas de l'algorithme PC. Dans le cas discret, il a été démontré qu'il était plus efficace que la méthode PC (Verny et al. 2017) et nous l'étendons ici aux données continues.

Le point de départ de l'algorithme est la fonction de vraisemblance qui doit être légèrement adaptée au cas continu :

$$\mathcal{L}(\mathcal{D}|\mathcal{G}) = \prod_{m=1}^{M} f_{\mathcal{G}}(\boldsymbol{x}[m]) = \exp\left(-MH(\hat{P}_M, P_{\mathcal{G}})\right)$$

où $H(\hat{P}_M, P_{\mathcal{G}}) = -\int_{\Omega_{\boldsymbol{X}}} \log f_{\mathcal{G}}(\boldsymbol{x})\mathrm{d}\hat{F}_M$ est l'entropie croisée entre la distribution empirique dont la fonction de répartition $\hat{F}_M$ s'écrit :

$$\hat{F}_M(\boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}_{\boldsymbol{x}[m]\leq\boldsymbol{x}}$$

et la distribution du modèle se factorisant sur $\mathcal{G}$. L'ordre est ensuite dérivé de la décomposition de l'entropie croisée sur la structure $\mathcal{G}$ et par le calcul du rapport de la fonction de vraisemblance. Seule son expression est rapportée ici et le lecteur intéressé pourra trouver les détails sur son origine dans (Affeldt, Verny, and Isambert 2016). L'ordre est basé sur la probabilité que le triplet (X_i, X_j, X_k) ne soit pas une

Algorithm 1: Algorithme MIIC (Verny et al. 2017)

Input: Échantillon de données $\mathcal{D}$
Result: DAG $\mathcal{G}$

$\mathcal{G} \leftarrow$ graphe non-dirigé complet sur $\mathbf{X}$;
// Recherche du squelette
forall *Lien* (X_i, X_j) **do**
  if $I'(X_i; X_j) < 0$ **then**
    Supprimer le lien $X_i - X_j$ de G;
    $\mathbf{Sepset}(\mathbf{X_i}, \mathbf{X_j}) \leftarrow \{\}$;
  else
    $X_k \leftarrow \arg\max_{\boldsymbol{Adj}(X_i, X_j)} r(X_i, X_j; X_k|\{\})$;
while *Il existe un lien* (X_i, X_j) *avec* $r(X_i, X_j; X_k|\mathbf{U})$ **do**
  for (X_i, X_j) *avec le plus haut rang* $r(X_i, X_j; X_k|\mathbf{U})$ **do**
    Augmenter l'ensemble contributif : $\mathbf{U} \leftarrow \mathbf{U} \cup \{X_k\}$;
    if $I'(X_i, X_j|\mathbf{U}) \leq 0$ **then**
      Supprimer le lien $X - Y$ de G;
      $\mathbf{Sepset}(\mathbf{X_i}, \mathbf{X_j}) \leftarrow \mathbf{U}$;
    else
      $X_k \leftarrow \arg\max_{\boldsymbol{Adj}(X_i, X_j)} r(X_i, X_j; X_k|\mathbf{U})$;
    Trier la liste des rangs $r(X_i, X_j; X_k|\mathbf{U})$;
// Recherche des v-structures
Trier la liste L des triplets $X_i - X_k - X_j$ par ordre décroissant de $|I'(X_i; X_j; X_k|\mathbf{U})|$;
repeat
  Prendre $(X_i, X_k, X_j) \in L$ avec la plus grande valeur de $|I'(X_i; X_j; X_k|\mathbf{U})|$ sur lequel la règle R_0 ou R_1 peut être appliquée;
  if $I'(X_i; X_j; X_k|\mathbf{U}) < 0$ **then**
    Si (X_i, X_k, X_j) n'a pas d'orientation divergente, Appliquer $R_0 : \{X_i - X_k - X_j \,\&\, \mathrm{not}(X_i - X_j) \& X_k \notin \mathbf{Sepset}(\mathbf{X_i}, \mathbf{X_j})\} \Rightarrow \{X_i \rightarrow X_k \leftarrow X_j\}$
  else
    Si (X_i, X_k, X_j) a une orientation convergente, appliquer R_1 : $\{X_i \rightarrow X_k - X_j \,\&\, \mathrm{not}(X_i - X_j)\} \Rightarrow \{X_k \rightarrow X_j\}$
  Appliquer une nouvelle orientations à tous les autres $(X'_i, X'_k, X'_j) \in L$;
until *Aucune orientation additionnelle ne peut être obtenue*;

v-structure conditionnée par $\mathbf{U}$:

$$P_{\mathrm{nv}}(X_i; X_j; X_k|\mathbf{U}) = \left(1 + e^{-MI(X_i;X_j;X_k|\mathbf{U})}\right)^{-1}$$

et la probabilité que sa base soit X_i et X_j :

$$P_{\mathrm{b}}(X_i, X_j; X_k|\mathbf{U}) = \frac{1}{1 + \frac{e^{-MI(X_i;X_k|\mathbf{U})}}{e^{-MI(X_i;X_j|\mathbf{U})}} + \frac{e^{-MI(X_j;X_k|\mathbf{U})}}{e^{-MI(X_i;X_j|\mathbf{U})}}}$$

En combinant ces deux probabilités, les paires de nœuds (X_i, X_j) avec la contribution la plus probable d'un troisième nœud X_k peuvent être ordonnées en fonction de :

$$r(X_i, X_j; X_k|\mathbf{U}) = \max_{X_k \in \boldsymbol{X}} \left(\min\left[P_{\mathrm{nv}}(X_i; X_j; X_k|\mathbf{U}), P_{\mathrm{b}}(X_i, X_j; X_k|\mathbf{U})\right]\right).$$

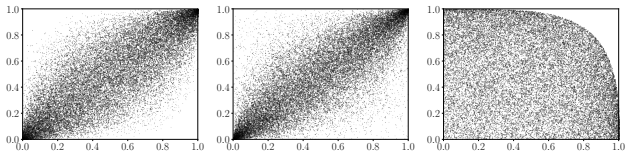
L'implémentation de MIIC est détaillée dans l'algorithme 1. Les termes d'information mutuelle conditionnelle à deux et trois points apparaissant dans les probabilités précédentes sont calculés en utilisant les équations (12, 14) et l'estimateur de l'entropie de la copule (équation (16)). Les estimateurs étant calculés sur des ensembles de données de tailles finies, ils sont donc biaisés. Pour cette raison, dans le cas discret, des corrections basées sur des critères tels que le maximum de vraisemblance normalisé (NML pour *Normalized Maximum Likelihood*) (Shtar'kov 1987), la longueur minimale de description (MDL pour *Minimal Description Length*) (Rissanen 1978) ou le BIC (Koller and Friedman 2009) sont utilisés pour effacer ce biais. Cependant, les deux premiers critères ne peuvent pas être étendus aux variables continues car ils divergent dans la limite continue. Quant au score BIC, il ne peut être appliqué à notre cas puisqu'il n'est défini que pour des modèles paramétriques. Par conséquent, nous avons décidé d'utiliser un paramètre α tel que $I'(X_i; X_j|\mathbf{U}) = I(X_i; X_j|\mathbf{U}) - \alpha$ et $I'(X_i; X_j; X_k|\mathbf{U}) = I(X_i; X_j; X_k|\mathbf{U}) + \alpha$. Le fait que nous ajoutons α dans le cas d'une information à trois points signifie que nous favorisons une non v-structure par rapport à une v-structure puisqu'une valeur négative de l'information à trois points conduit à une v-structure dans le graphe. Ce paramètre peut être considéré comme un seuil de confiance : plus α diminue (et plus notre test est précis), plus il faut de données pour décider de l'indépendance. La valeur $\alpha = 0.01$ s'est avérée expérimentalement être un bon compromis entre la taille d'échantillon nécessaire nécessaire pour apprendre les indépendances et la confiance du test.

5 Résultats expérimentaux

Cette section compare les résultats obtenus par les méthodes CBIC, CPC et CMIIC. Deux modèles de copules, Gaussien et Bernstein, sont utilisés pour estimer l'entropie de la copule avec CMIIC. Cela conduit à deux versions de l'algorithme qui seront dénommées G-CMIIC et B-CMIIC. Elles sont comparées en utilisant des métriques structurelles, le F-score et la distance de Hamming structurelle et en mesurant leurs temps d'exécution. Ces expériences ont été menées en utilisant les bibliothèques aGrUM (Ducamp, Gonzales, and Wuillemin 2020) et OpenTURNS (Baudin et al. 2016) pour respectivement construire les modèles graphiques et modéliser les distributions continues multivariées.

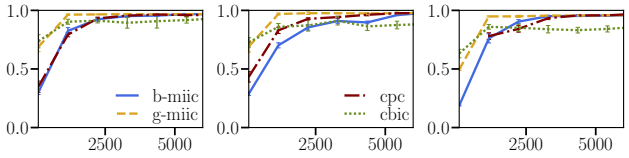
Protocole expérimental

Les algorithmes ont été testés sur des données générées soit à partir de la structure du réseau ALARM (Beinlich et al. 1989) afin d'avoir une structure provenant d'un cas applicatif, soit à partir de réseaux bayésiens aléatoires pour plus de généralité. Les structure aléatoires ont été générés en suivant la méthode de Ide and Cozman (2002) proposant de construire une *Monte-Carlo Markov Chain* (MCMC) qui converge vers une distribution uniforme sur l'ensemble des DAGs ayant le nombre désiré de nœuds et d'arcs. Pour une dimension D donnée, une structure aléatoire contient $1,2 \times D$ arcs, ce qui correspond au ratio de la structure ALARM. Une fois la structure sélectionnée (ALARM ou aléatoire), les copules locales du CBN sont paramétrées à l'aide de trois modèles paramétriques : gaussien, Student et Dirichlet. Ces modèles ont été choisis afin de construire les scénarios les plus défavorables pour notre algorithme et de comparer ses performances avec les méthodes d'apprentissage paramétriques lorsqu'on traite des données gaussiennes



(a) Gaussien (b) Student (c) Dirichlet

FIGURE 3 – Échantillons de densités de copules gaussiennes, de Student et de Dirichlet. Le paramètre de corrélation de la copule gaussienne et de Student est fixé à $\rho = 0.8$, le nombre de degré de liberté de la copule de Student est fixé à $\nu = 5$, les paramètres de la copule de Dirichlet ont été choisis tels que $\boldsymbol{\alpha} = (1/3, 2/3, 1)$.



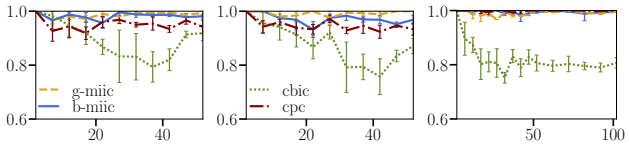
(a) Gaussien (b) Student (c) Dirichlet

FIGURE 4 – Évolution du F-score pour les méthodes CBIC, CPC, G-CMIIC et B-CMIIC en fonction de la taille de l'ensemble de données. La moyenne des résultats est calculée sur 5 réinitialisations avec différents ensembles de données générés à partir de la structure du réseau ALARM.

ou de Student. La copule de Dirichlet, quant à elle, a été choisie pour mettre notre algorithme l'épreuve en raison de son support restreint (voir figure (3)). Les trois modèles ont été paramétrés de manière à ce qu'ils induisent de fortes corrélations entre les variables (matrices de corrélation dont les paramètres hors diagonale sont fixés à $0,8$, copule de Dirichlet avec $\boldsymbol{\alpha} = (1/D, 2/D, \ldots, 1)$). La figure (3) montre des échantillons bidimensionnels obtenus à l'aide de ces paramétrages. Les CBNs sont ensuite échantillonnés à l'aide de la procédure *push-forward* décrite dans Koller and Friedman (2009).

Performances pour la reconstruction du squelette

Les performances des quatre algorithmes d'apprentissage ont été mesurées en comparant le squelette du graphe appris avec le squelette de la structure de référence utilisée pour générer les données. La précision (P) est la proportion de liens apprise qui sont effectivement dans la structure de référence, tandis que le rappel (R) est la proportion de liens qui sont dans la structure de référence qui ont bien été retrouvés. Le F-score est alors défini comme : $\mathrm{F} = 2\mathrm{PR}/(\mathrm{P} + \mathrm{R})$. Si le squelette de référence a été parfaitement retrouvé, la valeur du F-score est de 1. Les figures 4 et 5 donnent l'évolution du F-score en fonction de la taille de l'échantillon pour la structure ALARM et en fonction du nombre de nœuds pour les structures générées par MCMC. Comme on peut l'observer, G-CMIIC converge plus rapidement que les autres algorithmes, mais B-CMIIC et CPC convergent approximativement vers la même valeur. De manière surprenante, G-CMIIC conserve de bons résultats



(a) Gaussien (b) Student (c) Dirichlet

FIGURE 5 – Évolution du F-score pour les méthodes CBIC, CPC, G-CMIIC et B-CMIIC en fonction de la dimension des graphes aléatoires. Les résultats sont moyennés sur 2 différents graphes aléatoires de même dimension et sur 5 ensembles de données différents de taille $M = 10000$.

même pour les données Dirichlet. La méthode CBIC est cependant moins performante que les trois autres, quel que soit le modèle génératif.

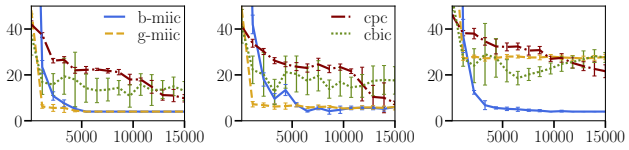
Performances pour la reconstruction du CPDAG

Pour évaluer la structure orientée, nous utilisons la distance de Hamming structurelle (SHD) (Colombo and Maathuis 2014). Cette métrique se calcule sur les graphes acycliques partiellement dirigés complétés (CPDAG pour *Completed Partially Directed Acyclic Graph*) qui représentent la classe d'équivalence de Markov d'un DAG. Pour un DAG $\mathcal{G}$ donné, cette dernière est composée de tous les graphes qui représentent le même ensemble d'indépendances conditionnelles que $\mathcal{G}$ (Koller and Friedman 2009). La SHD correspond alors au nombre d'opérations élémentaires nécessaires pour retrouver le CPDAG de la structure originale à partir du CPDAG de la structure estimée. Ces opérations sont l'ajout, la suppression et le retournement d'un lien ou d'un arc. Les figures 6 et 7 montrent l'évolution de la SHD en fonction de la taille de l'échantillon dans le cas de la structure ALARM et du nombre de nœud dans le cas des structures aléatoires.

L'interprétation de ces résultats est similaire à celles que nous avons faite pour la reconstruction de la squelette. En effet, la méthode G-CMIIC retrouve presque parfaitement le GAPDC dans le cas où le modèle ayant généré les données est gaussien ou de Student. De plus, elle a besoin de moins de données pour converger que les autres techniques. Cependant, ses performances sont assez pauvres dans le cas de données de Dirichlet. En revanche, B-CMIIC a de bonnes performances quel que soit le modèle génératif, ce qui illustre l'avantage d'une méthode non- paramétrique sur des méthodes paramétriques. Dans le cas de structures avec peu de nœuds, CPC semble converger vers la même valeur que les méthodes CMIIC mais nécessite pour cela beaucoup plus de données. En outre, ses performances diminuent lorsque la taille des structures augmente. Quant à la méthode CBIC, ses résultats sont moindres par rapport aux autres techniques et de plus, ses performances baissent pour des dimensions élevées.

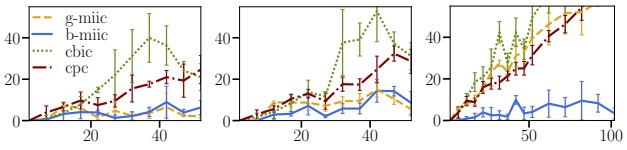
Complexité temporelle

Les temps d'exécution pour les différentes méthodes d'apprentissage ont été calculés en fonction de la dimension des structures aléatoires et pour des échantillon de taille $M = 10000$. Les résultats sont présentés sur la figure 8. Le



(a) Gaussien (b) Student (c) Dirichlet

FIGURE 6 – Évolution de la SHD pour les méthodes CBIC, CPC, G-CMIIC et B-CMIIC en fonction de la taille de l'ensemble de données. La moyenne des résultats est calculée sur 5 réinitialisations avec différents ensembles de données générés à partir de la structure du réseau ALARM.



(a) Gaussien (b) Student (c) Dirichlet

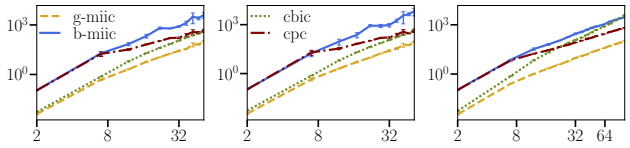
FIGURE 7 – Évolution de la SHD pour les méthodes CBIC, CPC, G-CMIIC et B-CMIIC en fonction de la dimension des graphes aléatoires. Les résultats sont moyennés sur 2 différents graphes aléatoires de même dimension et sur 5 ensembles de données différents de taille $M = 10000$.

temps d'apprentissage de l'algorithme B-CMIIC est le plus important malgré ses bonnes performances pour la reconstruction de la structure. G-CMIIC, quant à lui, est le plus rapide et, à ce titre, devrait être utilisé lorsque l'on sait que l'hypothèse gaussienne est valable. En revanche, lorsqu'aucune information n'est disponible, l'algorithme B-CMIIC devrait être utilisé en raison de la généralité de ses résultats.

6 Conclusion et perspectives

Le modèle des CBNs utilise des fonctions de copules pour paramétrer les DPCs d'un BN continu. La représentation du BN permet, d'autre part, de limiter la taille des fonctions copules qui peuvent être difficiles à manipuler pour des dimensions élevées. De plus, la structure d'un CBN possède la même langage graphique qu'un BN classique. Avec quelques adaptations, cela permet d'utiliser les mêmes techniques d'apprentissage que celles utilisées pour des données discrètes. À cet égard, nous avons proposé un algorithme MIIC continu se situant à mi-chemin entre les méthodes basées sur une fonction de score et celles basées sur des contraintes. Pour cela, nous avons du étendre le lien entre l'information mutuelle et l'entropie de la copule à l'information conditionnelle et l'information à trois points. Cette extension nous a permis la construction d'estimateurs non-paramétriques pour ces quantités à partir de la copule empirique de Bernstein. Le volet expérimental a illustré la généralité de la méthode non-paramétrique G-CMIIC par rapport aux méthodes paramétriques CBIC et G- CMIIC De plus, l'absence d'un ordre arbitraire a permet à G-CMIIC d'obtenir de meilleurs résultats avec moins de données que la méthode CPC. Cependant, comme c'est souvent le cas, les méthodes non-paramétriques sont plus coûteuses en temps d'exécution et si nous avons des indices sur le modèle qui a généré les données, les méthodes paramétriques peuvent être préférées comme l'illustrent les performances structurelles et le temps d'exécution de la version gaussienne de CMIIC. Malgré tout, le couplage des CBNs avec des copules de Bernstein empirique nous a permis d'apprendre des modèles non-paramétriques de grandes dimensions ce qui n'était pas possible auparavant. Tous les fichiers sources pour manipuler et apprendre les CBNs avec la méthode CMIIC peuvent être trouvés au sein d'un plugin expérimental, appelé otagrum, faisant partie de la bibliothèque OpenTURNS et utilisant la bibliothèque aGrUM.

Bien que ces résultats soient très encourageants, le fondement théorique du paramètre correctif α n'est pas satisfaisant. En remplacement, une pénalité de score continue pourrait être utilisée mais les pénalités discrètes sont soit divergentes dans la limite continue (NML, MDL) soit seulement extensibles pour des modèles paramétriques (BIC). Une idée plus prometteuse serait d'étendre l'estimateur de l'information mutuelle introduit dans (Belalia et al. 2017) à l'information conditionnelle et à l'information à trois points. Cet estimateur étant distribué selon une loi normale dans la la limite des grands échantillons, des p-values pourraient être utilisées pour quantifier la confiance dans les indépendances. Enfin, ces résultats ont été obtenus par l'utilisation de données générées et doivent être complétés par des données réelles.



(a) Gaussien (b) Student (c) Dirichlet

FIGURE 8 – Temps d'apprentissage en secondes pour les méthodes CBIC, CPC, G-CMIIC et B-CMIIC en fonction de la dimension des graphes aléatoires. Les résultats sont moyennés sur 2 différents graphes aléatoires de même dimension et sur 5 ensembles de données différents de taille $M = 10000$.

Disponibilité du code

Les codes sources de nos algorithmes et de nos tests sont respectivement accessibles sur les répertoires GitHub openturns/otagrum et MLasserre/otagrum-experiments.

Références

Affeldt, S. ; Verny, L. ; and Isambert, H. 2016. 3off2 : A network reconstruction algorithm based on 2-point and 3-point information statistics. In *BMC bioinformatics*, volume 17, S12. Springer.

Baudin, M. ; Dutfoy, A. ; Iooss, B. ; and Popelin, A.-L. 2016. *OpenTURNS : An Industrial Software for Uncertainty*

Quantification in Simulation, 1–38. Cham : Springer International Publishing.

Bedford, T.; Cooke, R. M.; et al. 2002. Vines–a new graphical model for dependent random variables. *The Annals of Statistics* 30(4) : 1031–1068.

Beinlich, I. A.; Suermondt, H. J.; Chavez, R. M.; and Cooper, G. F. 1989. The ALARM monitoring system : A case study with two probabilistic inference techniques for belief networks. In *AIME 89*, 247–256. Springer.

Belalia, M.; Bouezmarni, T.; Lemyre, F.; and Taamouti, A. 2017. Testing independence based on Bernstein empirical copula and copula density. *Journal of Nonparametric Statistics* 29(2) : 346–380.

Colombo, D.; and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research* 15(1) : 3741–3782.

Cover, T. M.; and Thomas, J. A. 2012. *Elements of information theory*. John Wiley & Sons.

Czado, C. 2010. Pair-copula constructions of multivariate copulas. In *Copula theory and its applications*, 93–109. Springer.

Ducamp, G.; Gonzales, C.; and Wuillemin, P.-H. 2020. aGrUM/pyAgrum : a toolbox to build models and algorithms for Probabilistic Graphical Models in Python. In *10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, 609–612. Skørping, Denmark.

Elidan, G. 2010. Copula bayesian networks. In *Advances in neural information processing systems*, 559–567.

Genest, C.; and Favre, A.-C. 2007. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering* 12(4) : 347–368.

Glover, F.; and Laguna, M. 1998. Tabu search. In *Handbook of combinatorial optimization*, 2093–2229. Springer.

Huang, J. C. 2009. *Cumulative distribution networks : Inference, estimation and applications of graphical models for cumulative distribution functions*. Citeseer.

Ide, J. S.; and Cozman, F. G. 2002. Random generation of Bayesian networks. In *Brazilian symposium on artificial intelligence*, 366–376. Springer.

Koller, D.; and Friedman, N. 2009. *Probabilistic graphical models : principles and techniques*. MIT press.

Langseth, H.; Nielsen, T. D.; Rumı, R.; and Salmerón, A. 2012. Mixtures of truncated basis functions. *International Journal of Approximate Reasoning* 53(2) : 212–227.

Lasserre, M.; Lebrun, R.; and Wuillemin, P.-H. 2020. Constraint-Based Learning for Non-Parametric Continuous Bayesian Networks. In *FLAIRS 33 - 33rd Florida Artificial Intelligence Research Society Conference*, 581–586. Miami, United States : AAAI. URL https://hal.archives-ouvertes.fr/hal-02615379.

Lasserre, M.; Lebrun, R.; and Wuillemin, P.-H. 2021a. Constraint-based learning for non-parametric continuous bayesian networks. *Annals of Mathematics and Artificial Intelligence* 1–18.

Lasserre, M.; Lebrun, R.; and Wuillemin, P.-H. 2021b. Learning Continuous High-Dimensional Models using Mutual Information and Copula Bayesian Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 12139–12146.

Lauritzen, S. L.; and Wermuth, N. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The annals of Statistics* 31–57.

Lindskog, F.; McNeil, A.; and Schmock, U. 2003. Kendall's tau for elliptical distributions. In *Credit Risk*, 149–156. Springer.

Ma, J.; and Sun, Z. 2011. Mutual information is copula entropy. *Tsinghua Science & Technology* 16(1) : 51–54.

McGill, W. 1954. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory* 4(4) : 93–111.

Neapolitan, R. E.; et al. 2004. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ.

Nelsen, R. B. 2007. *An introduction to copulas*. Springer Science & Business Media.

Pearl, J. 2014. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Elsevier.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14(5) : 465–471.

Romero, V.; Rumí, R.; and Salmerón, A. 2006. Learning hybrid Bayesian networks using mixtures of truncated exponentials. *International Journal of Approximate Reasoning* 42(1-2) : 54–68.

Sancetta, A.; and Satchell, S. 2004. The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory* 20(03) : 535–562.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics* 6(2) : 461–464.

Shenoy, P. P.; and West, J. C. 2011. Inference in hybrid Bayesian networks using mixtures of polynomials. *International Journal of Approximate Reasoning* 52(5) : 641–657.

Shtar'kov, Y. M. 1987. Universal sequential coding of single messages. *Problemy Peredachi Informatsii* 23(3) : 3–17.

Sklar, A. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* 8 : 229–231.

Spirtes, P.; Glymour, C. N.; Scheines, R.; Heckerman, D.; Meek, C.; Cooper, G.; and Richardson, T. 2000. *Causation, prediction, and search*. MIT press.

Verny, L.; Sella, N.; Affeldt, S.; Singh, P. P.; and Isambert, H. 2017. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS computational biology* 13(10) : e1005662.