THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Spécialité Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Marvin LASSERRE

Pour obtenir le grade de

DOCTEUR de SORBONNE UNIVERSITÉ

Apprentissages dans les Réseaux Bayésiens à Base de Copules Non-Paramétriques

---

Soutenue le 11 mars 2022 devant le jury composé de :

Rapporteurs :	Sébastien Destercke	Chargé de recherche, CNRS, UTC
	Simon de Givry	Chargé de recherche, CNRS, INRAE
Examinateurs :	Grégory Nuel	Directeur de recherche, CNRS, Sorbonne Université
	Patrice Perny	Professeur, Sorbonne Université
	Clémentine Prieur	Professeur, Université Grenoble Alpes
Directeur :	Christophe Gonzales	Professeur, Université Aix-Marseille
Encadrants :	Pierre-Henri Wuillemin	Maitre de Conférence, Sorbonne Université
	Régis Lebrun	Senior scientist, Airbus Central Research & Technology

À Audrey

Remerciements

Je tiens tout d'abord à remercier chaleureusement Sébastien Destercke et Simon de Givry pour l'intérêt qu'ils ont porté à mes travaux en acceptant d'être les rapporteurs de ma thèse. Je remercie également Gregory Nuel, Patrice Perny et Clémentine Prieur d'avoir accepté d'évaluer ma thèse en tant qu'examinateurs. Je remercie aussi Nicolas Maudet d'avoir fait partie de mon comité de suivi aux côtés de Gregory Nuel.

En ce qui concerne mon encadrement, je voudrais commencer par remercier Christophe Gonzales d'avoir accepté d'être mon directeur de thèse. Je remercie ensuite Régis Lebrun pour ses explications, ses relectures, ses conseils et sa bienveillance tout au long de la thèse. En dehors de nos réunions, j'ai grandement apprécié nos discussions autour d'un repas au Buisson Ardent où les sujets abordés pouvaient passer rapidement des mathématiques à la construction de structures en Lego (avec parfois vidéos à l'appui !) pour revenir tout aussi rapidement à la physique ou l'informatique. Enfin, je ne saurais comment exprimer toute la gratitude que j'ai envers Pierre-Henri Wuillemin sans qui cette thèse n'aurait jamais pu aboutir. Tout comme pour Régis, je le remercie de son soutien et de son investissement puisqu'il a toujours été là pour répondre à mes questions et ce, même à des heures très tardives. Je le remercie également de la patience légendaire dont il a fait preuve pendant l'écriture de ce manuscrit mais aussi lorsque j'ai fait preuve de phobie administrative. Je lui serai à jamais reconnaissant de ce qu'il a fait pour moi alors que rien ne l'y obligeait.

Je veux à présent remercier tous les doctorants qui ont partagé le même bureau que moi à commencer par Santiago qui m'a accueilli lors de mon premier jour dans l'équipe et qui m'a permis de bien m'intégrer. Même s'il n'a pas souvent été là, ce fut toujours un plaisir de discuter avec David Wu. Je remercie ensuite Cassandre d'avoir été ma coéquipière de badminton d'un jour mais surtout, et ce malgré ses cris lorsque je ratais un but, d'avoir été ma coéquipière sur Rocket League. Je remercie Clara qui, avec Gaspard, a été présente pour me soutenir moralement durant les moments les plus difficiles de ma thèse. Le repas quasiment traditionnel du mercredi midi avec elle et Thibaut me manqueront. Je tiens aussi à remercier Margot qui, arrivée à la fin de ma thèse, a réussi à ramener une bonne ambiance dans un bureau qui se faisait vide après le passage du COVID et plusieurs départs. Enfin, j'ai gardé le meilleur (ou le pire ?) pour la fin : Gaspard. Son humour noir, son ironie permanente, sa culture internet, nos débats pleins de mauvaise foi (surtout de son côté), notre marathon *Seigneur des anneaux*, nos aventures sur Minecraft (ce qui s'est passé dans Minecraft reste dans Minecraft) et bien d'autres moments auront réussi à égayer ma thèse. Je suis très heureux d'avoir pu partager cette expérience avec lui, car il a rendu ma thèse encore plus inoubliable.

Je remercie également tous les autres membres de l'équipe DECISION, de l'équipe RO et de l'équipe SMA avec qui j'ai partagé plein de moments de convivialité lors des pauses-café, des repas au restaurant du CROUS, ou encore autour d'un verre. Plus particulièrement, je remercie Alexandre, qui m'a lui aussi soutenu avec Gaspard et Clara lors des moments difficiles, Anne-Elisabeth, qui aura partagé avec moi certains de ses paniers de l'AMAP, David Saulpic, toujours prompt à parler de politique, Ismaïl, toujours prompt à parler, Kostas, with whom I had very deep philosophical discussions, Nadjet, qui a réussi à supporter mes questions parfois indiscrètes, Parham, avec qui nous avons pu nous mettre d'accord sur le fait que le cinéma de Christopher Nolan était surcoté, Hugo, qui pourra en témoigner, Thibaut, qui m'a prêté, malgré lui, sa chaise de bureau durant ma thèse et dont je m'excuse de ne pouvoir lui rendre en bon état, Nawal, qui a toujours été de bon conseil, Patrice, un adversaire redoutable aux échecs, Olivier, pour son humour, ses charades et ses jeux de société, Fanny, qui j'espère aura apprécié le thé offert lors d'un secret santa, Bruno, avec qui j'ai beaucoup aimé

discuter, Adèle, avec qui j'ai partagé de bonnes bières sur les quais de Seine, Arnaud et ses sacs énormes d'arachides pour avoir des protéines, Pierre, toujours prêt à louer les mérites de l'Auvergne et enfin Franco à qui je n'ai malheureusement pas pu dire au revoir avant son retour au Chili.

S'il existe une autre personne que Pierre-Henri sans qui cette thèse n'aurait jamais vu le jour, c'est bien Audrey. Depuis le lycée elle aura toujours été là pour me pousser à me dépasser et à (essayer) de la dépasser. En effet, notre compétition pour obtenir les meilleures notes et qui a débuté au lycée nous aura emmenés bien loin. Malgré tout, je n'ai pas de mal à avouer que c'est elle qui a toujours fini devant que ce soit pour le baccalauréat, la licence ou le master. La thèse n'aura pas fait exception puisqu'elle a réussi à l'obtenir la première. Cependant, qu'elle ne se repose pas sur ses lauriers car je ne perds pas espoir pour le futur ! Plus sérieusement, je la remercie d'avoir toujours été présente pour moi, de m'avoir soutenue de manière inconditionnelle tout au long de ces douze dernières années et de faire de moi une personne meilleure jour après jour.

Pour terminer, je remercie ma famille de m'avoir donné le goût d'apprendre mais aussi de m'avoir permis de faire mes études. Je remercie également tous mes amis d'avoir été là pour moi et en particulier Damien et Nicolas qui me supportent depuis bien longtemps.

En résumé : un grand merci à tous !

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations et notations

$\mathbb{N}$	Ensemble des entiers naturels
$\mathbb{R}$	Ensemble des réels
$\overline{\mathbb{R}}$	Droite réelle achevée
$\mathbb{I}$	Ensemble unité $[0, 1]$
$[\![1, n]\!]$	Ensemble des entiers entre 1 et n
$\lvert E\rvert$	Cardinal d'un ensemble E
$\lvert\lvert \boldsymbol{v} \rvert\rvert$	Norme d'un vecteur $\boldsymbol{v}$
$\mathbb{1}_A$	Fonction indicatrice d'un ensemble A
μ, ν, ρ	Mesures
λ	Mesure de Lebesgue
γ	Mesure de comptage
$B(.,.)$	Fonction bêta
$\Gamma(.)$	Fonction gamma
Ω	Univers
$\mathcal{A}, \mathcal{B}, \mathcal{T}$	Tribus
$\mathbb{P}$	Mesure de probabilité
$(\Omega, \mathcal{T}, \mathbb{P})$	Espace probabilisé
X, Y, Z	Variables aléatoires
Ω_X	Valeurs d'une variable aléatoire X
$\mathbb{P}_X$	Distribution d'une variable aléatoire X
$\mathcal{B}(\mathbb{R})$	Tribu de Borel sur $\mathbb{R}$
F_X	Fonction de répartition de X
f_X	Fonction densité de X
$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}$	Dérivée de Radon-Nikodym
p_X	Fonction de masse
$\mathbb{E}_{\mathbb{P}}[X]$	Espérance d'une variable aléatoire X
$\mathbb{V}_{\mathbb{P}}[X]$	Variance d'une variable aléatoire X
V_H	H-volume d'une fonction H
$\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$	Vecteurs aléatoires
$\lvert\boldsymbol{M}\rvert$	Déterminant d'une matrice $\boldsymbol{M}$
F_i	Distribution marginale
$\mathrm{Cov}(X, Y)$	Covariance entre X et Y
$H(X)$	Entropie d'une variable aléatoire X

$D(\mathbb{P}||\mathbb{Q})$ . . . . . .Entropie relative
$H_\rho(\mathbb{P}||\mathbb{Q})$ . . . . . .Entropie croisée
$I(X;Y)$ . . . . . .Information mutuelle
H_0, H_1 . . . . . .Hypothèses statistiques
B_{01} . . . . . .Facteur de Bayes entre H_0 et H_1
G . . . . . .Structure de réseau bayésien
$X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$ . . . . . .Indépendance
$\mathcal{I}(\mathbb{P}_X)$ . . . . . .Indépendances vérifiées par $\mathbb{P}_X$
$\mathcal{I}(\mathcal{G})$ . . . . . .Indépendances encodées par la structure G
$\mathbf{Ne}_v^G$ . . . . . .Voisins de v dans le graphe G
$\mathbf{Pa}_v^G$ . . . . . .Parents de v dans le graphe G
$\mathbf{ND}_v^G$ . . . . . .Non-descendants de v dans le graphe G
C . . . . . .Copule
c . . . . . .Copule densité
$\mathscr{C}_n$ . . . . . .Ensemble des copules de dimension n
W_n . . . . . .Borne inférieure de Fréchet-Hoeffding
M_n . . . . . .Borne supérieure de Fréchet-Hoeffding
$\rho(X,Y)$ . . . . . .Rho de Spearman entre X et Y
$\tau(X,Y)$ . . . . . .Tau de Kendall entre X et Y
Π . . . . . .Copule indépendante
$\hat{C}_m$ . . . . . .Copule empirique
$\hat{C}_{K,m}^B$ . . . . . .Copule de Bernstein empirique
$\hat{c}_{K,m}^B$ . . . . . .Copule densité de Bernstein empirique

Introduction

La science moderne se base sur l'observation de phénomènes empiriques qu'elle cherche à expliquer par le biais de modèles mathématiques. De ses débuts au XVI^{e} siècle, jusque vers la fin du XIX^{e} siècle, les modèles utilisés étaient pour la plupart déterministes et s'exprimaient, par exemple, sous la forme d'équations différentielles. Cependant, de par la nature même de certains problèmes ou bien par manque d'information, il arrive souvent que l'on soit confrontés à des incertitudes. Bien que plusieurs théories aient vu le jour au siècle dernier pour quantifier les incertitudes tels que les ensembles flous (Zadeh 1996), la théorie des possibilités (Dubois et al. 1988), la théorie des croyances de Dempster-Shafer (Dempster 1968 ; Shafer 1976) ou encore les probabilités imprécises (Walley 1990), la théorie dominante reste encore aujourd'hui celle des probabilités. Pour ces raisons et avec l'avènement de l'âge de l'information, l'apprentissage statistique qui consiste à construire des modèles probabilistes à partir de données d'observation joue un rôle toujours plus important en science.

Dans les domaines d'application tel que l'ingénierie, la biologie, la finance, etc., il est commun que les problèmes que l'on cherche à modéliser fassent intervenir un grand nombre de variables aléatoires continues interagissant entre elles. Parce que tout calcul numérique s'effectue avec des ressources finies, un compromis doit être fait entre la précision du modèle et sa complexité. Les réseaux bayésiens, issus de l'intelligence artificielle (Pearl 1988), permettent de réduire cette complexité en tirant parti des indépendances conditionnelles entre les variables aléatoires pour factoriser la distribution jointe, de grande dimension, comme un produit de distributions conditionnelles de moindres dimensions. De plus, cette factorisation est encodée au sein d'un graphe qui permet de visualiser l'ensemble des indépendances conditionnelles et qui facilite donc les discussions avec les experts du domaine. Deux questions importantes émergent alors : (i) Comment choisir la structure du graphe, c'est-à-dire comment déterminer les indépendances conditionnelles vérifiées entre les variables aléatoires ? (ii) Étant donnée une structure, comment choisir les paramètres des distributions conditionnelles ? Les structures peuvent être construites manuellement avec l'aide d'experts mais cela implique potentiellement un biais. Pour cette raison, de nombreux travaux se sont intéressés à l'acquisition automatique de connaissances à partir de données d'observations, tout en permettant l'incorporation de connaissances expertes (Neapolitan 2004). D'un point de vue purement statistique, l'apprentissage de la structure correspond à de la sélection de modèle tandis que l'apprentissage des paramètres correspond à de l'estimation paramétrique. La complexité de ces tâches augmente exponentiellement avec la dimension du problème et les calculs deviennent donc impossibles pour de grandes dimensions. Cependant, en exploitant les indépendances entre les variables aléatoires, cette complexité est grandement diminuée ce qui permet tout de même d'implémenter des algorithmes d'apprentissage pour de grandes dimensions. Malheureusement, elles restent principalement limitées au cas où les variables sont toutes discrètes et ce pour plusieurs raisons. La première est tout simplement l'absence, dans le cas continu, d'un modèle par défaut comme l'est la loi catégorielle dans le cas discret. En effet, les lois discrètes s'expriment

naturellement en donnant la probabilité de réalisation de chaque événement. Deux solutions apparaissent pour le traitement du continu : soit nous prenons un modèle paramétrique continu, ce qui impose de faire une hypothèse supplémentaire contraignante sur le modèle, soit nous utilisons des modèles non-paramétriques. Dans les deux cas, nous sommes confrontés à des restrictions supplémentaires puisque le modèle utilisé doit être stable sous un ensemble d'opérations (Shenoy 1997) afin de pouvoir mener des inférences exactes dans un réseau bayésien. Parmi les modèles continus, la loi normale (Lauritzen et Wermuth 1989) vérifie ces propriétés mais celle-ci manque d'expressivité puisqu'elle ne permet pas d'encoder de dépendances non-linéaires ou bien de simuler des événements rares. Plusieurs modèles non-paramétriques ont été proposés pour aller au-delà du modèle gaussien tout en vérifiant les propriétés de stabilités (Moral et al. 2001 ; Shenoy et West 2011 ; Langseth et al. 2012) mais ces modèles sont difficiles à apprendre et restent par conséquent limités à de petites dimensions (Romero et al. 2006). Une autre solution simple en apparence serait de discrétiser les variables continues afin de pouvoir utiliser des réseaux bayésiens discrets. Cependant, ceci soulève de nouvelles problématiques liées à la méthode de discrétisation. La complexité des algorithmes d'inférence étant exponentielle par rapport au nombre maximal de valeurs que prend une variable aléatoire, nous sommes limités dans la précision que l'on peut atteindre. Ceci peut donc être un problème lorsque nous avons par exemple des données distribuées selon une loi multimodale. Quand bien-même nous ne serions pas limités dans le nombre de classes pour la discrétisation, la question de savoir comment placer les points de discrétisation de façon à limiter la perte d'information n'est pas triviale. Enfin, il arrive que dans certaines applications, nous voulions échantillonner la distribution et la discrétisation n'est donc pas une bonne solution pour ces cas là. Toutes ces raisons font qu'il n'existe à l'heure actuelle aucun modèle qui soit pleinement satisfaisant pour le cas continu. Toutefois, il existe une approche intéressante mais encore peu développée qui consiste à utiliser la théorie des copules pour paramétrer et apprendre la structure du graphe.

La théorie des copules prend ses origines dans les années 50 avec, dans un premier temps, les travaux de Fréchet sur les distributions multivariées à marginales données (Fréchet 1951) puis avec ceux de Sklar (Sklar 1959) introduisant la notion de *copule* et le théorème qui porte désormais son nom. Durant plusieurs années, les copules restent confinées à la théorie des espaces métriques et peu de sources sur le sujet existent. Ce n'est que dans les années 90, avec l'intérêt qu'il leur est porté dans le domaine de la finance (Genest, Gendron et al. 2009), qu'elles vont peu à peu gagner en popularité avec notamment la publication de deux ouvrages (Joe 1997 ; Nelsen 2007) qui sont aujourd'hui considérés comme des références dans le domaine. Depuis, leur application s'est étendue à d'autres sujets comme par exemple l'hydrologie (Salvadori et al. 2007) ou les sciences environnementales (Bhatti et al. 2019). Concrètement, une copule est une distribution multivariée définie sur $[0, 1]^n$, avec n la dimension, et dont les marginales unidimensionnelles sont uniformes sur $[0, 1]$. Le théorème de Sklar permet alors la décomposition d'une distribution jointe en une copule et l'ensemble de ses marginales unidimensionnelles. Si, comme dans le cas qui nous intéresse, ces dernières sont continues, la copule est en plus unique. Ce théorème est un résultat central de la théorie des copules puisqu'il exprime le fait que la copule d'une distribution contient toute l'information sur les dépendances entre ses variables tandis que les marginales encodent leurs comportements individuels. Du point de vue de la modélisation statistique, cette décomposition nous offre la liberté de choisir indépendamment le modèle de dépendance de celui des marginales. Nous pouvons alors créer des modèles multivariés plus riches que les distributions classiques. Toutefois, la plupart des copules sont définies seulement pour le cas bidimensionnel et les rares copules définies pour n'importe

quelle dimension sont dérivées des distributions multivariées classiques en inversant le théorème de Sklar. Des procédures ont été proposées pour la construction de copules multidimensionnelles à partir de copules bivariées (Czado 2010) mais ces constructions deviennent difficiles à manipuler pour de grandes dimensions. Nous avons justement vu plus haut que les réseaux bayésiens permettaient de factoriser une distribution jointe en un produit de distributions conditionnelles de moindres dimensions permettant ainsi de réduire la complexité. Les modèles graphiques semblent donc être une bonne solution pour résoudre les difficultés rencontrées lors de la construction de copules multidimensionnelles. Inversement, comme la copule contient la dépendance entre les variables aléatoires, cela en fait un bon outil pour l'apprentissage de la structure d'un réseau bayésien.

C'est sur ces observations qu'ont été proposés les *réseaux bayésiens à base de copules* (CBNs pour *Copula Bayesian Networks*) (Elidan 2010). Tout comme pour les réseaux bayésiens, ils utilisent un graphe orienté acyclique pour encoder l'ensemble des indépendances conditionnelles vérifiées par la distribution multivariée. Mais, au lieu de factoriser la distribution jointe dans son ensemble, c'est la copule elle même qui est décomposée en un ensemble de copules locales de moindres dimensions. Ainsi, à chaque nœud du graphe est associée une de ces copules locales ainsi que la marginale unidimensionnelle de la variable. Cette paramétrisation donne alors un degré de liberté supplémentaire pour la modélisation puisqu'en plus de pouvoir choisir indépendamment le modèle de dépendance de celui des marginales, nous pouvons choisir un modèle différent pour chaque copule locale. D'autres propositions de modèles graphiques pour les copules ont été faites (Czado 2010 ; Bedford et al. 2002) mais celles-ci n'utilisent pas le même langage graphique que les réseaux bayésiens contrairement aux CBNs. Or, cette propriété est intéressante pour l'apprentissage de la structure puisque nous pouvons alors adapter les mêmes méthodes qui sont utilisées dans le cas discret. C'est ce que nous nous proposons de faire dans cette thèse afin d'apprendre des modèles continus de grandes dimensions. Afin de répondre aux problématiques que nous avons soulevées dans le cadre des réseaux bayésiens continus, nous faisons le choix d'utiliser la copule de Bernstein empirique (Sancetta et al. 2004) pour obtenir une modélisation non-paramétrique des copules locales du CBN. De plus, nous nous servirons de cette dernière pour dériver des tests d'indépendances non-paramétriques que nous utiliserons pour la reconstruction des graphes. Ainsi, le même modèle sera utilisé pour l'apprentissage de la structure et pour l'estimation des copules locales. Bien que nous nous concentrions uniquement sur l'apprentissage, remarquons que notre choix de modèle ne répond pas *a priori* aux critères pour pouvoir mener des inférences exactes. Lorsque cela n'est pas possible, il existe tout de même des méthodes approchées qui permettent de mener des inférences en échantillonnant la distribution (Koller et al. 2009). En conclusion de la thèse, nous donnerons des perspectives pour réussir à mener des inférences dans le cas continu à partir des algorithmes d'inférence pour les réseaux bayésiens discrets.

Organisation du manuscrit et contributions

Cette thèse a pour objectif l'apprentissage de distributions multivariées continues à l'aide de modèles graphiques probabilistes et de la théorie des copules. Comme nous entendons nous adresser aux deux communautés, nous nous plaçons dans le cadre général de l'inférence statistique. En cela, notre introduction des réseaux bayésiens est quelque peu différente des ouvrages classiques tels que Koller et al. (2009) ou Darwiche (2009) et s'inspire, dans l'esprit, de ce qui est fait dans Neapolitan (2004). De plus, nous nous intéressons au cas où les variables du problème sont toutes continues et uti-

lisons donc la théorie de la mesure pour obtenir un cadre rigoureux. Cela nous permet également de donner un cadre unifié aux réseaux bayésiens discrets et continus.

Pour ces raisons, la première des trois parties qui composent le manuscrit porte sur l'ensemble des prérequis nécessaires à cette présentation. Bien que certaines des notions qui y sont abordées peuvent être connues du lecteur, nous avons préféré ne pas reporter cette partie en annexe afin de donner un socle commun aux deux domaines. Malgré tout, le lecteur le désirant pourra éventuellement ignorer certains passages ou bien s'y reporter uniquement lorsque c'est nécessaire. La deuxième partie introduit ensuite l'état de l'art portant sur les réseaux bayésiens et les copules tandis que la troisième partie discute des méthodes d'apprentissage que nous avons implémentées durant la thèse et qui ont mené à des publications. Chacune de ces parties se divisent en plusieurs chapitres dont nous donnons maintenant un résumé.

Partie I : Préliminaires

Cette partie est composées de trois chapitres portant respectivement sur la théorie des probabilités, la théorie de l'information et les statistiques bayésiennes. Comme nous l'avons dit, les probabilités sont introduites dans le cadre de la théorie de la mesure ce qui nous permet d'unifier cas discret et cas continu. Cela sera également le cas pour la théorie de l'information. La théorie de la mesure n'étant en général pas abordée dans la littérature des réseaux bayésiens nous recommandons ces deux chapitres à un public venant des modèles graphiques probabilistes. Le chapitre sur les statistiques bayésiennes sera quant à lui plutôt recommandé pour un public venant du domaine des copules.

- **Chapitre 1 – Théorie des probabilités**. La notion importante de ce chapitre est celle de densité d'une distribution définie comme sa dérivée de Radon-Nikodym relativement à une mesure de référence. C'est elle qui par la suite nous permet de traiter les réseaux bayésiens discrets et continus de manière unifiée. Le reste du chapitre présente les notions de base de théorie des probabilités. Le lecteur voulant ignorer ce chapitre pourra remplacer dans le cas discret le terme de densité par celui de fonction de masse.
- **Chapitre 2 – Théorie de l'information**. Nous donnons les concepts de base de théorie de l'information en utilisant le cadre de la théorie de la mesure pour unifier les cas discrets et continus. En effet, la plupart du temps l'entropie est étendue au cas continu par le biais de l'entropie différentielle qui ne conserve quasiment aucune des bonnes propriétés de l'entropie discrète. Nous verrons que la bonne quantité s'avère être l'entropie relative.
- **Chapitre 3 – Statistiques bayésiennes**. Nous nous focalisons dans ce chapitre sur l'approche bayésienne de l'inférence statistique en présentant l'estimation paramétrique et la sélection de modèle. Un des concepts importants pour la suite est celui de facteur de Bayes que nous utiliserons pour l'apprentissage de la structure d'un réseau bayésien. Bien que la majorité du chapitre se concentre sur l'approche bayésienne, nous présentons également l'approche classique de la sélection de modèle qui est utilisée dans les méthodes d'apprentissage par contraintes.

Partie II : État de l'art

La partie sur l'état de l'art se compose elle aussi de trois chapitres. Les deux premiers introduisent les réseaux bayésiens et les méthodes d'apprentissage les concernant tandis que le troisième porte sur la théorie des copules.

- **Chapitre 4 – Les réseaux bayésiens**. L'objectif de cette thèse étant l'apprentissage de modèles probabilistes, nous introduisons les réseaux bayésiens dans le

contexte de l'inférence statistique. Pour cela, nous les voyons comme l'extension du chapitre 3 au cas multivarié pour lequel nous avons besoin de faire des hypothèses d'indépendances afin de simplifier le modèle et pouvoir mener les calculs. La partie graphique émerge alors de la volonté d'avoir un moyen compact de représenter ces indépendances. Pour cela, nous introduisons la d-séparation et les modèles d'indépendance graphique en suivant l'approche originale de Pearl (2014).

- **Chapitre 5 – Apprentissage des réseaux bayésiens**. Ce chapitre présente les diverses méthodes d'apprentissage des réseaux bayésiens et s'appuie pour cela sur le chapitre 3. Il se divise en deux sections dont une portant sur l'apprentissage des paramètres et l'autre sur l'apprentissage de la structure. Pour l'apprentissage de la structure, nous donnons trois algorithmes classiques qui seront ensuite adaptés aux réseaux bayésiens à base de copules dans les chapitres 7 et 8. Le premier est un algorithme de recherche locale maximisant un score bayésien ce qui, comme nous le montrons, revient à utiliser les facteurs de Bayes. Le deuxième est l'algorithme PC (Spirtes et al. 2000) pour lequel nous avons une discussion rarement abordée sur l'obtention d'un *Directed Acyclic Graph* à partir du *Completed Partially Directed Acyclic Graph*. Enfin, nous introduisons l'algorithme MIIC (Affeldt et Isambert 2015) pour lequel nous donnons une réécriture des différentes équations afin qu'elles soient plus en adéquation avec l'approche NML de la sélection de modèle que nous présenterons à cette occasion.
- **Chapitre 6 – Théorie des copules**. Nous introduisons les notions de base de la théorie des copules pour la modélisation de dépendances. Les deux points importants de ce chapitre sont le théorème de Sklar, qui permet de séparer la modélisation des dépendances de celle des marginales, et la copule de Bernstein empirique, qui donne un estimateur non-paramétrique de la copule. Dans la partie III, la copule de Bernstein sera utilisée afin de paramétrer les réseaux bayésiens à base de copules ainsi que pour dériver des tests d'indépendances conditionnelles non-paramétriques.

Partie III : Contributions à l'apprentissage de modèles continus

La troisième partie décrit les différents algorithmes pour l'apprentissage de la structure d'un réseau bayésien à base de copules que nous avons implémentés dans la bibliothèque *otagrum* et qui ont mené à des publications.

- **Chapitre 7 – CPC : un algorithme basé sur la distance de Hellinger**. Ce chapitre commence par introduire le modèle des réseaux bayésiens à base de copules (Elidan 2010) que nous utilisons afin de construire des distributions multivariées. Pour en apprendre la structure, nous proposons un algorithme PC utilisant un test d'indépendance non-paramétrique dont la statistique de test est basée sur la distance de Hellinger et la copule de Bernstein empirique. Nous comparons les performances de ce nouvel algorithme à celles de l'algorithme proposé par Elidan (2010) et d'un algorithme de recherche locale pour les réseaux bayésiens gaussiens (Geiger et Heckerman 1994). Le module *otagrum* qui est utilisé pour faire ces comparaisons y est discuté plus en détails.

  Publications associées :

  — Lasserre, M. et al. (2020). « Constraint-Based Learning for Non-Parametric Continuous Bayesian Networks ». In : *FLAIRS 33 - 33rd Florida Artificial Intelligence Research Society Conference*. Miami, United States : AAAI, p. 581-586

— Lasserre, M. et al. (2021a). « Constraint-based learning for non-parametric continuous bayesian networks ». In : *Annals of Mathematics and Artificial Intelligence*, p. 1-18

- **Chapitre 8 – CMIIC : un algorithme basé sur l'information mutuelle**. Nous généralisons la dérivation de tests d'indépendances non-paramétriques avec la copule de Bersntein empirique en introduisant la divergence de Csizar. Ainsi, en choisissant différentes fonctions génératrices, nous pouvons obtenir différentes statistiques de tests. Dans la suite du chapitre, nous faisons le choix de l'entropie relative pour des raisons que nous motivons. La statistique de test qui en découle étant l'information mutuelle conditionnelle, nous proposons l'algorithme CMIIC qui est l'adaptation de l'algorithme MIIC aux réseaux bayésiens à base de copules. Cette statistique nous permet également de donner une amélioration, en terme de vitesse, de l'algorithme proposé par Elidan (2010). Les performances de ces deux algorithmes sont alors comparées avec celles de l'algorithme CPC du chapitre 7. Enfin, nous utilisons ces algorithmes afin d'étudier l'échantillon *wine quality* provenant d'un cas d'application.

  Publications associées :

  — Lasserre, M. et al. (2021b). « Learning Continuous High-Dimensional Models using Mutual Information and Copula Bayesian Networks ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 35. 13, p. 12139-12146

Références

Affeldt, S. et Isambert, H. (2015). « Robust Reconstruction of Causal Graphical Models based on Conditional 2-point and 3-point Information. » In : *ACI@ UAI*, p. 1-29 (cf. p. 5, 90, 91, 93).

Bedford, T. et Cooke, R. M. (2002). « Vines–a new graphical model for dependent random variables ». In : *The Annals of Statistics* 30.4, p. 1031-1068 (cf. p. 3, 124).

Bhatti, M. I. et Do, H. Q. (2019). « Recent development in copula and its applications to the energy, forestry and environmental sciences ». In : *International Journal of Hydrogen Energy* 44.36, p. 19453-19473 (cf. p. 2).

Czado, C. (2010). « Pair-copula constructions of multivariate copulas ». In : *Copula theory and its applications*. Springer, p. 93-109 (cf. p. 3, 124).

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge university press (cf. p. 3, 75).

Dempster, A. P. (1968). « A generalization of Bayesian inference ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 30.2, p. 205-232 (cf. p. 1).

Dubois, D. et Prade, H. (1988). « Possibility Theory - An Approach to Computerized Processing of Uncertainty ». In : (cf. p. 1).

Elidan, G. (2010). « Copula bayesian networks ». In : *Advances in neural information processing systems*, p. 559-567 (cf. p. 3, 5, 6, 124, 125, 127, 135, 137, 141, 142, 149, 161).

Fréchet, M. (1951). « Sur les tableaux de corrélation dont les marges sont données ». In : *Ann. Univ. Lyon, 3ˆ e serie, Sciences, Sect. A* 14, p. 53-77 (cf. p. 2).

Geiger, D. et Heckerman, D. (1994). « Learning gaussian networks ». In : *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., p. 235-243 (cf. p. 5, 80, 135, 161).

Genest, C., Gendron, M. et Bourdeau-Brien, M. (2009). « The advent of copulas in finance ». In : *The European journal of finance* 15.7-8, p. 609-618 (cf. p. 2).

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press (cf. p. 2, 98).

Koller, D. et Friedman, N. (2009). *Probabilistic graphical models : principles and techniques*. MIT press (cf. p. 3, 71, 72, 75, 76, 80, 131, 154).

Langseth, H., Nielsen, T. D., Rumi, R. et Salmerón, A. (2012). « Mixtures of truncated basis functions ». In : *International Journal of Approximate Reasoning* 53.2, p. 212-227 (cf. p. 2, 124).

Lasserre, M., Lebrun, R. et Wuillemin, P.-H. (2020). « Constraint-Based Learning for Non-Parametric Continuous Bayesian Networks ». In : *FLAIRS 33 - 33rd Florida Artificial Intelligence Research Society Conference*. Miami, United States : AAAI, p. 581-586 (cf. p. 5, 124).

Lasserre, M., Lebrun, R. et Wuillemin, P.-H. (2021a). « Constraint-based learning for non-parametric continuous bayesian networks ». In : *Annals of Mathematics and Artificial Intelligence*, p. 1-18 (cf. p. 6, 124).

Lasserre, M., Lebrun, R. et 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*. T. 35. 13, p. 12139-12146 (cf. p. 6, 142).

Lauritzen, S. L. et Wermuth, N. (1989). « Graphical models for associations between variables, some of which are qualitative and some quantitative ». In : *The annals of Statistics*, p. 31-57 (cf. p. 2, 124).

Moral, S., Rumií, R. et Salmerón, A. (2001). « Mixtures of truncated exponentials in hybrid Bayesian networks ». In : *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer, p. 156-167 (cf. p. 2, 124).

Neapolitan, R. E. (2004). *Learning bayesian networks*. T. 38. Pearson Prentice Hall Upper Saddle River, NJ (cf. p. 1, 3, 75).

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media (cf. p. 2, 98, 102, 104, 108, 116).

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan kaufmann (cf. p. 1).

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Elsevier (cf. p. 5, 70, 75).

Romero, V., Rumí, R. et Salmerón, A. (2006). « Learning hybrid Bayesian networks using mixtures of truncated exponentials ». In : *International Journal of Approximate Reasoning* 42.1-2, p. 54-68 (cf. p. 2, 124).

Salvadori, G. et De Michele, C. (2007). « On the use of copulas in hydrology : theory and practice ». In : *Journal of Hydrologic Engineering* 12.4, p. 369-380 (cf. p. 2).

Sancetta, A. et Satchell, S. (2004). « The Bernstein copula and its applications to modeling and approximations of multivariate distributions ». In : *Econometric theory* 20.3, p. 535-562 (cf. p. 3, 111, 113, 116).

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton university press (cf. p. 1).

Shenoy, P. P. (1997). « Binary join trees for computing marginals in the Shenoy-Shafer architecture ». In : *International Journal of approximate reasoning* 17.2-3, p. 239-263 (cf. p. 2, 164).

Shenoy, P. P. et West, J. C. (2011). « Inference in hybrid Bayesian networks using mixtures of polynomials ». In : *International Journal of Approximate Reasoning* 52.5, p. 641-657 (cf. p. 2, 124).

Sklar, A. (1959). « Fonctions de répartition à n dimensions et leurs marges ». In : *Publ. Inst. Statist. Univ. Paris* 8, p. 229-231 (cf. p. 2).

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G. et Richardson, T. (2000). *Causation, prediction, and search*. MIT press (cf. p. 5, 85, 86, 161).

Walley, P. (1990). « Statistical Reasoning with Imprecise Probabilities ». In : (cf. p. 1).

Zadeh, L. A. (1996). « Fuzzy sets ». In : *Fuzzy sets, fuzzy logic, and fuzzy systems : selected papers by Lotfi A Zadeh*. World Scientific, p. 394-432 (cf. p. 1).

Partie I

Préliminaires

Chapitre 1

Théorie des probabilités

Sommaire

Nous introduisons ici la théorie des probabilités dans le cadre naturel de la théorie de la mesure permettant d'unifier les cas discrets, continus et mixtes. Le but de cette section est d'introduire les différentes définitions et notations qui seront utilisées tout au long de cette thèse. Il existe de nombreux ouvrages de références et le lecteur intéressé peut se référer par exemple à Candelpergher (2013), Ouvrard (2004), Grimmett et al. (2020), Williams (1991) et Rosenthal (2006) pour une introduction plus détaillée. En particulier, les théorèmes et propositions qui sont présentés ici sont démontrés dans ces références.

1.1 Variable aléatoire

La construction théorique des probabilités, proposée par Kolmogorov et al. (2018), repose sur la notion d'espace probabilisé :

Définition 1.1.1 (Espace probabilisé)**.** Un espace probabilisé est un espace mesuré $(\Omega, \mathcal{A}, \mathbb{P})$ dont la mesure $\mathbb{P}$ est positive et vérifie :

$$\mathbb{P}(\Omega) = 1. \tag{1.1}$$

Ω est appelé univers ou espace des éventualités, les éléments de la tribu $\mathcal{A}$ des événements et la mesure $\mathbb{P}$ est qualifiée de mesure de probabilité.

Il est rare cependant de travailler directement avec l'ensemble Ω, soit parce que cet ensemble est trop grand (complexe) soit parce que nous n'y avons tout simplement pas accès. À la place, nous travaillons avec des variables aléatoires :

Définition 1.1.2 (Variable aléatoire)**.** Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé. Une variable aléatoire X à valeurs dans l'espace probabilisable $(\Omega_X, \mathcal{A}_X)$ est une application mesurable de $(\Omega, \mathcal{A})$ vers $(\Omega_X, \mathcal{A}_X)$, c'est-à-dire une application vérifiant :

$$\forall A_X \in \mathcal{A}_X, \quad X^{-1}(A_X) \in \mathcal{A}.$$

Lorsque l'ensemble Ω_X est dénombrable, on parle de variable aléatoire discrète, tandis que lorsqu'il est non-dénombrable on parle de variable aléatoire continue. Si $\Omega_X \subseteq \mathbb{R}$, on parle de variable aléatoire réelle.

Dans la suite de la thèse, l'ensemble des variables aléatoires que nous considérons sont réelles. La variable aléatoire X nous permet alors de définir une mesure image, appelée distribution de la variable aléatoire, sur l'espace probabilisable $(\Omega_X, \mathcal{A}_X)$:

Définition 1.1.3 (Distribution d'une variable aléatoire)**.** La distribution ou loi d'une variable aléatoire X sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ est la mesure image de $\mathbb{P}$ par X :

$$\mathbb{P}_X = \mathbb{P} \circ X^{-1}. \tag{1.2}$$

La figure 1.1 illustre la définition d'une variable aléatoire entre deux espaces probabilisés $(\Omega, \mathcal{A}, \mathbb{P})$ et $(\Omega_X, \mathcal{A}_X, \mathbb{P}_X)$. Dans la pratique, nous travaillons sur $(\Omega_X, \mathcal{A}_X, \mathbb{P}_X)$ et souvent il n'est pas fait mention de l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ sous-jacent.

***Exemple* 1.1.1.** Considérons un paquet de 52 cartes et l'expérience selon laquelle on tire une carte au hasard. L'univers Ω de l'espace probabilisé associé à cette expérience est l'ensemble des 52 cartes, sa tribu $\mathcal{A}$ est l'ensemble des parties de Ω et la mesure de probabilité est donnée par :

$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} = \frac{1}{52}$$

pour tout $\omega \in \Omega$. Plusieurs variables aléatoires peuvent être définies sur $(\Omega, \mathcal{A}, \mathbb{P})$: l'application X_1 qui à une carte associe sa valeur, l'application X_2 qui à une carte associe sa couleur, etc. Imaginons que nous ayons une balance assez précise, nous pourrions même avoir la variable aléatoire X_3 associant sa masse à une carte [a]. Si nous prenons le cas de X_1, elle prend ses valeurs dans $\Omega_{X_1} = [\![1, 13]\!]$ et la tribu associée est $\mathcal{A}_{X_1} = \mathcal{P}(\Omega_{X_1})$. Finalement, la distribution de X_1 est donnée par :

$$\mathbb{P}_{X_1}(\{x_1\}) = \mathbb{P} \circ X_1^{-1}(\{x_1\}) = \mathbb{P}\left(\left\{ \boxed{\spadesuit\ x_1\ \spadesuit}, \boxed{\clubsuit\ x_1\ \clubsuit}, \boxed{\heartsuit\ x_1\ \heartsuit}, \boxed{\diamondsuit\ x_1\ \diamondsuit} \right\}\right) = \frac{1}{13}$$

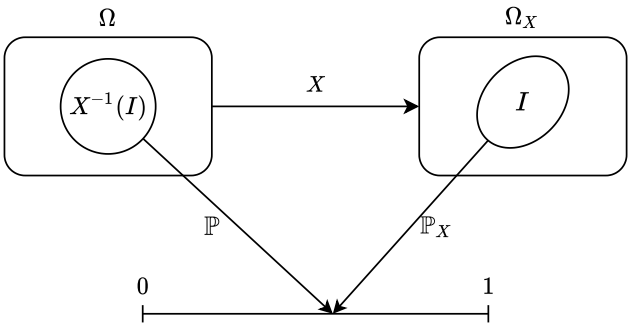


FIGURE 1.1 – Illustration d'une variable aléatoire X et de sa mesure image $\mathbb{P}_X$.

pour tout $x_1 \in \Omega_{X_1}$, puisque pour chaque valeur il existe quatre carte (trèfle, pique, carreau, cœur).

a. Même si les valeurs de X_3 sont dans $\mathbb{R}$, la variable reste discrète puisque Ω_{X_3} est dénombrable.

1.2 Fonctions de répartition et densité

Dans l'exemple précédent, nous avons donné plusieurs exemples de variables aléatoires discrètes. Nous avons vu que dans ce cas là, la tribu associée était l'ensemble des parties sur l'ensemble d'arrivée de la variable aléatoire. En effet, nous pouvons montrer que cet ensemble vérifie les propriétés d'une tribu. Dans le cas des variables continues en revanche, la tribu utilisée est la tribu de Borel sur $\mathbb{R}$ [1].

Définition 1.2.1 (Tribu de Borel sur $\mathbb{R}$)**.** La tribu de Borel sur $\mathbb{R}$, notée $\mathcal{B}(\mathbb{R})$, est la tribu engendrée par les intervalles $]-\infty, x]$, avec $x \in \mathbb{R}$.

Lorsque nous avons défini la mesure de probabilité, nous nous sommes contentés de spécifier la valeur qu'elle prenait pour les singletons. En effet, ceux-ci engendrent la tribu et en utilisant les propriétés de la mesure de probabilité, nous pouvons alors évaluer sa valeur pour n'importe quel élément de la tribu. Dans le cas d'une variable continue, les intervalles $]-\infty, x]$ engendrant la tribu de Borel forment un continuum et les valeurs de la mesure de probabilité sur ces ensembles définissent la fonction de répartition :

Définition 1.2.2 (Fonction de répartition d'une variable aléatoire)**.** Soit X une variable aléatoire définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ et prenant ses valeurs dans $\Omega_X \subseteq \mathbb{R}$. La fonction de répartition $F_X : \mathbb{R} \to [0,1]$ associée à $\mathbb{P}_X$ est définie par :

$$F_X(x) = \mathbb{P}_X(X \leq x) = \mathbb{P}_X(]-\infty, x]) = \mathbb{P} \circ X^{-1}(]-\infty, x]) \tag{1.3}$$

Cette fonction vérifie un certain nombre de propriétés :

Théorème 1.2.1. Soit X une variable aléatoire et soit $\mathbb{P}_X$ sa distribution La fonction de répartition de $\mathbb{P}_X$ vérifie les propriétés suivantes :

i) $\forall x \in \mathbb{R}, F(x) \geq 0$,

1. Voir CANDELPERGHER (2013, p.44) pour une raison de ce choix.

ii) F est croissante,

iii) $\lim_{x\to-\infty} F_X(x) = 0$ et $\lim_{x\to+\infty} F_X(x) = 1$,

iv) F est continue à droite, c'est-à-dire $\forall c \in \mathbb{R}$, $\lim_{x\to c^+} F(x) = F(c)$.

De plus, étant données deux distributions $\mathbb{P}_X$ et $\mathbb{Q}_X$ ayant pour fonctions de répartition respectives F_X et G_X, nous avons :

$$F_X = G_X \iff \mathbb{P}_X = \mathbb{Q}_X. \tag{1.4}$$

D'après la relation 1.4, la fonction de répartition détermine donc la mesure de probabilité. Quant aux propriétés vérifiées par la fonction de répartition, elles servent d'axiomes pour la définition d'une fonction de répartition indépendamment d'une mesure de probabilité :

Définition 1.2.3 (Fonction de répartition)**.** Une fonction $F : \mathbb{R} \to [0, 1]$ est une fonction de répartition si elle vérifie les propriétés i)-iv).

De la même manière qu'étant donnée une mesure de probabilité nous pouvons définir une fonction de répartition associée, le théorème suivant montre que la relation inverse est également valable :

Théorème 1.2.2. Soit $F : \mathbb{R} \to [0, 1]$ une fonction de répartition. L'application $\sigma_F : \mathcal{B}(\mathbb{R}) \to [0, 1]$ définie par :

$$\begin{aligned} \sigma_F(]a, b]) &= F(b) - F(a), \\ \sigma_F(]-\infty, b]) &= F(b), \\ \sigma_F(]a, +\infty[) &= 1 - F(a) \end{aligned}$$

est une mesure sur la tribu de Borel [a]. La mesure ainsi définie est appelée mesure de Stieltjes associée à la fonction de répartition F.

a. À proprement parler, σ_F se prolonge en une mesure (Candelpergher 2013).

Ainsi, il existe une bijection entre l'espace des mesures de probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et l'espace des fonctions de répartition sur $\mathbb{R}$. Dans la suite nous utiliserons donc indifféremment $\mathbb{P}_X$ ou F_X pour dénoter la distribution d'une variable aléatoire. Pour cette raison, l'intégrale sur un ensemble $E \subseteq \mathbb{R}$ par rapport à une mesure de probabilité $\mathbb{P}_X$ d'une fonction h sur un ensemble $E \subseteq \mathbb{R}$ peut se noter :

$$\int_E h(x)\mathrm{d}F(x). \tag{1.5}$$

Une distribution peut également être définie à partir de sa fonction densité :

Définition 1.2.4 (Fonction densité)**.** Soit X une variable aléatoire à valeurs dans l'espace probabilisable $(\Omega_X, \mathcal{A}_X)$, soit $\mathbb{P}_X$ sa distribution et soit μ une mesure sur ce même espace probabilisable. La distribution $\mathbb{P}_X$ est dite distribution à densité par rapport à μ s'il existe une fonction f_X mesurable positive définie sur Ω_X et vérifiant :

$$\forall x \in \Omega_X, \quad F_X(x) = \int_{\{u\leq x\}} f_X(u)\mathrm{d}\mu. \tag{1.6}$$

f_X est alors appelée la fonction de densité de la variable aléatoire X.

La mesure $\mathbb{P}_X$ est dite absolument continue par rapport à μ, ce que l'on note $\mathbb{P}_X \ll \mu$, si $\forall A_X \in \mathcal{A}_X$ tel que $\mu(A_X) = 0$, on a $\mathbb{P}_X(A_X) = 0$. Sous cette condition, le théorème de Radon-Nikodym nous assure l'existence d'une fonction f_X mesurable positive vérifiant la relation 1.6. f_X est alors appelée la dérivée de Radon-Nikodym de $\mathbb{P}_X$ par rapport à la mesure μ et notée $\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mu}$. Lorsque la mesure μ est la mesure de Lebesgue λ, la densité est la fonction densité habituelle dans le cas d'une variable aléatoire continue. Dans le cas où la variable aléatoire est discrète, elle possède une densité par rapport à la mesure de comptage γ qui est la fonction de masse de la distribution définie habituellement comme :

$$p_X(x) = \mathbb{P}_X(X = x), \quad x \in \mathbb{R}. \tag{1.7}$$

Dans la suite nous ne ferons pas de distinction entre les deux cas et emploierons le terme densité indifféremment.

L'intégrale d'une fonction par rapport à une distribution de probabilité peut s'exprimer avec la fonction de densité :

Proposition 1.2.3 (Intégrale par rapport à une distribution à densité)**.** Soit $\mathbb{P}_X$ une distribution sur $(\Omega_X, \mathcal{A})$ de fonction de répartition F_X et de densité f_X par rapport à une mesure μ. Soit h une fonction mesurable sur $(\Omega_X, \mathcal{A})$, alors :

$$\int_{\Omega_X} h(x)\mathrm{d}F_X(x) = \int_{\Omega_X} h(x)f_X(x)\mathrm{d}\mu(x) \tag{1.8}$$

Enfin, lorsqu'une variable aléatoire réelle Y s'exprime comme fonction d'une autre variable aléatoire réelle X, sa densité f_Y s'exprime en fonction de celle de X :

Théorème 1.2.4. Soient $\mathcal{U}$ et $\mathcal{V}$ deux ouverts de $\mathbb{R}$ et $\phi : \mathcal{U} \to \mathcal{V}$ un difféomorphisme, c'est-à-dire une fonction bijective, différentiable et d'inverse différentiable. Soit X une variable aléatoire définie sur l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ possédant une densité f_X et telle que $X(\Omega) \subseteq \mathcal{U}$. La variable aléatoire $Y = \phi(X)$ à valeurs dans $Y(\Omega) \subseteq \mathcal{V}$ a pour densité :

$$f_Y(y) = \left|\frac{\mathrm{d}\phi^{-1}(y)}{\mathrm{d}y}\right| f_X\left(\phi^{-1}(y)\right). \tag{1.9}$$

1.3 Moments d'une variable aléatoires

Une caractéristique importante d'une variable aléatoire est, si elle existe, son espérance. Un exemple classique est que si les valeurs de la variables aléatoires sont associées à un gain, alors son espérance représente le gain moyen. L'espérance d'une variable aléatoire permet donc de déterminer la valeur autour de laquelle cette dernière est répartie.

Définition 1.3.1 (Espérance d'une variable aléatoire)**.** Soit X une variable aléatoire définie sur l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ telle que $X \in \mathcal{L}^1(\mathbb{P})$. L'espérance ou la moyenne de X est définie par :

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\Omega} X\mathrm{d}\mathbb{P} = \int_{\Omega_X} x\mathrm{d}\mathbb{P}_X. \tag{1.10}$$

Lorsqu'il n'y aura pas d'ambiguïté possible sur la mesure de probabilité, nous noterons simplement $\mathbb{E}[X]$. Si la variable aléatoire possède en plus une densité

$f = \frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mu}$ par rapport à une mesure μ, alors :

$$\mathbb{E}[X] = \int_{\Omega_X} x f(x) \mathrm{d}\mu \tag{1.11}$$

d'après le théorème 1.2.3.

L'espérance d'une variable aléatoire est un cas particulier du concept plus général de moment :

Définition 1.3.2 (Moments et moments centrés)**.** Soit une variable aléatoire $X \in \mathcal{L}^k(\mathbb{P})$, où $k > 0$. Son moment d'ordre k est défini par :

$$m_k = \int_{\Omega} X^k \mathrm{d}\mathbb{P} = \int_{\Omega_X} x^k \mathrm{d}\mathbb{P}_X. \tag{1.12}$$

Si $k > 1$, on appelle moment centré d'ordre k la quantité :

$$s_k = \int_{\Omega} (X - m_1)^k \mathrm{d}\mathbb{P} = \int_{\Omega_X} (x - m_1)^k \mathrm{d}\mathbb{P}_X. \tag{1.13}$$

Les moments donnent de l'information sur la forme de la distribution de la variable aléatoire et permettent ainsi de la résumer par un ensemble de scalaires. Dans certains cas, les moments d'une loi peuvent la caractériser de façon unique. C'est par exemple le cas de la distribution gaussienne qui est entièrement déterminée par sa moyenne $m_1 = \mathbb{E}[X]$ et sa variance, c'est-à-dire son moment centré d'ordre 2 $s_2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$.

On appelle échantillon de taille m la suite de variables aléatoires identiquement distribuées $\boldsymbol{D} = (X[1], \ldots, X[m])$ et on note $\boldsymbol{d} = (x[1], \ldots, x[m])$ une réalisation de celle-ci. Dans la suite, les variables $X[i]$ des échantillons que nous considérons seront supposées indépendantes et on parle alors de variables *i.i.d* pour *indépendantes et identiquement distribuées*. Si la distribution de l'échantillon a une espérance μ et une variance $\sigma^2 > 0$, alors le théorème central limite (TCL) stipule que la suite $(\bar{X}_m)$ définie par $\bar{X}_m = \frac{1}{m}\sum_{i=1}^{m} X[i]$ converge en loi vers une distribution normale $N\left(\mu, \frac{\sigma^2}{m}\right)$. Ceci explique donc le rôle important joué par les deux premiers moments d'une loi d'un point de vue statistique. On peut alors être intéressé par l'estimation de la valeur de ces paramètres à partir d'un échantillon de loi inconnue [2]. On peut pour cela utiliser ce que l'on appelle la moyenne et la variance empiriques d'expression respectives :

$$\bar{\mu} = \frac{1}{m}\sum_{i=1}^{m} x[i] \quad \text{et} \quad \bar{\nu} = \frac{1}{m}\sum_{i=1}^{m} (x[i] - \bar{\mu})^2 \tag{1.14}$$

Ces quantités peuvent être vues comme la moyenne et la variance de la distribution empirique de l'échantillon :

Définition 1.3.3 (Distribution empirique)**.** Soit X une variable aléatoire et soit $\boldsymbol{d} = \{x[1], \ldots, x[m]\}$ un ensemble de m réalisations de cette variable. La distribution empirique associée à $\boldsymbol{d}$ est définie par :

$$\hat{F}_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{x[i] \leq x}. \tag{1.15}$$

2. L'estimation de paramètres est abordée plus en détails dans le chapitre 3.

1.4 Lois classiques

Nous rappelons à présent plusieurs lois classiques qui seront utilisées par la suite.

1.4.1 Lois discrètes

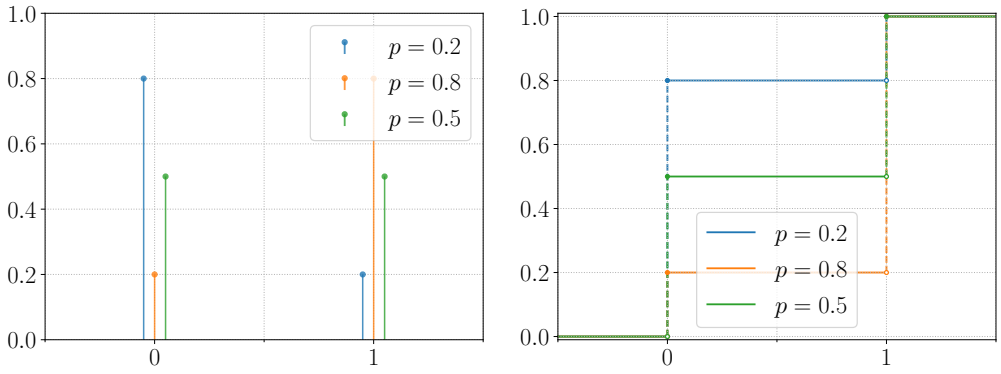
Loi de Bernoulli $B(p)$

Soit X une variable aléatoire prenant ses valeurs dans $\Omega_X = \{0, 1\}$. La loi de X est une loi de Bernoulli de paramètre p si elle a pour densité :

$$f(x) = \begin{cases} p & \text{si } x = 1, \\ 1 - p & \text{si } x = 0 \end{cases} = p^x(1-p)^{1-x}. \tag{1.16}$$

Elle a pour espérance et variance :

$$\mathbb{E}[X] = p, \quad \mathbb{V}[X] = p(1-p) \tag{1.17}$$



(a) Densité (b) Distribution

FIGURE 1.2 – Loi de Bernoulli pour différentes valeurs du paramètre p.

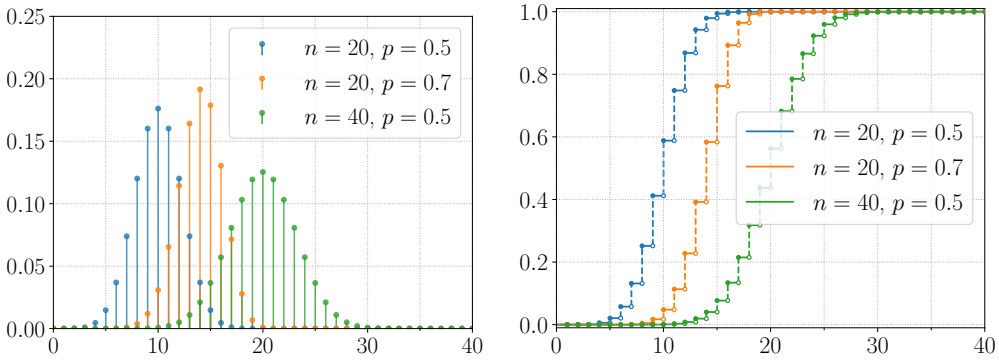
Loi binomiale $B(n, p)$

La variable aléatoire X définie comme la somme de n variables aléatoires $\{X_i\}_{1 \leq i \leq n}$ indépendantes (voir 1.6.5) et suivant toutes une même loi de Bernoulli de paramètre p, suit une loi binomiale de paramètres (n, p). Elle prend ses valeurs dans $\Omega_X = [\![1, n]\!]$ et a pour densité :

$$f(x) = \binom{n}{k} p^x (1-p)^{n-x} \tag{1.18}$$

Elle a pour espérance et variance :

$$\mathbb{E}[X] = np, \quad \mathbb{V}[X] = np(1-p) \tag{1.19}$$



(a) Densité (b) Distribution

FIGURE 1.3 – Loi binomiale pour différentes valeurs des paramètres n et p.

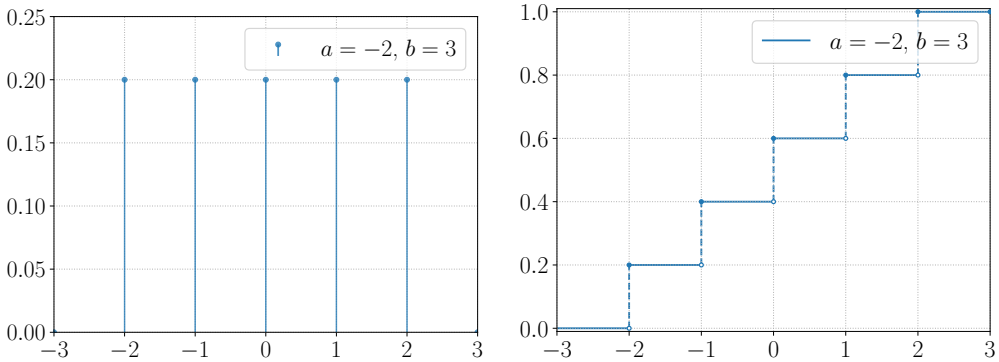
Loi uniforme discrète $U(a, b)$

Soit X une variable aléatoire prenant ses valeurs dans $\Omega_X = [\![a, b]\!]$, un ensemble fini de cardinal $|\Omega_X| = n = b - a + 1$. Cette variable suit une loi uniforme (discrète) de paramètres (a, b) si elle a pour densité :

$$f(x) = \frac{1}{n}, \forall x \in \Omega_X. \tag{1.20}$$

Elle a pour espérance et variance :

$$\mathbb{E}[X] = \frac{a+b}{2}, \quad \mathbb{V}[X] = \frac{n^2-1}{12} \tag{1.21}$$



(a) Densité (b) Distribution

FIGURE 1.4 – Loi uniforme discrète pour différentes valeurs des paramètres a et b.

1.4.2 Lois absolument continues

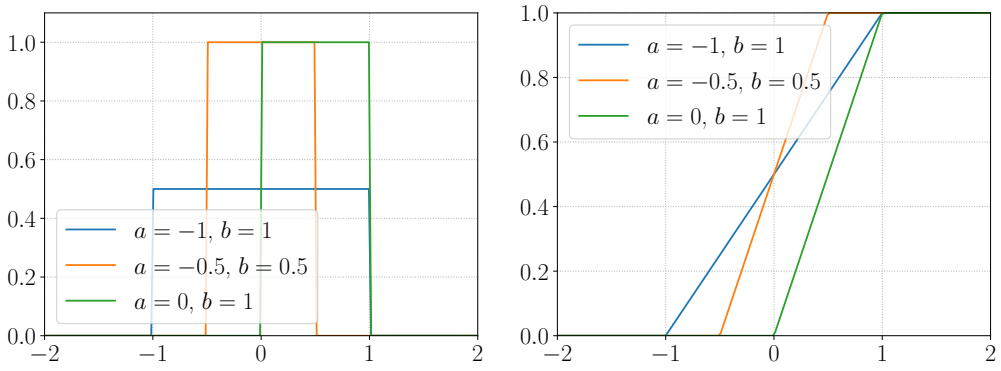
Loi uniforme continue $U(a, b)$

Une variable aléatoire prenant ses valeurs dans $\Omega_X = [a, b] \subset \mathbb{R}$ suit une une loi uniforme (continue) de paramètres (a, b) si elle a pour densité :

$$f(x) = \frac{1}{|b-a|} \mathbb{1}_{[a,b]}(x). \tag{1.22}$$

Elle a pour espérance et variance :

$$\mathbb{E}[X] = \frac{a+b}{2}, \quad \mathbb{V}[X] = \frac{(b-a)^2}{12} \tag{1.23}$$



(a) Densité (b) Distribution

FIGURE 1.5 – Loi uniforme continue pour différentes valeurs des paramètres a et b.

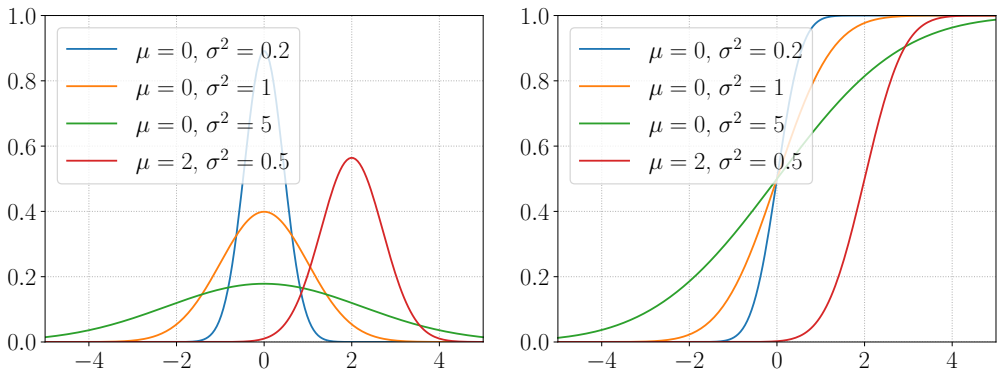
Loi normale $N(\mu, \sigma)$

Une variable aléatoire réelle suit une loi normale (ou gaussienne) de paramètres (μ, σ) si elle a pour densité :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{1.24}$$

Elle a pour espérance et variance :

$$\mathbb{E}[X] = \mu, \quad \mathbb{V}[X] = \sigma^2 \tag{1.25}$$



(a) Densité (b) Distribution

FIGURE 1.6 – Loi normale pour différentes valeurs des paramètres μ et σ^2.

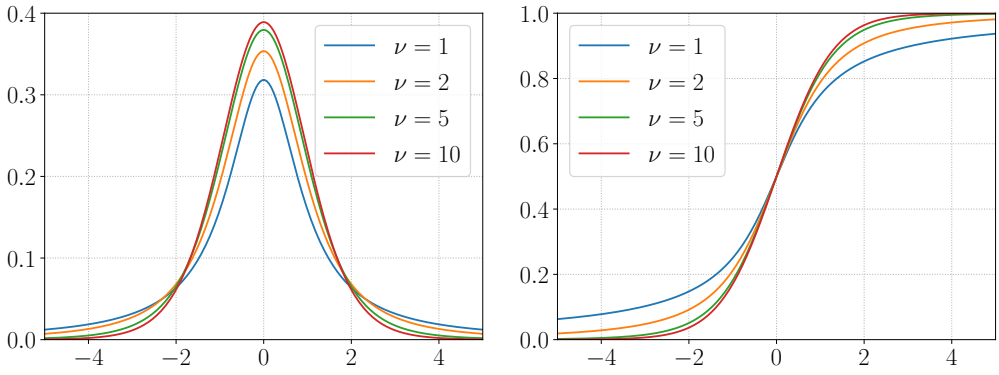
Loi de Student

Une variable aléatoire réelle X suit une distribution de Student (généralisée) de paramètres (ν, μ, σ) si elle a pour densité :

$$f(x) = \frac{1}{\sqrt{\pi\nu\sigma^2}B(\frac{1}{2}, \frac{\nu}{2})}\left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \tag{1.26}$$

Elle a pour espérance et variance :

$$\mathbb{E}[X] = \mu \text{ pour } \nu > 1, \quad \mathbb{V}[X] = \frac{\nu}{\nu-2}\sigma^2 \text{ pour } \nu > 2 \tag{1.27}$$



(a) Densité (b) Distribution

FIGURE 1.7 – Loi de Student généralisée pour différentes valeurs des paramètres ν, μ et σ.

Loi Beta(α, β)

Une variable aléatoire réelle X à valeurs dans $[0, 1]$ suivant une loi bêta de paramètres (α, β) a pour densité la fonction :

$$f(x) = \frac{1}{B(\alpha, \beta)}x^{\alpha-1}(1-x)^{\beta-1} \tag{1.28}$$

où B est la fonction bêta définie comme :

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \qquad (x, y \geq 0) \tag{1.29}$$

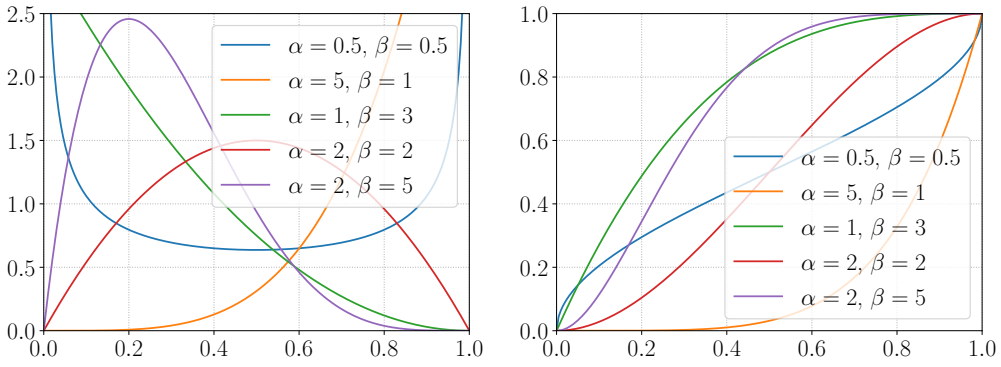
avec $\Gamma(x) = \int_0^{+\infty} \exp(-t)t^{x-1}\mathrm{d}t$, $x \geq 0$, la fonction Gamma. Elle a pour espérance et variance :

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad \mathbb{V}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{1.30}$$

Loi Inv-Gamma(α, β)

Une variable aléatoire réelle à valeurs dans $]0, +\infty[$ suit une loi inverse-gamma de paramètres (α, β) si elle a pour densité la fonction :

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{x}\right)^{\alpha+1}\exp\left(\frac{-\beta}{\alpha}\right). \tag{1.31}$$

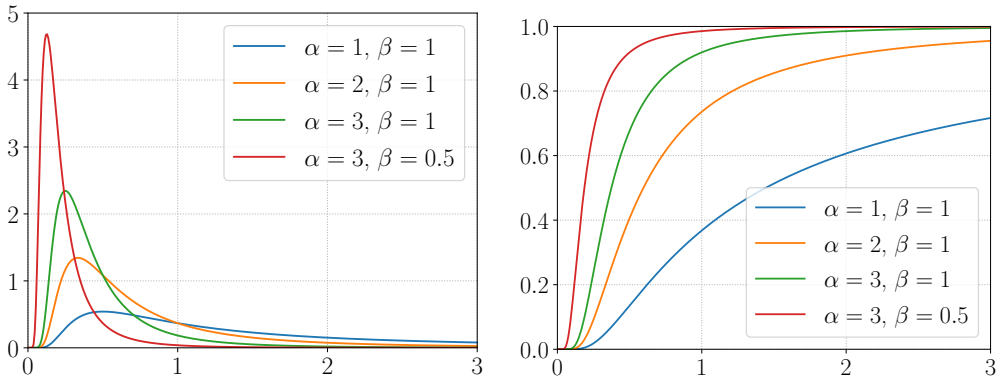


(a) Densité (b) Distribution

FIGURE 1.8 – Loi bêta pour différentes valeurs des paramètres α et β.

Elle a alors pour espérance et variance :

$$\mathbb{E}[X] = \frac{\beta}{\alpha - 1}, \quad \mathbb{V}[X] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \tag{1.32}$$



(a) Densité (b) Distribution

FIGURE 1.9 – Loi inverse-gamma pour différentes valeurs de α et β.

1.5 Échantillonnage d'une variable aléatoire

Afin de mener des expériences numériques, il est commun de vouloir simuler un échantillon provenant d'une variable aléatoire X distribuée selon une loi $\mathbb{P}_X$. Pour cela, plusieurs méthodes plus ou moins efficaces existent selon la distribution de X. Nous présentons ici la méthode d'échantillonnage par inversion de la fonction de répartition qui s'applique à n'importe quelle distribution et qui tire parti du théorème suivant :

Théorème 1.5.1. Soit X une variable aléatoire définie sur l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et soit F_X sa fonction de répartition.

1. Si F_X est continue, alors $F_X(X)$ est distribuée uniformément sur $[0, 1]$.
2. Si U est une variable aléatoire distribuée uniformément sur $[0, 1]$, alors $F_X^{-1}(U)$ a pour fonction de répartition F_X.

La transformation 2, permet donc d'obtenir une variable distribuée selon n'importe quelle distribution à partir d'une variable uniforme. Étant donné un échantillon $\{u[j]\}_{1\leq j\leq m}$ distribué uniformément sur $[0,1]$, obtenu par exemple avec un générateur de nombre pseudo-aléatoires (Matsumoto et al. 1998), l'échantillon $\left\{F_X^{-1}(u[i])\right\}_{1\leq i\leq m}$ est distribué selon $\mathbb{P}_X$. Lorsque la variable aléatoire X n'est pas continue, la méthode peut être appliquée en utilisant la fonction quantile de F_X à la place de sa fonction inverse :

Définition 1.5.1 (Fonction quantile)**.** La fonction *quantile* ou *inverse généralisée* d'une distribution $F_X : \overline{\mathbb{R}} \to [0,1]$ est la fonction F_X^{-1} définie sur $[0,1]$ et ayant pour expression

$$F^{-1}(y) = \inf\{x | F(x) \geq y\} = \sup\{x | F(x) \leq y\}.$$

Lorsque F_X est continue et strictement croissante, la fonction quantile correspond à la fonction inverse.

Bien que la méthode que nous venons de présenter soit applicable à n'importe quelle distribution, la fonction de répartition et la fonction quantile n'ont pas toujours d'expression analytique. Nous devons alors avoir recours à des méthodes numériques, pouvant être coûteuses, afin de les évaluer. C'est par exemple le cas de la loi gaussienne dont la fonction de répartition ne possède pas de forme analytique. Il existe cependant une méthode plus efficace appelée méthode de Box-Mueller (Box 1958) mais qui est spécifique à la distribution gaussienne.

1.6 Vecteurs aléatoires

Nous finissons ce chapitre en présentant le cas multidimensionnel qui sera le cadre de cette thèse. Nous introduisons pour la suite de la thèse la notation $\mathbb{I}$ pour le segment $[0,1]$ et pour une fonction F définie sur $\mathbb{R}$, nous noterons $\lim_{x\to+\infty} F(x) = F(+\infty)$ et $\lim_{x\to-\infty} F(x) = F(-\infty)$. La définition de variable aléatoire peut s'étendre à une fonction vectorielle qu'on appelle vecteur aléatoire :

Définition 1.6.1 (Vecteur aléatoire)**.** Un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ est une application vectorielle dont les composantes X_i sont des variables aléatoires définies sur un même espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. Ses composantes étant à valeurs dans $(\Omega_{X_i}, \mathcal{A}_{X_i})$, le vecteur aléatoire prend ses valeurs dans $(\Omega_{\mathbf{X}} = \bigtimes_{i=1}^n \Omega_{X_i}, \mathcal{A}_{\mathbf{X}} = \bigotimes_{i=1}^n \mathcal{A}_i)$ où $\mathcal{A}_{\mathbf{X}}$ est la tribu produit engendrée par les ensembles $\{A = \bigtimes_{i=1}^n A_i | \, A_i \in \mathcal{A}_i\}$. On appelle distribution *jointe* sa distribution $\mathbb{P}_{\mathbf{X}}$ et *distributions marginales* (unidimensionnelles) les distributions $\mathbb{P}_{X_i}$ de ses composantes. Un vecteur aléatoire sera qualifié de réel si $\mathbf{X}(\Omega) \subseteq \mathbb{R}^n$.

Si, comme nous allons le voir plus bas, la loi jointe nous permet de déterminer les lois marginales la réciproque est en général fausse.

1.6.1 Fonctions de répartition et densité

Nous pouvons étendre la définition de fonction de répartition au cas multivarié mais cela nécessite au préalable l'introduction du H-volume d'une fonction H :

Définition 1.6.2 (H-Volume)**.** Soit $\mathbf{a}$ et $\mathbf{b}$ deux points de $\overline{\mathbb{R}}^n$ tels que pour tout $i \in [\![1, n]\!]$, $a_i < b_i$. Le pavé de dimension n engendré par $(\boldsymbol{a}, \boldsymbol{b})$, noté $[\boldsymbol{a}, \boldsymbol{b}]$,

est l'ensemble défini comme le produit cartésien $\bigtimes_{i=1}^{d}[a_i, b_i]$. Étant donnée une fonction H définie sur un sous-ensemble de $\overline{\mathbb{R}}^n$ contenant $[\boldsymbol{a}, \boldsymbol{b}]$, le H-volume de $[\boldsymbol{a}, \boldsymbol{b}]$ est la quantité :

$$V_H([\boldsymbol{a}, \boldsymbol{b}]) = \sum_{\mathbf{v}\in\mathcal{V}} (-1)^{N(\mathbf{v})} H(\mathbf{v})$$

où $\mathcal{V} = \bigtimes_{i=1}^{n}\{a_i, b_i\}$ est l'ensemble des sommets du pavé et $N(\mathbf{v}) = \text{card}\,\{i|v_i = a_i\}$.

***Exemple* 1.6.1.** Soit $H : (u, v) \to uv$ une fonction définie sur $\mathbb{R}^2$ et soient $\mathbf{a} = (a_1, a_2)$ et $\mathbf{b} = (b_1, b_2)$ deux points de $\mathbb{R}^2$. Ils définissent le pavé $[a_1, b_1] \times [a_2, b_2]$ dont le H-volume est donné par

$$\begin{aligned} V_H([a_1, b_1] \times [a_2, b_2]) &= H(a_1, a_2) + H(b_1, b_2) - H(a_1, b_2) - H(b_1, a_2) \\ &= a_1a_2 + b_1b_2 - a_1b_2 - a_2b_1 \\ &= (a_1 - b_1)(a_2 - b_2), \end{aligned}$$

qui est l'aire du pavé. Plus généralement, le H-volume de la fonction $H : \boldsymbol{u} \to \prod_{i=1}^{d} u_i$ définie sur $\mathbb{R}^n$ rejoint la notion du volume d'un pavé dans un espace euclidien.

Dans le cas $d = 1$, le H-volume est donné par $V_H([a, b]) = H(b) - H(a)$ et la propriété $V_H([a, b]) \geq 0$ signifie alors que la fonction H est croissante. Pour $d > 1$, la positivité du H-volume peut donc être considéré comme une extension de la définition de croissance pour les fonctions multivariées. Nous donnons à présent une propriété que nous utiliserons plus tard dans le cadre de la théorie des copules.

Proposition 1.6.1. Si la dérivée partielle $h = \frac{\partial^n H}{\partial x_1 \ldots \partial x_n}$ existe, alors

$$V_H([\boldsymbol{a}, \boldsymbol{b}]) = \int_{[\boldsymbol{a},\boldsymbol{b}]} h(\mathbf{x})\mathrm{d}\mathbf{x} \tag{1.33}$$

En particulier, si $\forall \mathbf{x} \in \overline{\mathbb{R}}, h(\mathbf{x}) \geq 0$, alors $V_H([\boldsymbol{a}, \boldsymbol{b}]) \geq 0$.

Enfin, nous donnons la définition d'une fonction de répartition multivariée :

Définition 1.6.3 (Fonction de répartition multivariée)**.** La fonction de répartition $F_{\mathbf{X}} : \overline{\mathbb{R}}^n \to \mathbb{I}$ d'un vecteur aléatoire $\mathbf{X}$ défini sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ est donnée par :

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = \mathbb{P}_{\mathbf{X}}\left(X_1 \leq x_1, \ldots, X_n \leq x_n\right),$$

Indépendamment de toute distribution, une fonction H est une fonction de répartition si elle satisfait les propriétés suivantes :

1. H est continue à droite pour chacune de ses composantes,
2. H est n-croissante : pour chaque pavé $[\boldsymbol{a}, \boldsymbol{b}] \subseteq \overline{\mathbb{R}}^n$, $V_H([\boldsymbol{a}, \boldsymbol{b}]) \geq 0$,
3. $H(x_1, \ldots, x_n) = 0$ s'il existe i tel que $x_i = -\infty$,
4. $H(+\infty, \ldots, +\infty) = 1$.

La fonction de densité jointe $f_{\mathbf{X}}$, quant à elle, est définie comme pour le cas unidimensionnel (1.2), c'est-à-dire comme la dérivée de Radon-Nikodym de $\mathbb{P}_{\mathbf{X}}$ par rapport à une mesure de référence μ. Les densités f_{X_i} des composantes X_i sont appelées densités marginales unidimensionnelles. Tout comme pour la fonction de répartition, nous pouvons retrouver les densités marginales à partir de la densité jointe mais la réciproque n'est pas vraie en général. Enfin, le théorème 1.2.4 peut également être étendu au cas multidimensionnel :

Théorème 1.6.2. Soient $\mathcal{U}$ et $\mathcal{V}$ deux ouverts de $\mathbb{R}^n$ et $\phi : \mathcal{U} \to \mathcal{V}$ un difféomorphisme. Soit $\mathbf{X}$ un vecteur aléatoire défini sur l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ possédant une densité $f_{\mathbf{X}}$ et tel que $\mathbf{X}(\Omega) \subseteq \mathcal{U}$. Le vecteur aléatoire $\mathbf{Y} = \phi(\mathbf{X})$ à valeurs dans $\mathbf{Y}(\Omega) \subseteq \mathcal{V}$ a alors pour densité :

$$f_{\mathbf{Y}}(\mathbf{y}) = \left|\det(\boldsymbol{J}_{\phi^{-1}})(\mathbf{y})\right| f_{\mathbf{X}}\left(\phi^{-1}(\mathbf{y})\right) \tag{1.34}$$

1.6.2 Marginales

Étant donné un vecteur aléatoire $\mathbf{X}$, il peut arriver que nous ne soyons intéressés que par un sous-ensemble de ses composantes. À cet effet, nous pouvons déterminer la loi de n'importe quel sous-vecteur à partir de la loi jointe par l'opération dite de marginalisation. La loi ainsi obtenue est appelée la loi marginale du sous-vecteur et généralise la définition de marginale unidimensionnelle que nous avons vu plus haut.

Définition 1.6.4 (Distribution marginale)**.** Soit $\mathbf{X}$ un vecteur aléatoire à n dimensions, de distribution $\mathbb{P}_{\mathbf{X}}$ et de fonction de répartition $F_{\mathbf{X}}$. Soit $I = [\![1, n]\!]$ l'ensemble des indices et soit $\boldsymbol{j} = (j_1, \ldots, j_k)$ un k-uplet d'indices distincts avec $1 \leq k \leq n$. Le vecteur aléatoire marginal à k dimensions, $\mathbf{X}_{\boldsymbol{j}}$, est défini sur le même espace probabilisé que $\mathbf{X}$ et a pour fonction de répartition la fonction $F_{\mathbf{X}_{\boldsymbol{j}}} : \mathbb{R}^k \to \mathbb{I}$ définie par :

$$F_{\mathbf{X}_{\boldsymbol{j}}}(x_1, \ldots, x_k) = F_{\mathbf{X}}(y_1, \ldots, y_n), \tag{1.35}$$

où $y_i = x_i$ si $i \in \{j_1, \ldots, j_k\}$, et $y_i = +\infty$ sinon. Cette fonction est appelée la $\boldsymbol{j}$-marginale de la fonction de répartition $F_{\mathbf{X}}$.

La marginale unidimensionnelle de la i-ème composante est alors retrouvée avec la relation $F_{\mathbf{X}}(+\infty, \ldots, x_i, \ldots, +\infty)$. Sauf mention contraire, le terme marginale désignera par la suite les marginales unidimensionnelles. Si la distribution jointe possède une densité par rapport à une mesure μ, alors la relation 1.35 peut être réécrite en terme de fonctions densités :

Définition 1.6.5 (Densité marginale)**.** Soit $\mathbf{X}$ un vecteur aléatoire à n dimensions et soit $\mathbb{P}_{\mathbf{X}}$ sa distribution qui possèdent une densité jointe $f_{\mathbf{X}} = \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}}}{\mathrm{d}\mu}$ par rapport à une mesure μ. Soit $\mathbf{X}_{\boldsymbol{j}}$ une vecteur aléatoire marginal à k dimensions, sa densité $f_{\mathbf{X}_{\boldsymbol{j}}}$ est définie par :

$$f_{\mathbf{X}_{\boldsymbol{j}}}(x_1, \ldots, x_k) = \int_{\Omega_{\mathbf{X}_{I\setminus\boldsymbol{j}}}} f_{\mathbf{X}}(x_1, \ldots, x_n) \mathrm{d}\mathbf{x}_{I\setminus\boldsymbol{j}} \tag{1.36}$$

La densité $f_{\mathbf{X}_{\boldsymbol{j}}}$ est appelée la $\boldsymbol{j}$-marginale de la fonction densité jointe $f_{\mathbf{X}}$.

1.6.3 Moyenne et covariance

Nous n'étendons ici que la moyenne et la variance pour un vecteur aléatoire. Pour la moyenne, celle-ci n'est autre que le vecteur composé par la moyenne de chaque composante du vecteur aléatoire :

Définition 1.6.6 (Moyenne d'un vecteur aléatoire)**.** Soit $\mathbf{X}$ un vecteur aléatoire définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ tel que ses composantes X_i appartiennent toutes à $\mathcal{L}^1(\mathbb{P})$. La moyenne (ou espérance) d'un vecteur aléatoire est le vecteur défini par :

$$\mathbb{E}_{\mathbb{P}}[\mathbf{X}] = (\mathbb{E}_{\mathbb{P}}[X_1], \ldots, \mathbb{E}_{\mathbb{P}}[X_n]). \tag{1.37}$$

En revanche, l'extension de la variance au cas multidimensionnel, appelée covariance aboutit à une matrice :

Définition 1.6.7 (Covariance d'un vecteur aléatoire)**.** Soit X_i et X_j deux variables aléatoires définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et appartenant à $\mathcal{L}^2(\mathbb{P})$. La covariance entre ces deux variables aléatoires est donnée par :

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] \tag{1.38}$$

Pour un vecteur aléatoire $\mathbf{X}$, on peut alors définir une matrice symétrique $\boldsymbol{C}$ appelée matrice de covariance et dont les éléments sont données par $\boldsymbol{C}_{ij} = \mathrm{Cov}(X_i, X_j)$, avec X_i et X_j deux composantes de $\mathbf{X}$.

Remarquons que les élément diagonaux C_{ii} de la matrice de covariance sont les variances des variables X_i. Tout comme pour la variance dans le cas unidimensionnel, la matrice de covariance est importante puisqu'elle paramétrise, avec le vecteur moyenne, la gaussienne multivariée.

1.6.4 Lois classiques

Nous présentons ici l'extension multivariée de la loi gaussienne, de Student et bêta que nous avons présentées plus haut. Celles-ci seront utilisées à plusieurs reprises au cours de cette thèse.

1.6.4.1 Loi gaussienne

La densité de la loi gaussienne à n dimensions, de paramètres $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avec $\boldsymbol{\mu}$ un vecteur moyenne et $\boldsymbol{\Sigma}$ une matrice de covariance symétrique semi-définie positive, a pour expression :

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{-1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{1.39}$$

1.6.4.2 Loi de Student

La loi de Student (généralisée) à n dimensions de paramètres $(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, avec ν le nombre de degrés de liberté et $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ définis comme pour la distribution gaussienne, a pour densité la fonction :

$$f(\mathbf{x}) = \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\frac{\nu}{2})\nu^{n/2}\pi^{n/2}(\det \Sigma)^{1/2}} \left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\nu+n)/2} \tag{1.40}$$

1.6.4.3 Loi de Dirichlet

La loi de Dirichlet à n dimensions de paramètres $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n+1})$, avec $\alpha_i > 0$ pour tout $i \in [\![1, n+1]\!]$, est l'extension multidimensionnelle de la loi bêta. Sa fonction densité a pour expression :

$$f(\mathbf{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{n+1} x_i^{\alpha_i - 1} \mathbb{1}_{\Delta_n}(\mathbf{x}) \tag{1.41}$$

où B est la fonction bêta généralisée définie comme :

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{n+1} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{n+1} \alpha_i)} \tag{1.42}$$

et $\mathbf{x}$ appartient au n-simplexe $\Delta_n = \{\mathbf{x} \in \mathbb{R}_+^n | \sum_{i=1}^n x_i \leq 1, \forall i \in [\![1, n]\!]\}$.

1.6.5 Indépendance

Nous avons évoqué plus haut le fait que la loi jointe n'était, en général, pas déterminée par ses marginales. Pour que cela soit le cas, les composantes du vecteur aléatoire doivent être indépendantes entres elles :

Définition 1.6.8 (Indépendance entre variables aléatoires)**.** Soit $\{X_i\}_{1 \leq i \leq n}$ une famille de n variables aléatoires à valeurs dans $\{(\Omega_{X_i}, \mathcal{A}_{X_i})\}_{1 \leq i \leq n}$. On dit qu'elles sont mutuellement indépendantes si pour tout $(A_1, \ldots, A_n) \in \bigtimes_{i=1}^n \Omega_{X_i}$

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i) \tag{1.43}$$

Autrement dit, la loi jointe du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ s'écrit comme la mesure produit des lois marginales :

$$\mathbb{P}_{\mathbf{X}} = \bigotimes_{i=1}^n \mathbb{P}_{X_i}. \tag{1.44}$$

Dans le cas où la loi jointe possède une densité par rapport à une mesure, la factorisation de la loi jointe a pour conséquence la propriété suivante :

Proposition 1.6.3. Soit $\mathbf{X}$ un vecteur aléatoire et soit $\mathbb{P}_{\mathbf{X}}$ sa distribution. Si les composantes de $\mathbf{X}$ sont indépendantes et que $\mathbb{P}_{\mathbf{X}}$ possède une densité $f_{\mathbf{X}} = \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}}}{\mathrm{d}\mu}$ par rapport à une mesure μ alors

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^n f_{X_i}(x_i). \tag{1.45}$$

Nous verrons dans le chapitre relatif à la théorie des copules (chapitre 6), que la relation entre loi jointe et marginales peut être étendue au cas où les variables ne sont pas indépendantes via le théorème de Sklar.

1.6.6 Distribution et densité conditionnelle

La définition de distribution conditionnelle dans le cadre général de la théorie de la mesure fait appel à la notion de noyau de Markov [3] :

Définition 1.6.9 (Noyau de Markov)**.** Soit $(\Omega_1, \mathcal{A}_1)$ et $(\Omega_2, \mathcal{A}_2)$ deux espace probabilisables. Une application $\kappa : \Omega_1 \times \mathcal{A}_2 \to [0, 1]$ est appelée noyau de Markov si :

1. $\forall \omega_1 \in \Omega_1, \kappa(\omega_1, \cdot)$ est une mesure de probabilité sur $(\Omega_2, \mathcal{A}_2)$,
2. $\forall A_2 \in \mathcal{A}_2, \kappa(\cdot, A_2)$ est $\mathcal{A}_1$-mesurable.

***Exemple* 1.6.2.**

1. Soit $(E, \mathcal{T})$ un espace probabilisable et $\mathbb{P}_X$ une distribution sur $(\Omega_X, \mathcal{A}_X)$. L'application $\kappa : E \times \mathcal{A}_X \to [0, 1]$, définie par $\kappa(\cdot, A_X) = \mathbb{P}_X(A_X), A_X \in \mathcal{A}_X$, est un noyau de Markov.
2. La distribution $\kappa(x, \cdot) = N(0, x)$ est un noyau de Markov sur $]0, +\infty[\times \mathcal{B}(\mathbb{R})$.

Comme le montre le théorème suivant, la donnée d'un noyau de Markov et d'une distribution de probabilité permet de définir une distribution sur l'espace produit :

Théorème 1.6.4. Soit $(\Omega_1, \mathcal{A}_1)$ et $(\Omega_2, \mathcal{A}_2)$ deux espace probabilisables. Soit μ une mesure de probabilité sur $(\Omega_1, \mathcal{A}_1)$ et soit κ un noyau de Markov sur $\Omega_1 \times \mathcal{A}_2$. L'application $\mu \cdot \kappa$ définie sur $\mathcal{A}_1 \otimes \mathcal{A}_2$ par

$$\mu \cdot \kappa(A \times B) = \int_{A_1} \kappa(x, A_2) \mathrm{d}\mu(x), \tag{1.46}$$

est une mesure de probabilité [a] sur l'espace mesurable $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$.

a. Plus rigoureusement, l'application $\mu \cdot \kappa$ peut être prolongée de manière unique en une mesure de probabilité.

***Exemple* 1.6.3.** Pour le noyau $\kappa(\cdot, A_X) = \mathbb{P}_X(A_X)$ de l'exemple précédent (1.6.2), $\mu \cdot \kappa$ correspond à la mesure produit $\mu \otimes \mathbb{P}_X$.

Le dernier théorème nous permet de définir la loi conditionnelle comme un noyau de Markov reliant la loi jointe à une de ses marginales :

Définition 1.6.10 (Loi conditionnelle)**.** Soient X et Y deux variables aléatoires à valeurs dans $(\Omega_X, \mathcal{A}_X)$ et $(\Omega_Y, \mathcal{A}_Y)$. La loi (ou distribution) conditionnelle de Y sachant X est un noyau sur $\Omega_X \times \mathcal{A}_Y$, noté $\mathbb{P}_{Y|X}$ et tel que :

$$\mathbb{P}_{(X,Y)} = \mathbb{P}_X \cdot \mathbb{P}_{Y|X}. \tag{1.47}$$

***Exemple* 1.6.4.** En reprenant le noyau de l'exemple précédent (1.6.3) et en prenant $\mu = \mathbb{P}_Y$, la mesure $\mathbb{P}_Y \cdot \mathbb{P}_{X|Y}$ correspond à la mesure produit $\mathbb{P}_X \otimes \mathbb{P}_Y$ et les variables aléatoires sont indépendantes. De manière générale, lorsque X

3. Il existe une autre construction passant par la définition d'espérance conditionnelle. Voir par exemple le chapitre 6 de Kallenberg et al. 1997.

et Y sont indépendantes, ce noyau, appelé noyau constant, est une distribution conditionnelle.

Cette définition de la loi conditionnelle nous permet d'unifier au sein d'une même relation le lien qui existe entre la densité conditionnelle et la densité jointe pour les variables discrètes et continues :

Proposition 1.6.5. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé et soient $(\Omega_X, \mathcal{A}_X, \mu)$ et $(\Omega_Y, \mathcal{A}_Y, \nu)$ deux espaces de mesures finies. Soit $X : \Omega \to \Omega_X$ et $Y : \Omega \to \Omega_Y$ deux variables aléatoires et $\mathbb{P}_{(X,Y)}$ leur distribution jointe. Si $\mathbb{P}_{(X,Y)}$ est absolument continue par rapport à la mesure $\mu \otimes \nu$, ce que l'on note $\mathbb{P}_{(X,Y)} \ll \mu \otimes \nu$, il existe une densité jointe $f_{(X,Y)}$ et une densité marginale f_X. La distribution conditionnelle $\mathbb{P}_{Y|X}$, si elle existe, a pour densité par rapport à ν la fonction

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x,y)}{f_X(x)}. \tag{1.48}$$

Nous pouvons à présent introduire le théorème de Bayes qui est au centre de la statistique bayésienne que nous introduisons dans le chapitre 3.

Théorème 1.6.6 (Théorème de Bayes (Schervish 2012))**.** Soit X et Y deux variables aléatoires définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $(\Omega_X, \mathcal{A}_X)$ et $(\Omega_Y, \mathcal{A}_Y)$. Elles ont pour distributions marginales respectives $\mathbb{P}_X$ et $\mathbb{P}_Y$. Supposons que $\mathbb{P}_Y \ll \mu$ pour tout $y \in \Omega_Y$ et soit $f_{X|Y}$ la densité conditionnelle par rapport à μ de X sachant que $Y = y$. Soit $\mathbb{P}_{Y|X}$ la distribution conditionnelle de Y sachant que $X = x$. Alors $\mathbb{P}_{Y|X} \ll \mathbb{P}_Y$ et la dérivée de Radon-Nikodym est

$$\frac{\mathrm{d}\mathbb{P}_{Y|X}}{\mathrm{d}\mathbb{P}_Y}(y|x) = \frac{f_{X|Y}(x|y)}{\int_{\Omega_Y} f_{X|Y}(x|y)\mathrm{d}\mathbb{P}_Y(y)} \tag{1.49}$$

pour l'ensemble des x tels que le dénominateur n'est ni nul ni infini. Le cas contraire, la probabilité est nulle.

Étant donnée une mesure μ telle que $\mathbb{P}_{Y|X} \ll \mu$, nous retrouvons alors la formule classique :

$$f_{Y|X}(y|x) = \frac{\mathrm{d}\mathbb{P}_{Y|X}}{\mathrm{d}\mu}(y|x) = \frac{\mathrm{d}\mathbb{P}_{Y|X}}{\mathrm{d}\mathbb{P}_Y}(y|x)\frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mu}(y) = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{\Omega_Y} f_{X|Y}(x|y)\mathrm{d}\mathbb{P}_Y(y)}$$

Les définitions précédentes peuvent s'étendre au cas où $\mathbf{X}$ et $\mathbf{Y}$ sont deux vecteurs aléatoires. Par application successive du théorème 1.6.4, on peut alors en déduire la règle de chaine :

Proposition 1.6.7 (Règle de chaine)**.** Soit $\mathbf{X}$ un vecteur aléatoire et soit $\mathbb{P}_{\mathbf{X}}$ sa distribution. Celle-ci se factorise de la manière suivante :

$$\mathbb{P}_{\mathbf{X}} = \bigotimes_{i=1}^{n} \mathbb{P}_{X_i|X_1,\ldots,X_{i-1}}. \tag{1.50}$$

De même, si les densités conditionnelles existent, nous avons :

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i|X_1,\ldots,X_{i-1}}(x_i|x_1,\ldots,x_{i-1}). \tag{1.51}$$

Enfin une notion centrale pour le reste de la thèse est celle d'indépendance conditionnelle :

Définition 1.6.11 (Indépendance conditionnelle)**.** Soit $\boldsymbol{V}$ un vecteur aléatoire définit sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et soient $\mathbf{X} = (V_{i_1}, \dots, V_{i_p})$, $\mathbf{Y} = (V_{j_1}, \dots, V_{j_q})$ et $\mathbf{Z} = (V_{k_1}, \dots, V_{k_r})$ trois vecteurs extraits de $\boldsymbol{V}$ tels que les ensembles d'indices $\{i_1, \dots, i_p\}$, $\{j_1, \dots, j_q\}$ et $\{k_1, \dots, k_r\}$ sont deux à deux disjoints. $\mathbf{X}$ et $\mathbf{Y}$ sont indépendants conditionnellement à $\mathbf{Z}$, ce que l'on note $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}$, si la distribution conditionnelle $\mathbb{P}_{(\mathbf{X},\mathbf{Y})|\mathbf{Z}}$ s'écrit comme le produit des distributions conditionnelles $\mathbb{P}_{\mathbf{X}|\mathbf{Z}}$ et $\mathbb{P}_{\mathbf{Y}|\mathbf{Z}}$:

$$\mathbb{P}_{(\mathbf{X},\mathbf{Y})|\mathbf{Z}} = \mathbb{P}_{\mathbf{X}|\mathbf{Z}} \otimes \mathbb{P}_{\mathbf{Y}|\mathbf{Z}}. \tag{1.52}$$

De manière équivalente, nous avons dans ce cas :

$$\mathbb{P}_{\mathbf{X}|(\mathbf{Y},\mathbf{Z})} = \mathbb{P}_{\mathbf{X}|\mathbf{Z}} \quad \text{et} \quad \mathbb{P}_{\mathbf{Y}|(\mathbf{X},\mathbf{Z})} = \mathbb{P}_{\mathbf{Y}|\mathbf{Z}}. \tag{1.53}$$

Dans ce cadre, la notion d'indépendance que nous avons vu plus haut est appelée indépendance marginale et correspond par convention au cas où $\mathbf{Z} = \emptyset$. Enfin, les relations de la définition précédente s'étendent aux densités conditionnelles lorsqu'elles existent :

$$f_{(\mathbf{X},\mathbf{Y})|\mathbf{Z}}(\mathbf{x}, \mathbf{y}|\mathbf{z}) = f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}, \mathbf{z}) f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}, \mathbf{z}), \qquad f_{\mathbf{X}|(\mathbf{Y},\mathbf{Z})}(\mathbf{x}|\mathbf{y}, \mathbf{z}) = f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z}). \tag{1.54}$$

Nous avons présenté dans ce chapitre les principales définitions et propriétés de la théorie des probabilités qui vont être utilisées au cours de cette thèse. Un point important pour la suite est celui de la définition de densité relativement à une mesure qui unifie les concepts de fonction de masse et de fonction densité au sein d'un même objet. Toutefois, le lecteur peu familier avec la théorie de la mesure pourra retrouver les définitions habituelles en remplaçant le terme de densité par fonction de masse lorsque les variables aléatoires sont discrètes. Nous allons à présent étendre ce cadre à la théorie de l'information qui jouera un rôle important lorsque nous nous intéresserons à l'apprentissage des réseaux bayésiens dans le chapitre 5.

Références

Box, G. E. P. (1958). « A note on the generation of random normal deviates ». In : *Ann. Math. Statist.* 29, p. 610-611 (cf. p. 22).

Candelpergher, B. (2013). *Théorie des probabilités* (cf. p. 11, 13, 14).

Grimmett, G. et Stirzaker, D. (2020). *Probability and random processes*. Oxford university press (cf. p. 11).

Kallenberg, O. et Kallenberg, O. (1997). *Foundations of modern probability*. T. 2. Springer (cf. p. 27).

Kolmogorov, A. N. et Bharucha-Reid, A. T. (2018). *Foundations of the theory of probability : Second English Edition*. Courier Dover Publications (cf. p. 11).

Matsumoto, M. et Nishimura, T. (1998). « Mersenne twister : a 623-dimensionally equidistributed uniform pseudo-random number generator ». In : *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 8.1, p. 3-30 (cf. p. 22).

Ouvrard, J.-Y. (2004). « Probabilités : Tome II ». In : *Master-Agrégation, Cassini* (cf. p. 11).

Rosenthal, J. S. (2006). *First Look At Rigorous Probability Theory, A*. World Scientific Publishing Company (cf. p. 11).

Schervish, M. J. (2012). *Theory of statistics*. Springer Science & Business Media (cf. p. 28).

Williams, D. (1991). *Probability with martingales*. Cambridge university press (cf. p. 11).

Chapitre 2

Théorie de l'information

Sommaire

La théorie de l'information est à l'interface de plusieurs domaines scientifiques : la physique, les télécommunications, la statistique, etc. Elle définit plusieurs objets permettant de quantifier l'incertitude d'une distribution au travers d'une valeur scalaire. Nous rappelons ici les concepts de base de théorie de l'information et portons une attention particulière à l'entropie et l'information mutuelle qui seront plus tard au centre de nos méthodes d'apprentissage des réseaux bayésiens. Pour une approche plus exhaustive du sujet, nous renvoyons le lecteur à Gray (2011) et Cover et Thomas (2012).

2.1 L'entropie

La fonction d'entropie pour une distribution discrète a été introduite dans le but de quantifier l'incertitude de celle-ci. Pour cela, il a été axiomatisé que cette fonction, notée H, doit vérifier les propriétés suivantes :

1. H est maximale lorsque X est distribuée uniformément,
2. Soit k le nombre de valeurs que peut prendre la variable aléatoire X et soit X' la variable aléatoire prenant $k+1$ valeurs telle que $\mathbb{P}(X' = x'_i) = \mathbb{P}(X = x_i)$ pour $1 \leq i \leq k$ et $\mathbb{P}(X' = x'_{k+1}) = 0$. Dans ce cas, X et X' ont la même entropie : $H(X') = H(X)$.
3. Soit $H(X, Y)$ l'entropie de la distribution jointe $\mathbb{P}_{(X,Y)}$ et soit $H_{X=x}(Y)$ l'entropie de la distribution conditionnelle $\mathbb{P}_{(Y|X=x)}$, alors :

$$H(X, Y) = H(X) + \sum_{x \in \Omega_X} \mathbb{P}(X = x) H_{X=x}(Y) \tag{2.1}$$

Il a été démontré (Khinchin 1957) que la seule fonction (à une constante positive multiplicative près) vérifiant ces propriétés était la fonction définie par :

$$H(X) = -\sum_{\mathbf{x}\in\Omega_{\mathbf{X}}} f_{\mathbf{X}}(\mathbf{x}) \log f_{\mathbf{X}}(\mathbf{x}).$$

Cependant, cette définition ne s'étend pas facilement au cas continu. En effet, l'extension classique appelée entropie différentielle, n'est pas invariante d'échelle. Pour résoudre ce problème, Jaynes (1963) a défini l'entropie pour une distribution relativement à une mesure. Cette version de l'entropie, appelée entropie générale ou encore entropie de Gibbs-Jaynes, est ici introduite aux côtés des entropies relative et croisée qui en sont des variations. Ces dernières, nous le verrons plus tard, interviennent dans l'expression de la fonction de vraisemblance d'un échantillon de données.

2.1.1 Entropie générale

Définition 2.1.1 (Entropie générale)**.** Soit $\mathbf{X}$ un vecteur aléatoire de distribution $\mathbb{P}_{\mathbf{X}}$ et soit ρ une mesure définie sur $\Omega_{\mathbf{X}}$ telle que $\mathbb{P}_{\mathbf{X}} \ll \rho$. L'entropie de $\mathbf{X}$ par rapport à la mesure ρ, notée H_ρ, est donnée par

$$H_\rho(\mathbb{P}_{\mathbf{X}}) = -\mathbb{E}_{\mathbb{P}_{\mathbf{X}}}\left[\log \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}}}{\mathrm{d}\rho}\right] \tag{2.2}$$

où $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}$ est la dérivée de Radon-Nikodym de $\mathbb{P}$ par rapport à ρ. Lorsqu'il n'y a pas de confusion possible sur la distribution, nous noterons l'entropie $H(\mathbf{X})$.

Dans le cas où les composantes du vecteur aléatoire $\mathbf{X}$ sont discrètes, nous retrouvons l'entropie de Shannon en utilisant la mesure de comptage comme mesure de référence :

$$H_\gamma(\mathbf{X}) = -\int_{\Omega_{\mathbf{X}}} \log \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}}}{\mathrm{d}\gamma} \mathrm{d}\mathbb{P}_{\mathbf{X}} = -\int_{\Omega_{\mathbf{X}}} \frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\gamma} \log \frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\gamma} \mathrm{d}\gamma = -\sum_{\mathbf{x}\in\Omega_{\mathbf{X}}} f_{\mathbf{X}}(\mathbf{x}) \log f_{\mathbf{X}}(\mathbf{x})$$

avec $f_{\mathbf{X}} = \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}}}{\mathrm{d}\gamma}$ la densité de $\mathbb{P}_{\mathbf{X}}$ par rapport à la mesure de comptage, c'est-à-dire la fonction de masse de $\mathbf{X}$. Par prolongement en 0, la convention $0 \log 0 = 0$ est adoptée.

***Exemple* 2.1.1.** Soit une variable aléatoire X à valeurs dans $\Omega_X = \{x_i\}_{1\leq i\leq r}$. Si le résultat de cette variable aléatoire est certain, c'est-à-dire si sa densité est un delta de Kronecker δ défini comme :

$$f(x) = \delta(x, x_j) = \begin{cases} 1 & \text{si } x = x_j,\ j \in [\![1, r]\!] \\ 0 & \text{sinon} \end{cases}, \tag{2.3}$$

alors son entropie vaut :

$$H_\gamma(X) = -\sum_{x\in\Omega_X} \delta(x, x_j) \log \delta(x, x_j) = 1 \cdot \log 1 = 0. \tag{2.4}$$

En revanche, si la distribution de X est uniforme, c'est-à-dire si :

$$f(x) = \frac{1}{r} \quad \forall x \in \Omega_X, \tag{2.5}$$

son entropie vaut :

$$H_\gamma(X) = -\sum_{x\in\Omega_X} \frac{1}{r} \log \frac{1}{r} = \log r \tag{2.6}$$

et augmente donc avec la taille de l'ensemble Ω_X. En résumé, lorsque le résultat est certain l'entropie est nulle et lorsque le résultat est uniformément aléatoire, l'entropie augmente avec le nombre de valeurs possibles que peut prendre la variable : plus le nombre de résultats possibles est grand, plus l'incertitude est grande. Cette observation rejoint bien l'interprétation de l'entropie de Shannon comme une mesure d'incertitude.

Cette interprétation ne peut pas être étendue à l'entropie générale puisque celle-ci n'est pas forcément positive. C'est également le cas de l'entropie différentielle d'un vecteur aléatoire $\mathbf{X}$ réel, qui en dérive en utilisant la mesure de Lebesgue comme mesure de référence :

$$H_\lambda(\mathbf{X}) = -\int_{\Omega_\mathbf{X}} \log \frac{\mathrm{d}\mathbb{P}_\mathbf{X}}{\mathrm{d}\lambda} \mathrm{d}\mathbb{P}_\mathbf{X} = -\int_{\Omega_\mathbf{X}} \frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\lambda} \log \frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\lambda} \mathrm{d}\lambda = -\int_{\Omega_\mathbf{X}} f(\mathbf{x}) \log f(\mathbf{x}) \mathrm{d}\mathbf{x}$$

où $f_\mathbf{X} = \frac{\mathrm{d}\mathbb{P}_\mathbf{X}}{\mathrm{d}\gamma}$ est la densité de $\mathbb{P}_\mathbf{X}$ par rapport à la mesure de Lebesgue. Nous pouvons vérifier aisément que l'entropie différentielle n'est pas invariante d'échelle puisque étant donnée une variable aléatoire $Y = aX$, où a est le facteur d'échelle, nous avons $H_\lambda(\mathbf{Y}) = H_\lambda(\mathbf{X}) + \log|a|$.

2.1.2 Entropie conditionnelle

Définition 2.1.2 (Entropie conditionnelle)**.** Soit $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ un vecteur aléatoire et soient $\mathbb{P}_\mathbf{X}$ sa distribution jointe et $\mathbb{P}_{\mathbf{X}_1}$ et $\mathbb{P}_{\mathbf{X}_2}$ ses marginales. L'entropie conditionnelle de $\mathbf{X}_2$ sachant $\mathbf{X}_1$ relativement à une mesure ρ est donnée par :

$$H_\rho(\mathbf{X}_2|\mathbf{X}_1) = -\mathbb{E}_{\mathbb{P}_\mathbf{X}}\left[\log \mathbb{P}_{\mathbf{X}_2|\mathbf{X}_1}\right] \tag{2.7}$$

Si les deux variables sont indépendantes, on a :

$$H_\rho(\mathbf{X}_2|\mathbf{X}_1) = H_\rho(\mathbf{X}_2), \tag{2.8}$$

Dans le cas de l'entropie de Shannon, l'entropie conditionnelle nous renseigne sur l'incertitude du vecteur aléatoire $\mathbf{X}_2$ étant donnée la réalisation du vecteur aléatoire $\mathbf{X}_1$. L'équation (2.8) est en accord avec cette interprétation puisque dans ce cas le fait de connaître l'issue $\mathbf{X}_1$ ne nous renseigne aucunement sur l'issue de $\mathbf{X}_2$. L'entropie conditionnelle nous permet de définir une règle de chaine pour l'entropie :

Proposition 2.1.1 (Règle de chaine pour l'entropie)**.** L'entropie d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ vérifie la propriété de chaine suivante :

$$H_\rho(\mathbf{X}) = \sum_{i=1}^{n} H_\rho(X_i|X_1, \ldots, X_{i-1}) \tag{2.9}$$

avec par définition, $H_\rho(X_1|\emptyset) = H_\rho(X_1)$.

Démonstration. À partir de la règle de chaine pour les probabilités et de la linéarité de

l'espérance, nous avons :

$$H_\rho(\mathbf{X}) = -\mathbb{E}_{\mathbb{P}_\mathbf{X}}\left[\log \frac{\mathrm{d}\mathbb{P}_\mathbf{X}}{\mathrm{d}\rho}\right] = -\mathbb{E}_{\mathbb{P}_\mathbf{X}}\left[\sum_{i=1}^{n} \log \frac{\mathrm{d}\mathbb{P}_{X_i|X_1,\dots,X_{i-1}}}{\mathrm{d}\rho}\right]$$
$$= -\sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}_\mathbf{X}}\left[\log \frac{\mathrm{d}\mathbb{P}_{X_i|X_1,\dots,X_{i-1}}}{\mathrm{d}\rho}\right] = \sum_{i=1}^{n} H_\rho(X_i|X_1,\dots,X_{i-1}).$$

■

2.1.3 Entropie relative

Bien que l'entropie de Gibbs-Jaynes nous permette d'unifier les notations pour le cas discret et le cas continu, nous avons vu que pour la mesure de Lebesgue, celle-ci n'était pas invariante d'échelle. De plus, elle n'est pas positive et ne permet pas la même interprétation en terme de mesure d'incertitude comme dans le cas discret. La solution à ces deux problèmes et alors d'utiliser l'entropie de Gibs-Jaynes relativement à une mesure de probabilité. Cela nous mène naturellement vers la définition d'entropie relative ou *divergence de Kullback-Liebler*, qui est utilisée par Gray (2011) afin d'unifier les cas discrets et continus :

Définition 2.1.3 (Entropie relative)**.** Soit $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilité sur un même espace probabilisable $(\Omega, \mathcal{A})$, telles que $\mathbb{P} \ll \mathbb{Q}$. L'entropie relative entre $\mathbb{P}$ et $\mathbb{Q}$ est définie par :

$$D\left(\mathbb{P}||\mathbb{Q}\right) = \mathbb{E}_{\mathbb{P}}\left[\log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right] \tag{2.10}$$

Si $\mathbb{P}$ n'est pas absolument continue par rapport à $\mathbb{Q}$ alors $D\left(\mathbb{P}||\mathbb{Q}\right) = +\infty$.

Soient $\mathbb{P}$ et $\mathbb{Q}$ deux distributions sur $(\Omega_\mathbf{X}, \mathcal{A}_\mathbf{X})$ et soit μ une mesure telle que $\mathbb{P} \ll \mu \ll \mathbb{Q}$. On peut alors écrire l'entropie relative comme :

$$D\left(\mathbb{P}_\mathbf{X}||\mathbb{Q}_\mathbf{X}\right) = \int_{\Omega_\mathbf{X}} \log \frac{\mathrm{d}\mathbb{P}_\mathbf{X}}{\mathrm{d}\mathbb{Q}_\mathbf{X}} \mathrm{d}\mathbb{P}_\mathbf{X} = \int_{\Omega_\mathbf{X}} \frac{\mathrm{d}\mathbb{P}_\mathbf{X}}{\mathrm{d}\mu} \log \frac{\mathrm{d}\mathbb{P}_\mathbf{X}}{\mathrm{d}\mu} \frac{\mathrm{d}\mu}{\mathrm{d}\mathbb{Q}_\mathbf{X}} \mathrm{d}\mu = \int_{\Omega_\mathbf{X}} f \log \frac{f}{g} \mathrm{d}\mu$$

où f et g sont les dérivées de $\mathbb{P}_\mathbf{X}$ et $\mathbb{Q}_\mathbf{X}$ par rapport à la mesure μ. En utilisant la mesure de comptage dans le cas où la variable X est discrète et la mesure de Lebesgue dans le cas où elle est continue, nous avons :

$$D\left(\mathbb{P}_\mathbf{X}||\mathbb{Q}_\mathbf{X}\right) = \begin{cases} \sum\limits_{\mathbf{x}\in\Omega_\mathbf{X}} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} & \text{Cas discret,} \\ \int_{\Omega_\mathbf{X}} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} \mathrm{d}\mathbf{x} & \text{Cas continu.} \end{cases} \tag{2.11}$$

Notons que dans les deux cas, nous utilisons les conventions $0 \log \frac{0}{0} = 0$ et $0 \log \frac{0}{q} = 0$ par continuité en 0. Bien que l'entropie générale et l'entropie relative semblent être égales à un signe négatif près, nous insistons sur le fait que l'entropie relative est contrainte à ce que la mesure de référence soit une mesure de probabilité. Cette remarque est importante puisque c'est cette propriété qui confère à l'entropie relative sa positivité et son invariance d'échelle :

Proposition 2.1.2. Soit $(\Omega, \mathcal{A})$ un espace probabilisable. Pour toutes mesures de probabilité $\mathbb{P}$ et $\mathbb{Q}$ définies sur cet espace, alors

$$D\left(\mathbb{P}||\mathbb{Q}\right) \geq 0, \tag{2.12}$$

avec égalité si et seulement si $\mathbb{P} = \mathbb{Q}$ presque partout.

Démonstration. D'après l'inégalité de Jensen, $\mathbb{E}_{\mathbb{P}}\left[-\log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right] \geq -\log \mathbb{E}_{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right] = -\log 1 = 0$. La deuxième partie découle de la condition d'égalité pour l'inégalité de Jensen. ■

Contrairement à l'entropie générale, l'entropie relative est invariante par changement de variable comme le montre la proposition suivante :

Proposition 2.1.3. Soit $\mathbf{X}$ un vecteur aléatoire défini sur un ensemble $\Omega_{\mathbf{X}}$ et soient deux distributions $\mathbb{P}_{\mathbf{X}}$ et $\mathbb{Q}_{\mathbf{X}}$. Si $\mathbf{Y} = \phi(\mathbf{X})$ où $\phi : \Omega_{\mathbf{X}} \to \Omega_{\mathbf{Y}}$ est une fonction bijective différentiable alors :

$$D\left(\mathbb{P}_{\mathbf{X}} || \mathbb{Q}_{\mathbf{X}}\right) = D\left(\mathbb{P}_{\mathbf{Y}} || \mathbb{Q}_{\mathbf{Y}}\right) \tag{2.13}$$

Démonstration. Soient $f_{\mathbf{X}}$ et $g_{\mathbf{X}}$ les densités de $\mathbb{P}_{\mathbf{X}}$ et $\mathbb{Q}_{\mathbf{X}}$. En faisant le changement de variable $y = \phi(x)$ et en utilisant le théorème 1.6.2, nous avons

$$\begin{aligned} D\left(\mathbb{P}_{\mathbf{X}} || \mathbb{Q}_{\mathbf{X}}\right) &= \int_{\Omega_{\mathbf{X}}} f_{\mathbf{X}}(\mathbf{x}) \log\left(\frac{f_{\mathbf{X}}(\mathbf{x})}{g_{\mathbf{X}}(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \\ &= \int_{\Omega_{\mathbf{Y}}} f_{\mathbf{X}}(\phi^{-1}(\mathbf{y})) \log\left(\frac{f_{\mathbf{X}}(\phi^{-1}(\mathbf{y}))}{g_{\mathbf{X}}(\phi^{-1}(\mathbf{y}))} \times \frac{\left|\det \boldsymbol{J}_{\phi^{-1}}\right|}{\left|\det \boldsymbol{J}_{\phi^{-1}}\right|}\right) \left|\det \boldsymbol{J}_{\phi^{-1}}\right| \mathrm{d}\mathbf{y} \\ &= \int_{\Omega_{\mathbf{Y}}} f_{\mathbf{Y}}(\mathbf{y}) \log\left(\frac{f_{\mathbf{Y}}(\mathbf{y})}{g_{\mathbf{Y}}(\mathbf{y})}\right) \mathrm{d}\mathbf{y} = D\left(\mathbb{P}_{\mathbf{Y}} || \mathbb{Q}_{\mathbf{Y}}\right) \end{aligned}$$

■

De par sa positivité, l'entropie relative est souvent considérée comme une distance sur l'ensemble des distributions bien qu'elle ne vérifie ni la condition de symétrie, ni l'inégalité triangulaire. De la même manière que pour l'entropie, une entropie relative conditionnelle et une règle de chaine peuvent être définies :

Définition 2.1.4 (Entropie relative conditionnelle)**.** Soit $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ un vecteur aléatoire et soient $\mathbb{P}_{\mathbf{X}} << \mathbb{Q}_{\mathbf{X}}$ deux distributions jointes définies sur un même espace mesurable. L'entropie relative conditionnelle est donnée par :

$$D\left(\mathbb{P}_{\mathbf{X}_2|\mathbf{X}_1} || \mathbb{Q}_{\mathbf{X}_2|\mathbf{X}_1} | \mathbb{P}_{\mathbf{X}_1}\right) = \mathbb{E}_{\mathbb{P}_{\mathbf{X}_1}}\left[D\left(\mathbb{P}_{\mathbf{X}_2|\mathbf{X}_1} || \mathbb{Q}_{\mathbf{X}_2|\mathbf{X}_1}\right)\right] \tag{2.14}$$

Proposition 2.1.4 (Règle de chaine pour l'entropie relative)**.** L'entropie relative d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ vérifie la propriété de chaine suivante :

$$D\left(\mathbb{P}_{\mathbf{X}} || \mathbb{Q}_{\mathbf{X}}\right) = \sum_{i=1}^{n} D\left(\mathbb{P}_{X_i|X_1,\ldots,X_{i-1}} || \mathbb{Q}_{X_i|X_1,\ldots,X_{i-1}} | \mathbb{P}_{X_1,\ldots,X_{i-1}}\right) \tag{2.15}$$

Contrairement au choix de la mesure de comptage ou de Lebesgue qui apparaît naturel pour l'entropie d'une variable discrète ou continue, le choix d'une mesure de probabilité peut paraître arbitraire. Cependant, comme le montre la propriété suivante, l'entropie de Shannon est implicitement définie comme l'entropie relative par rapport à une distribution uniforme :

Proposition 2.1.5. Soit $\mathbf{X}$ un vecteur aléatoire sur un domaine Ω_X discret et soit $\mathbb{P}_{\mathbf{X}}$ sa distribution. Son entropie vérifie la relation suivante :

$$H_\gamma(\mathbf{X}) = \log|\Omega_{\mathbf{X}}| - D(\mathbb{P}_{\mathbf{X}}||U) \tag{2.16}$$

où U est la distribution uniforme sur Ω_X.

Démonstration. Soient $f_{\mathbf{x}}$ et u les densité de $\mathbb{P}_{\mathbf{X}}$ et U par rapport à la mesure de comptage. L'entropie relative entre $\mathbb{P}_{\mathbf{X}}$ et U a pour expression :

$$D(\mathbb{P}_{\mathbf{X}}||U) = \sum_{\mathbf{x}\in\Omega_{\mathbf{X}}} f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{u(\mathbf{x})}\right) = \sum_{\mathbf{x}\in\Omega_{\mathbf{X}}} f(\mathbf{x}) \log f(\mathbf{x}) - \log\frac{1}{|\Omega_X|}$$

■

Ceci explique alors le statut particulier du cas discret qui ne peut être étendu au cas continu que lorsque le domaine de la variable aléatoire est fini. Dans ce cas, nous avons $H_\lambda(\mathbf{X}) = \log\lambda(\Omega_{\mathbf{X}}) - D(\mathbb{P}_{\mathbf{X}}||U)$. D'un point de vue qualitatif, il paraît satisfaisant que l'incertitude d'une distribution soit relative à une distribution de référence. Dans le cadre de l'inférence bayésienne, introduite dans le chapitre 3, nous pouvons faire le parallèle entre la distribution de référence par rapport à laquelle l'incertitude est mesurée et la distribution *a priori*. L'entropie différentielle définie en utilisant la mesure de Lebesgue comme référence fait alors figure d'un *a priori* impropre.

2.1.4 Entropie croisée

Définition 2.1.5 (Entropie croisée)**.** Soient $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilité définies sur un même espace probabilisable $(\Omega, \mathcal{A})$, et soit ρ une mesure telle que $\mathbb{Q} \ll \rho$. L'entropie croisée (*cross-entropy* en anglais) entre $\mathbb{P}$ et $\mathbb{Q}$ relativement à ρ a pour expression :

$$H_\rho(\mathbb{P}||\mathbb{Q}) = -\mathbb{E}_{\mathbb{P}}\left[\log\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\rho}\right] \tag{2.17}$$

Elle va jouer un rôle important dans le cadre de l'estimation statistique puisque, comme nous le verrons plus tard, l'entropie croisée entre le modèle et la distribution empirique est exactement égale à la log-vraisemblance négative [1]. On montre facilement à partir des différentes définitions introduites plus haut que l'entropie croisée vérifie la propriété suivante :

Proposition 2.1.6. Soient $\mathbb{P}$ et $\mathbb{Q}$ deux distributions définies sur un même espace probabilisable et soit ρ une mesure telle que $\mathbb{Q} \ll \rho \ll \mathbb{P}$. L'entropie croisée vérifie alors la relation suivante :

$$H_\rho(\mathbb{P}||\mathbb{Q}) = H_\rho(\mathbb{P}) + D(\mathbb{P}||\mathbb{Q}) \tag{2.18}$$

1. Pour plus de détails sur le lien entre théorie de l'information et l'estimation statistique, voir Kulhavỳ (1996).

2.2 L'information mutuelle

Définition 2.2.1 (Information mutuelle)**.** Soit $\mathbf{X}$ un vecteur aléatoire de distribution jointe $\mathbb{P}_{\mathbf{X}}$ et soient $\mathbf{X}_1$ et $\mathbf{X}_2$ deux sous-vecteurs tels que $\mathbf{X}_1 \cap \mathbf{X}_2 = \emptyset$ et dont les distributions marginales sont $\mathbb{P}_{\mathbf{X}_1}$ et $\mathbb{P}_{\mathbf{X}_2}$. L'*information mutuelle* de $\mathbf{X}_1$ et $\mathbf{X}_2$ est définie par :

$$I(\mathbf{X}_1;\mathbf{X}_2) = D(\mathbb{P}_{\mathbf{X}}||\mathbb{P}_{\mathbf{X}_1}\mathbb{P}_{\mathbf{X}_2}) \tag{2.19}$$

L'information mutuelle mesure l'information commune entre deux vecteurs aléatoires $\mathbf{X}_1$ et $\mathbf{X}_2$ ou, autrement dit, leur dépendance. En effet, il est aisé de voir qu'elle est nulle si et seulement si les deux variables aléatoires sont indépendantes.

D'après la relation (2.19) et la proposition 2.1.2, l'information mutuelle est symétrique et positive. De plus, la règle de chaîne permet de relier l'information mutuelle à l'entropie :

Proposition 2.2.1. Soient $\mathbf{X}_1$ et $\mathbf{X}_2$ deux vecteurs aléatoires et soit ρ une mesure. Leur information mutuelle vérifie la relation suivante :

$$I(\mathbf{X}_1;\mathbf{X}_2) = H_\rho(\mathbf{X}_1) - H_\rho(\mathbf{X}_1|\mathbf{X}_2) = H_\rho(\mathbf{X}_1) + H_\rho(\mathbf{X}_2) - H_\rho(\mathbf{X}_1,\mathbf{X}_2) \tag{2.20}$$

Démonstration.

$$\begin{aligned} I(\mathbf{X}_1;\mathbf{X}_2) &= D(\mathbb{P}_{\mathbf{X}}||\mathbb{P}_{\mathbf{X}_1}\mathbb{P}_{\mathbf{X}_2}) = D(\mathbb{P}_{\mathbf{X}_2}||\mathbb{P}_{\mathbf{X}_2}) + D\left(\mathbb{P}_{\mathbf{X}_1|\mathbf{X}_2}||\mathbb{P}_{\mathbf{X}_1}\right) \\ &= \mathbb{E}_{\mathbb{P}_{\mathbf{X}}}\left[\log \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}_1|\mathbf{X}_2}}{\mathrm{d}\rho}\right] - \mathbb{E}_{\mathbb{P}_{\mathbf{X}}}\left[\log \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}_1}}{\mathrm{d}\rho}\right] \\ &= H_\rho(\mathbf{X}_1) - H_\rho(\mathbf{X}_1|\mathbf{X}_2) \end{aligned}$$

La deuxième relation est obtenue en utilisant la règle de chaîne pour l'entropie. ■

Cette dernière relation montre que l'information mutuelle représente le gain en information sur un vecteur aléatoire $\mathbf{X}_1$ étant donnée l'observation d'un autre vecteur aléatoire $\mathbf{X}_2$. De par la relation 2.8, nous pouvons voir de nouveau que cette quantité est nulle si et seulement si les variables aléatoires sont indépendantes.

2.2.1 Information conditionnelle

Comme pour l'entropie, on peut définir une information mutuelle conditionnelle :

Définition 2.2.2 (Information conditionnelle)**.** Soit $\mathbf{X}$ un vecteur aléatoire de distribution $\mathbb{P}_{\mathbf{X}}$ et soient $\mathbf{X}_1$, $\mathbf{X}_2$ et $\mathbf{U}$ trois sous-vecteurs disjoints. L'information mutuelle entre $\mathbf{X}_1$ et $\mathbf{X}_2$ conditionnellement à $\mathbf{U}$ est définie par :

$$I(\mathbf{X}_1;\mathbf{X}_2|\mathbf{U}) = D\left(\mathbb{P}_{\mathbf{X}_1,\mathbf{X}_2|\mathbf{U}}||\mathbb{P}_{\mathbf{X}_1|\mathbf{U}}\mathbb{P}_{\mathbf{X}_2|\mathbf{U}}\right). \tag{2.21}$$

De la même manière que pour la proposition 2.2.1, nous pouvons montrer que l'information mutuelle conditionnelle vérifie la relation suivante :

$$\begin{aligned} I(\mathbf{X}_1;\mathbf{X}_2|\mathbf{U}) &= H_\rho(\mathbf{X}_1|\mathbf{U}) - H_\rho(\mathbf{X}_1|\mathbf{X}_2,\mathbf{U}) \\ &= -H_\rho(\mathbf{U}) + H_\rho(\mathbf{X}_1,\mathbf{U}) + H_\rho(\mathbf{X}_2,\mathbf{U}) - H_\rho(\mathbf{X}_1,\mathbf{X}_2,\mathbf{U}). \end{aligned} \tag{2.22}$$

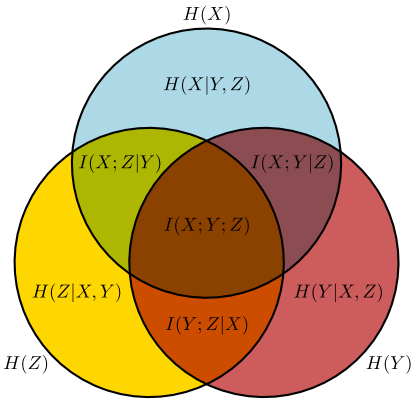


FIGURE 2.1 – Représentation sous forme de diagramme de Venn des principales relations entre entropie et information mutuelle. En appliquant le principe d'inclusion-exclusion, on retrouve facilement les formules 2.9, 2.21 et 2.24.

2.2.2 Information multivariée

L'information mutuelle peut être étendue à plus de deux vecteurs aléatoires et on la qualifie alors de *multivariée*. Plusieurs extensions existent et parmi celles-ci, c'est ici la version de McGILL (1954) qui nous intéresse. Elle est définie à partir d'un principe d'inclusion-exclusion généralisant la formule (2.20) :

Définition 2.2.3 (Information mutuelle multivariée)**.** Soit $\mathbf{X}$ une variable aléatoire et soit $\boldsymbol{T} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ un ensemble de n sous-vecteurs disjoints deux à deux. L'information mutuelle multivariée de $\boldsymbol{T}$ est définie par :

$$I(\boldsymbol{T}) = -\sum_{\boldsymbol{S}\subseteq\boldsymbol{T}} (-1)^{|\boldsymbol{S}|} H(\boldsymbol{S}), \tag{2.23}$$

Par définition, $I(\mathbf{X}_1) = H(\mathbf{X}_1)$.

L'information mutuelle multivariée est parfois appelée information mutuelle à n-points où n est la dimension de $\boldsymbol{T}$. En particulier, l'information à 3-points sera importante dans la suite :

$$\begin{aligned} I(\mathbf{X}_1;\mathbf{X}_2;\mathbf{X}_3) = {} & H(\mathbf{X}_1) + H(\mathbf{X}_2) + H(\mathbf{X}_3) - H(\mathbf{X}_1,\mathbf{X}_2) - H(\mathbf{X}_1,\mathbf{X}_3) \\ & - H(\mathbf{X}_2,\mathbf{X}_3) + H(\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3). \end{aligned} \tag{2.24}$$

La positivité de l'information mutuelle n'est valable que pour $n \leq 2$ (TE SUN 1980) et l'information à 3-points peut donc être négative. Cette propriété est fondamentale dans l'implémentation de l'algorithme MIIC pour la sélection de modèle que nous aborderons au chapitre 4. En revanche, elle reste invariante par n'importe quelle permutation de ses variables. En utilisant les formules (2.19) et (2.22), on peut montrer qu'elle peut s'écrire comme :

$$I(\mathbf{X}_1;\mathbf{X}_2;\mathbf{X}_3) = H(\mathbf{X}_1) - H(\mathbf{X}_1|\mathbf{X}_2) - H(\mathbf{X}_1|\mathbf{X}_3) + H(\mathbf{X}_1|\mathbf{X}_2,\mathbf{X}_3) \tag{2.25}$$

$$= I(\mathbf{X}_1;\mathbf{X}_2) - I(\mathbf{X}_1;\mathbf{X}_2|\mathbf{X}_3) \tag{2.26}$$

À partir de cette formule, nous définissons l'information à 3-points conditionnelle :

$$I(\mathbf{X}_1;\mathbf{X}_2;\mathbf{X}_3|\mathbf{U}) = I(\mathbf{X}_1;\mathbf{X}_2|\mathbf{U}) - I(\mathbf{X}_1;\mathbf{X}_2|\mathbf{X}_3,\mathbf{U}) \tag{2.27}$$

En résumé de ce chapitre, l'ensemble des relations entre l'entropie et l'information mutuelle que nous avons vues peuvent être résumées par un diagramme de Venn comme

sur la figure 2.1. Par exemple, l'union des deux ensembles représentant $H(X)$ et $H(Y)$ correspond à $H(X,Y)$ et d'après le principe d'inclusion-exclusion, nous retrouvons alors la formule

$$H(X,Y) = H(X) + H(Y) - [I(X;Y|Z) + I(X;Y;Z)] = H(X) + H(Y) - I(X;Y).$$

Références

Cover, T. M. et Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons (cf. p. 31).

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media (cf. p. 31, 34).

Jaynes, E. (1963). *Brandeis Summer Institute Lectures in Theoretical Physics : Statistical Physics* (cf. p. 32).

Khinchin, A. Y. (1957). « Mathematical foundations of information theory ». In : (cf. p. 32).

Kulhavỳ, R. (1996). *Recursive nonlinear estimation : a geometric approach*. T. 216. Springer (cf. p. 36, 145).

McGill, W. J. (1954). « Multivariate information transmission ». In : *Psychometrika* 19.2, p. 97-116 (cf. p. 38).

Te Sun, H. (1980). « Multiple mutual informations and multiple interactions in frequency data ». In : *Info. Control* 46.26-45, p. 4 (cf. p. 38).

Chapitre 3

Statistiques bayésiennes

Sommaire

La théorie statistique consiste à développer et étudier des méthodes pour inférer les propriétés d'une distribution liée à un processus aléatoire à partir de données d'observation. Elle trouve son application dans de nombreux domaines de la science tels que la physique, l'économie ou encore la médecine. Il existe plusieurs manières de modéliser la distribution sous-jacente. Si l'hypothèse est faite que la distribution ou sa densité appartiennent à une famille de fonctions indexée par un ensemble de paramètres $\boldsymbol{\theta}$, on parle de méthodes *paramétriques*. Cette contrainte peut-être relâchée en ne spécifiant que certaines propriétés de la distribution tels que ses moments : on parle dans ce cas de méthodes *semi-paramétriques*. Enfin, si la distribution est considérée comme étant la plus générale possible on parle de méthodes *non-paramétriques*. Un exemple de méthode non-paramétrique que l'on a vu plus tôt est celui de la distribution empirique (1.3.3) qui évalue la distribution uniquement à partir des données d'observation. Il est à noter que les méthodes non-paramétriques sont en général plus expressives mais ceci au prix d'une grande complexité et de problèmes de généralisation.

L'inférence paramétrique qui est au centre de ce chapitre, a pour but d'obtenir de l'information sur les paramètres du modèle à partir de données d'observation. En particulier, nous nous intéressons à deux types de problèmes non-exclusifs : *l'estimation des paramètres* et le *test d'hypothèses*. Dans le premier cas, c'est la meilleure approximation possible des paramètres qui est recherchée alors que dans le deuxième nous voulons valider ou infirmer une hypothèse faite sur les paramètres comme son appartenance à un certain domaine. Avant d'étudier ces méthodes nous commençons ce chapitre par une courte discussion sur l'interprétation des probabilités qui est laissée ouverte par la construction axiomatique des probabilités que nous avons vue dans le chapitre précédent. Cette question est importante car elle mène à des méthodes statistiques différentes.

3.1 Inférence fréquentiste et bayésienne

Les deux interprétations dominantes des probabilités et qui sont celles discutées ici sont l'*objectivisme* et le *subjectivisme* (Freedman 1997) mais de nombreuses autres existent (Salmon 2017).

Pour les objectivistes, les probabilités sont une propriété inhérente du système étudié (on parle parfois de probabilité physique). Dans ce cas, la probabilité d'un événement est la limite théorique de sa fréquence d'apparition dans un ensemble d'expériences lorsque le nombre d'expériences tend vers l'infini. C'est de cette manière que les probabilités des exemples que nous avons vus précédemment sont fixées. Si nous reprenons l'exemple du lancers d'une pièce, la probabilité que le résultat soit « face » est donnée par la proportion de « face » pour un nombre infini de lancers. Ainsi, il est important de noter que pour un fréquentiste, la vraie probabilité de la pièce a sa propre existence indépendamment des données et il se refusera donc à définir une distribution de probabilité sur ce paramètre. Bien évidemment, dans la pratique nous sommes limités à un nombre fini de lancer et ce sont les méthodes statistiques qui nous permettent de pouvoir tout de même tirer des conclusions basées sur cet échantillon. Pour les raisons que nous venons d'évoquer, les méthodes statistiques issues de cette interprétation des probabilités sont appelées méthodes fréquentistes.

Les subjectivistes, quant à eux, considèrent la valeur d'une probabilité comme le degré de croyance ou l'état de connaissance d'un individu dans la réalisation d'un événement. Ainsi, à toute incertitude peut être liée une probabilité et le subjectiviste s'autorisera à définir une distribution de probabilité sur un paramètre. Pour reprendre l'exemple du lancer de pièce, l'expérimentateur, de par son expérience, peut avoir un *a priori* sur la probabilité que le résultat soit « face » . En effet, de par nos expériences passées, nous avons normalement rencontré des pièces majoritairement équilibrées et nous aurons tendance à *croire* que cette probabilité est $\frac{1}{2}$ ou en tous cas proche de cette valeur. Cependant, si après dix lancers le résultat est toujours pile nous serons enclin à mettre à jour cet *a priori* et de considérer que la pièce est truquée. D'un point de vue mathématique, nous allons voir que cette mise à jour fait intervenir le théorème de Bayes et pour cette raison, les méthodes statistiques basées sur cette interprétation des probabilités sont dites bayésiennes. Pour finir, l'interprétation subjectiviste permet d'associer une distribution de probabilité à des événements qui ne sont pas reproductibles ce que ne permet pas l'interprétation objectiviste. Un exemple pourrait être « Quelle est la probabilité que la pandémie de COVID-19 soit terminée d'ici la fin de l'écriture de cette thèse ? » .

Ces deux interprétations font l'objet de nombreux débats et nous n'entendons pas prendre partie. Nous nous plaçons dans le cadre subjectiviste car c'est le cadre naturel des réseaux bayésiens qui sont au centre de cette thèse. De plus, certaines méthodes comme l'estimateur de maximum de vraisemblance peuvent être considérées du point de vue bayésien et nous verrons qu'en général, pour de grands échantillons de données, ces méthodes ont tendance à s'accorder.

3.2 Définitions

La distribution encodant l'incertitude sur les paramètres est appelée distribution a priori et sa densité est notée π. Étant donné un échantillon de m réalisations $i.i.d$ $\boldsymbol{d} = \{x[1], \ldots, x[m]\}$, nous voulons déterminer la densité dite *a posteriori* des paramètres conditionnellement à cet échantillon. Ceci nécessite au préalable de se fixer un modèle paramétrique :

Définition 3.2.1 (Modèle paramétrique)**.** Soit $\mathbf{X}$ une variable aléatoire, soit F^* sa distribution et soit $\boldsymbol{d}$ un échantillon de données obtenu à partir de cette distribution. Un modèle paramétrique pour F^* est une famille de densités

$$\mathcal{F} = \left\{ f(.|\boldsymbol{\theta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\} \tag{3.1}$$

où $\boldsymbol{\Theta}$ est le domaine [a] de variation du paramètre $\boldsymbol{\theta}$.

a. Dans la suite, nous considérons sans perte de généralité que $\boldsymbol{\Theta}$ est continu.

La densité *a posteriori*, que l'on note ρ, est alors reliée à la densité *a priori* via le théorème de Bayes :

$$\rho(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{f(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{d})} \tag{3.2}$$

où $f(\boldsymbol{d}|\boldsymbol{\theta})$ est la vraisemblance et $f(\boldsymbol{d})$ la vraisemblance marginale. Cette dernière peut souvent être ignorée puisqu'elle ne dépend pas des paramètres. L'échantillon étant *i.i.d*, la fonction de vraisemblance peut s'écrire comme un produit :

$$f(\boldsymbol{d}|\boldsymbol{\theta}) = \prod_{i=1}^{m} f(x[i]|\boldsymbol{\theta}). \tag{3.3}$$

***Exemple* 3.2.1** (Loi catégorielle)**.** Soit X une variable aléatoire discrète pouvant prendre k valeurs et $\boldsymbol{d} = (x[1], \ldots, x[m])$ un échantillon de m réalisations provenant de X. Le modèle paramétrique considéré ici est l'ensemble des lois catégorielles à k paramètres $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. La log-vraisemblance s'écrit alors comme :

$$\begin{aligned}\log f(\boldsymbol{d}|\boldsymbol{\theta}) &= \sum_{i=1}^{m} \log f(x[i]|\boldsymbol{\theta}) = \sum_{i=1}^{m}\sum_{j=1}^{k} \delta\left(x[i], j\right) \log f(j|\boldsymbol{\theta}) \\ &= \sum_{j=1}^{k} \log \theta_j \sum_{i=1}^{m} \delta\left(x[i], j\right) = \sum_{j=1}^{k} m_j \log \theta_j.\end{aligned}$$

où δ est le delta de Kronecker et $m_j = \sum_{i=1}^{m} \delta\left(x[i], j\right)$ est le nombre de fois où $X = j$ dans l'échantillon. En repassant à la vraisemblance, on peut de plus observer que :

$$f(\boldsymbol{d}|\boldsymbol{\theta}) \propto \text{Dirichlet}(\boldsymbol{\theta}; \boldsymbol{m}) \tag{3.4}$$

où $\boldsymbol{m} = (m_1, \ldots, m_k)$.

***Exemple* 3.2.2** (Loi normale)**.** Soit X une variable aléatoire réelle et $\boldsymbol{d}$ un échantillon de m réalisations provenant de X. Nous considérons comme modèle paramétrique l'ensemble des lois normales de paramètres $\boldsymbol{\theta} = (\mu, \nu) \in \boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}_+^*$. La vraisemblance s'écrit comme :

$$f(\boldsymbol{d}|\boldsymbol{\theta}) = \prod_{i=1}^{m} f(x[i]|\boldsymbol{\theta}) = \frac{1}{(2\pi\nu)^{m/2}} \exp\left(-\frac{1}{2\nu} \sum_{i=1}^{m} (x[i] - \mu)^2 \right). \tag{3.5}$$

En faisant apparaître la moyenne empirique, le terme quadratique peut être

réécrit :

$$\sum_{i=1}^{m}(x[i]-\mu)^2 = \sum_{i=1}^{m}[(x[i]-\bar{\mu})+(\bar{\mu}-\mu)]^2 = m(\bar{\nu}+(\bar{\mu}-\mu)^2)$$

et la vraisemblance a finalement pour expression :

$$f(\boldsymbol{d}|\boldsymbol{\theta}) \propto \nu^{-[(m-3)/2]+1}\exp(-\frac{m\bar{\nu}}{2}\nu^{-1})\nu^{-1/2}\exp(-\frac{1}{2\nu}(\mu-\bar{\mu})^2). \tag{3.6}$$

La vraisemblance est donc proportionnelle au produit d'une loi $\Gamma^{-1}(\nu;\frac{m-3}{2},\frac{m\bar{\nu}}{2})$ et d'une loi normale $N\left(\mu;\bar{\mu},\frac{\nu}{m}\right)$. Cette loi bivariée est appelée une loi normale-inverse-gamma et est notée $\mathrm{N}\Gamma^{-1}$. Elle est paramétrée par $\boldsymbol{\theta}=(\mu_0,\lambda,\alpha,\beta)$ et $\boldsymbol{\Theta}=\mathbb{R}\times\mathbb{R}^3_+$. Elle peut être étendue à un vecteur $\boldsymbol{\mu}$ et une matrice de covariance $\boldsymbol{\Sigma}$ et on parle dans ce cas de loi normale-inverse-Wishart (DeGroot 2005).

Au choix du modèle paramétrique se rajoute celui d'une densité *a priori*. Celle-ci permet l'ajout au modèle d'information non-issue de l'échantillon et provenant d'experts ou d'expériences similaires. Néanmoins, cette information prend rarement la forme d'une densité et son choix est alors arbitraire. Si nous reprenons l'exemple du lancer de pièce, comment incorporer dans le modèle l'*a priori* sur le fait que la probabilité que le résultat soit face est $\frac{1}{2}$? Nous pourrions imaginer utiliser par exemple une loi gaussienne de moyenne $\mu=\frac{1}{2}$ et de variance ν dont la valeur quantifierait notre certitude. Mais nous pourrions tout aussi bien utiliser une loi uniforme sur le sous ensemble $[\frac{1}{2}-a,\frac{1}{2}+a]$ où a est un nombre réel. Parmi les familles de densités possibles, il en existe une qui permet de mener les calculs analytiquement appelée famille conjuguée.

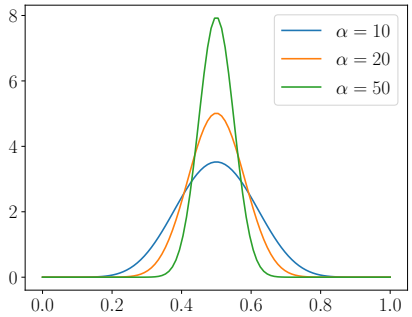
Définition 3.2.2 (Famille conjuguée)**.** Une famille de densité de probabilité $\mathcal{F}$ est dite conjuguée si pour tout $\pi\in\mathcal{F}$, la densité *a posteriori* ρ appartient à la même famille $\mathcal{F}$. Dans ce cas, l'*a priori* est appelé densité conjuguée de $f(.|\boldsymbol{\theta})$.

Pour pouvoir déterminer ces familles conjuguées, remarquons que dans les exemples précédents la vraisemblance est proportionnelle à une densité sur les paramètres. Il s'avère que ces densités sont stables par multiplication, c'est-à-dire que leur produit est proportionnel à une troisième densité de la même famille.

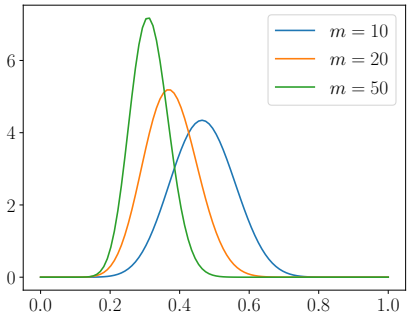
***Exemple* 3.2.3.**
- $\mathrm{Dirichlet}(\mathbf{x};\boldsymbol{\theta})\times\mathrm{Dirichlet}(\mathbf{x};\boldsymbol{\alpha})\propto\mathrm{Dirichlet}(\mathbf{x};\boldsymbol{\theta}+\boldsymbol{\alpha})$.
- $\Gamma^{-1}(\mathbf{x};\alpha_1,\beta_1)\times\Gamma^{-1}(\mathbf{x};\alpha_2,\beta_2)\propto\Gamma^{-1}(\mathbf{x};\alpha_1+\alpha_2+1,\beta_1+\beta_2)$.
- $N\left(x;\mu_1;\nu_1\right)N\left(x;\mu_2;\nu_2\right)=N\left(\mu_2-\mu_1;0,\nu_1+\nu_2\right)N\left(x;\frac{\nu_1\mu_2+\nu_2\mu_1}{\nu_1+\nu_2},\frac{\nu_1\nu_2}{\nu_1+\nu_2}\right)$.

Par conséquent, la densité d'une loi de Dirichlet est la conjuguée d'une densité d'une loi catégorielle et la densité d'une loi Normale-Inverse-Gamma est la conjuguée d'une densité d'une loi normale. Le raisonnement précédent peut être généralisé à n'importe quel type de modèle paramétrique possédant ce qu'on appelle une statistique suffisante (DeGroot 2005). L'ensemble des densités que nous avons vues jusqu'à présent font partie d'une famille conjuguée.

***Exemple* 3.2.4.** Revenons à présent à l'exemple du lancer de la pièce. Nous voulions encoder notre *a priori* selon lequel la probabilité que le résultat soit « face » est $\frac{1}{2}$. Le modèle paramétrique utilisé ici est l'ensemble des lois de Bernoulli dont la conjuguée est la loi beta [a]. Les paramètres de cette loi $Beta(\alpha,\beta)$



(a) Densité *a priori*



(b) Densité *a posteriori*

FIGURE 3.1 – La figure de gauche représente plusieurs densités *a priori* de loi $Beta(\alpha, \alpha)$ pour différentes valeurs de α. La figure de droite représente plusieurs densités *a posteriori* pour différentes tailles d'échantillon provenant d'une loi de $Bernoulli(\frac{1}{4})$ et en utilisant une densité *a priori* $Beta(10, 10)$.

doivent être égaux ($\alpha = \beta$) pour que le maximum de la densité corresponde avec $p = \frac{1}{2}$. Plus la valeur de α augmente plus la densité sera centrée sur cette valeur comme le montre la figure 3.1a. En d'autres termes, plus α est grand plus nous sommes certains que $p = \frac{1}{2}$. Supposons à présent que la pièce est truquée et que la vraie probabilité est $p = \frac{1}{4}$. Dans ce cas, notre information *a priori* est incorrecte mais plus le nombre de données obeservées augmente plus la distribution *a posteriori* va être centrée sur la bonne valeur du paramètre comme l'illustre la figure 3.1b.

a. Ces lois sont le cas à une dimension de la loi catégorielle et de la loi de Dirichlet.

L'utilisation d'un *a priori* conjugué est pratique lorsque nous avons de l'information à incorporer au modèle. En revanche, quand aucune information n'est disponible il est difficile de justifier le choix des paramètres. Dans ce cas, nous utilisons ce qu'on appelle un *a priori* non-informatif. L'exemple le plus simple auquel nous nous limitons est une distribution uniforme sur le domaine des paramètres [1] : chaque paramètre est équiprobable. Cependant, si le domaine des paramètres n'est pas fini, l'utilisation d'un *a priori* uniforme est impossible. Dans ce cas, nous avons recours à un *a priori* impropre :

Définition 3.2.3 (A priori impropre)**.** Si l'*a priori* π n'est pas une densité, c'est-à-dire si son intégrale sur $\boldsymbol{\Theta}$ n'est pas finie, il est qualifié d'impropre.

***Exemple* 3.2.5.** Supposons que nous voulions estimer la moyenne μ d'une loi gaussienne dont la variance ν est connue. Le domaine de μ étant l'ensemble des réels, il n'est pas possible de définir un *a priori* uniforme. À la place, nous utilisons l'*a priori* impropre $\pi(\mu) = c$. Dans ce cas, la densité *a posteriori*, s'écrit :

$$\rho(\theta|\boldsymbol{d}) = \frac{f(\boldsymbol{d}|\mu) \times c}{\int_{\mathbb{R}} f(\boldsymbol{d}|\mu) \times c \mathrm{d}\mu}. \tag{3.7}$$

En réutilisant l'expression de la vraisemblance vue plus haut, il est aisé de voir

1. Notons qu'un *a priori* uniforme n'est pas à proprement parler non-informatif puisque cela suppose d'une échelle pour le paramètre. En effet, la densité uniforme n'est pas invariante par changement de variable. Pour une discussion détaillée sur les *a priori* non-informatifs, voir ROBERT (2007).

que :

$$\rho(\theta|\boldsymbol{d}) = N(\mu; \bar{\mu}, \frac{\nu}{m}) \tag{3.8}$$

3.3 Estimation paramétrique

L'estimation paramétrique consiste à résumer la densité *a posteriori* en une seule valeur scalaire ou vectorielle appelée estimateur. Supposons que nous connaissions la valeur théorique $\boldsymbol{\theta}$ des paramètres étudiés, nous quantifions l'écart entre cette valeur et celle estimée à l'aide d'une fonction positive appelée fonction de coût.

Définition 3.3.1 (Estimateur bayésien)**.** L'estimateur bayésien $\hat{\boldsymbol{\theta}}$ associé à un *a priori* π et une fonction de coût L est l'estimateur obtenu en minimisant le coût moyen :

$$\mathbb{E}\left[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})|\boldsymbol{d}\right] = \int_{\boldsymbol{\Theta}} \rho(\boldsymbol{\theta}|\boldsymbol{d}) L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \mathrm{d}\boldsymbol{\theta}. \tag{3.9}$$

Nous présentons ici trois fonctions de coût classiques et les estimateurs bayésiens associés. Il est à noter cependant que cette fonction de coût peut être définie de manière plus générale à partir d'une fonction d'utilité dans le cadre de la théorie de la décision (Robert 2007). Quand cette fonction d'utilité n'est pas disponible comme dans notre cas, nous faisons appel à des fonctions de coût traditionnelles. La plus connue est sûrement la fonction de coût quadratique ou tout simplement la norme L2 :

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right\|_2 \tag{3.10}$$

Par annulation du gradient du coût moyen, on montre facilement que l'estimateur bayésien est donné dans ce cas par :

$$\hat{\theta}_i^{\mathrm{L2}} = \mathbb{E}\left[\theta_i|\boldsymbol{d}\right] = \int_{\Theta_i} \theta_i \rho(\boldsymbol{\theta}_i|\boldsymbol{d}) \mathrm{d}\theta_i, \qquad i \in [\![1, n]\!] \tag{3.11}$$

De la même manière, on montre que l'estimateur associé à la fonction de coût :

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right\|_1 \tag{3.12}$$

est donné par :

$$\hat{\theta}_i^{\mathrm{L1}} = \mathrm{median}[\theta_i|\boldsymbol{d}]. \tag{3.13}$$

Enfin, la dernière fonction de coût que nous introduisons est la fonction de coût 0-1 définie comme :

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = 1 - \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}). \tag{3.14}$$

Dans ce cas, l'estimateur bayésien associé est

$$\hat{\boldsymbol{\theta}}^{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} \rho(\boldsymbol{\theta}|\boldsymbol{d}) \tag{3.15}$$

appelé estimateur de maximum *a posteriori* (MAP). Ce dernier nous permet de retrouver l'estimateur de maximum de vraisemblance (ML) lorsque l'*a priori* utilisé est non-informatif. En effet, dans ce cas nous avons :

$$\hat{\boldsymbol{\theta}}^{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} \rho(\boldsymbol{\theta}|\boldsymbol{d}) = \arg\max_{\boldsymbol{\theta}} f(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} f(\boldsymbol{d}|\boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}^{\mathrm{ML}}.$$

Nous illustrons à présent ces différents estimateurs sur plusieurs exemples qui seront réutilisés lorsque nous discuterons de l'apprentissage des réseaux bayésiens.

***Exemple* 3.3.1** (Estimateurs pour la loi catégorielle)**.** Reprenons l'exemple de la loi catégorielle à k paramètres $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ pour laquelle nous avons vu que la vraisemblance s'écrivait comme :

$$f(\boldsymbol{d}|\boldsymbol{\theta}) = \prod_{i=1}^{k} \theta_i^{m_i} \propto \text{Dirichlet}(\boldsymbol{\theta}; \boldsymbol{m}). \tag{3.16}$$

Nous avons également vu qu'en utilisant son *a priori* conjugué Dirichlet$(\boldsymbol{\theta}; \boldsymbol{\alpha})$, la distribution a posteriori s'écrivait comme :

$$\rho(\boldsymbol{\theta}|\boldsymbol{d}) = \text{Dirichlet}(\boldsymbol{\theta}; \boldsymbol{m} + \boldsymbol{\alpha}). \tag{3.17}$$

En utilisant les propriétés de la loi de Dirichlet (voir (1.6)), il est alors facile de dériver les différents estimateurs bayésiens que nous venons de voir :

$$\begin{aligned}
\hat{\theta}_i^{\text{L2}} &= \mathbb{E}\left[\theta_i|\boldsymbol{d}\right] = \frac{\alpha'_i}{\sum_{i=1}^{k} \alpha'_i} = \frac{m_i + \alpha_i}{m + \alpha} \\
\hat{\theta}_i^{\text{L1}} &= \mathbb{M}\left[\theta_i|\boldsymbol{d}\right] = I^{-1}_{\frac{1}{2}}(\alpha'_i, \alpha' - \alpha'_i) \approx \frac{\alpha'_i - \frac{1}{3}}{\alpha' - \frac{2}{3}} = \frac{m_i + \alpha_i - \frac{1}{3}}{m + \alpha - \frac{2}{3}} \\
\hat{\theta}_i^{\text{MAP}} &= \arg\max_{\theta_i} \rho(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{\alpha'_i - 1}{\sum_{i=1}^{k} \alpha'_i - k} = \frac{m_i + \alpha_i - 1}{m + \alpha - k}
\end{aligned}$$

Pour la loi catégorielle, le domaine des paramètres $\boldsymbol{\Theta} = k$ est fini ce qui nous permet d'utiliser un *a priori* uniforme. Dans ce cas, la densité *a posteriori* est celle d'une Dirichlet$(\boldsymbol{\theta}; \boldsymbol{m})$. Ainsi, l'estimateur ML est donné par :

$$\hat{\theta}_i^{\text{ML}} = \arg\max_{\theta_i} \rho(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{m_i}{m}. \tag{3.18}$$

Nous pouvons voir sur cet exemple simple que les différents estimateurs bayésiens s'accordent entre eux pour les grands échantillons :

$$\hat{\theta}_i = \hat{\theta}_i^{\text{ML}} + O(m^{-1}). \tag{3.19}$$

Pour cette raison, quand la taille de l'échantillon est suffisante, l'estimateur ML peut être utilisé. Enfin, en comparant l'expression de l'estimateur bayésien pour la norme L2 avec celle de l'estimateur ML, on remarque que lorsque α_i est entier, l'estimateur bayésien peut être interprété comme un estimateur ML calculé sur l'échantillon auquel ont été rajoutées α *pseudo-observations* de la variable aléatoire, dont α_i, pour lesquelles elle prend la valeur i. L'ensemble de ces pseudo-observations forme ce que l'on appelle l'échantillon virtuel $\boldsymbol{d}'$. Cette interprétation sera plus tard utilisée pour la définition d'un score équivalent dans le cadre de l'apprentissage réseaux bayésiens (cf. 5).

***Exemple* 3.3.2** (Estimateurs pour la gaussienne linéaire)**.** Enfin, prenons comme modèle paramétrique le cas de la gaussienne linéaire $f(x|\mathbf{u}; \boldsymbol{\theta}) = \mathcal{N}(\mu + \beta^T \mathbf{u}; \nu)$. Ainsi, $\boldsymbol{\theta} = (\mu, \beta, \nu)$. Dans ce cas, la fonction de vrai-

semblance est donnée par :

$$\begin{aligned}\ell(\boldsymbol{\theta};\boldsymbol{d}) &= \sum_{i=1}^{m} \log f(x[i]|u[i]) \\ &= -\frac{m}{2}\log(2\pi\nu) - \tfrac{1}{2}\frac{1}{\nu}\sum_{i=1}^{m}\left[(x[i]-\mu-\beta^T\mathbf{u}[i])^2\right]\end{aligned}$$

L'annulation du gradient de cette fonction nous donne alors le système d'équations :

$$\begin{cases} \frac{\partial \ell}{\partial \mu} = -\frac{1}{\nu}\sum\limits_{i=1}^{m}(x[i]-\mu-\beta^T\mathbf{u}[i]) = 0 \\ \frac{\partial \ell}{\partial \beta_j} = -\frac{1}{\nu}\sum\limits_{i=1}^{m}(x[i]-\mu-\beta^T\mathbf{u}[i])u_j[i] = 0 \\ \frac{\partial \ell}{\partial \nu} = -\frac{m}{2\nu} + \frac{1}{2\nu^2}\sum\limits_{i=1}^{m}\left[(x[i]-\mu-\beta^T\mathbf{u}[i])^2\right] = 0 \end{cases} \tag{3.20}$$

En utilisant la première équation, on obtient :

$$\bar{x} = \mu + \beta^T\bar{\mathbf{u}} \tag{3.21}$$

où $\bar{\mathbf{u}} = (\bar{u_1},\ldots,\bar{u_k})$. La deuxième équation nous donne :

$$x\bar{u}_j = \mu\bar{u}_j + \sum_{i=1}^{k}\beta_i u_i\bar{u}_j \tag{3.22}$$

qui en lui soustrayant le terme $\bar{x}\bar{u_j}$ et en utilisant l'équation (3.21) nous donne :

$$\mathrm{Cov}(X;U_j) = \sum_{i=1}^{k}\beta_i\mathrm{Cov}(U_i;U_j). \tag{3.23}$$

Nous pouvons réécrire l'équation sous forme vectorielle comme :

$$\mathrm{Cov}(X;\mathbf{U})^T = \beta^T\boldsymbol{C}(\mathbf{U}) \tag{3.24}$$

où $\mathrm{Cov}(X;\mathbf{U}) = (\mathrm{Cov}(X,U_1),\ldots,\mathrm{Cov}(X,U_k))^T$ et $\boldsymbol{C}(\mathbf{U})$ est la matrice de corrélation de la variable $\mathbf{U}$. Finalement, la troisième équation nous donne :

$$\nu = \frac{1}{m}\sum_{i=1}^{m}\left[(x[i]-\mu-\beta^T\mathbf{u}[i])^2\right] \tag{3.25}$$

qui en utilisant (3.21), peut se réécrire comme :

$$\nu = \mathrm{Cov}(X;X) - \sum_{i=1}^{k}\sum_{j=1}^{k}\beta_i\beta_j\mathrm{Cov}(U_i;U_j) \tag{3.26}$$

soit en écriture vectorielle :

$$\nu = \mathrm{Cov}(X;X) - \beta^T\boldsymbol{C}(\mathbf{U})\beta \tag{3.27}$$

Finalement, les paramètres de MLE sont donnés par :

$$\begin{cases} \hat{\mu} = \bar{x} - \beta^T\bar{\mathbf{u}} \\ \hat{\beta} = \boldsymbol{C}(\mathbf{U})^{-1}\mathrm{Cov}(X;\mathbf{U}) \\ \hat{\nu} = \mathrm{Cov}(X;X) - \mathrm{Cov}(X;\mathbf{U})^T\boldsymbol{C}(\mathbf{U})^{-1}\mathrm{Cov}(X;\mathbf{U}) \end{cases} \tag{3.28}$$

3.4 Tests d'hypothèse

Outre l'estimation paramétrique que nous venons de voir, une autre branche importante de l'inférence statistique concerne le *test d'hypothèses statistiques*. Ces méthodes permettent de juger de la validité d'une hypothèse faite par un expérimentateur ou un analyste à partir d'un ensemble de données d'observation. Les hypothèses formulées peuvent être variées et aller de savoir si un dé est truqué, à savoir si un vaccin est responsable de cas atypiques de thrombose chez certains patients, jusqu'à savoir s'il existe un boson de Higgs (Aad et al. 2012).

Bien que nous nous soyons placés jusqu'à présent dans un cadre bayésien, certains algorithmes d'apprentissage des BNs utilisent l'approche classique des tests d'hypothèse. Au contraire de ce que l'on a vu pour l'estimation paramétrique, celle-ci ne peut pas être vue comme un cas limite de l'approche bayésienne [2]. Pour ces raisons, avant de présenter l'approche bayésienne nous faisons une parenthèse fréquentiste [3] afin de présenter les principales définitions et résultats nécessaires à la compréhension de la suite de cette thèse. Pour une introduction plus détaillée, le lecteur peut se référer par exemple à Lehmann et Romano (2006), Mittelhammer (2013) et Hogg et al. (2005).

3.4.1 Approche classique

Formellement, une *hypothèse statistique* est un ensemble de distributions candidates pour avoir généré les données observées. Si l'hypothèse statistique est composée d'une seule distribution on parle d'hypothèse *simple* sinon on parle d'hypothèse *composite*. Étant donné un modèle paramétrique [4] $\mathcal{F}$ pour les observations, l'hypothèse statistique peut être plus simplement définie par un ensemble de paramètres, ce que l'on note :

$$H_0 : \theta \in \Theta_0 \tag{3.29}$$

où Θ_0 est un sous-ensemble de Θ. L'hypothèse H_0, appelée hypothèse *nulle*, est testée en concurrence avec l'hypothèse $H_1 : \theta \in \Theta_1 \subseteq \bar{\Theta}_0$, appelée hypothèse *alternative*. Dans la suite, nous nous plaçons dans le cas particulier où $H_1 = \bar{H}_0$.

***Exemple* 3.4.1.** Pour illustrer les définitions que nous venons de donner, supposons que nous lancions une pièce 100 fois et que celle-ci atterrisse 66 fois sur face. Nous voulons savoir à partir de cet ensemble d'observations $\boldsymbol{d}$ si la pièce est truquée. Le modèle paramétrique naturel pour cette expérience est une loi de Bernoulli et l'hypothèse nulle est alors donnée par :

$$H_0 = \left\{ f(D|\theta) = \theta^8(1-\theta)^2, \theta \in \left\{ \tfrac{1}{2} \right\} \right\} \tag{3.30}$$

Comme nous l'avons dit, dans le cadre paramétrique nous pouvons réécrire les hypothèses comme des ensembles de paramètres. Ainsi, nous avons :

$$H_0 : \theta = \tfrac{1}{2} \quad \text{versus} \quad H_1 : \theta \neq \tfrac{1}{2} \tag{3.31}$$

L'hypothèse nulle est simple tandis que l'hypothèse alternative est composite.

2. Voir le chapitre 5 de Robert (2007) et Berger et Sellke (1987) pour une discussion sur ce sujet.

3. Par consistance, nous gardons la notation $f(x|\theta)$ bien que nous devrions écrire $f(x;\theta)$ puisque θ est un paramètre dans ce cas.

4. Nous rappelons que nous nous restreignons dans ce chapitre à des méthodes paramétriques.

	H_0 est vraie	H_1 est vraie
H_0 est rejetée	Erreur de type I	Correct
H_0 est choisie	Correct	Erreur de type II

TABLE 3.1 – Un test statistique peut mener à deux types d'erreurs.

Un test statistique est une règle de décision basée sur un ensemble d'observations $\boldsymbol{d}$ pour déterminer l'hypothèse la plus crédible entre H_0 et H_1. L'ensemble des échantillons possibles $\mathcal{D}$ se divise alors en deux régions $\mathcal{C}$ et $\bar{\mathcal{C}}$ appelées respectivement *région de rejet* et *région de non-rejet* de l'hypothèse nulle. La règle de décision est entièrement spécifiée par la *fonction critique* ϕ définie par :

$$\phi(\boldsymbol{d}) = \mathbb{1}_{\mathcal{C}}(\boldsymbol{d}) = \begin{cases} 1 & \text{si } \boldsymbol{d} \in \mathcal{C} \quad (H_0 \text{ est rejetée}) \\ 0 & \text{sinon} \quad\quad (H_0 \text{ n'est pas rejetée}) \end{cases} \tag{3.32}$$

En pratique, la décision n'est pas faite directement à partir de l'échantillon mais d'une statistique de test :

Définition 3.4.1 (Statistique de test)**.** Soit $\mathcal{C}$ la région de rejet associée à un test statistique. On appelle *statistique de test* une variable aléatoire réelle $T = T(\boldsymbol{D})$ permettant d'écrire la région de rejet comme $\mathcal{C} = \{\boldsymbol{d} | t(\boldsymbol{d}) \in \mathcal{C}_T\}$ où $\mathcal{C}_T$ est un ensemble d'observations $t \in \mathbb{R}$ de T. L'ensemble $\mathcal{C}_T$ est appelé région de rejet de la statistique de test T.

Comme en général une observation $\mathbf{x}[i]$ est en accord avec les deux hypothèses, il n'existe pas de test parfait permettant de rejeter ou d'accepter avec certitude l'hypothèse nulle. Il peut donc arriver que le test rejette H_0 alors qu'elle est vraie, on parle dans ce cas d'erreur de type I, ou bien qu'il l'accepte alors qu'elle est fausse et on parle dans ce cas d'erreur de type II. Le tableau 3.1 résume ces définitions. Bien que la minimisation simultanée des deux types d'erreur soit enviable, ces deux objectifs sont en contradiction et un compromis doit donc être trouvé entre chaque type d'erreur. Il est usuel pour cela de fixer une borne supérieure α à la probabilité d'erreur de type I, appelée *niveau de signification* du test, et de minimiser la probabilité d'erreur de type II sous cette contrainte [5] :

$$\min_{\mathcal{C}_T} \mathbb{P}(t \in \mathcal{C}_T | \theta \in H_1) \quad \text{sous la contrainte} \quad \mathbb{P}(t \in \mathcal{C}_T | \theta \in H_0) \leq \alpha,$$

ce qui en général se réduit à $\sup_{\mathcal{C}_T} \mathbb{P}(t \in \mathcal{C}_T) = \alpha$ où $\sup_{\mathcal{C}_T} \mathbb{P}(t \in \mathcal{C}_T)$ est appelé la *taille du test*. De manière équivalente, minimiser l'erreur de type II revient à maximiser sur la *puissance du test* β :

$$\beta = \mathbb{P}_{\mathbf{X}}(\boldsymbol{d} \in \mathcal{C}_T | \theta \in H_1). \tag{3.33}$$

Cette dernière peut-être vue comme une fonction de $\theta \in \Theta$, appelée fonction de puissance du test et qui encode l'information sur les erreurs de type I et II :

Définition 3.4.2 (Fonction de puissance d'un test)**.** La fonction de puissance d'un test paramétrique est définie par

$$\beta(\boldsymbol{\theta}) = \mathbb{P}(\boldsymbol{d} \in \mathcal{C}) = \int_{\boldsymbol{d} \in \mathcal{C}} f(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x} \tag{3.34}$$

5. On parle alors de test *uniformément le plus puissant* (UPP). Pour plus de détails voir le chapitre 3 de LEHMANN et ROMANO (2006).

La valeur du niveau de signification α est arbitraire et est en général fixée à $0,01$ où $0,05$ bien qu'aucune justification théorique ne justifie ce choix. Pour éviter ce choix arbitraire, la *p-value* d'un test est utilisée lorsque les régions de rejet sont imbriquées lorsque α varie

$$\mathcal{C}(\alpha) \subset \mathcal{C}(\alpha') \quad \text{si} \quad \alpha < \alpha' \tag{3.35}$$

où $\mathcal{C}(\alpha)$ est la région de rejet associée au test de niveau de signification α. Dans ce cas là, la *p-value* est la taille minimale pour laquelle l'hypothèse nulle serait rejetée :

$$\text{p-value} = \min_{\alpha}\left(\sup_{\theta\in H_0} \mathbb{P}(t \in \mathcal{C}_T(\alpha)|\theta) \quad \text{tel que} \quad t \in \mathcal{C}_T(\alpha)\right) \tag{3.36}$$

Nous finissons cette section avec un exemple illustrant toutes les notions que nous venons d'introduire.

***Exemple* 3.4.2.** Soit $\boldsymbol{D} = \{X[1], \ldots, X[n]\}$ un ensemble de m observations provenant de plusieurs lancers de pièces de taille n et soit $T(\boldsymbol{D}) = \sum_{j=1}^{m} X[j]$ le nombre de fois où le résultat est *face*. Afin de simplifier la discussion et sans perte de généralité, nous considérons dans la suite le cas où m est pair. Sous l'hypothèse nulle, les variables aléatoires $X[j]$ suivent toutes une même loi de Bernoulli de paramètre $p = \frac{1}{2}$ et la variable aléatoire T suit donc une loi binomiale de paramètres $(m, \frac{1}{2})$. Le nombre moyen de lancers dont le résultat est *face* est $\frac{m}{2}$. Afin de déterminer si la pièce est équilibrée, nous construisons un test dont la statistique de test est T et dont la région de non-rejet de H_0 est centrée autour de cette valeur moyenne :

$$\mathcal{C}^T = [0, \frac{m}{2} - c] \cup [\frac{m}{2} + c, m] \tag{3.37}$$

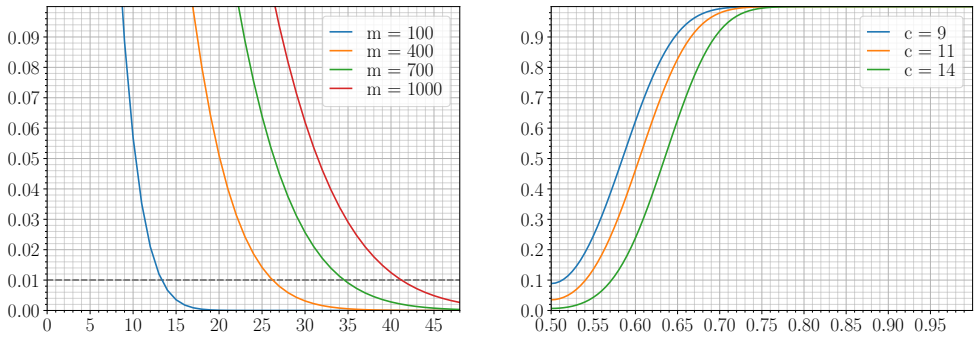
L'hypothèse nulle étant simple, la taille du test α est identique à la probabilité d'erreur de type I que l'on exprime en fonction de c :

$$\alpha(c) = \mathbb{P}(t \in \mathcal{C}^T) = 2\sum_{t=0}^{\frac{m}{2}-c} f\left(t; m, \tfrac{1}{2}\right) = \frac{1}{2^{m-1}} \sum_{t=0}^{\frac{m}{2}-c} \binom{m}{t}$$

La deuxième égalité dans l'équation précédente utilise le fait que $f\left(t; m, \frac{1}{2}\right)$ est symétrique par rapport à $\frac{m}{2}$. Pour obtenir la valeur de c pour une taille de test α donnée, il nous faut inverser la relation précédente. Cette dernière ne possédant pas de forme explicite, nous résolvons cette équation numériquement à l'aide de la figure 3.2a représentant l'évolution de c en fonction de α pour plusieurs tailles d'échantillon.
Dans l'exemple 3.4.1, nous avions $m = 100$ et $t = 66$. Pour $\alpha = 0.01$, la figure 3.2a nous indique que la valeur critique correspondante est $c = 14$. Ainsi, la région de rejet est $\mathcal{C}^T = [0, 36] \cup [64, 100]$ et l'hypothèse nulle est dans ce cas rejetée. La p-value pour $t = 66$ quant à elle prend une valeur de :

$$\begin{aligned}\text{p-value} &= \min_{\alpha}\{\mathbb{P}(t \in \mathcal{C}_T) \text{ tel que } c(\alpha) \geq 16\} \\ &= \frac{1}{2^{99}} \sum_{t=0}^{34} \binom{100}{t} \approx 1,790.10^{-3}\end{aligned}$$

Pour finir, la figure 3.2b représente la fonction de puissance γ_c du test pour différentes régions de rejet. Celle-ci étant symétrique par rapport à l'axe $\theta = \frac{1}{2}$, nous n'avons représenté que l'ensemble $[\frac{1}{2}, 1]$. Cette figure illustre le fait que réduire le risque d'erreur de type I augmente inévitablement le risque maximal d'erreur de type II donné par $1 - \inf_{\theta\in\Theta_1} \gamma_c(\theta)$ soit $1 - \alpha$ dans notre cas.



(a) (b)

FIGURE 3.2 – Évolution de la valeur critique c en fonction de la taille du test α pour différentes tailles d'échantillon m (à gauche) et fonction de puissance du test pour plusieurs valeurs critiques c (à droite).

Dans l'exemple que nous venons de voir, nous avons défini la statistique de test de manière heuristique. Il existe cependant plusieurs méthodes permettant de la définir de manière systématique et lui assurant plusieurs propriétés d'optimalité [6]. L'une de ces statistiques de test est le rapport de vraisemblance :

$$R(\boldsymbol{d}) = \frac{\sup_{\theta \in \Theta} f(\boldsymbol{d}|\theta)}{\sup_{\theta \in \Theta_0} f(\boldsymbol{d}|\theta)} \tag{3.38}$$

Malgré tout, il peut être difficile de déterminer la distribution de ces tests et nous avons recours dans ce cas à la distribution asymptotique, c'est-à-dire la distribution obtenue lorsque $m \to \infty$, pour pouvoir calculer des *p-values*. L'une des applications de cette approximation est dans la dérivation d'un test de qualité d'ajustement (*goodness of fit* en anglais). Ce genre de test permet de vérifier si le modèle paramétrique utilisé et que nous avons jusqu'à présent choisi de manière arbitraire, est en adéquation avec les données d'observation. Le test classique pour une variable aléatoire discrète est celui du χ^2 introduit par Karl Pearson en 1900 (RAO 2002). Soit un ensemble d'observations issu d'une loi catégorielle de paramètres $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, l'hypothèse nulle testée est :

$$H_0 : \theta_i = p_i \in [0, 1]. \tag{3.39}$$

Étant donné $m[i]$ le nombre de fois où le résultat i est observé, on peut obtient l'expression de la vraisemblance de l'échantillon :

$$f(\boldsymbol{d}|\boldsymbol{\theta}) = m! \prod_{j=1}^{k} \frac{\theta_j^{m[j]}}{m[j]!} \tag{3.40}$$

grâce à laquelle nous pouvons dériver l'expression du rapport de vraisemblances :

$$\log R(\boldsymbol{d}) = m \sum_{j=1}^{k} \frac{m[j]}{m} \log\left(\frac{m[j]}{mp_j}\right) \tag{3.41}$$

En utilisant alors un développement limité de $x \log(x/x_0)$ en x_0 à l'ordre 2, on montre que $2 \log R(\boldsymbol{d}) \simeq W(\boldsymbol{d})$ où W est la statistique de test du χ^2 :

$$W(\boldsymbol{d}) = \sum_{j=1}^{k} \frac{(m[j] - mp_j)^2}{mp_j}. \tag{3.42}$$

6. Voir le chapitre 10 de MITTELHAMMER (2013) pour plus de détails.

Ainsi, sous l'hypothèse H_0, $2\log(R(\boldsymbol{d}))$ et $W(\boldsymbol{d})$ ont la même distribution asymptotique qui, comme démontré dans Benhamou et al. (2018), est une distribution χ^2_k. Lorsque m est assez grand, nous pouvons donc utiliser cette approximation et calculer la p-value d'un échantillon de données à l'aide de la fonction quantile :

$$\text{p-value} = 1 - F_{\chi^2_k}(w) \tag{3.43}$$

Un cas particulier de ce test et qui nous intéresse par la suite est celui du test d'indépendance entre deux variables X_1 et X_2 conditionnellement à une troisième variable X_3. Dans ce cas, l'hypothèse nulle est :

$$f(x_1^i, x_2^j, x_3^l|\boldsymbol{\theta}) = f(x_3^l|\theta_{x_3^l}) f(x_1^i|x_3^l, \theta_{x_1^i|x_3^l}) f(x_2^j|x_3^l, \theta_{x_2^j|x_3^l}) \tag{3.44}$$

pour laquelle le test χ^2 prend la forme suivante :

$$W(\boldsymbol{d}) = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2}\sum_{l=1}^{k_3} \frac{\left(m[x_1^i, x_2^j, x_3^l] - m\hat{\theta}_{x_3^l}\hat{\theta}_{x_1^i|x_3^l}\hat{\theta}_{x_2^j|x_3^l}\right)^2}{m\hat{\theta}_{x_3^l}\hat{\theta}_{x_1^i|x_3^l}\hat{\theta}_{x_2^j|x_3^l}} \tag{3.45}$$

où k_1, k_2 et k_3 sont respectivement le nombre de valeurs que peuvent prendre les variables X_1, X_2 et X_3 et $\hat{\theta}_{x_3^l} = \frac{m[x_3^l]}{m}$, $\hat{\theta}_{x_1^i|x_3^l} = \frac{m[x_1^i, x_3^l]}{m}$ et $\hat{\theta}_{x_2^j|x_3^l} = \frac{m[x_2^j, x_3^l]}{m}$ sont les estimateurs ML de chaque paramètre. Cette formule s'étend facilement au cas où l'on teste l'indépendance conditionnellement à un ensemble $\mathbf{U}$ de variables.

Dans le cas d'une variable aléatoire continue, un test de qualité d'ajustement porte sur la fonction de répartition F de la variable et l'hypothèse nulle s'écrit alors $H_0 : F = F_0$. La distribution F_0 sera en général la distribution uniforme $U(0,1)$ puisque le cas général peut toujours être ramené à celui-ci en utilisant la transformation $U = F(X)$. Une statistique de test est alors construite à partir de la fonction de répartition empirique $\hat{F}_n$ de l'échantillon et d'une métrique sur l'espace des distributions afin de la comparer à F_0. La métrique la plus souvent utilisée dans ce cas est celle de Kolmogorov-Smirnov (Massey Jr 1951). Pour ce qui est de construire un test d'indépendance conditionnelle continue, nous repoussons la discussion jusqu'au chapitre 4 où nous aurons introduit la notion de copule qui généralise l'idée derrière la transformation $U = F(X)$ au cas multidimensionnel.

3.4.2 Approche bayésienne

Tout comme pour l'estimation paramétrique, l'approche bayésienne pour le test d'hypothèse est basée sur la densité *a posteriori*. À partir de celle-ci, un estimateur bayésien $\hat{\varphi}$ de la fonction critique (équation 3.32) est dérivé en utilisant la fonction de coût $0-1$. L'estimateur a pour expression :

$$\hat{\varphi} = \begin{cases} 1 & \text{si } \mathbb{P}(\theta \in \Theta_0|\boldsymbol{d}) > \frac{1}{2} \\ 0 & \text{sinon} \end{cases} \tag{3.46}$$

Bien qu'il soit équivalent à la densité *a posteriori* de l'hypothèse nulle, l'utilisation du facteur de Bayes est préférée :

Définition 3.4.3 (Facteur de Bayes)**.** Le facteur de Bayes, défini comme le rapport :

$$\mathrm{B}_{01}(\boldsymbol{d}) = \frac{\rho(H_0|\boldsymbol{d})}{\rho(H_1|\boldsymbol{d})} \Big/ \frac{\pi(H_0)}{\pi(H_1)}, \tag{3.47}$$

mesure l'effet des observations sur la probabilité relative entre l'hypothèse H_0 et

l'hypothèse H_1. Il n'est défini que lorsque les probabilités *a priori* des hypothèses, $\pi(H_0)$ et $\pi(H_1)$, sont non-nulles.

Si le facteur de Bayes est supérieur à 1 (resp. inférieur à 1) cela signifie que les données d'observation font pencher en faveur de l'hypothèse nulle (resp. l'hypothèse alternative). Dans le cas où $\mathrm{B}_{01}(\boldsymbol{d}) = 1$, les données ne permettent pas de décider quelle hypothèse est la plus crédible. N'étant défini que si $\pi(H_i) > 0$, nous ne pouvons donc utiliser de densité *a priori* continue dans le cas ou l'une des deux hypothèses est simple. Dans le cas où ces probabilités *a priori* sont égales, le facteur de Bayes se réécrit :

$$\mathrm{B}(\boldsymbol{d}) = \frac{\int_{\Theta_0} f(\boldsymbol{d}|\theta)\pi(\theta)\mathrm{d}\theta}{\int_{\Theta_1} f(\boldsymbol{d}|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{f_0(\boldsymbol{d})}{f_1(\boldsymbol{d})} \tag{3.48}$$

où $f_i(\boldsymbol{d})$ est la vraisemblance marginale sous l'hypothèse H_i.

***Exemple* 3.4.3.** Reprenons l'exemple du lancer de pièce dans le cadre bayésien. Nous devons d'abord définir un *a priori* sur l'ensemble des paramètres. L'hypothèse nulle étant simple, nous ne pouvons pas utiliser une densité continue pour θ. Soit $p_0 = \pi(\theta \in \Theta_0)$ et soit $g_1(\theta)$ une densité sur Θ_1, nous définissons la densité *a priori* comme :

$$\pi(\theta) = p_0 \mathbb{1}_{\Theta_0}(\theta) + (1 - p_0) g_1(\theta) \mathbb{1}_{\Theta_1}(\theta) \tag{3.49}$$

La probabilité *a posteriori* de l'hypothèse H_0 s'écrit alors comme :

$$\mathbb{P}(\theta \in \Theta_0) = \frac{f(\boldsymbol{d}|\theta_0)\pi(\theta_0)}{\int_{\Theta} f(\boldsymbol{d}|\theta)\pi(\theta)\mathrm{d}\theta} = \left[1 + \frac{1 - p_0}{p_0} \frac{1}{\mathrm{B}_{01}(\boldsymbol{d})}\right]^{-1} \tag{3.50}$$

Le modèle paramétrique étant une loi de Bernoulli, le facteur de Bayes a pour expression :

$$\mathrm{B}_{01}(\boldsymbol{d}) = \frac{f_0(\boldsymbol{d})}{f_1(\boldsymbol{d})} \Big/ \frac{p_0}{1 - p_0} = \frac{(\frac{1}{2})^m}{\int_0^1 \theta^k (1 - \theta)^{m-k} \mathrm{d}\theta} = \frac{(m+1)!}{2^m k!(m-k)!} \tag{3.51}$$

où k est le nombre de fois où la pièce est tombée sur *face* durant l'expérience. En particulier, pour $p_0 = \frac{1}{2}$, $g_1(\theta) = 1$, $m = 100$ et $k = 66$, on obtient un facteur de Bayes d'une valeur de 5.14×10^{-18} : l'hypothèse nulle est rejetée.

Un cas particulier de test d'hypothèse est celui de la sélection de modèle, c'est-à-dire le choix d'une densité de probabilité f pour les données d'observation. L'approche bayésienne s'étend facilement au cas où plus de deux modèles sont en compétition :

$$M_i : \mathbf{X} \sim f(\mathbf{x}|\boldsymbol{\theta}_i, M_i), \qquad \boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i, \quad i \in I \tag{3.52}$$

où I est l'ensemble des indices. Comme nous avons une incertitude sur le modèle, cela se traduit par une densité *a priori* sur l'ensemble des modèles. Ajoutons qu'une densité *a priori* sur les paramètres de chaque modèle doit être en plus spécifiée rendant la tâche de trouver un *a priori* plus complexe. Cela pose donc des difficultés dans la recherche d'*a priori* non-informatifs.

Afin de trouver le modèle le plus adapté nous utilisons les facteurs de Bayes entre deux modèles i et j :

$$\mathrm{B}_{ij}(\boldsymbol{d}) = \frac{\mathbb{P}(M_i|\boldsymbol{d})}{\mathbb{P}(M_j|\boldsymbol{d})} \Big/ \frac{\mathbb{P}(M_i)}{\mathbb{P}(M_j)} = \frac{f(\boldsymbol{d}|M_i)}{f(\boldsymbol{d}|M_j)} \tag{3.53}$$

où $f(\boldsymbol{d}|M_i)$ est la vraisemblance marginale du modèle M_i :

$$f(\boldsymbol{d}|M_i) = \int_\theta f(\boldsymbol{d}|\theta, M_i)\pi(\theta|M_i)\mathrm{d}\theta \tag{3.54}$$

Les facteurs de Bayes étant transitifs, c'est-à-dire vérifiant la propriété $\mathrm{B}_{ij} = \mathrm{B}_{ik}\mathrm{B}_{kj}$, nous pouvons comparer les modèles par paires afin de sélectionner le plus adapté aux données d'observation.

***Exemple* 3.4.4** (Test d'indépendance)**.** Dans le prochain chapitre nous allons nous intéresser à la recherche d'indépendances entre variables. En terme de sélection de modèle, cela revient à fixer un modèle pour la loi jointe et comparer le cas où la densité jointe se factorise (indépendance) à celui où elle ne se factorise pas (dépendance). Considérons ici le cas simple d'un vecteur aléatoire discret $\mathbf{X}$ dont les composantes X_1 et X_2 sont des variables binaires et pour lesquelles nous avons un échantillon d'observations $\boldsymbol{d}$ de taille m. Nous choisissons comme modèle paramétrique pour la densité jointe une loi catégorielle et les deux modèles correspondants s'écrivent alors :

$$\begin{cases} M_0 : \mathbf{X} \sim f(x_1^j, x_2^k|\boldsymbol{\theta}^{M_0}) = f(x_1^j|\boldsymbol{\theta}^{M_0})f(x_2^k|\boldsymbol{\theta}^{M_0}) = \theta^{M_0}_{x_1^j}\theta^{M_0}_{x_2^k} \\ M_1 : \mathbf{X} \sim f(x_1^j, x_2^k|\boldsymbol{\theta}^{M_1}) = f(x_1^j|\boldsymbol{\theta}^{M_1})f(x_2^k|x_1^j, \boldsymbol{\theta}^{M_1}) = \theta^{M_1}_{x_1^j}\theta^{M_1}_{x_2^k|x_1^j} \end{cases}$$

où M_0 correspond au cas où X_1 et X_2 sont indépendantes et M_1 à celui où elles ne le sont pas. Nous notons $\boldsymbol{\theta}^{M_i}$ l'ensemble des paramètres correspondant au modèle M_i. Soit $m[x_1^j, x_2^k]$ le nombre de fois où la configuration (x_1^j, x_2^k) apparaît dans l'échantillon d'observations. La vraisemblance sous le modèle M_0 s'écrit :

$$f(\boldsymbol{d}|\boldsymbol{\theta}^0, M_0) = \left[\theta^{M_0}_{x_1^0}\right]^{m[x_1^0]}\left[\theta^{M_0}_{x_1^1}\right]^{m[x_1^1]}\left[\theta^{M_0}_{x_2^0}\right]^{m[x_2^0]}\left[\theta^{M_0}_{x_2^1}\right]^{m[x_2^1]} \tag{3.55}$$

Nous avons vu précédemment que la conjuguée de la loi catégorielle était la loi de Dirichlet et choisissons pour cette raison la densité *a priori* :

$$\pi(\theta^{M_0}) = \pi(\theta^{M_0}_{X_1})\pi(\theta^{M_0}_{X_2}) \tag{3.56}$$

telle que $\pi(\theta^{M_0}_{X_i}) = \text{Dirichlet}(\boldsymbol{\alpha}^{M_0}_{X_i})$ où $\boldsymbol{\alpha}^{M_0}_{X_i} = (\alpha^{M_0}_{x_i^0}, \alpha^{M_0}_{x_i^1})$. Ainsi, nous pouvons calculer la vraisemblance marginale pour le modèle M_0 :

$$\begin{aligned} f(\boldsymbol{d}|M_0) &= \frac{1}{B(\boldsymbol{\alpha}^{M_0}_{X_1})}\int_{\boldsymbol{\Theta}^{M_0}_{X_1}}\left(\left[\theta^{M_0}_{x_1^0}\right]^{m[x_1^0]+\alpha^{M_0}_{x_1^0}}\left[\theta^{M_0}_{x_1^1}\right]^{m[x_1^1]+\alpha^{M_0}_{x_1^1}}\right)\mathrm{d}\boldsymbol{\theta}^{M_0}_{X_1} \\ &\quad\times\frac{1}{B(\boldsymbol{\alpha}^{M_0}_{X_2})}\int_{\boldsymbol{\Theta}^{M_0}_{X_1}}\left(\left[\theta^{M_0}_{x_2^0}\right]^{m[x_2^0]+\alpha^{M_0}_{x_2^0}}\left[\theta^{M_0}_{x_2^1}\right]^{m[x_2^1]+\alpha^{M_0}_{x_2^1}}\right)\mathrm{d}\boldsymbol{\theta}^{M_0}_{X_2} \\ &= \frac{B(\boldsymbol{\alpha}^{M_0\prime}_{X_1})}{B(\boldsymbol{\alpha}^{M_0}_{X_1})}\frac{B(\boldsymbol{\alpha}^{M_0\prime}_{X_2})}{B(\boldsymbol{\alpha}^{M_0}_{X_2})} \end{aligned}$$

où $\boldsymbol{\alpha}^{M_0\prime}_{X_i} = (m[x_i^0] + \alpha^{M_0}_{x_i^0}, m[x_i^1] + \alpha^{M_0}_{x_i^1})$. En prenant comme densité *a priori*

$$\pi(\theta^{M_1}) = \pi(\theta^{M_1}_{X_1})\pi(\theta^{M_1}_{X_2|x_1^0})\pi(\theta^{M_1}_{X_2|x_1^1}) \tag{3.57}$$

où $\pi(\theta^{M_1}_{X_1}) = \text{Dirichlet}(\boldsymbol{\alpha}^{M_1}_{X_1})$ et $\pi(\theta^{M_1}_{X_2|x_1^j}) = \text{Dirichlet}(\boldsymbol{\alpha}^{M_1}_{X_2|x_1^j})$ avec $\boldsymbol{\alpha}^{M_1}_{X_1} = (\alpha^{M_1}_{x_1^0}, \alpha^{M_1}_{x_1^1})$ et $\boldsymbol{\alpha}^{M_1}_{X_2|x_1^j} = (\alpha^{M_1}_{x_2^0|x_1^j}, \alpha^{M_1}_{x_2^1|x_1^j})$, on dérive de la même manière la vraisemblance marginale pour le modèle M_1 :

$$f(\boldsymbol{d}|M_1) = \frac{B(\boldsymbol{\alpha}^{M_1\prime}_{X_1})}{B(\boldsymbol{\alpha}^{M_1}_{X_1})}\frac{B(\boldsymbol{\alpha}^{M_1}_{X_2|x_1^0}{}')}{B(\boldsymbol{\alpha}^{M_1}_{X_2|x_1^0})}\frac{B(\boldsymbol{\alpha}^{M_1}_{X_2|x_1^1}{}')}{B(\boldsymbol{\alpha}^{M_1}_{X_2|x_1^1})} \tag{3.58}$$

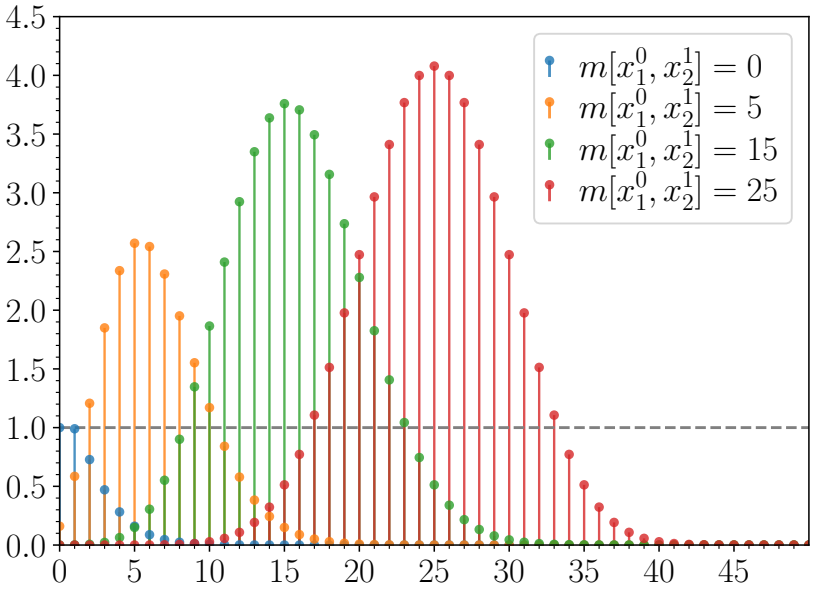


FIGURE 3.3 – Évolution du facteur de Bayes en fonction de $m[x_1^0, x_2^0]$ pour plusieurs valeurs de $m[x_1^1, x_2^0]$ et pour $m[x_1^0] = 50$.

Il reste encore à fixer la valeur des paramètres des densités *a priori* pour pouvoir calculer la valeur du facteur de Bayes entre les deux modèles. Il existe plusieurs possibilités qui seront discutées plus en détails dans le prochain chapitre et choisissons ici de fixer l'ensemble des paramètres à 1. Une question reste en suspend jusqu'à présent : comment fixer les paramètres des densités *a priori*. Le facteur de Bayes s'écrit finalement :

$$\mathrm{B}_{01}(\boldsymbol{d}) = \frac{(m[x_1^0]+1)!(m[x_1^1]+1)!}{(m+1)!} \frac{m[x_2^0]!}{m[x_1^0, x_2^0]!m[x_1^1, x_2^0]!} \frac{m[x_2^1]!}{m[x_1^0, x_2^1]!m[x_1^1, x_2^1]!} \tag{3.59}$$

Sachant que $m[x_1^0, x_2^0]+m[x_1^1, x_2^0]+m[x_1^0, x_2^1]+m[x_1^1, x_2^1] = m$, le facteur de Bayes est fonctions de 3 paramètres indépendants. La figure 3.3 représente l'évolution du facteur de Bayes en fonction de $m[x_1^0, x_2^0]$ pour plusieurs valeurs de $m[x_1^0, x_2^1]$ et pour $m[x_1^0] = 50$. Nous voyons que dans ce cas le facteur de Bayes atteint une valeur maximale $\mathrm{B}_{01}(\boldsymbol{d}) = 4.07$ lorsque $m[x_1^0, x_2^0] = m[x_1^1, x_2^0] = 25$. En effet, dans ce cas les 4 configurations possibles sont toutes présentes en même proportion ce qui est une évidence forte pour l'indépendance entre ces variables. Nous verrons dans le prochain chapitre comment généraliser cette analyse au cas général d'un vecteur aléatoire discret de dimension n.

Dans l'exemple précédent, nous avons pu calculer les vraisemblances marginales $f(\boldsymbol{d}|M_i)$ de manière analytique mais cela n'est pas toujours le cas. Pour ces raisons, nous devons parfois avoir recours soit à des méthodes numériques pour calculer ces intégrales, soit à des approximations. L'approximation la plus utilisée à cet effet est l'approximation de Laplace (BUTLER 2007). Soit g une fonction régulière possédant un unique maximum en $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ et soit h une fonction régulière positive sur $\boldsymbol{\Theta}$, on a alors :

$$\int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta}) \exp\left(m g(\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta} = h(\boldsymbol{\theta}^*) \exp\left(m g(\boldsymbol{\theta}^*)\right) \left(\frac{2\pi}{m}\right)^{p/2} \left|\tilde{H}_g(\boldsymbol{\theta}^*)\right|^{-\frac{1}{2}} + \mathcal{O}(m^{-1}),$$

où p est la dimension de $\boldsymbol{\Theta}$ et $\tilde{H}_g = -H_g$ la matrice hessienne négative de g. La vraisemblance marginale pour le modèle M peut être approximée en prenant $h_1(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|M)$ et $g_1(\boldsymbol{\theta}) = \frac{1}{m}\log f(\boldsymbol{d}|\boldsymbol{\theta}, M)$:

$$f(\boldsymbol{d}|M) = f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, M)\pi(\hat{\boldsymbol{\theta}}|M)(2\pi)^{p/2}\left|\mathcal{J}(\hat{\boldsymbol{\theta}})\right|^{-\frac{1}{2}} + \mathcal{O}(m^{-1}) \tag{3.60}$$

où $\hat{\boldsymbol{\theta}}$ est l'estimateur ML et $\mathcal{J}(\hat{\boldsymbol{\theta}})$ est la matrice de l'information de Fisher observée [7] évaluée en $\hat{\boldsymbol{\theta}}$:

$$\left[\mathcal{J}(\hat{\boldsymbol{\theta}})\right]_{ij} = -\frac{\partial^2 \log f(\boldsymbol{d}|\theta)}{\partial\theta_i\partial\theta_j}\Big|_{\theta=\hat{\theta}} \tag{3.61}$$

Remarquons que nous aurions également pu choisir $h_2(\boldsymbol{\theta}) = 1$ et $g_2(\boldsymbol{\theta}) = \frac{1}{m}\log f(\boldsymbol{d}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)$ et obtenir :

$$f(\boldsymbol{d}|M) = f(\boldsymbol{d}|\tilde{\boldsymbol{\theta}}, M)\pi(\tilde{\boldsymbol{\theta}}|M)(2\pi)^{p/2}\left|\mathcal{J}(\tilde{\boldsymbol{\theta}}) + \tilde{H}_\pi(\tilde{\boldsymbol{\theta}})\right|^{-\frac{1}{2}} + \mathcal{O}(m^{-1}) \tag{3.62}$$

où $\tilde{\boldsymbol{\theta}}$ correspond à l'estimateur MAP. Bien que cette approximation soit en général plus précise que la première, elles s'accordent asymptotiquement. En effet, d'une part nous avons vu que dans ce cas les estimateurs ML et MAP étaient identiques. D'autre part, nous pouvons voir que plus m est grand plus la fonction $\exp(mg(\boldsymbol{\theta}))$ est piquée autour de $\hat{\boldsymbol{\theta}}$ traduisant le fait que $\mathcal{I}(\tilde{\boldsymbol{\theta}})$ augmente avec m. Comme $\tilde{H}_\pi(\tilde{\boldsymbol{\theta}})$ ne dépend pas de m, le premier terme l'emporte. Plus précisément, en utilisant la loi faible des grands nombres on peut voir que :

$$\tfrac{1}{m}\left[\mathcal{J}(\boldsymbol{\theta})\right]_{ij} = -\frac{1}{m}\frac{\partial^2 \log f(\boldsymbol{d}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} = -\frac{1}{m}\sum_{j=1}^{m}\frac{\partial^2 \log f(\mathbf{x}[j]|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} \xrightarrow{p} \left[\mathcal{I}(\boldsymbol{\theta})\right]_{ij} \tag{3.63}$$

où $\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}[\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log f(X|\boldsymbol{\theta})]$ est la matrice de l'information de Fisher attendue. Ainsi, lorsque m est assez grand, nous avons :

$$\log f(\boldsymbol{d}|M) \simeq \log f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, M) + \log\pi(\hat{\boldsymbol{\theta}}|M) + \tfrac{p}{2}\log 2\pi - \tfrac{p}{2}\log m - \tfrac{1}{2}\log\left|\mathcal{I}(\hat{\boldsymbol{\theta}})\right| \tag{3.64}$$

En ne gardant que les termes dépendant de m dans la dernière expression, on définit alors le score BIC (pour *Bayesian Information Criterion*)(Schwarz 1978) du modèle M :

$$\mathcal{S}_{BIC}(M; \boldsymbol{d}) = \log f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, M) - \frac{p}{2}\log m \tag{3.65}$$

Le logarithme du facteur de Bayes peut alors être approximé par la différence des scores BIC des modèles M_0 et M_1.

$$\log \mathrm{B}_{01} \simeq \mathcal{S}_{BIC}(M_0; \boldsymbol{d}) - \mathcal{S}_{BIC}(M_1; \boldsymbol{d}) = \log\frac{f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}^{M_0})}{f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}^{M_1})} - \frac{p_0 - p_1}{2}\log m$$

Les approximations qui sont présentées ici sont liées à l'approximation gaussienne de la densité *a posteriori*. Pour plus de détails, le lecteur peut se reporter à la page 224 de Berger (2013).

***Exemple* 3.4.5.** Reprenons l'exemple précédent où nous avons dérivé le facteur de Bayes entre un modèle M_0 où les deux variables X_1 et X_2 sont indépendantes et un modèle M_1 où elles ne le sont pas. La log-vraisemblance pour le modèle

7. Nous reprenons la nomenclature de Efron et al. (1978)

M_0 s'écrit :

$$\log f(\boldsymbol{d}|\boldsymbol{\theta}^{M_0}, M_0) = \left(m[x_1^0]\log\theta_{x_1^0} + m[x_1^1]\log\theta_{x_1^1}\right) + \left(m[x_2^0]\log\theta_{x_2^0} + m[x_2^1]\log\theta_{x_2^1}\right)$$

Quant à celle du modèle M_1, elle s'écrit :

$$\log f(\boldsymbol{d}|\boldsymbol{\theta}^{M_1}, M_0) = \left(m[x_1^0]\log\theta_{x_1^0} + m[x_1^1]\log\theta_{x_1^1}\right) + \left(m[x_1^0, x_2^0]\log\theta_{x_2^0|x_1^0} + m[x_1^0, x_2^1]\log\theta_{x_2^1|x_1^0}\right) + \left(m[x_1^1, x_2^0]\log\theta_{x_2^0|x_1^1} + m[x_1^1, x_2^1]\log\theta_{x_2^1|x_1^1}\right)$$

La différence entre les log-vraisemblances s'exprime comme :

$$\begin{aligned}
\log\frac{f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}^{M_1})}{f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}^{M_0})} &= \sum_{i=0}^{1}\sum_{j=0}^{1} m[x_1^i, x_2^j]\log\hat{\theta}_{x_2^j|x_1^i} - \sum_{j=0}^{1} m[x_2^j]\log\hat{\theta}_{x_2^j} \\
&= \sum_{i=0}^{1}\sum_{j=0}^{1} m[x_1^i, x_2^j]\log\frac{m[x_1^i, x_2^j]}{m[x_1^i]} - \sum_{j=0}^{1} m[x_2^j]\log\frac{m[x_2^j]}{m} \\
&= m\left(\sum_{i=0}^{1}\sum_{j=0}^{1}\frac{m[x_1^i, x_2^j]}{m}\log\frac{m[x_1^i, x_2^j]}{m} - \right. \\
&\qquad \left. \sum_{i=0}^{1}\frac{m[x_1^i]}{m}\log\frac{m[x_1^i]}{m} - \sum_{j=0}^{1}\frac{m[x_2^j]}{m}\log\frac{m[x_2^j]}{m}\right) \\
&= m\left(H_{\hat{f}}(X_1) + H_{\hat{f}}(X_2) - H_{\hat{f}}(X_1, X_2)\right) = mI_{\hat{f}}(X_1; X_2)
\end{aligned}$$

où $\hat{f}$ est la densité empirique définie à partir de la fréquence d'occurrence de chaque configuration dans l'ensemble des observations :

$$\hat{f}(x_1^i, x_2^j) = \frac{m[x_1^i, x_2^j]}{m}. \tag{3.66}$$

Finalement, le facteur de Bayes a pour expression :

$$B_{01}(\boldsymbol{d}) = m^{\frac{1}{2}}\exp\left(-mI_{\hat{f}}(X_1, X_2)\right) + \mathcal{O}(1) \tag{3.67}$$

Cette approximation est d'autant plus fidèle que le nombre d'observations est grand.

D'autres approximations de la vraisemblance marginale ont été proposées telles que les scores AIC (pour *Akaike Information Criterion*), NML (pour *Normalized Maximum Likelihood*), la déviance Bayésienne, etc. Celles-ci sont traitées par Robert (2007) et une étude comparative de plusieurs d'entre elles peut être trouvée dans Chickering et Heckerman (1997).

Nous avons vu dans de ce chapitre les notions fondamentales de statistiques bayésiennes pour l'estimation de paramètres et pour le test d'hypothèse. Nous nous sommes cependant limité au cas unidimensionel. Comme nous allons voir dans le prochain chapitre, la complexité des modèles augmente fortement avec la dimension. Les BNs, en plus de permettre de réduire cette complexité permettent également une représentation graphique des relations entre variables aléatoires du problème.

Références

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Khalek, S. A., Abdelalim, A., Aben, R., Abi, B., Abolins, M., Abouzeid, O. et al. (2012). « Combined search for the Standard Model Higgs boson in p p collisions at s= 7 TeV with the ATLAS detector ». In : *Physical Review D* 86.3, p. 032003 (cf. p. 49).

Benhamou, E. et Melot, V. (2018). « Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation ». In : *arXiv preprint arXiv :1808.09171* (cf. p. 53).

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media (cf. p. 57).

Berger, J. O. et Sellke, T. (1987). « Testing a point null hypothesis : The irreconcilability of p values and evidence ». In : *Journal of the American statistical Association* 82.397, p. 112-122 (cf. p. 49).

Butler, R. W. (2007). *Saddlepoint approximations with applications*. T. 22. Cambridge University Press (cf. p. 56).

Chickering, D. M. et Heckerman, D. (1997). « Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables ». In : *Machine learning* 29.2, p. 181-212 (cf. p. 58).

DeGroot, M. H. (2005). *Optimal statistical decisions*. T. 82. John Wiley & Sons (cf. p. 44).

Efron, B. et Hinkley, D. V. (1978). « Assessing the accuracy of the maximum likelihood estimator : Observed versus expected Fisher information ». In : *Biometrika* 65.3, p. 457-483 (cf. p. 57).

Freedman, D. (1997). « Some issues in the foundation of statistics ». In : *Topics in the Foundation of Statistics*. Springer, p. 19-39 (cf. p. 42).

Hogg, R. V., McKean, J. et Craig, A. T. (2005). *Introduction to mathematical statistics*. Pearson Education (cf. p. 49).

Lehmann, E. L. et Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media (cf. p. 49, 50).

Massey Jr, F. J. (1951). « The Kolmogorov-Smirnov test for goodness of fit ». In : *Journal of the American statistical Association* 46.253, p. 68-78 (cf. p. 53).

Mittelhammer, R. C. (2013). *Mathematical Statistics for Economics and Business*. Springer Science & Business Media (cf. p. 49, 52).

Rao, C. (2002). « Karl Pearson chi-square test the dawn of statistical inference ». In : *Goodness-of-fit tests and model validity*. Springer, p. 9-24 (cf. p. 52).

Robert, C. (2007). *The Bayesian choice : from decision-theoretic foundations to computational implementation*. Springer Science & Business Media (cf. p. 45, 46, 49, 58).

Salmon, W. C. (2017). *The foundations of scientific inference*. University of Pittsburgh Press (cf. p. 42).

Schwarz, G. (1978). « Estimating the dimension of a model ». In : *The annals of statistics* 6.2, p. 461-464 (cf. p. 57).

Partie II

État de l'art

Chapitre 4

Les réseaux bayésiens

Sommaire

Nous avons introduit dans le chapitre précédent plusieurs méthodes de statistiques bayésiennes. Bien que celles-ci n'aient pas de limitations théoriques vis à vis de la dimension du problème, dans la pratique elles sont limitées à un nombre restreint de variables aléatoires à cause d'une trop grande complexité des calculs. Le modèle des réseaux bayésien, qui fait l'objet de ce chapitre, utilise un ensemble d'hypothèses d'indépendances vérifiées par la distribution jointe du modèle afin de réduire sa complexité. Ces indépendances sont encodées au sein d'un graphe appelé la structure du réseau bayésien qui confère une grande interprétabilité au modèle et facilite donc les échanges avec les experts du domaine auquel il est appliqué.

Ce chapitre s'ouvre avec un exemple introductif motivant l'utilisation d'hypothèses d'indépendance pour simplifier un modèle multivarié. Cet exemple est suivi par un bref rappel de théorie des graphes fournissant les définitions de base nécessaires à l'introduction des réseaux bayésiens qui conclut ce chapitre.

4.1 Exemple introductif

Prenons le cas simple d'un vecteur aléatoire $\mathbf{X}$ à n dimensions dont les composantes X_i sont des variables discrètes de domaine $\Omega_i = \{0, 1\}$. Le modèle paramétrique par défaut dans ce cas de figure est une loi catégorielle à d dimensions : $f(\mathbf{x}|\boldsymbol{\theta}) = \boldsymbol{\theta}_{\mathbf{x}}$ tel que $\sum_{\mathbf{x}} \boldsymbol{\theta}_{\mathbf{x}} = 1$.

Commençons par observer que cette paramétrisation de la densité est mal adaptée pour l'incorporation d'information. En effet, dans le cas multivarié celle-ci est souvent disponible sous la forme de probabilités conditionnelles plutôt que sous la forme de

probabilités jointes. Par exemple, un médecin sera plus à même de donner la probabilité qu'une femme développe un cancer du sein sachant qu'elle fume plutôt que la probabilité que cette personne fume et qu'elle développe un cancer. Pour cela, nous pouvons reparamétrer la densité en utilisant la règle de chaîne :

$$f(\mathbf{x}|\boldsymbol{\theta}') = \prod_{i=1}^{d} f(x_i|x_1, \ldots, x_{i-1}, \boldsymbol{\theta}') \tag{4.1}$$

où $\boldsymbol{\theta}' = (\boldsymbol{\theta}_{x_1}, \boldsymbol{\theta}_{x_2|x_1}, \ldots, \boldsymbol{\theta}_{x_d|x_1,\ldots,x_{d-1}})$.

Toutefois, le nombre de paramètres indépendants associé à cette distribution, $|\boldsymbol{\theta}| = 2^d - 1$, étant exponentiel par rapport à la dimension, le nombre d'estimation à mener devient rapidement ingérable. À supposer que l'on puisse quand même faire ces estimations dans un temps raisonnable, la plupart des paramètres estimés seraient nuls ou auraient une trop grande variance par manque de données. Pour le voir, définissons $\theta_{min} = \min_{\mathbf{x}} \boldsymbol{\theta}_{\mathbf{x}} \leq (2^d - 1)^{-1}$ le paramètre dont la valeur théorique est la plus faible. Il faudra en moyenne $\theta_{min}^{-1} \geq 2^d - 1$ observations pour que l'événement associé $\mathbf{x}_{min}$ se réalise au moins une fois. Ainsi, non seulement le nombre de paramètres est exponentiel mais le nombre d'observations nécessaires l'est également.

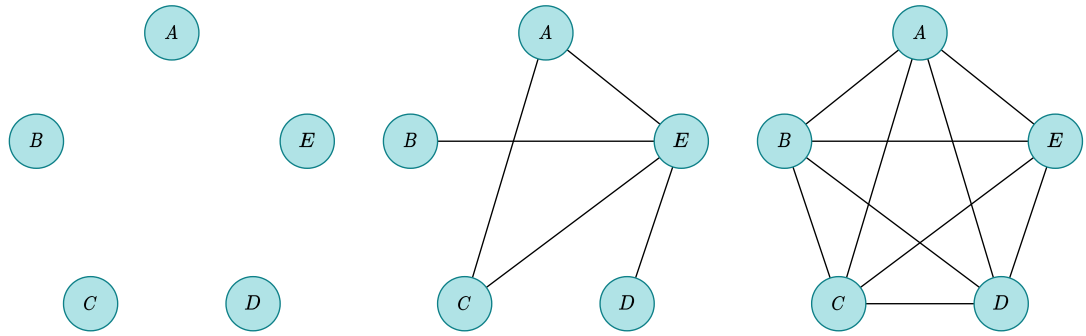
Pour diminuer le nombre de paramètres, nous pouvons ajouter un certain nombre d'hypothèses d'indépendance au modèle qui vont contraindre les paramètres. Supposons que le modèle vérifie les indépendances $X_i \perp\!\!\!\perp \{X_1, \ldots, X_{i-2}\} \mid X_{i-1}$, dans ce cas la densité jointe s'écrit comme [1] :

$$f(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} f(x_i|x_{i-1}, \boldsymbol{\theta}). \tag{4.2}$$

où $\boldsymbol{\theta} = (\boldsymbol{\theta}_{x_1}, \boldsymbol{\theta}_{x_2|x_1}, \ldots, \boldsymbol{\theta}_{x_d|x_{d-1}})$. La complexité du modèle est alors réduite puisqu'il contient maintenant $|\boldsymbol{\theta}| = 2d - 1$ paramètres : nous sommes passés d'une complexité exponentielle à une complexité linéaire au prix d'hypothèses fortes. Nous avons fourni ici l'ensemble des indépendances de façon *ad hoc* mais de la même manière que les paramètres elles peuvent être données par un expert. Si par exemple nous connaissons le résultat de la mammographie d'une patiente, l'information que cette dernière fume devient superflue. Cependant, à mesure que le nombre de variables augmente il devient de plus en plus difficile de déterminer ces indépendances avec exactitude. Pour une meilleure visualisation, nous allons donc utiliser un graphe pour les encoder. Il n'est pas rare cependant que l'ensemble des indépendances soit inconnu. Dans ce cas, nous devons nous reposer sur les données pour apprendre le modèle d'indépendance. Comme pour les paramètres, cette incertitude va se traduire par une distribution de probabilité sur l'ensemble des modèles. Cet ensemble étant très grand, l'utilisation de graphes va également nous permettre l'implémentation d'algorithmes efficaces. Ce sujet sera abordé en détail lors du prochain chapitre.

Toutes ces observations sont à la base des réseaux bayésiens qui vont encoder une densité jointe sous la forme d'un *DAG* et d'un ensemble de densités conditionnelles. Nous allons voir que ce *DAG* représente à la fois un ensemble d'indépendances et la factorisation d'une densité jointe. Enfin, notons que la structure graphique confère aux réseaux bayésiens une interprétabilité qui peut faire défaut à d'autres modèles tels que les réseaux de neurones. Pour ces raisons, nous introduisons dans la prochaine section plusieurs notions de théorie des graphes qui serons utilisées pour la définition des réseaux bayésiens.

1. Ce modèle est équivalent à une chaîne de Markov non-homogène.



(a) Graphe vide (b) Graphe quelconque (c) Graphe complet

FIGURE 4.1 – Le graphe vide, un graphe quelconque et le graphe complet pour l'ensemble de nœuds $V = \{A, B, C, D, E\}$.

4.2 Notions de théorie des graphes

Les notions de théorie des graphes qui sont abordées dans cette section sont généralement connues du lecteur et servent principalement à introduire les notations que nous allons utiliser par la suite. Pour une introduction plus exhaustive, le lecteur peut se référer aux ouvrages suivants : BRETTO et al. (2012), DIESTEL (2005), BONDY et al. (1976) et BOLLOBÁS (2013).

4.2.1 Graphes non-orientés

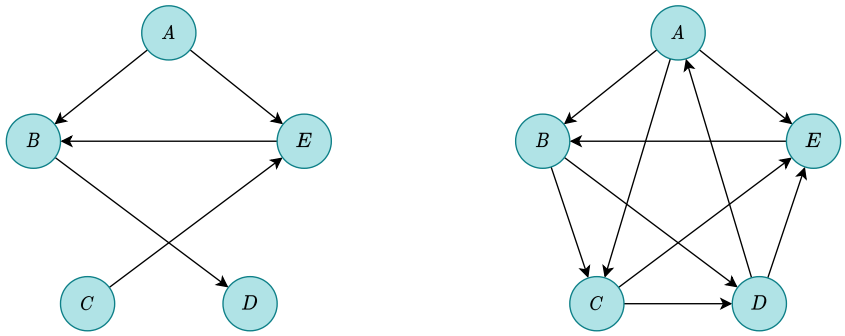
Soit V un ensemble fini et $[V]^k$ l'ensemble de ses parties à k éléments. Nous donnons la définition de graphe non-orienté :

Définition 4.2.1 (Graphe non-orienté)**.** Un graphe non-orienté (UG pour *Undirected Graph*) est un couple $G = (V, E)$ où V est un ensemble fini dont les éléments sont appelés *nœuds* du graphe et où $E \subseteq [V]^2$ est un ensemble de parties à deux éléments [a] de V dont les éléments sont appelés *liens* du graphe.

a. Notre définition de l'ensemble des liens écarte les graphes contenant des boucles ou contenant plusieurs liens différents entre deux mêmes nœuds. Les graphes que nous considérons sont appelés graphes simples.

La ***taille*** d'un graphe correspond au nombre de nœuds $|V|$ qu'il contient. Les liens $\{v_1, v_2\}$ d'un graphe sont également notés $v_1 - v_2$ et v_1 et v_2 sont appelés les ***extrémités*** du lien. Lorsque v_1 et v_2 sont les extrémités d'un même lien ils sont dits ***voisins*** et on note $\mathbf{Ne}(v)$ l'ensemble des voisins d'un nœud v donné. Le graphe ne contenant aucun lien ($E = \emptyset$) est appelé le graphe ***vide*** et le graphe contenant tous les liens possibles ($E = [V]^2$) le graphe ***complet***. La représentation d'un graphe non-orienté se fait en traçant un point pour chaque nœud et deux de ces points sont reliés par un trait s'il existe un lien entre les deux nœuds correspondants. Plusieurs exemples de graphes non-orientés sont donnés sur la figure 4.1. Pour un nombre de nœuds $d = |V|$ fixé, il existe $2^{d(d-1)/2}$ graphes non-orientés sans boucle : ce nombre est super-exponentiel par rapport à la dimension du graphe (voir figure 4.3).

Un ***chemin*** de longueur l entre deux nœuds v_0 et v_l est une suite de nœuds $(v_0, \ldots, v_i, \ldots, v_l)$ telle que deux nœuds successifs sont voisins. Un chemin pour lequel $v_0 = v_l$ est appelé un ***cycle***.



(a) Graphe quelconque (b) Graphe complet

FIGURE 4.2 – Un graphe orienté quelconque et un graphe orienté complet pour l'ensemble de nœuds $V = \{A, B, C, D, E\}$.

4.2.2 Graphes orientés

Les graphes orientés rajoutent la notion d'orientation aux liens d'un graphe non-orienté $G = (V, E)$. Celle-ci est une fonction $f_o^G : E \to V \times V$ telle que pour un lien $\{x, y\} \in E$, on a $f_o^G(\{x, y\}) = (x, y)$ ou $f_o^G(\{x, y\}) = (y, x)$. Nous donnons à présent la définition de graphe orienté.

Définition 4.2.2 (Graphe orienté)**.** Un graphe orienté (DiG pour *Directed Graph*) est un triplet $G = (V, E, f_o)$ où V est l'ensemble des nœuds du graphe et $E \subseteq [V]^2$ est l'ensemble des liens et f_o est l'orientation des liens.

Dans la suite nous simplifierons parfois la définition d'un graphe orienté comme étant un couple $G = (V, A)$ où $A = f_o^G(E)$ est un ensemble de couples d'éléments de V qu'on appelle ***arcs***. L'ajout de l'orientation enrichit les définitions qui ont été données pour les graphes non-orientés. Les arcs (v_1, v_2) d'un graphe sont également notés $v_1 \to v_2$ et v_1 et v_2 sont respectivement appelés extrémités ***initiale*** et ***finale*** de l'arc. Lorsque v_1 et v_2 sont les extrémités initiale et finale d'un même arc, v_1 est un ***parent*** de v_2 et v_2 est un ***enfant*** de v_1. L'ensemble des parents d'un nœud v dans le graphe G est noté $\mathbf{Pa}_v^G$. La représentation d'un graphe orienté se fait en traçant un point pour chaque nœud et deux de ces points sont reliés par une flèche s'il existe un arc entre les deux nœuds correspondants. La base de la flèche est associée au nœud correspondant à l'extrémité initiale de l'arc et la pointe de la flèche au nœud correspondant à l'extrémité finale. Plusieurs exemples de graphes orientés sont donnés sur la figure 4.2.

Pour un nombre de nœuds $d = |V|$ fixé, il existe $3^{d(d-1)/2}$ graphes orientés. Le ***squelette*** d'un graphe orienté (V, E, f_o^G) est le graphe non-orienté (V, E) obtenu en ne tenant pas compte de l'orientation des arcs. Un ***chemin orienté*** de longueur l entre deux nœuds v_0 et v_l est une suite de nœuds $(v_0, \ldots, v_i, \ldots, v_l)$ telle que deux nœuds successifs v_i et v_{i+1} sont respectivement parent et enfant l'un de l'autre. S'il existe un chemin orienté partant d'un nœud v_1 vers un nœud v_2 alors v_2 est appelé un ***descendant*** de v_1 sinon il est appelé un ***non-descendant*** de v_1. Pour un nœud v donné, on note $\mathbf{ND}_v^G$ l'ensemble de ses non-descendants dans le graphe G[2]. Un chemin orienté pour lequel $v_0 = v_l$ est appelé un ***cycle orienté***. Si un chemin ne tient pas compte des orientations, on parle de chemin non-orienté.

2. Dans la suite, quand il ne peut y avoir de confusion possible sur le graphe utilisé, nous omettons l'exposant pour les parents et les non-descendants d'un nœud afin de simplifier les notations.

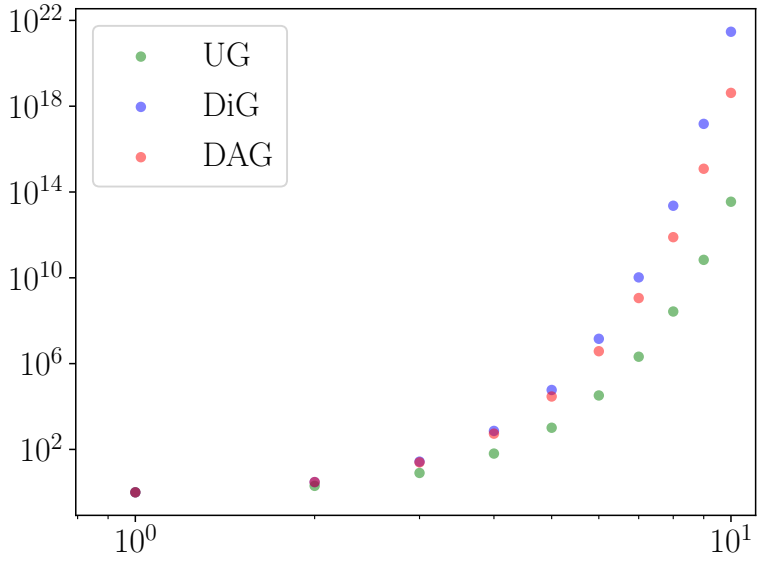


FIGURE 4.3 – Le nombre de graphes non-orientés (UG), orientés (DiG) et orientés acycliques (DAG) possibles en fonction du nombre de nœuds est super-exponentiel.

Définition 4.2.3 (DAG). Un graphe orienté ne contenant pas de cycle orienté est appelé graphe orienté ***acyclique*** ou *DAG* (pour *Directed Acyclic Graph*).

Le graphe de la figure 4.2a est un DAG tandis que celui de la figure 4.2b n'en est pas un puisqu'il existe un cycle orienté $B \to C \to E \to B$. ROBINSON (1977) a démontré que le nombre de DAG N_d pour un nombre de nœuds d donné vérifiait la relation de récurrence suivante :

$$N_d = \sum_{k=1}^{d} (-1)^{k-1} \binom{d}{k} 2^{k(d-k)} N_{d-k} \tag{4.3}$$

L'évolution du nombre de graphes possibles en fonction du nombre de nœuds est représentée sur la figure 4.3 pour chacune des catégories de graphes que nous avons vues jusqu'à présent (UG, DiG et DAG).

Nous pouvons définir la fonction d'orientation f_o sur un sous-ensemble de liens $E' \subset E$. Dans ce cas, certains liens possèdent une orientation alors que d'autres restent non-orientés. On parle dans ce cas de graphe ***mixte***. Une sous-catégorie de graphes mixtes qui sera importante pour la représentation des classes d'équivalence de Markov (voir 4.6) est celle des graphes acycliques partiellement dirigés :

Définition 4.2.4 (PDAG). Un graphe acyclique partiellement dirigé (PDAG) est un graphe mixte ne contenant pas de cycles orientés.

Enfin, nous allons définir dans la prochaine section un concept d'indépendance graphique appelé *d-séparation* faisant intervenir la notion de v-structure :

Définition 4.2.5 (V-structure). On appelle *v-structure* un triplet de nœuds (X, Z, Y) tel que $X \to Z \leftarrow Y$ et tel que X et Y ne soient pas voisins. Le nœud Z forme ce que l'on appelle le *sommet* de la v-structure tandis que les nœuds X et Y forment ce que l'on appelle la *base* de la v-structure.

Les nœuds A, E et C de la figure 4.2a forment une v-structure dont le sommet est le nœud E. En revanche, bien que B soit un enfant commun de A et E, ces trois nœuds ne forment pas une v-structure puisque A et E sont voisins.

4.3 Modèle d'indépendance

Pour encoder les indépendances du modèle au travers d'un graphe, nous devons d'abord définir une notion d'indépendance graphique possédant les mêmes propriétés que l'indépendance probabiliste. Pour cette raison, Pearl et Paz (1985) ont axiomatisé ces propriétés en une structure appelée graphoïde :

Définition 4.3.1 (Modèle d'indépendance)**.** Soit E un ensemble fini. Un modèle d'indépendance I sur E est un ensemble de triplets (X, Z, Y) où X, Y et Z sont des sous-ensembles disjoints de E. Les éléments sont notés $I(X, Z, Y)$ et X et Y sont dits indépendants conditionnellement à Z.

Définition 4.3.2 (Graphoïde)**.** Un modèle d'indépendance I sur E est un graphoïde si pour tout sous-ensembles $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{W}$ de U les propriétés suivantes sont vérifiées :

i) Indépendance triviale : $I(\mathbf{X}, \mathbf{Z}, \emptyset)$

ii) Symétrie : $I(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \iff I(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$,

iii) Décomposition : $I(\mathbf{X}, \mathbf{Z}, \mathbf{W} \cup \mathbf{Y}) \implies I(\mathbf{X}, \mathbf{Z}, \mathbf{W})$ et $I(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$,

iv) Union : $I(\mathbf{X}, \mathbf{Z}, \mathbf{W} \cup \mathbf{Y}) \implies I(\mathbf{X}, \mathbf{W} \cup \mathbf{Z}, \mathbf{Y})$,

v) Contraction : $I(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ et $I(\mathbf{X}, \mathbf{Y} \cup \mathbf{Z}, \mathbf{W}) \implies I(\mathbf{X}, \mathbf{Z}, \mathbf{W} \cup \mathbf{Y})$

vi) Intersection : $I(\mathbf{X}, \mathbf{W} \cup \mathbf{Z}, \mathbf{Y})$ et $I(\mathbf{X}, \mathbf{Y} \cup \mathbf{Z}, \mathbf{W}) \implies I(\mathbf{X}, \mathbf{Z}, \mathbf{W} \cup \mathbf{Y})$.

Dans le cas où la dernière propriété n'est pas vérifiée, on parle de semi-graphoïde.

À partir de la définition d'indépendance conditionnelle vue dans la section 1.6.6, on peut vérifier que l'ensemble des indépendances $\mathcal{I}(\mathbb{P}_{\mathbf{X}})$ d'une distribution quelconque possède une structure de semi-graphoïde. Si la distribution est en plus positive, $\mathcal{I}(\mathbb{P}_{\mathbf{X}})$ a une structure de graphoïde. Verma et al. (1988) ont montré que dans le cas des DAGs, une bonne notion d'indépendance graphique est la d-séparation :

Définition 4.3.3 (Chemin actif)**.** Soit $G = (\mathbf{V}, \mathbf{E})$ un DAG, $\mathbf{U} \subset \mathbf{V}$ un sous-ensemble de nœuds et soit $v_1 - \cdots - v_k$ un chemin dans G. Ce chemin est actif étant donné $\mathbf{U}$ si tout nœud v le long du chemin vérifie l'une des conditions suivantes :

i) v est le sommet d'une v-structure et soit v, soit un de ses descendants, appartient à $\mathbf{U}$,

ii) v n'est pas le sommet d'une v-structure et $v \notin \mathbf{U}$.

Dans le cas contraire, on dit que le chemin est bloqué.

Définition 4.3.4 (d-séparation)**.** Soit G un DAG et soient $\mathbf{X}$, $\mathbf{Y}$ et $\mathbf{U}$ trois ensembles disjoints de nœuds de G. $\mathbf{X}$ et $\mathbf{Y}$ sont d-séparés par $\mathbf{U}$, ce que l'on note $\langle \mathbf{X} | \mathbf{U} | \mathbf{Y} \rangle_G$, s'il n'existe pas de chemin actif entre un nœud de $\mathbf{X}$ et un nœud de $\mathbf{Y}$ étant donné $\mathbf{U}$.

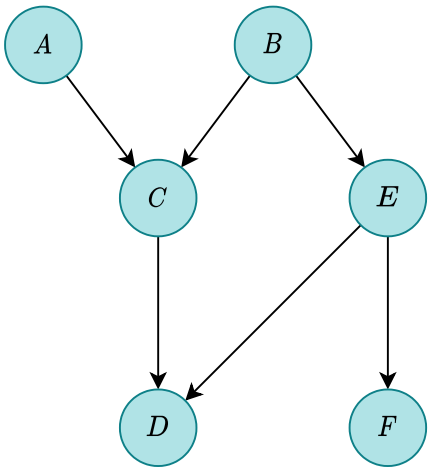


FIGURE 4.4 – Les variables A et B sont d-séparées conditionnellement à l'ensemble vide mais ne le sont pas conditionnellement à l'ensemble $\{D\}$. De même A et D sont d-séparés par $\{B, C\}$ mais ne le sont plus si B est supprimé de l'ensemble conditionnant.

***Exemple* 4.3.1.** Soit G le DAG de la figure 4.4. Si nous prenons $\mathbf{U} = \emptyset$, nous avons $\langle A|\emptyset|B\rangle_G$. En effet, le chemin $A - C - B$ est bloqué puisque C est le sommet d'une v-structure et ni C ni un de ses descendants n'appartient à $\mathbf{U}$. Si en revanche $\mathbf{U} = \{D\}$, le chemin $A - C - B$ est actif car D est un descendant de C. De même, le chemin $A - C - D - E - B$ est actif puisque D est le sommet d'une v-structure. Prenons $\mathbf{X} = \{A\}$, $\mathbf{Y} = \{D\}$ et $\mathbf{U} = \{C\}$. D'après la condition *ii)*, le chemin $A - C - D$ est bloqué. Cependant, C est le sommet d'une v-structure et le chemin $A - C - B - E - D$ est actif : A et D ne sont pas d-séparés par C. Si le nœud B est ajouté à $\mathbf{U}$, le chemin $A - C - B - E - D$ est cette fois bloqué et donc $\langle A|\{B, C\}|D\rangle_G$.

Étant donné un DAG $G = (V, E)$ muni de la d-séparation, le modèle d'indépendance sur V associé :

$$\mathcal{I}(G) = \left\{(\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3) \in \mathcal{P}(V)^3 \,|\mathbf{V}_i \cap \mathbf{V}_j = \emptyset, \langle \mathbf{V}_1|\mathbf{V}_3|\mathbf{V}_2\rangle_G\right\} \tag{4.4}$$

possède une structure de graphoïde. Comme G détermine de façon unique $\mathcal{I}(G)$ via la d-séparation, G est aussi appelé modèle d'indépendance par abus de terminologie.

4.4 I-map

La question est à présent de savoir si étant donné un ensemble d'indépendances $\mathcal{I}(\mathbb{P}_\mathbf{X})$, il existe toujours un DAG $G = (\mathbf{X}, E)$ tel que $\mathcal{I}(G) = \mathcal{I}(\mathbb{P}_\mathbf{X})$. En effet, lorsque nous allons déterminer les indépendances d'une distribution à partir d'un ensemble d'observations, nous voulons pouvoir toutes les encoder au sein du graphe que l'on va construire. Comme le montre l'exemple suivant, ce n'est pas toujours le cas :

***Exemple* 4.4.1.** Soit $\mathbf{X} = (A, B, C, D)$ un vecteur aléatoire et soit $\mathbb{P}_\mathbf{X}$ sa distribution vérifiant $A \perp\!\!\!\perp B \,|\, \{C, D\}$ et $C \perp\!\!\!\perp D \,|\, \{A, B\}$. Supposons qu'il existe un modèle d'indépendance G qui vérifie ces indépendances. Dans ce cas, il ne contient pas d'arc entre A et B ni entre C et D. Pour pouvoir encoder la première indépendance dans le graphe, ni C ni D ne doit être le sommet d'une v-structure. Cela laisse donc le choix entre les deux structures de la figure 4.5. Ainsi, A ou

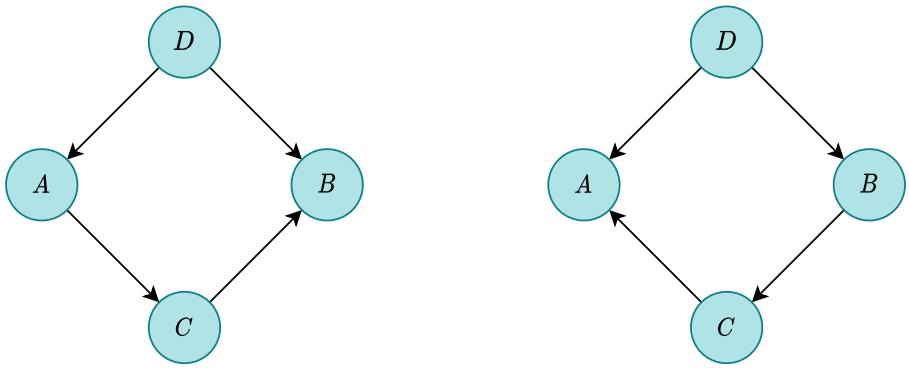


FIGURE 4.5 – Pour encoder l'indépendance $A \perp\!\!\!\perp B \,|\, \{C, D\}$ en utilisant la d-séparation, la création d'une v-structure en A ou B est inévitable.

B est contraint d'être le sommet d'une v-structure puisque nous ne pouvons pas avoir de cycle orienté. Ceci est en contradiction avec la deuxième indépendance et par conséquent il n'existe pas de DAG associé vérifiant simultanément ces indépendances.

Comme par la suite nous allons uniquement nous servir du graphe pour déduire les indépendances vérifiées par la distribution, nous voulons qu'à défaut d'être un P-map le graphe encode un ensemble d'indépendances effectivement présentes dans $\mathcal{I}(\mathbb{P}_{\mathbf{X}})$.

Définition 4.4.1 (I-map, D-map et P-map)**.** Soit $\mathbb{P}_{\mathbf{X}}$ une distribution et soit $\mathcal{I}(\mathbb{P}_{\mathbf{X}})$ l'ensemble de ses indépendances. Un DAG $G = (\mathbf{X}, E)$ est un :

- I-map (pour *independence map*) de $\mathbb{P}_{\mathbf{X}}$ si $\mathcal{I}(G) \subseteq \mathcal{I}(\mathbb{P}_{\mathbf{X}})$,
- D-map (pour *dependence map*) de $\mathbb{P}_{\mathbf{X}}$ si $\mathcal{I}(G) \supseteq \mathcal{I}(\mathbb{P}_{\mathbf{X}})$,
- P-map (pour *perfect map*) de $\mathbb{P}_{\mathbf{X}}$ si $\mathcal{I}(G) = \mathcal{I}(\mathbb{P}_{\mathbf{X}})$.

Lorsque la suppression (resp. l'ajout) d'un arc quelconque fait perdre sa propriété d'I-map ou (resp. D-map) au graphe G, il est qualifié d'I-map (resp. D-map) minimal.

Un DAG complet G_c étant un I-map pour n'importe quelle distribution puisque $\mathcal{I}(G_c) = \emptyset$, il existe toujours un I-map minimal associé à une distribution.

4.5 Réseau bayésien

Définition 4.5.1 (Structure d'un réseau bayésien)**.** Un DAG G est la structure d'un réseau bayésien pour une distribution $\mathbb{P}_{\mathbf{X}}$ si c'est un I-map.

Étant donnée une distribution et un DAG, on peut vérifier si ce dernier est un I-map minimal en utilisant le théorème suivant :

Théorème 4.5.1 (PEARL (2014))**.** Soit $\mathbb{P}_{\mathbf{X}}$ une distribution et soit G un DAG. G est un I-map de $\mathbb{P}_{\mathbf{X}}$ si et seulement si

$$\{X_i \perp\!\!\!\perp \mathbf{ND}_{X_i} \,|\, \mathbf{Pa}_{X_i}\} \subseteq \mathcal{I}(\mathbb{P}_{\mathbf{X}}) \tag{4.5}$$

C'est ce que l'on appelle la propriété de Markov locale.

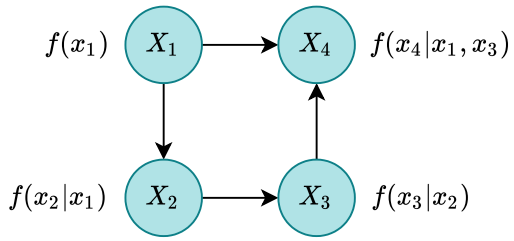


FIGURE 4.6 – Exemple d'une structure de réseau bayésien et de sa paramétrisation.

Dans la suite, nous condenserons les notations $\mathbf{ND}_{X_i}$ et $\mathbf{Pa}_{X_i}$ en $\mathbf{ND}_i$ et $\mathbf{Pa}_i$. La structure d'un réseau bayésien étant un I-map minimal pour une distribution, elle vérifie par conséquent la condition de Markov. Ces indépendances permettent de montrer que la distribution se factorise selon la structure du réseau bayésien (KOLLER et al. 2009) :

Définition 4.5.2 (Factorisation sur G)**.** Soit G une structure de réseau bayésien sur $\mathbf{X}$. Une distribution $\mathbb{P}_{\mathbf{X}}$ se factorise sur G si sa densité peut s'écrire :

$$f(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} f(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}) \tag{4.6}$$

Théorème 4.5.2 (GEIGER et PEARL (1990))**.** Si G est une structure de réseau bayésien pour une distribution $\mathbb{P}_{\mathbf{X}}$ cette dernière se factorise sur G.

Inversement, la factorisation d'une distribution sur un graphe est une condition suffisante pour que le graphe soit un I-map de la distribution :

Théorème 4.5.3 (GEIGER et PEARL (1990))**.** Soit $G = (\mathbf{X}, E)$ un DAG et soit $\mathbb{P}_{\mathbf{X}}$ une distribution. Si $\mathbb{P}_{\mathbf{X}}$ se factorise sur G alors ce dernier est une structure de réseau bayésien pour la distribution.

Ces deux derniers théorèmes sont importants puisqu'ils nous permettent de paramétrer la densité jointe comme un ensemble de d densités conditionnelles locales aux variables et à leurs parents. Ceci nous mène naturellement à la définition d'un réseau bayésien :

Définition 4.5.3 (Réseau bayésien)**.** Un réseau bayésien est une paire $\mathcal{B} = (G, \mathbb{P}_{\mathbf{X}})$ où G est une structure de réseau bayésien pour $\mathbb{P}_{\mathbf{X}}$ qui se décompose alors sur G :

$$f(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} f(x_i|\mathbf{pa}_i, \boldsymbol{\theta}) \tag{4.7}$$

Dans ce contexte, l'équation précédente est appelée règle de chaîne pour les réseaux bayésiens. La distribution $\mathbb{P}_{\mathbf{X}}$ est spécifiée par les densités conditionnelles $f(x_i|\mathbf{pa}_i, \boldsymbol{\theta})$ qui sont associées au nœud X_i dans le graphe.

Nous donnons à présent les deux modèles de réseaux bayésiens classiques pour le cas de variables aléatoires toutes discrètes et le cas de variables aléatoires toutes continues.

***Exemple* 4.5.1** (Réseau bayésien multinomial (MBN))**.** Un réseau bayésien multinomial (MBN) est un réseau bayésien dont les densités conditionnelles sont

des distributions catégorielles. Soit la structure représentée sur la figure 4.6. Chaque densité conditionnelle définit un tenseur où chaque ligne représente les paramètres $\boldsymbol{\theta}_{X_i|\mathbf{pa}_i}$ dont la somme doit être 1 :

$$\boldsymbol{\theta}_{X_1} = \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \qquad \boldsymbol{\theta}_{X_2|X_1} = \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix} \qquad \boldsymbol{\theta}_{X_3|X_2} = \begin{pmatrix} 0.4 & 0.6 \\ 0.5 & 0.5 \end{pmatrix}$$

$$\boldsymbol{\theta}_{X_4|X_1,x_3^0} = \begin{pmatrix} 0.6 & 0.4 \\ 0.8 & 0.2 \end{pmatrix} \qquad \boldsymbol{\theta}_{X_4|X_1,x_3^1} = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}$$

***Exemple* 4.5.2** (Réseau bayésien gaussien (GBN))**.** Un réseau bayésien gaussien (GBN) est un réseau bayésien dont les densités conditionnelles sont des distributions gaussiennes linéaires définies comme :

$$f(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}) = N\left(x_i; \mu_i + \boldsymbol{b}_i^T \mathbf{pa}_i, \nu_i\right) \tag{4.8}$$

où $\boldsymbol{b} = (b_1, \ldots, b_k)$ est un vecteur de paramètres et $\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i} = \left\{\mu_i, \boldsymbol{b}_i^T, \nu_i\right\}$. Un GBN de paramètres $(\boldsymbol{\mu}', \boldsymbol{b}, \boldsymbol{\nu})$ est équivalent à une distribution gaussienne multivariée de paramètres $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Koller et al. 2009). La transformation pour passer d'une paramétrisation à une autre est donnée par $\boldsymbol{\mu} = \boldsymbol{\mu}'$ et la relation de récurrence suivante (Shachter et al. 1989) :

$$\boldsymbol{\Sigma}_0 = \nu_1\,; \qquad \boldsymbol{\Sigma}_{i+1} = \begin{pmatrix} \boldsymbol{\Sigma}_i + \frac{\boldsymbol{b}_{i+1}\boldsymbol{b}_{i+1}^T}{\nu_{i+1}} & -\frac{\boldsymbol{b}_{i+1}}{\nu_{i+1}} \\ -\frac{\boldsymbol{b}_{i+1}}{\nu_{i+1}} & \frac{1}{\nu_{i+1}} \end{pmatrix} \tag{4.9}$$

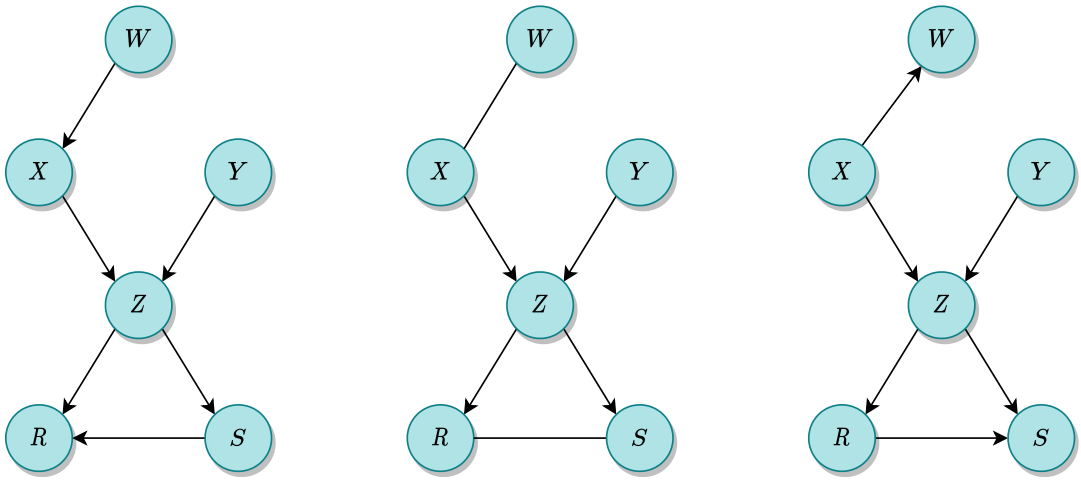
4.6 Équivalence de Markov

Maintenant que nous avons introduit les réseaux bayésiens, revenons à une remarque que nous avons fait plus haut. Nous avons dit qu'il existe toujours un I-map pour une distribution quelconque qui est le DAG complet G_c. Or il existe autant de DAGs complets de taille d qu'il existe de permutations de d objets, soit $d!$. Cela veut donc dire qu'il existe plusieurs DAGs encodant les mêmes indépendances. Cette observation ne s'arrête pas à ce cas particulier puisque dans le cas simple où $\mathcal{I}(\mathbb{P}_{\mathbf{X}}) = \{A \perp\!\!\!\perp C \,|\, B\}$ les DAGs $A \to B \to C$, $A \leftarrow B \to C$, $A \leftarrow B \leftarrow C$ sont tous des I-map et ne sont pourtant pas complets. Nous définissons alors la relation d'équivalence suivante :

Définition 4.6.1 (Markov équivalence)**.** Soit G_1 et G_2 deux DAGs. Ils sont dits Markov équivalents si $\mathcal{I}(G_1) = \mathcal{I}(G_2)$.

L'ensemble des DAGs étant à présent muni d'une relation d'équivalence, nous pouvons définir des classes d'équivalence. Soit class (G) la classe d'équivalence de Markov de G, on dit qu'un arc est contraint s'il a la même orientation pour tout DAG G' appartenant à class (G). Dans le cas contraire, il est dit réversible. Verma et al. (1990) ont démontré le théorème suivant :

Théorème 4.6.1 (Verma et al. (1990))**.** Deux DAGs G_1 et G_2 sont Markov équivalents si et seulement si ils ont le même squelette et les mêmes v-structures.



(a) DAG 1 (b) CPDAG (c) DAG 2

FIGURE 4.7 – Les deux DAGs sont équivalents et leur classe d'équivalence est représentée par l'unique CPDAG correspondant. $X \to Z \leftarrow Y$ forme une v-structure et par conséquent les arcs $X \to Z$ et $Y \to Z$ sont contraints. De même, les arcs $Z \to R$ et $Z \to S$ sont contraints puisque si l'un de ces arcs est inversé une nouvelle v-structure est créée aboutissant à une nouvelle indépendance. Enfin, les arcs $W \to X$ et $R \to S$ sont quant à eux réversibles et correspondent donc à des liens dans le CPDAG.

Par conséquent, tout arc participant à une v-structure est contraint. La réciproque n'est pas vraie puisque l'arc $Z \to R$ de la figure 4.7a est contraint mais n'appartient pas à une v-structure. Le *PDAG complété* (CPDAG) ou encore graphe essentiel est le PDAG dont les arcs correspondent à l'ensemble des arcs contraints entre les DAGs d'une même classe et dont les liens correspondent aux arcs réversibles. Il existe un unique CPDAG par classe d'équivalence, c'est-à-dire que deux DAGs de la même classe ont le même CPDAG. La figure 4.7 résume les notions que nous venons d'aborder. L'existence de classes d'équivalences et leur représentation sera importante lorsque nous nous intéresserons à l'apprentissage par contraintes.

Nous venons d'introduire le modèle des réseaux bayésiens qui encode une distribution jointe de manière compacte via sa factorisation sur un graphe. Ce graphe est l'élément central du modèle puisqu'il permet non seulement la factorisation de la distribution jointe mais aussi la lecture des indépendances qu'elle vérifie. Dans le prochain chapitre nous allons voir qu'il permet en plus l'implémentation de méthodes efficaces pour l'apprentissage des indépendances de la distribution. Ces méthodes d'apprentissage vont se diviser en deux grandes catégories en fonction du type de méthode utilisé. En effet, dans le chapitre précédent nous avons introduit des méthodes de test d'hypothèse à la fois dans le cadre fréquentiste et bayésien. Dans le premier cas les tests sont binaires et ne permettent pas de tester plusieurs modèles à la fois. Pour cela, plusieurs tests vont être menés successivement et de manière à rendre la recherche efficace : on parle dans ce cas de méthodes par contraintes. Dans le cadre bayésien en revanche, nous avons vu que cette méthode s'étendait facilement à plusieurs modèles en parallèle au travers des facteurs de Bayes. Cependant, l'espace de recherche étant trop large pour pouvoir trouver un maximum global, nous devrons avoir recours à des heuristiques : on parle dans ce cas de méthodes de scores.

Références

Bollobás, B. (2013). *Modern graph theory*. T. 184. Springer Science & Business Media (cf. p. 65).

Bondy, J. A. et Murty, U. S. R. (1976). *Graph theory with applications*. T. 290. Macmillan London (cf. p. 65).

Bretto, A., Faisant, A. et Hennecart, F. (2012). *Éléments de théorie des graphes*. Springer (cf. p. 65).

Diestel, R. (2005). « Graph theory 3rd ed ». In : *Graduate texts in mathematics* 173 (cf. p. 65).

Geiger, D. et Pearl, J. (1990). « On the logic of causal models ». In : *Machine Intelligence and Pattern Recognition*. T. 9. Elsevier, p. 3-14 (cf. p. 71).

Koller, D. et Friedman, N. (2009). *Probabilistic graphical models : principles and techniques*. MIT press (cf. p. 3, 71, 72, 75, 76, 80, 131, 154).

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Elsevier (cf. p. 5, 70, 75).

Pearl, J. et Paz, A. (1985). *Graphoids : A graph-based logic for reasoning about relevance relations*. University of California (Los Angeles). Computer Science Department (cf. p. 68).

Robinson, R. W. (1977). « Counting unlabeled acyclic digraphs ». In : *Combinatorial mathematics V*. Springer, p. 28-43 (cf. p. 67).

Shachter, R. D. et Kenley, C. R. (1989). « Gaussian influence diagrams ». In : *Management science* 35.5, p. 527-550 (cf. p. 72).

Verma, T. et Pearl, J. (1988). *Influence diagrams and d-separation*. UCLA, Computer Science Department (cf. p. 68).

Verma, T. et Pearl, J. (1990). *Equivalence and synthesis of causal models*. UCLA, Computer Science Department (cf. p. 72).

Chapitre 5

Apprentissage des réseaux bayésiens

Sommaire

Nous présentons maintenant l'application des méthodes d'inférence statistique vues dans le chapitre 3. Tout d'abord, notons que le vocabulaire change quelque peu dans ce cadre puisqu'on parle d'apprentissage plutôt que d'inférence, ce dernier terme étant réservé à ce qui est appelé prédiction en statistique [1]. Bien que cet aspect prédictif peut justifier à lui seul l'intérêt des réseaux bayésiens, il n'est pas abordé dans cette thèse qui se focalise uniquement sur la construction du modèle et non sur son application. Le lecteur intéressé peut toutefois se référer à la littérature classique sur les réseaux bayésiens (Pearl 2014 ; Koller et al. 2009 ; Darwiche 2009 ; Neapolitan 2004) pour une introduction. Fondamentalement, l'application de ces techniques ne change pas mais la propriété de factorisation de la densité jointe permet de décomposer le problème en plusieurs sous-problèmes de moindre complexité qui restent malgré tout NP-difficile (Chickering 1996).

Nous commençons par supposer que la structure du réseau bayésien est connue et présentons l'apprentissage des paramètres dans les réseaux bayésiens. Nous allons voir en particulier que la factorisation de la densité jointe implique la factorisation de la vraisemblance. Nous relâcherons ensuite l'hypothèse de la connaissance de la structure du réseau bayésien et verrons les méthodes existantes pour pouvoir la construire à partir d'un échantillon de donnée. Nous supposerons que cet échantillon est *i.i.d* et complet, c'est-à-dire qu'aucune donnée n'est manquante et qu'il n'existe pas de variables cachées.

1. Ce changement de terminologie s'explique du fait que les réseaux bayésiens prennent leur origine dans le domaine de l'intelligence artificielle.

5.1 Apprentissage des paramètres

Nous commençons par rappeler les notations utilisées pour les paramètres du BN : $\boldsymbol{\theta}$ représente l'ensemble de ses paramètres et $\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}$ l'ensemble des paramètres de la densité conditionnelle associée à la variable X_i. Dans la suite, nous faisons l'hypothèse que chaque densité conditionnelle est paramétrée avec des paramètres $\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}$ qui sont indépendants entre eux [2].

***Exemple* 5.1.1.** Dans le cas de deux variables aléatoires discrètes binaires dont la structure est $X_1 \to X_2$ (cas de l'hypothèse H_1 dans l'exemple 3.4.4) nous avons 6 paramètres $\boldsymbol{\theta} = (\boldsymbol{\theta}_{X_1} = (\theta_{x_1^0}, \theta_{x_1^1}), \boldsymbol{\theta}_{X_2|x_1^0} = (\theta_{x_2^0|x_1^0}, \theta_{x_2^1|x_1^0}), \boldsymbol{\theta}_{X_2|x_1^1} = (\theta_{x_2^0|x_1^1}, \theta_{x_2^1|x_1^1}))$ dont seulement 3 sont indépendants puisque $\theta_{x_1^0} + \theta_{x_1^1} = \theta_{x_2^0|x_1^0} + \theta_{x_2^1|x_1^0} = \theta_{x_2^0|x_1^1} + \theta_{x_2^1|x_1^1} = 1$.

Montrons à présent que la factorisation de la densité jointe entraîne la factorisation de la vraisemblance :

$$\begin{aligned} f(\boldsymbol{d}|\boldsymbol{\theta}, G) &= \prod_{j=1}^{m} f(\mathbf{x}[j]|\boldsymbol{\theta}, G) = \prod_{i=1}^{n} \left(\prod_{j=1}^{m} f(x_i[j]|\mathbf{pa}_i[j], \boldsymbol{\theta}_{X_i|\mathbf{pa}_i}) \right) \\ &= \prod_{i=1}^{n} f_{X_i|\mathbf{Pa}_i}(\boldsymbol{d}_i|\boldsymbol{\theta}_{X_i|\mathbf{pa}_i}) \end{aligned} \tag{5.1}$$

où $\boldsymbol{d}_i$ ($\boldsymbol{d}_i^G$ lorsque c'est nécessaire) est la restriction de l'échantillon $\boldsymbol{d}$ aux variables $X_i \cup \mathbf{Pa}_i$ et où $f_{X_i|\mathbf{Pa}_i}(\boldsymbol{d}_i|\boldsymbol{\theta}_{X_i|\mathbf{pa}_i})$ est la vraisemblance locale associée à la variable X_i. De la même manière, la densité *a posteriori* peut se factoriser sur la structure du BN mais pour cela il doit en être de même pour la densité *a priori*. On dit dans ce cas que la densité *a priori* vérifie la propriété d'indépendance globale des paramètres :

Définition 5.1.1 (Indépendance globale des paramètres)**.** Soit $\mathcal{B} = (G, \boldsymbol{\theta})$ un BN avec $\boldsymbol{\theta} = (\boldsymbol{\theta}_{X_1|\mathbf{Pa}_1}, \ldots, \boldsymbol{\theta}_{X_n|\mathbf{Pa}_n})$. Un *a priori* π sur les paramètres vérifie la propriété d'indépendance globale des paramètres si :

$$\pi(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_i(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}). \tag{5.2}$$

Ainsi, sous l'hypothèse de l'indépendance globale des paramètres, la distribution *a posteriori* se factorise également sur la structure du BN :

$$\begin{aligned} \rho(\boldsymbol{\theta}|\boldsymbol{d}, G) &= \frac{f(\boldsymbol{d}|\boldsymbol{\theta}, G)\pi(\boldsymbol{\theta}|G)}{f(\boldsymbol{d}|G)} = \frac{\prod_{i=1}^{n} f_{X_i|\mathbf{Pa}_i}(\boldsymbol{d}_i|\boldsymbol{\theta}_{X_i|\mathbf{pa}_i}) \prod_{i=1}^{n} \pi_i(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i})}{\prod_{i=1}^{n} f_{X_i|\mathbf{Pa}_i}(\boldsymbol{d}_i)} \\ &= \prod_{i=1}^{n} \rho_i(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}|\boldsymbol{d}_i) \end{aligned}$$

La recherche d'un estimateur bayésien pour $\boldsymbol{\theta}$ peut alors se faire en trouvant de manière indépendante un estimateur pour chaque $\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}$ à partir des densités *a posteriori* ρ_i. Pour cela, nous devons néanmoins définir une densité *a priori* pour chaque groupe de paramètres $\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}$ ce qui peut être fastidieux si le nombre de paramètres est grand.

2. Cette hypothèse peut toutefois être relâchée comme le montre la section 17.5 de Koller et al. (2009).

Pour simplifier cette tâche, nous pouvons prendre des *a priori* uniformes ce qui, comme nous l'avons vu dans le chapitre 3, revient à utiliser la méthode du maximum de vraisemblance. Une autre solution est d'utiliser des *a priori* conjugués mais dans ce cas, la valeur des paramètres reste à déterminer. Dans le cas des MBNs, rappelons que la loi conjuguée de $\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}$ est la loi de Dirichlet de paramètres $\boldsymbol{\alpha}_{X_i|\mathbf{Pa}_i}$ (sous-section 1.6.4). Il existe plusieurs manières de fixer ces paramètres qui aboutissent à plusieurs *a priori* classiques. La plus triviale est de fixer tous les paramètres égaux à une même constante c. Le cas $\alpha_{x_i^j|\mathbf{pa}_i^{\boldsymbol{k}}} = 1$ correspond à ce que l'on appelle l'*a priori* K2 (Cooper et Herskovits 1992). Nous allons cependant voir dans la prochaine section que cet *a priori* ne nous permet pas d'obtenir un score structurel équivalent (définition 5.2.1) et pour cette raison, l'*a priori* BDe (pour *Bayesian Dirichlet equivalent*, (Heckerman et al. 1995)) lui est parfois préféré pour l'apprentissage de la structure. Ce dernier repose sur la notion d'échantillon d'observation virtuel $\boldsymbol{d}'$ que nous avions mentionné lors de l'exemple 3.3.1 : le paramètre $\alpha_{x_i^j|\mathbf{pa}_i^{\boldsymbol{k}}}$ peut être vu comme le nombre de fois, noté $\alpha[x_i^j, \mathbf{pa}_i^{\boldsymbol{k}}]$, où la configuration $(x_i^j, \mathbf{pa}_i^{\boldsymbol{k}})$ apparaît dans $\boldsymbol{d}'$. Plutôt que d'utiliser le nombre d'occurrence, on peut utiliser la taille de l'échantillon virtuel α et la densité virtuelle $f'(x_i, \mathbf{pa}_i)$ définie comme la fréquence d'occurrence de chaque configuration dans $\boldsymbol{d}'$. Ainsi si on définit une densité virtuelle f' les paramètres sont définis comme

$$\alpha_{x_i^j|\mathbf{pa}_i^{\boldsymbol{k}}} = \alpha f'(x_i^j, \mathbf{pa}_i^{\boldsymbol{k}}) \tag{5.3}$$

Dans le cas où la densité f' est uniforme, c'est-à-dire lorsque :

$$f'(x_i^j, \mathbf{pa}_i^{\boldsymbol{k}}) = \frac{1}{|\boldsymbol{\Theta}_{X_i}|\,|\boldsymbol{\Theta}_{\mathbf{Pa}_i}|}, \tag{5.4}$$

les $\alpha_{x_i^j|\mathbf{pa}_i^{\boldsymbol{k}}}$ prennent la même valeur quelques soient j et $\boldsymbol{k}$ et on parle d'*a priori* BDeu (pour *Bayesian Dirichlet equivalent uniform*, (Buntine 1991)).

***Exemple* 5.1.2.** Le problème pouvant être décomposé, il nous suffit d'appliquer les formules dérivées dans le chapitre précédent pour chaque paramètre. Dans le cas des MBNs, nous utilisons les résultats de l'exemple 3.3.1 et nous avons :

$$\hat{\theta}^{ML}_{x_i^k|\mathbf{pa}_i^{\boldsymbol{l}}} = \frac{m[x_i^k, \mathbf{pa}_i^{\boldsymbol{l}}]}{m[\mathbf{pa}_i^{\boldsymbol{l}}]} \tag{5.5}$$

pour l'estimateur ML et

$$\hat{\theta}^{\mathrm{L2}}_{x_i^k|\mathbf{pa}_i^{\boldsymbol{l}}} = \frac{m[x_i^k, \mathbf{pa}_i^{\boldsymbol{l}}] + \alpha_{x_i^k|\mathbf{pa}_i^{\boldsymbol{l}}}}{m[\mathbf{pa}_i^{\boldsymbol{l}}] + \alpha_{\mathbf{pa}_i^{\boldsymbol{l}}}}$$

$$\hat{\theta}^{\mathrm{MAP}}_{x_i^k|\mathbf{pa}_i^{\boldsymbol{l}}} = \frac{m[x_i^k, \mathbf{pa}_i^{\boldsymbol{l}}] + \alpha_{x_i^k|\mathbf{pa}_i^{\boldsymbol{l}}} - 1}{m[\mathbf{pa}_i^{\boldsymbol{l}}] + \alpha_{\mathbf{pa}_i^{\boldsymbol{l}}} - \left|\boldsymbol{\Theta}_{X_i|\mathbf{Pa}_i}\right|}$$

pour les estimateurs bayésiens moyens et MAP avec un *a priori* de Dirichlet. Il suffit ensuite d'injecter les valeurs des paramètres selon le type d'*a priori* utilisé. Dans le cas des GBNs que nous avons présentés plus tôt, les paramètres peuvent être appris en utilisant l'estimateur de maximum de vraisemblance (voir 3.3.2) ou bien en utilisant l'équivalence entre GBNs et distributions gaussiennes multivariées et en utilisant par exemple un a priori Normal-Inverse-Wishart.

5.2 Apprentissage de la structure

Nous nous plaçons à présent dans le cas où la structure du BN n'est pas connue et nous intéressons aux méthodes permettant sa reconstruction à partir des données. Ces méthodes reposent sur le test d'hypothèses et plus particulièrement sur la sélection de modèle. Elles se divisent en deux grandes catégories : les méthodes basées sur une fonction de score et les méthodes basées sur des contraintes. Les premières reposent sur une approche bayésienne permettant la comparaison simultanée de plusieurs modèles et sur une recherche locale dans l'espace des DAGs tandis que les deuxièmes reposent en général sur l'approche classique du test d'hypothèses pour la recherche d'un CPDAG. Nous présentons dans cette section plusieurs méthodes classiques d'apprentissage des réseaux bayésiens qui seront plus tard étendues aux réseaux bayésiens à base de copules qui font l'objet du prochain chapitre.

5.2.1 Apprentissage basé sur une fonction de score

L'apprentissage avec fonction de score consiste à appliquer les méthodes de la section 3.4.2 au contexte des BNs. En particulier, nous avions vu avec l'exemple 3.4.4 comment tester l'indépendance entre deux variables via la factorisation de la densité jointe. De la même manière, nous pouvons utiliser la factorisation d'une densité induite par la structure d'un BN pour comparer plusieurs structures candidates et sélectionner la plus adaptée aux données.

5.2.1.1 Score bayésien

Pour cela, nous n'utilisons pas directement les facteurs de Bayes mais le score bayésien d'une structure G,

$$\mathcal{S}_B(G;\boldsymbol{d}) = \log f(\boldsymbol{d}|G) + \log P(G), \tag{5.6}$$

qui est, à une constante près, le logarithme de $P(G|\boldsymbol{d})$. L'utilisation d'autres scores est possible, comme le score BIC, lorsque la vraisemblance marginale ne possède pas d'expression analytique.

***Exemple* 5.2.1** (Score bayésien pour un MBN)**.** Soit $\mathbf{X}$ un vecteur aléatoire discret de dimension n et dont les composantes X_i prennent leurs valeurs dans Ω_{X_i}. Dans le cas où sa densité jointe est modélisée par un MBN, la vraisemblance d'un ensemble de réalisations $\boldsymbol{d}$ s'écrit :

$$\begin{aligned} f(\boldsymbol{d}|\boldsymbol{\theta}^G, G) &= \prod_{i=1}^{n} \left(\prod_{j=1}^{m} f(x_i[j]|\mathbf{pa}_i[j], \boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}) \right) \\ &= \prod_{i=1}^{n} \prod_{\mathbf{pa}_i^{\boldsymbol{l}} \in \Omega_{\mathbf{Pa}_i}} \left(\prod_{x_i^k \in \Omega_{X_i}} \theta_{x_i^k|\mathbf{pa}_i^{\boldsymbol{l}}}^{m[x_i^k, \mathbf{pa}_i^{\boldsymbol{l}}]} \right) \end{aligned}$$

Dans le cas des MBNs les paramètres $\boldsymbol{\theta}_{X_i|\mathbf{pa}_i^{\boldsymbol{j}}}$ sont indépendants pour différentes valeurs de $\mathbf{pa}_i^{\boldsymbol{j}}$ ce qui nous permet une décomposition supplémentaire du problème. Pour cela, nous devons faire l'hypothèse qu'en plus de vérifier la propriété de décomposition globale, la densité *a priori* vérifie la propriété de décomposition locale :

$$\pi(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}) = \prod_{\mathbf{pa}_i^{\boldsymbol{j}} \in \Omega_{\mathbf{Pa}_i}} \pi(\boldsymbol{\theta}_{X_i|\mathbf{pa}_i^{\boldsymbol{j}}}) \tag{5.7}$$

En choisissant alors des densités *a priori* selon une loi de Dirichlet (cf. 1.6.4.3),

$$\pi(\boldsymbol{\theta}_{X_i|\mathbf{pa}_i^j}) = \frac{1}{B(\boldsymbol{\alpha}_{X_i|\mathbf{pa}_i^j})} \prod_{x_i^k \in \Omega_{X_i}} \left[\theta_{x_i^k|\mathbf{pa}_i^j}\right]^{\alpha_{x_i^k|\mathbf{pa}_i^j}-1} \tag{5.8}$$

la vraisemblance marginale pour la structure G a pour expression

$$\begin{aligned} f(\boldsymbol{d}|G) &= \int_{\boldsymbol{\Theta}} f(\boldsymbol{d}|\boldsymbol{\theta}, G)\pi(\boldsymbol{\theta}|G)\mathrm{d}\boldsymbol{\theta} \\ &= \prod_{i=1}^{n} \prod_{\mathbf{pa}_i^l \in \Omega_{\mathbf{Pa}_i}} \frac{1}{B(\boldsymbol{\alpha}_{X_i|\mathbf{pa}_i^j})} \int_{\boldsymbol{\Theta}_{X_i|\mathbf{pa}_i^j}} \prod_{x_i^k \in \Omega_{X_i}} \theta_{x_i^k|\mathbf{pa}_i^l}^{m[x_i^k,\mathbf{pa}_i^l]+\alpha_{x_i^k|\mathbf{pa}_i^j}-1} \mathrm{d}\boldsymbol{\theta}_{X_i|\mathbf{pa}_i^l} \\ &= \prod_{i=1}^{n} \prod_{\mathbf{pa}_i^l \in \Omega_{\mathbf{Pa}_i}} \frac{B(\boldsymbol{\alpha}'_{X_i|\mathbf{pa}_i^l})}{B(\boldsymbol{\alpha}_{X_i|\mathbf{pa}_i^j})} \end{aligned}$$

où $\boldsymbol{\alpha}'^k_{X_i|\mathbf{pa}_i^l} = m[x_i^k, \mathbf{pa}_i^l] + \alpha_{x_i^k|\mathbf{pa}_i^j}$. Finalement, le score bayésien pour un MBN avec des densités provenant de la loi de Dirichlet, appelé score BD (pour *Bayesian Dirichlet*) s'écrit

$$\mathcal{S}_B(G_i; \boldsymbol{d}) = \sum_{i=1}^{n} \sum_{\mathbf{pa}_i^l \in \Omega_{\mathbf{Pa}_i}} \left(\log B(\boldsymbol{\alpha}'_{X_i|\mathbf{pa}_i^l}) - \log B(\boldsymbol{\alpha}_{X_i|\mathbf{pa}_i^j})\right). \tag{5.9}$$

L'écriture classique est retrouvée en injectant la relation entre fonction bêta et fonction gamma (cf. équation 1.29).

Certains DAG étant équivalents, une propriété souhaitable pour la fonction de score est que celle-ci prenne la même valeur pour de tels graphes :

Définition 5.2.1 (Équivalence)**.** Une fonction de score $\mathcal{S}$ vérifie la propriété d'équivalence si elle prend la même valeur pour deux graphes appartenant à la même classe d'équivalence.

Comme nous l'avons vu plus haut, pour que le score bayésien possède cette propriété, les densités *a priori* $\pi_i(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i}|G)$ et plus précisément leurs paramètres, doivent vérifier certaines contraintes. Le score BIC, quant à lui, néglige les effets des *a priori* et vérifie donc la propriété d'équivalence.

***Exemple* 5.2.2.** Dans l'exemple précédent, nous avions laissé la valeur des paramètres des densités *a priori* indéterminée. Nous avons vu précédemment qu'une solution était l'utilisation de l'*a priori* K2, fixant tous les paramètres à 1. Soient X_1 et X_2 deux variables binaires et soient $G_1 : X_1 \to X_2$ et $G_2 : X_2 \to X_1$ deux structures équivalentes. On voit avec ce cas simple que l'*a priori* K2 ne confère pas la propriété d'équivalence au score bayésien puisque :

- $f(\boldsymbol{d}|G_1) = \frac{m[x_1^0]!m[x_1^1]!}{(m+1)!} \frac{m[x_1^0,x_2^0]!m[x_1^0,x_2^1]!}{(m[x_1^0]+1)!} \frac{m[x_1^1,x_2^0]!m[x_1^1,x_2^1]!}{(m[x_1^1]+1)!}$,
- $f(\boldsymbol{d}|G_2) = \frac{m[x_2^0]!m[x_2^1]!}{(m+1)!} \frac{m[x_1^0,x_2^0]!m[x_1^1,x_2^0]!}{(m[x_2^0]+1)!} \frac{m[x_1^0,x_2^1]!m[x_1^1,x_2^1]!}{(m[x_2^1]+1)!}$,

au contraire de l'*a priori* BDe [a] (Heckerman et al. 1995) :

- $f(\boldsymbol{d}|G_1) = \frac{\Gamma(\alpha)}{\Gamma(m+\alpha)} \frac{\Gamma(m[x_1^0,x_2^0]+\frac{\alpha}{4})}{\Gamma(\frac{\alpha}{4})} \frac{\Gamma(m[x_1^0,x_2^1]+\frac{\alpha}{4})}{\Gamma(\frac{\alpha}{4})} \frac{\Gamma(m[x_1^1,x_2^0]+\frac{\alpha}{4})}{\Gamma(\frac{\alpha}{4})} \frac{\Gamma(m[x_1^1,x_2^1]+\frac{\alpha}{4})}{\Gamma(\frac{\alpha}{4})}$,

- $f(\boldsymbol{d}|G_2) = \frac{\Gamma(\alpha)}{\Gamma(m+\alpha)} \frac{\Gamma(m[x_1^0,x_2^0]+\frac{\alpha}{4})}{\Gamma(\frac{\alpha}{4})} \frac{\Gamma(m[x_1^1,x_2^0]+\frac{\alpha}{4})}{\Gamma(\frac{\alpha}{4})} \frac{\Gamma(m[x_1^0,x_2^1]+\frac{\alpha}{4})}{\Gamma(\frac{\alpha}{4})} \frac{\Gamma(m[x_1^1,x_2^1]+\frac{\alpha}{4})}{\Gamma(\frac{\alpha}{4})}$.

Dans le cas des GBNs, Geiger et Heckerman (1994) ont dérivé le score BGe (pour *Bayesian Gaussian equivalent*) possédant la propriété d'équivalence.

a. Pour simplifier nous prenons l'*a priori* BDeu vu plus haut.

5.2.1.2 Méthode de recherche locale

Rechercher l'estimateur MAP revient à maximiser le score bayésien sur l'ensemble des DAGs. Rappelons cependant que la taille de cet ensemble est super-exponentielle par rapport au nombre de variables n sur lesquelles les DAGs sont définis (Figure 4.3). Il est donc en général difficile de trouver un maximum global et nous avons alors recours à la méthode de recherche locale du Hill-Climbing [3]. Certains des minima locaux peuvent être évités en réalisant plusieurs itérations de l'algorithme avec différentes conditions initiales ou bien en utilisant des méta-heuristiques comme celles de la recherche TABU (Glover et al. 1998) ou du recuit-simulé (Kirkpatrick et al. 1983).

La définition d'une recherche locale se fait par la donnée de :

- un ensemble d'états Φ,
- une fonction de score $\mathcal{S}$,
- un état initial ϕ_0
- un ensemble d'opérateurs O permettant d'obtenir les voisins d'un état.

Partant de l'état initial, ses voisins sont obtenus par application des opérateurs de O et leur score est calculé. Si certains voisins ont un score plus élevé que l'état initial, celui dont le score est maximal est sélectionné et la procédure est réitérée jusqu'à que le score ne puisse plus être amélioré. De cette manière, nous obtenons un maximum mais sans garantie que ce dernier soit global. Les principales différences entre les méthodes que nous avons citées résident dans leurs stratégies pour éviter d'être piégées au niveau d'un plateau ou d'un maximum local. La méthode générique que nous venons de décrire est résumée par l'algorithme 1.

Dans notre cas, l'espace des états est l'ensemble des DAGs sur $\mathbf{X}$ et la fonction de score est le score bayésien. En général, l'état initial est le graphe vide ou bien un DAG choisi aléatoirement. Enfin pour ce qui est du voisinage d'un DAG, il est obtenu en ajoutant, supprimant ou retournant un arc sous la contrainte que le graphe obtenu soit acyclique. Bien qu'en partant du graphe vide seul l'ajout d'un arc semble pertinent, la suppression et le retournement d'un arc permettent d'éviter certains des maxima locaux (Koller et al. 2009, chapitre 18).

Cette méthode étant heuristique, on ne peut pas avoir un estimation exacte de sa complexité. Nous pouvons toutefois faire une estimation grossière de la complexité d'une itération. Pour chacune de ces étapes, nous devons appliquer $\mathcal{O}(n^2)$ opérateurs pour trouver le voisinage du graphe considéré. Pour chaque voisin, nous devons vérifier qu'il est acyclique, $\mathcal{O}(nq)$, et calculer les statistiques suffisantes $\mathcal{O}(m)$. La complexité d'une étape est donc $\mathcal{O}(n^2(nq+m))$. Ainsi, on voit que ces algorithmes vont évaluer un grand nombre de fois la fonction de score sur des graphes qui sont proches. Cette complexité est grandement réduite si la fonction de score est décomposable :

3. Notons toutefois qu'il existe des méthodes de recherche exactes permettant l'apprentissage de structures pouvant contenir jusqu'à 100 variables (Bartlett et al. 2017 ; Trösser et al. 2021).

Algorithm 1: Recherche locale

Input: Φ un ensemble d'états, ϕ_0 un état initial , $\mathcal{S}$ une fonction de score, O un ensemble d'opérateurs
Result: Maximum ϕ^*

```
φ* ← φ0
Progress ← True
while Progress do
    φ ← φ*
    Progress ← False
    forall o ∈ O do
        φ' = o(φ)
        if φ' ∈ Φ then
            if S(φ') > S(φ*) then
                φ* ← φ'
                Progress ← True
            end
        end
    end
end
```

Définition 5.2.2 (Décomposabilité)**.** Une fonction de score $\mathcal{S}$ est décomposable si elle peut s'écrire comme la somme de fonctions de score locales s_i portant sur chaque nœud et ses parents dans le graphe G :

$$\mathcal{S}(G; \boldsymbol{d}) = \sum_{i=1}^{n} s_i(X_i, \mathbf{Pa}_i; \boldsymbol{d}). \tag{5.10}$$

En effet, le score d'un voisin G' n'a plus besoin d'être calculé sur l'entièreté du graphe mais seulement sur les familles, c'est-à-dire les couples $(X_i, \mathbf{Pa}_i)$, qui ont été modifiées par l'application de l'opérateur. Il suffit alors de calculer la variation du score associée à cette modification :

$$\begin{aligned}
\Delta\mathcal{S}_o &= \mathcal{S}(G'; \boldsymbol{d}) - \mathcal{S}(G; \boldsymbol{d}) \\
&= \sum_{i=1}^{n} \left[s_i(X_i, \mathbf{Pa}_i^{G'}; \boldsymbol{d}) - s_i(X_i, \mathbf{Pa}_i^{G}; \boldsymbol{d}) \right] \\
&= \sum_{i:\mathbf{Pa}_i^G \neq \mathbf{Pa}_i^{G'}} \left[s_i(X_i, \mathbf{Pa}_i^{G'}; \boldsymbol{d}) - s_i(X_i, \mathbf{Pa}_i^{G}; \boldsymbol{d}) \right]
\end{aligned}$$

Dans la dernière expression mais aussi avec les notations utilisées jusqu'ici, nous avons fait l'hypothèse implicite que la valeur des paramètres ne dépendait que de la structure du graphe, c'est-à-dire que $\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i^G} = \boldsymbol{\theta}_{X_i|\mathbf{Pa}_i^{G'}}$ si $\mathbf{Pa}_i^G = \mathbf{Pa}_i^{G'}$.

Définition 5.2.3 (Modularité des paramètres)**.** Soient (G, G') un couple quelconque de DAG sur $\mathbf{X}$. Si la variable X_i a les même parents dans les deux structures, $\mathbf{Pa}_i^G = \mathbf{Pa}_i^{G'}$, alors :

$$\pi(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i^G}) = \pi(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_i^{G'}}). \tag{5.11}$$

Pour les trois opérations utilisées, cette somme se réduit à un terme pour l'addition et la suppression d'un arc et à deux termes pour le retournement d'un arc :

Proposition 5.2.1. Soient G et $G' = o(G)$ deux DAGs et soit $\mathcal{S}$ une fonction de score décomposable.

- Si o correspond à l'ajout de l'arc $X_i \to X_j$, alors

$$\Delta\mathcal{S}_o = s_j(X_j, \mathbf{Pa}_j^G \cup \{X_i\}\,; \boldsymbol{d}) - s_j(X_j, \mathbf{Pa}_j^G; \boldsymbol{d}) \tag{5.12}$$

- Si o correspond à la suppression de l'arc $X_i \to X_j$, alors

$$\Delta\mathcal{S}_o = s_j(X_j, \mathbf{Pa}_j^G \setminus \{X_i\}\,; \boldsymbol{d}) - s_j(X_j, \mathbf{Pa}_j^G; \boldsymbol{d}) \tag{5.13}$$

- Si o correspond au retournement de l'arc $X_i \to X_j$, alors

$$\begin{aligned}\Delta\mathcal{S}_o =& s_i(X_i, \mathbf{Pa}_i^G \cup \{X_j\}\,; \boldsymbol{d}) + s_j(X_j, \mathbf{Pa}_j^G \setminus \{X_i\}\,; \boldsymbol{d}) \\ & - s_i(X_i, \mathbf{Pa}_i^G; \boldsymbol{d}) - s_j(X_k, \mathbf{Pa}_j^G; \boldsymbol{d})\end{aligned}$$

Notons au passage que si, comme ce sera le cas par la suite, l'*a priori* sur les structures est uniforme [4], la différence de score entre deux graphes G_i et G_j correspond au logarithme de leur facteur de Bayes :

$$\Delta\mathcal{S}_{ij} = \mathcal{S}_B(G_j; \boldsymbol{d}) - \mathcal{S}(G_i; \boldsymbol{d}) = \log \mathrm{B}_{ji}(\boldsymbol{d}).$$

***Exemple* 5.2.3** (Décomposition du score BIC)**.** Pour rappel, le score BIC est composé de deux parties. La première est la log-vraisemblance évaluée en son maximum $\hat{\boldsymbol{\theta}}$ et la deuxième est une pénalité en terme de complexité du modèle. Nous avons déjà vu que la vraisemblance se décomposait sur la structure mais dans le cas d'un MBN, nous pouvons exprimer celle-ci en fonction de l'information mutuelle empirique :

$$\begin{aligned}\log f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, G) &= \sum_{i=1}^{n} \sum_{\mathbf{pa}_i^l \in \Omega_{\mathbf{Pa}_i}} \left(\sum_{x_i^k \in \Omega_{X_i}} m[x_i^k, \mathbf{pa}_i^l] \log \hat{\theta}_{x_i^k|\mathbf{pa}_i^l} \right) \\ &= \sum_{i=1}^{n} \sum_{\mathbf{pa}_i^l \in \Omega_{\mathbf{Pa}_i}} \left(\sum_{x_i^k \in \Omega_{X_i}} m[x_i^k, \mathbf{pa}_i^l] \log \frac{m[x_i^k, \mathbf{pa}_i^l]}{m[\mathbf{pa}_i^l]} \right) \\ &= m \sum_{i=1}^{n} \sum_{\mathbf{pa}_i^l \in \Omega_{\mathbf{Pa}_i}} \left(\sum_{x_i^k \in \Omega_{X_i}} \hat{f}(x_i^k, \mathbf{pa}_i^l) \log \hat{f}(x_i^k|\mathbf{pa}_i^l) \right) \\ &= m \left(\sum_{i=1}^{n} I_{\hat{f}}(X_i; \mathbf{Pa}_i^G) - \sum_{i=1}^{n} H_{\hat{f}}(X_i) \right).\end{aligned}$$

où nous avons utilisé le fait que les estimateurs ML ont pour expression $\hat{\theta}_{x_i^k|\mathbf{pa}_i^l} = \frac{m[x_i^k, \mathbf{pa}_i^l]}{m[\mathbf{pa}_i^l]}$. On peut voir ici que le second terme de la décomposition faisant intervenir l'entropie ne dépend pas de la structure et n'interviendra donc pas dans le calcul de $\Delta\mathcal{S}_{ij}$. Cela généralise alors ce que nous avions vu dans l'exemple 3.4.5 lorsque nous avions exprimé le facteur de Bayes en fonction de l'information mutuelle de la densité empirique (équation (3.67)).

4. Puisque nous testerons plus tard ces méthodes sur des données dont nous connaissons la structure, l'ajout d'information n'aurait pas de sens et nous choisissons pour cela un *a priori* non-informatif.

Tout comme nous venons de le voir pour le score BIC, il est aisé de montrer que le score bayésien (et donc K2, BDe et BDeu) est décomposable à partir de l'équation (5.9), ce qui nous permet d'appliquer la méthode de recherche locale de manière efficace pour ces scores.

5.2.2 Apprentissage par contraintes

Nous venons de voir l'approche bayésienne pour l'apprentissage de la structure d'un BN. Celle-ci utilise des fonctions de score permettant de tester en simultané plusieurs modèles. Ci-dessous, nous montrons que les méthodes d'apprentissage par contraintes reposent en général sur l'approche classique des tests d'hypothèses ou du moins sur des tests binaires. De plus, l'espace de recherche n'est plus celui des DAGs mais celui des classes d'équivalence, c'est-à-dire des CPDAGs.

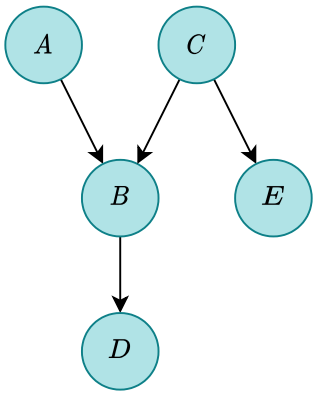
5.2.2.1 L'algorithme PC

En supposant qu'il existe un DAG G^* qui soit un P-map pour la distribution, les DAGs appartenant à sa classe d'équivalence possèdent le même squelette et les mêmes v-structures (théorème 4.6.1). Pour cette raison, les méthodes basées sur contraintes se décomposent de la manière suivante : 1) recherche du squelette, 2) recherche des v–structures et 3) propagation des contraintes. Ces trois étapes permettent de retrouver le CPDAG représentant la classe d'équivalence de G^* et n'importe quel DAG appartenant à cette classe pourra ensuite être sélectionné pour construire un BN. La recherche du squelette repose sur le fait que G^* étant un P-map de la distribution génératrice, il existe un lien entre deux variables X_i et X_j de G^* si et seulement si il n'existe pas d'ensemble *conditionnant* $\mathbf{U} \subseteq \mathbf{X}$, $X_i \perp\!\!\!\perp X_j \,|\, \mathbf{U}$. Ainsi en partant du graphe non-orienté complet, c'est-à-dire du modèle pour lequel il n'existe aucune indépendance, nous pouvons supprimer un lien entre X_i et X_j si nous trouvons un ensemble $\mathbf{U} \subseteq \mathbf{X}$ tel que $X_i \perp\!\!\!\perp X_j \,|\, \mathbf{U}$. Cependant, comme le montre l'exemple suivant, si $\mathbf{U}$ ne d-sépare pas X_i et X_j cela peut être le cas d'un sous-ensemble ou d'un sur-ensemble.

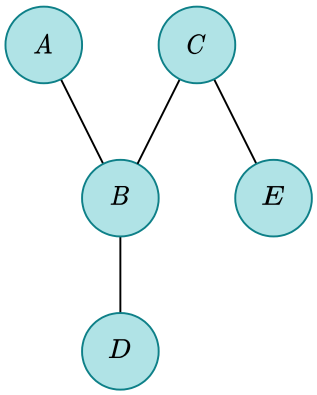
> ***Exemple* 5.2.4.** Soient trois variables aléatoires X, Y et Z. Considérons tout d'abord le cas où leur distribution jointe admet pour P-map la v-structure $X \to Z \leftarrow Y$. L'ensemble $\mathbf{U}_1 = \{Z\}$ ne d-sépare pas X et Y alors que c'est le cas de l'ensemble $\mathbf{U}_2 = \emptyset \subset \mathbf{U}_1$. Si à présent la distribution jointe admet pour P-map $X \to Z \to Y$ alors $\mathbf{U}_2$ ne d-sépare pas X et Y tandis que $\mathbf{U}_1 \supset \mathbf{U}_2$ les d-sépare.

Les tests d'indépendances étant plus robustes pour des ensembles conditionnant de petite taille, la recherche d'un ensemble d-séparant deux variables X_i et X_j est organisée par ensembles de taille croissante. L'ensemble ainsi obtenu, appelé ensemble *séparateur* et noté $\text{Sepset}(X_i, X_j)$, est minimal au sens qu'aucun de ses sous-ensembles ne d-sépare X_i et X_j. Il existe plusieurs ensembles séparateurs pour un couple de variables donné et l'ensemble séparateur sélectionné dépend de l'ordre dans lequel sont testés les ensembles conditionnant de même taille. Dans le cas où nous n'avons pas une information parfaite sur les indépendances de la distribution sous-jacente, nous allons voir que ceci a des conséquences sur le DAG appris. Pour cette raison, nous devons spécifier un ordre $\prec$ sur les variables en entrée de l'algorithme qui induit par la suite un ordre sur les liens et sur les ensembles conditionnant avec l'ordre lexicographique.

> ***Exemple* 5.2.5.** Soit $\mathbf{X} = (A, B, C, D, E)$ un vecteur aléatoire dont la distribution $\mathbb{P}_{\mathbf{X}}$ admet pour P-map le DAG donné sur la figure 5.1a. Nous appliquons la recherche du squelette que nous venons de détailler jusqu'à présent en choisissant



(a) P-map



(b) Squelette

FIGURE 5.1 – Le P-map d'une distribution et le squelette reconstruit à l'aide de la méthode par contrainte décrite plus haut en utilisant l'ordre $A \prec B \prec C \prec D \prec E$.

l'ordre $A \prec B \prec C \prec D \prec E$. L'algorithme démarre avec le graphe non-orienté complet sur $\mathbf{X}$. Le premier lien à être testé, étant donné $\prec$, est le lien $A - B$. La recherche d'une indépendance $A \perp\!\!\!\perp B \,|\, \mathbf{U}$ se fait par ensembles de taille $l = |\mathbf{U}|$ croissante et le premier à être testé est donc toujours l'ensemble vide ($l = 0$). Puis sont testés les ensembles de taille $l = 1$ en suivant l'ordre $\prec$, c'est-à-dire $\{C\}$, puis $\{D\}$ puis $\{E\}$. Ceci jusqu'à qu'une indépendance soit trouvée ou bien que l'ensemble maximal, ici $\mathbf{U} = \{C, D, E\}$, ait été testé. Comme A et B sont dépendants, tous les ensembles sont testés et le lien reste dans le squelette. Le deuxième lien testé est le lien $A - C$ qui est supprimé puisque $A \perp\!\!\!\perp C$ et donc $\text{Sepset}(A, C) = \emptyset$. De la même manière, les liens $A - D$, $A - E$, $B - E$, $C - D$, et $D - E$ sont supprimés avec respectivement pour ensembles séparateurs $\{B\}$, $\emptyset$, $\{C\}$, $\{B\}$ et $\{B\}$. Le squelette ainsi retrouvé est celui donné sur la figure 5.1b. Il correspond bien au squelette du P-map de la distribution.

Bien que dans l'exemple précédent l'algorithme ait permis de retrouver le squelette du graphe P-map de la distribution, celui-ci est inefficace dans le cas où deux variables sont dépendantes. En effet, l'ensemble des 2^{d-2} sous-ensembles possibles sont testés avant de conclure que les variables sont dépendantes. Ceci peut être évité grâce à l'observation suivante : si X_i et X_j ne sont pas voisins dans G^*, G^* étant un P-map (et donc un I-map), alors la propriété de Markov locale implique que soit $\mathbf{Pa}_i$ soit $\mathbf{Pa}_j$ les d-séparent. Vis-à-vis du squelette, cela se traduit par le fait que nous pouvons restreindre la recherche d'un ensemble séparateur parmi $\mathbf{Ne}(X_i) \setminus X_j$ ou $\mathbf{Ne}(X_j) \setminus X_i$. Afin de tirer parti de cette propriété, plutôt que de tester les ensembles par taille croissante pour un même lien, nous devons tester les ensembles de taille fixe l pour l'ensemble des liens. En procédant ainsi, si le nombre maximal de parent d'un nœud dans le graphe est q, nous n'aurons à tester que les ensembles conditionnant de taille inférieure ou égale à $q - 1$.

***Exemple* 5.2.6.** Dans l'exemple précédent le premier lien à être testé était le lien $A - B$ qui est effectivement dans le P-map de la distribution. Dans ce cas là, nous avons dû tester les 8 sous-ensembles possibles avant de conclure sur la dépendance entre les deux variables. Si au contraire nous avions testé tous les liens avec l'ensemble vide avant de tester l'indépendance conditionnelle de A et B avec des ensembles conditionnant de taille $l = 1$, nous aurions au préalable

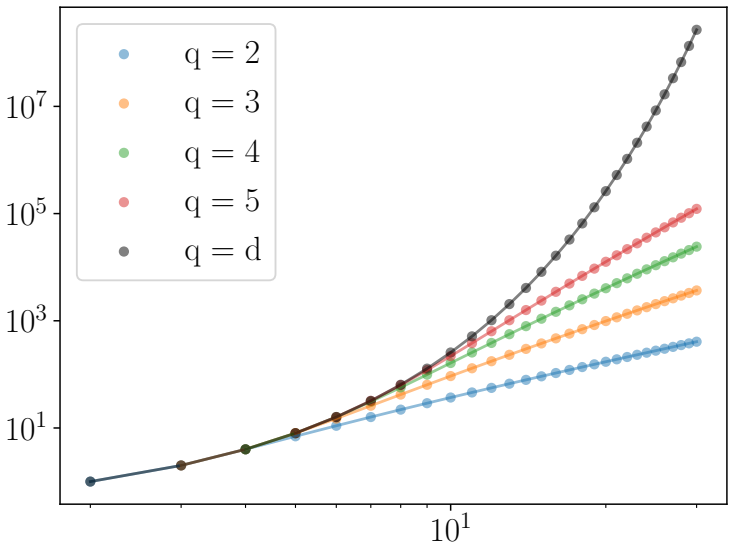


FIGURE 5.2 – Complexité pire cas pour la recherche de l'ensemble séparateur d'un lien avec la recherche du squelette selon PC.

supprimés les liens $A - C$ et $A - E$. Ainsi, comme nous pouvons nous restreindre aux voisins de A pour chercher un ensemble séparateur, nous pouvons éliminer tous ceux contenant C ou E faisant que pour $l = 1$ nous n'avons qu'à tester $A \perp\!\!\!\perp B \mid D$. De plus, le lien entre A et D est supprimé avant de passer à $l = 2$ puisque $A \perp\!\!\!\perp D \mid B$. Pour $l = 2$ il n'existe pas de sous-ensemble de cette taille parmi $\mathbf{Ne}(A) \setminus \{B\} = \{B\}$. Nous passons donc de 8 tests d'indépendance à 2 seulement.

La complexité pire cas pour un lien passe donc de 2^{d-2} à $\sum_{l=0}^{q} \binom{d-2}{l} \leq \frac{(d-2)^q}{(q-1)!}$. Bien que la complexité reste exponentielle, comme le montre la figure 5.2, celle-ci est réduite. La recherche du squelette telle qu'elle est présentée ici est celle proposée par l'algorithme PC (SPIRTES et al. 2000) qui est résumé par l'algorithme 2.

Une fois le squelette reconstruit, la deuxième étape commence par l'énumération des candidats pour la création d'une v-structure, c'est-à-dire l'ensemble des triplets de nœuds (X_i, X_k, X_j) tels que $X_i - X_k - X_j$ et tels que X_i et X_j ne sont pas voisins. G^* étant un P-map de la distribution, d'après la définition de la d-séparation, un candidat est orienté pour former une v-structure $X_i \to X_k \leftarrow X_j$ si et seulement si X_k n'appartient pas à l'ensemble séparateur de X_i et X_j.

Pour finir, les liens restants sont orientés suivant les trois règles suivantes :

- R1 : le lien $X_j - X_k$ est orienté en $X_j \to X_k$ quand il y a un lien $X_i \to X_j$ tel que X_i et X_j ne sont pas voisins (le cas contraire, une v-structure serait créée),
- R2 : le lien $X_i - X_j$ est orienté en $X_i \to X_j$ quand il existe un chemin orienté $X_i \to X_k \to X_j$ (le cas contraire, un cycle orienté serait créé),
- R3 : le lien $X_i - X_k$ est orienté en $X_i \to X_k$ quand il existe deux chemins $X_i - X_j \to X_k$ et $X_i - X_l \to X_k$ tels que X_j et X_l ne sont pas voisins (le cas contraire, une v-structure ou un cycle orienté serait créé).

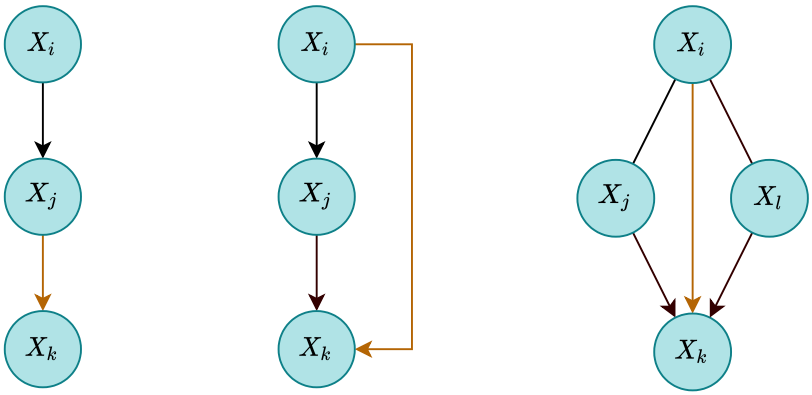
Elles sont représentées graphiquement sur la figure 5.3.

L'ensemble de ces étapes constitue l'algorithme PC qui est résumé par l'algorithme

Algorithm 2: Recherche du squelette selon PC

Input: Échantillon de données $\boldsymbol{d}$, Ordre $\prec$ sur les variables
Result: Squelette S

$S \leftarrow$ graphe non-dirigé complet sur $\mathbf{X}$
$l \leftarrow 0$
$L \leftarrow$ liste des couples (X_i, X_j) ordonnée selon $\prec$
while L *n'est pas vide* **do**
  forall $(X_i, X_j) \in L$ **do**
    $P \leftarrow$ liste des ensembles $\mathbf{U} \subseteq |\mathbf{Ne}(X_i) \setminus \{X_j\}|$ tels que $|\mathbf{U}| = l$ ordonnée selon $\prec$
    forall $\mathbf{U} \in P$ **do**
      if $X_i \perp\!\!\!\perp X_j \,|\, \mathbf{U}$ **then**
        Supprimer $X_i - X_j$ de S
        Sepset$(X_i, X_j) \leftarrow \mathbf{U}$
        break
      end
    end
  end
  $l \leftarrow l + 1$
  $L \leftarrow$ liste des couples (X_i, X_j) tels que $|\mathbf{Ne}(X_i) \setminus \{X_j\}| \geq l$ ordonnée selon $\prec$
end



(a) R1 (b) R2 (c) R3

FIGURE 5.3 – Les trois règles pour la propagation des contraintes. L'arc orange est celui qui a été orienté.

3. SPIRTES et al. (2000) ont démontré [5] que dans le cadre d'une information parfaite l'algorithme PC était consistant :

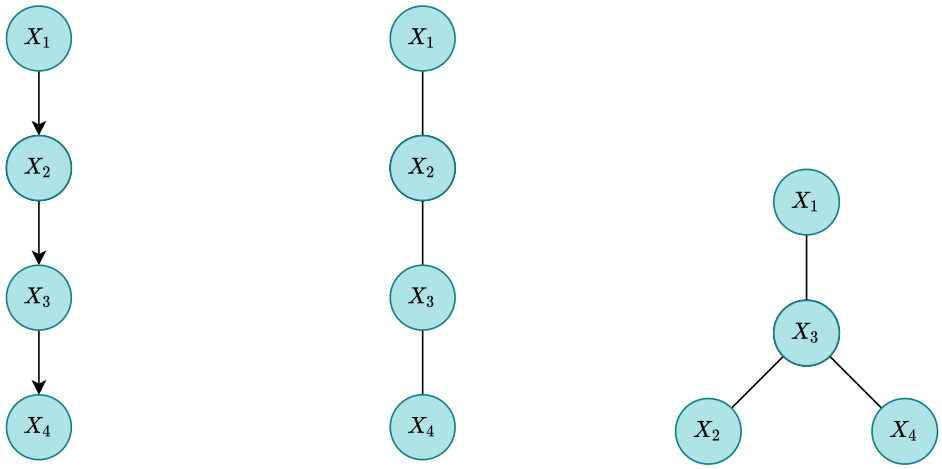
Théorème 5.2.2 (Spirtes et al.)**.** Supposons que la distribution de probabilité de $\mathbf{X}$ ait un P-map G^* et supposons que nous ayons une information parfaite sur les indépendances conditionnelles entre toutes les paires de nœuds (X_i, X_j) étant donné un sous-ensemble de $\mathbf{X} \setminus \{X_i, X_j\}$. Dans ce cas, le CPDAG en sortie de l'algorithme PC est celui de G^*.

Jusqu'à présent nous avons supposé que nous avions une information parfaite sur les indépendances conditionnelles de la distribution ayant généré les données. Deux cas sont possibles : soit nous connaissons le P-map de la distribution, soit nous avons un nombre infini d'observations. Dans un cas applicatif, à supposer qu'il en existe une, la « vraie » structure n'est pas connue et le nombre d'observations à disposition est limité. Il peut alors arriver que l'estimation des indépendances conditionnelles soit erronée à

5. Le lecteur intéressé peut se référer à la page 410.

Algorithm 3: Algorithme PC

Input: Échantillon de données $\boldsymbol{d}$
Result: DAG G
1 Recherche du squelette et des ensembles séparateurs d'après l'algorithme 2
2 Orientation des triplets $X_i - X_k - X_j$ en $X_i \to X_k \leftarrow X_j$ si $X_k \notin \text{Sepset}(X_i, X_j)$
3 Orientation des liens restants en suivant les règles R1, R2 et R3.



(a) DAG (b) Squelette pour $\prec_1$ et $\prec_3$ (c) Squelette pour $\prec_2$

FIGURE 5.4 – Différents ordres peuvent aboutir à différents squelettes.

cause du bruit statistique conduisant à des faux-positifs ou des faux-négatifs (erreurs de type II et I). Cela rend chaque étape de l'algorithme PC dépendante de l'ordre sur les variables qui est utilisé. Pour la recherche du squelette, ceci s'explique par le fait que celui-ci est mis à jour dès qu'une indépendance est trouvée et le voisinage des nœuds est donc modifié. Comme les ensembles conditionnant testés pour un lien donné sont des sous-ensembles du voisinage d'une de ses extrémités, si un lien est supprimé de manière incorrecte alors cela peut avoir des conséquences sur quelles indépendances sont testées par la suite et par conséquent sur le squelette appris.

***Exemple* 5.2.7.** Considérons le graphe de la figure 5.4a comme P-map de la distribution et les deux ordres $X_1 \prec_1 X_3 \prec_1 X_2 \prec_1 X_4$ et $X_1 \prec_2 X_2 \prec_2 X_3 \prec_2 X_4$. Supposons de plus que l'indépendance $X_1 \perp\!\!\!\perp X_2 \,|\, X_3$ soit détectée bien qu'elle ne soit pas vérifiée par la distribution. Le tableau 5.1a liste les indépendances conditionnelles testées pour $l = 1$. Les squelettes obtenus sont présentés sur les figures 5.4b et 5.4c. Leur différence tient du fait qu'avec $\prec_1$, l'indépendance $X_1 \perp\!\!\!\perp X_2 \,|\, X_3$ n'est pas testée puisque le lien $X_1 - X_3$ est d'abord supprimé et X_3 n'est plus voisin de X_1 lorsque l'indépendance entre X_1 et X_2 est testée.

Pour les mêmes raisons que précédemment, les ensembles séparateurs trouvés peuvent dépendre de l'ordre et par conséquent la recherche des v-structures peut aussi être impactée.

***Exemple* 5.2.8.** Reprenons pour P-map le DAG de la figure 5.4a et supposons cette fois-ci que $X_1 \perp\!\!\!\perp X_3 \,|\, X_4$ soit un faux-positif. Soit $\prec_1$ l'ordre introduit lors de l'exemple précédent et $\prec_3$ l'ordre $X_1 \prec_3 X_3 \prec_3 X_4 \prec_3 X_2$. Le tableau 5.1b liste les indépendances conditionnelles testées pour $l = 1$ lors de la recherche du squelette. L'ordre $\prec_1$ testant d'abord l'indépendance $X_1 \perp\!\!\!\perp X_3 \mid X_2$, le lien

	$\prec_1$	$\prec_2$
1	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \perp\!\!\!\perp X_2 \mid X_3$
2	$X_1 \perp\!\!\!\perp X_2 \mid X_4$	$X_1 \perp\!\!\!\perp X_3 \mid X_4$
3	$X_1 \perp\!\!\!\perp X_4 \mid X_2$	$X_1 \perp\!\!\!\perp X_4 \mid X_3$
4	$X_3 \perp\!\!\!\perp X_2 \mid X_4$	$X_2 \perp\!\!\!\perp X_3 \mid X_4$
5	$X_3 \perp\!\!\!\perp X_4 \mid X_2$	$X_2 \perp\!\!\!\perp X_4 \mid X_3$
6	$X_2 \perp\!\!\!\perp X_4 \mid X_1$	$X_3 \perp\!\!\!\perp X_4 \mid X_1$
7	$X_2 \perp\!\!\!\perp X_4 \mid X_3$	$X_3 \perp\!\!\!\perp X_4 \mid X_2$

(a) $X_1 \perp\!\!\!\perp X_2 \mid X_3$ est un faux-positif

	$\prec_1$	$\prec_3$
1	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \perp\!\!\!\perp X_3 \mid X_4$
2	$X_1 \perp\!\!\!\perp X_2 \mid X_4$	$X_1 \perp\!\!\!\perp X_4 \mid X_2$
3	$X_1 \perp\!\!\!\perp X_4 \mid X_2$	$X_3 \perp\!\!\!\perp X_4 \mid X_2$
4	$X_3 \perp\!\!\!\perp X_2 \mid X_4$	$X_3 \perp\!\!\!\perp X_2 \mid X_4$
5	$X_3 \perp\!\!\!\perp X_4 \mid X_2$	$X_4 \perp\!\!\!\perp X_2 \mid X_3$
6	$X_2 \perp\!\!\!\perp X_4 \mid X_1$	–
7	$X_2 \perp\!\!\!\perp X_4 \mid X_3$	–

(b) $X_1 \perp\!\!\!\perp X_3 \mid X_4$ est un faux-positif

TABLE 5.1 – Test d'indépendances menés par l'algorithme PC avec différents ordres et différents faux positifs.

entre X_1 et X_3 est supprimé et l'indépendance $X_1 \perp\!\!\!\perp X_3 \mid X_4$ n'est donc pas testée. Elle l'est en revanche pour l'ordre $\prec_3$ et permet de supprimer le lien entre X_1 et X_3. Malgré cette erreur, le squelette appris en utilisant $\prec_1$ et $\prec_3$ est identique et est celui de la figure 5.4b. La différence réside dans l'ensemble séparateur $\text{Sepset}(X_1, X_3)$ qui est $\{X_2\}$ pour $\prec_1$ et $\{X_4\}$ pour $\prec_3$: le CPDAG ne contiendra pas de v-structure dans le cas de $\prec_1$ mais contiendra la v-structure $X_1 \to X_2 \leftarrow X_3$ dans le cas de $\prec_3$. Ainsi, bien que les squelettes soient identiques, les CPDAG eux sont différents.

En plus d'affecter le voisinage et les ensembles séparateurs, l'ordre joue également un rôle dans l'orientation des liens lors de la recherche des v-structures et de la propagation des contraintes. En effet, il peut arriver qu'un lien soit en conflit avec deux orientations possibles et l'ordre dans lequel sont considérés les liens va déterminer artificiellement l'orientation choisie.

***Exemple* 5.2.9.** Supposons que la recherche du squelette aboutisse au graphe de la figure 5.5a et aux ensembles séparateurs $\text{Sepset}(X_1, X_3) = \text{Sepset}(X_2, X_4) = \emptyset$. Lors de la phase de recherche des v-structures, $X_1 - X_2 - X_3$ et $X_2 - X_3 - X_4$ doivent être orientés comme des v-structures puisque $X_2 \notin \text{Sepset}(X_1, X_3)$ et $X_3 \notin \text{Sepset}(X_2, X_4)$. Ainsi, le lien $X_2 - X_3$ est en conflit avec les orientations $X_2 \to X_3$ et $X_2 \leftarrow X_3$ et va dépendre de quel candidat à une v-structure est considéré en premier. Avec l'ordre $X_1 \prec_2 X_2 \prec_2 X_3 \prec_2 X_4$, c'est $X_1 - X_2 - X_3$ qui est d'abord orienté et $X_2 - X_3 - X_4$ n'est alors plus un candidat. Au contraire, avec l'ordre $X_4 \prec_4 X_3 \prec_4 X_2 \prec_4 X_1$, $X_2 - X_3 - X_4$ aurait été orienté en premier. Considérons à présent qu'après la recherche du squelette et des v-structures le graphe de la figure 5.5b soit obtenu. D'après la règle R1 de la propagation des contraintes, le lien $X_3 - X_6$ est en conflit avec les deux orientations $X_3 \to X_6$ et $X_3 \leftarrow X_6$. En effet, l'une où l'autre des orientations implique la création de deux nouvelles v-structures. Si $X_1 \to X_3 - X_6$ est considéré en premier, dans ce cas $X_3 - X_6$ est orienté en $X_3 \to X_6$ et deux nouvelles v-structures $X_3 \to X_6 \leftarrow X_4$ et $X_3 \to X_6 \leftarrow X_5$ sont créées. Pour éviter de rajouter les indépendances qui leurs sont associées, nous devons alors rajouter les arcs $X_3 \to X_4$ et $X_3 \to X_5$. De cette manière, le DAG final est un I-map de l'ensemble des indépendances trouvées lors de la première étape de l'algorithme PC. Si par contre $X_4 \to X_6 - X_3$ est considéré en premier, le lien est orienté en $X_3 \to X_6$ et dans ce cas ce sont les arcs $X_6 \to X_1$ et $X_6 \to X_3$ qui doivent être rajoutés pour éviter l'ajout de nouvelles

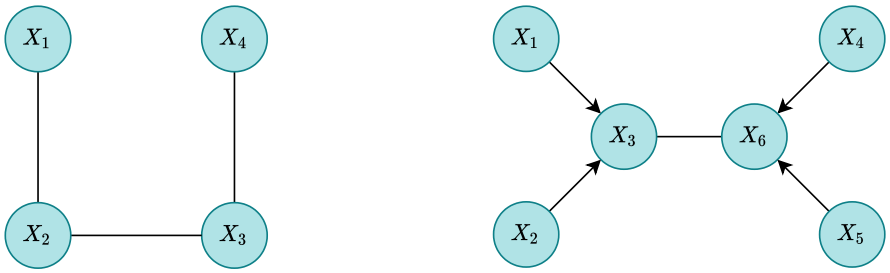


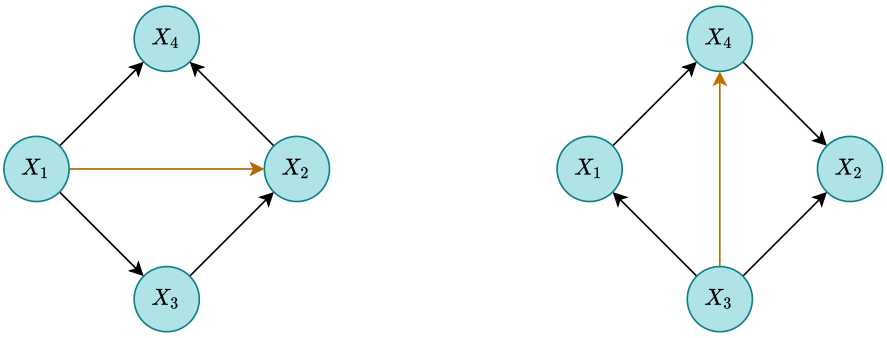
FIGURE 5.5 – L'orientation de certains liens peut être en conflit en fonction des séparateurs trouvés lors de la recherche du squelette et n'est résolue artificiellement qu'en fonction de l'ordre dans lequel les nœuds sont considérés. Par conséquent, le DAG obtenu dépend de cet ordre.

v-structures.

Pour finir, l'algorithme PC fait l'hypothèse qu'il existe un P-map pour la distribution ayant généré les données. Nous avons vu cependant avec l'exemple 4.4.1 qu'il y avait des distributions pour lesquelles il n'existe pas de tel DAG. Si c'est effectivement le cas de la distribution ou parce que des erreurs dans l'estimation des indépendances nous place dans ce cas, le DAG final va une nouvelle fois dépendre de l'ordre sur les variables.

***Exemple* 5.2.10.** Soit la distribution de l'exemple 4.4.1 vérifiant les indépendances $X_1 \perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$ et $X_3 \perp\!\!\!\perp X_4 \mid \{X_1, X_2\}$ et à laquelle nous appliquons l'algorithme PC. Le squelette appris est le cycle non-orienté $X_1 - X_2 - X_3 - X_4 - X_1$ et les ensembles séparateurs sont $\text{Sepset}(X_1, X_3) = \{X_2, X_4\}$ et $\text{Sepset}(X_2, X_4) = \{X_1, X_3\}$. Il n'existe pas de v-structure et l'étape de propagation des contraintes consiste donc à orienter le cycle. L'exemple 4.4.1 montrait que cela était impossible puisque pour ne pas créer un cycle orienté nous sommes obligés de créer une nouvelle v-structure et donc de rajouter une indépendance qui n'a pas été trouvée lors de la recherche du squelette. Par exemple, nous pouvons orienter le lien $X_1 - X_3$ en $X_1 \to X_3$ et, en suivant la règle R1, successivement orienter $X_3 - X_2$ et $X_2 - X_4$ en $X_3 \to X_2$ et en $X_2 \to X_4$. Pour ne pas créer de cycle orienté, le lien $X_1 - X_4$ doit être orienté en $X_1 \to X_4$ créant une v-structure $X_1 \to X_4 \leftarrow X_2$. Ce DAG encode les indépendances $X_1 \perp\!\!\!\perp X_2 \,|\, X_3$ et $X_3 \perp\!\!\!\perp X_4 \,|\, \{X_1, X_2\}$ et n'est donc pas un I-map des indépendances trouvées lors de la recherche du squelette. Pour supprimer la v-structure, il suffit de rajouter un arc entre les nœuds formant sa base. Dans notre cas nous devons rajouter un arc $X_1 \to X_2$ pour que le DAG devienne un I-map. Le DAG final est représenté sur la figure 5.6a. Si, dû à l'ordre, nous avions d'abord orienté $X_3 - X_1$ en $X_1 \to X_3$, nous aurions à la place obtenu le DAG de la figure 5.6b. Remarquons que ces deux DAGs sont des I-maps mais n'encodent pas la même indépendance. Le premier encode $X_3 \perp\!\!\!\perp X_4 \mid \{X_1, X_2\}$ tandis que le deuxième encode $X_1 \perp\!\!\!\perp X_2 \,|\, \{X_3, X_4\}$. Cet exemple se généralise facilement pour des cycles non-orientés de tailles supérieures.

Plusieurs modifications ont été proposées afin de résoudre le problème de la stabilité de l'algorithme PC. Par exemple, la recherche du squelette n'est plus dépendante de l'ordre dès lors que les liens sont supprimés après que l'ensemble des tests pour une même taille de conditionnement l ont été menés plutôt que directement après chacun



(a) (b)

FIGURE 5.6 – L'orientation d'un cycle non-orienté de longueur $l \geq 4$ après la phase d'orientation des v-structures aboutit à différents DAGs non-équivalents selon l'ordre dans lequel les nœuds sont considérés.

d'entre eux. Cet algorithme, appelé PC-stable, ainsi que plusieurs autres variations de l'algorithme PC sont discutées dans COLOMBO et al. (2014).

5.2.2.2 L'algorithme MIIC

Un autre point de vue porté par l'algorithme MIIC (AFFELDT et ISAMBERT 2015 ; AFFELDT, VERNY et al. 2016) est d'utiliser une fonction de rang pour déterminer l'ordre dans lequel les indépendances doivent être testées. De plus, cet algorithme utilise l'approche NML [6] (pour *Normalized Maximum Likelihood*) pour tester les indépendances du modèle mêlant ainsi méthode de score et méthode par contraintes. Étant donnés plusieurs modèles M, nous cherchons celui qui maximise la probabilité (J. J. RISSANEN 1996) :

$$f_{\text{NML}}(\boldsymbol{d}|M) = \frac{f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, M)}{Z(M)}, \quad \text{avec } Z(M) = \int_{\boldsymbol{d}\in\mathcal{D}} f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, M)\mathrm{d}\boldsymbol{d} \tag{5.14}$$

Pour certains modèles, la constante de normalisation $Z(\boldsymbol{d})$, dont le logarithme est appelé complexité, n'est finie que si le domaine $\mathcal{D}$ est restreint (ROOS et al. 2008). Dans le cas de variables aléatoires discrètes, qui est le cadre dans lequel se place l'algorithme MIIC, elle se réduit à la somme :

$$Z(M) = \sum_{\boldsymbol{d}\in\mathcal{D}} f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, M) \tag{5.15}$$

Le nombre de terme dans cette somme étant exponentiel vis à vis de la taille m et de la dimension n de l'échantillon, son évaluation est compliquée. Dans le cas d'une variable aléatoire catégorielle, KONTKANEN et al. (2007) ont dérivé une relation de récurrence permettant son calcul avec une complexité linéaire $\mathcal{O}(n+m)$. Son expression étant fastidieuse et n'étant pas importante pour la suite, nous renvoyons le lecteur à l'article correspondant pour en avoir les détails. Dans le contexte des BNs, ROOS et al. (2008) ont montré que la densité NML et en particulier la complexité, pouvaient se factoriser sur le graphe :

$$\log Z(G) = \sum_{i=1}^{n} \log Z_{X_i|\mathbf{Pa}_i} \tag{5.16}$$

6. Elle est un cas particulier du principe MDL (pour *Minimum Description Length*) qui est une approche basée sur la théorie de l'information de l'inférence statistique (J. RISSANEN 1983 ; GRÜNWALD et al. 2007).

permettant alors l'application de la formule de récurrence proposée par Kontkanen et al. (2007) [7]. La complexité peut également être approximée en utilisant l'information de Fischer (J. J. Rissanen 1996) :

$$\log Z(M) = \frac{p}{2}\log m + \log \int_{\boldsymbol{\Theta}} \sqrt{\det \mathcal{I}(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} + \mathcal{O}(1), \tag{5.17}$$

où p est la dimension de l'espace des paramètres. Remarquons que si seul le terme variant avec la taille de l'échantillon est pris en compte, nous retrouvons l'expression du score BIC. Nous pouvons comparer deux modèles M_1 et M_2 en évaluant le rapport de leurs densités NML. Pour la recherche du squelette, nous avons besoin de tester les indépendances conditionnelles $X_i \perp\!\!\!\perp X_j \mid \mathbf{U}$ avec $\mathbf{U} \subseteq \mathbf{X} \setminus \{X, Y\}$. Les deux modèles sont alors définis par :

$$\begin{cases} f(x_i, x_j, \mathbf{u}|\hat{\boldsymbol{\theta}}, M_1) &= f(\mathbf{u}|\hat{\boldsymbol{\theta}}) f(x_i|\mathbf{u}, \hat{\boldsymbol{\theta}}) f(x_j|\mathbf{u}, \hat{\boldsymbol{\theta}}) \\ f(x_i, x_j, \mathbf{u}|\hat{\boldsymbol{\theta}}, M_2) &= f(\mathbf{u}|\hat{\boldsymbol{\theta}}) f(x_i, x_j|\mathbf{u}, \hat{\boldsymbol{\theta}}) \end{cases} \tag{5.18}$$

Le maximum de vraisemblance pour le modèle M_1 a pour expression :

$$\begin{aligned} f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, M_1) &= \exp\left(m \sum_{k_i=1}^{r_i} \sum_{k_j=1}^{r_j} \sum_{k_{\boldsymbol{u}}=1}^{r_{\boldsymbol{u}}} \frac{m[k_i, k_j, k_{\boldsymbol{u}}]}{m} \log f(k_i, k_j, k_{\boldsymbol{u}}|\hat{\boldsymbol{\theta}}, M_1) \right) \\ &= \exp\left(m \sum_{k_i=1}^{r_i} \sum_{k_j=1}^{r_j} \sum_{k_{\boldsymbol{u}}=1}^{r_{\boldsymbol{u}}} \hat{f}(k_i, k_j, k_{\boldsymbol{u}}) \log \hat{f}(k_i, k_j, k_{\boldsymbol{u}}|M_1) \right) \\ &= \exp\left[-m \left(H_{\hat{f}}(\mathbf{U}) + H_{\hat{f}}(X_i|\mathbf{U}) + H_{\hat{f}}(X_j|\mathbf{U}) \right) \right] \end{aligned} \tag{5.19}$$

où r_i, r_j et $r_{\boldsymbol{u}}$ sont les tailles des ensembles Ω_{X_i}, Ω_{X_j} et $\Omega_{\mathbf{U}}$, c'est-à-dire le nombre de valeurs que peuvent prendre les variables X_i, X_j, $\mathbf{U}$. La deuxième égalité utilise le fait que pour un modèle catégoriel $f(.|\hat{\boldsymbol{\theta}}) = \hat{f}$. De la même manière, on dérive le maximum de vraisemblance pour le modèle M_2 :

$$f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, M_2) = \exp\left[-m \left(H_{\hat{f}}(\mathbf{U}) + H_{\hat{f}}(X_i, X_j|\mathbf{U}) \right) \right] \tag{5.20}$$

Finalement, le rapport des densités NML a pour expression :

$$\frac{f_{\mathrm{NML}}(\boldsymbol{d}|M_1)}{f_{\mathrm{NML}}(\boldsymbol{d}|M_2)} = \exp\left(-m\hat{I}_{\hat{f}}(X_i; X_j|\mathbf{U}) + q_{X_i;X_j|\mathbf{U}} \right) \tag{5.21}$$

où $q_{X_i;X_j|\mathbf{U}} = \log\left(Z(\boldsymbol{d}|M_2)/Z(\boldsymbol{d}|M_1)\right)$ peut être calculé selon les modalités décrites plus haut. Si le rapport est supérieur à 1, c'est-à-dire si $I'_{\hat{f}}(X_i; X_j|\mathbf{U}) = I_{\hat{f}}(X_i; X_j|\mathbf{U}) - \frac{q_{X_i;X_j|\mathbf{U}}}{m} < 0$, l'hypothèse d'indépendance est acceptée. Contrairement à l'algorithme PC, tous les ensembles conditionnant d'une taille l donnée ne sont pas testés mais uniquement ceux contenant celui de l'étape $l-1$. Autrement dit, étant donné l'ensemble conditionnant $\mathbf{U}_{l-1}$ d'un lien $X_i - X_j$, on cherche parmi les voisins $X_k \notin \mathbf{U}_{l-1}$ de X_i celui dont la contribution à l'indépendance est la plus importante. Le rapport des densité NML pour les deux modèles a pour expression :

$$\frac{f_{\mathrm{NML}}(\boldsymbol{d}|M_{X_i \perp X_j|\mathbf{U}\cup X_k})}{f_{\mathrm{NML}}(\boldsymbol{d}|M_{X_i \perp X_j|\mathbf{U}})} = \exp\left(mI_{\hat{f}}(X_i; X_j; X_k|\mathbf{U}) + q_{X_i;X_j;X_k|\mathbf{U}} \right) \tag{5.22}$$

7. Le détail des équations est explicité dans les notes supplémentaires de Affeldt et Isambert (2015).

Algorithm 4: Recherche du squelette selon MIIC

Input: Échantillon de données $\boldsymbol{d}$
Result: Squelette S
// Initialisation
$S_c \leftarrow$ graphe non-dirigé complet sur $\mathbf{X}$
forall *Lien* (X_i, X_j) **do**
  if $I'(X_i; X_j) < 0$ **then**
    Supprimer le lien $X_i - X_j$ de G
    Sepset$(\mathbf{X_i}, \mathbf{X_j}) \leftarrow \emptyset$
  end
  else
    $X_k \leftarrow \arg\max_{\mathbf{Ne}(X_i) \cup \mathbf{Ne}(X_j)} r(X_i, X_j; X_k | \{\})$;
  end
end
// Itération
while $\exists (X_i, X_j)$ *avec* $r(X_i, X_j; X_k | \mathbf{U}) > \frac{1}{2}$ **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
      Sepset$(\mathbf{X_i}, \mathbf{X_j}) \leftarrow \mathbf{U}$
    end
    else
      $X_k \leftarrow \arg\max_{\mathbf{Ne}(X_i) \cup \mathbf{Ne}(X_j)} r(X_i, X_j; X_k | \mathbf{U})$;
    end
    Trier la liste des rangs $r(X_i, X_j; X_k | \mathbf{U})$
  end
end

où $q_{X_i;X_j;X_k|\mathbf{U}} = q_{X_i;X_j|\mathbf{U}\cup X_k} - q_{X_i;X_j|\mathbf{U}}$. Si le rapport est supérieur à 1, c'est-à-dire si $I'_{\hat{f}}(X_i; X_j; X_k|\mathbf{U}) = I_{\hat{f}}(X_i; X_j; X_k|\mathbf{U}) + \frac{q_{X_i;X_j;X_k|\mathbf{U}}}{m} > 0$, alors X_k peut être ajouté à $\mathbf{U}$ et la probabilité associée est donnée par :

$$\begin{aligned} P_{\text{nv}}(X_i; X_j; X_k|\mathbf{U}) &= \frac{f_{\text{NML}}(\boldsymbol{d}|M_{X_i \perp X_j|\mathbf{U}\cup X_k})}{f_{\text{NML}}(\boldsymbol{d}|M_{X_i \perp X_j|\mathbf{U}\cup X_k}) + f_{\text{NML}}(\boldsymbol{d}|M_{X_i \perp X_j|\mathbf{U}})} \\ &= \left[1 + \exp\left(-m I'_{\hat{f}}(X_i; X_j; X_k|\mathbf{U})\right)\right]^{-1} \end{aligned}$$

Remarquons que lorsque $I'_{\hat{f}}(X_i; X_j; X_k|\mathbf{U})$ est négatif, X_k n'est pas ajouté à $\mathbf{U}$ témoignant de la présence d'une v-structure dont la probabilité associée est $P_{\text{v}}(X_i; X_j; X_k|\mathbf{U}) = 1 - P_{\text{nv}}(X_i; X_j; X_k|\mathbf{U})$. La probabilité P_{nv} peut également être interprétée comme étant celle de la suppression du lien $X_i - X_j$ conditionnellement à $\mathbf{U} \cup X_k$. Cependant, l'information mutuelle à trois variables étant symétrique par rapport à ses arguments, elle est aussi la probabilité de supprimer les liens $X_i - X_k$ et $X_j - X_k$ conditionnellement à $\mathbf{U} \cup X_j$ et $\mathbf{U} \cup X_i$. Toutefois, l'information mutuelle à trois variables s'exprime en fonction de celle à deux variables :

$$\begin{aligned} I'_{\hat{f}}(X_i; X_j; X_k|\mathbf{U}) &= I'_{\hat{f}}(X_i; X_j|\mathbf{U}) - I'_{\hat{f}}(X_i; X_j|\mathbf{U} \cup X_k) \\ &= I'_{\hat{f}}(X_i; X_k|\mathbf{U}) - I'_{\hat{f}}(X_i; X_k|\mathbf{U} \cup X_j) \\ &= I'_{\hat{f}}(X_j; X_k|\mathbf{U}) - I'_{\hat{f}}(X_j; X_k|\mathbf{U} \cup X_i) \end{aligned}$$

qui ne sont pas égales. En accord avec l'inégalité de traitement des données (Cover 1999, p.61) :

Algorithm 5: Recherche des v-structures selon MIIC

Input: Échantillon de données $\boldsymbol{d}$
Result: DAG G
`// 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 $R_0 : \{X_i - X_k - X_j \,\&\, \text{not}(X_i - X_j) \& X_k \notin \mathbf{Sepset}(\mathbf{X_i}, \mathbf{X_j})\} \Rightarrow \{X_i \to X_k \leftarrow X_j\}$
  end
  else
    Si (X_i, X_k, X_j) a une orientation convergente, appliquer $R_1 : \{X_i \to X_k - X_j \,\&\, \text{not}(X_i - X_j)\} \Rightarrow \{X_k \to X_j\}$
  end
  Appliquer une nouvelle orientations à tous les autres $(X'_i, X'_k, X'_j) \in L$
until *Aucune orientation additionnelle ne peut être obtenue*

Théorème 5.2.3 (Inégalité de traitement des données)**.** Si $X_i \to X_j \to X_k$, alors $I(X_i; X_j) \geq I(X_i; X_k)$.

le lien $X_i - X_j$ est supprimé si $I'_{\hat{f}}(X_i; X_j; X_k|\mathbf{U}) > 0$ et si de plus l'information mutuelle $I'_{\hat{f}}(X_i; X_j|\mathbf{U})$ est plus faible que celle des liens $X_i - X_k$ et $X_j - X_k$. On définit alors la probabilité que cette inégalité soit respectée :

$$P_{\text{dpi}}(X_i, X_j; X_k|\mathbf{U}) = \frac{f_{\text{NML}}(\boldsymbol{d}|M_{X_i \perp X_j|\mathbf{U}})}{f_{\text{NML}}(\boldsymbol{d}|M_{X_i \perp X_j|\mathbf{U}}) + f_{\text{NML}}(\boldsymbol{d}|M_{X_i \perp X_k|\mathbf{U}}) + f_{\text{NML}}(\boldsymbol{d}|M_{X_j \perp X_k|\mathbf{U}})}$$
$$= \left[1 + \frac{e^{-mI'_{\hat{f}}(X_i;X_k|\mathbf{U})}}{e^{-mI'_{\hat{f}}(X_i;X_j|\mathbf{U})}} + \frac{e^{-I'_{\hat{f}}(X_j;X_k|\mathbf{U})}}{e^{-mI'_{\hat{f}}(X_i;X_j|\mathbf{U})}}\right]^{-1}$$

qui avec la probabilité P_{nv} est utilisée pour la définition du rang :

$$r(X_i, X_j; X_k|\mathbf{U}) \coloneqq \max_{X_k \in \boldsymbol{X}} \left(\min\left[P_{\text{nv}}(X_i; X_j; X_k|\mathbf{U}), P_{\text{dpi}}(X_i, X_j; X_k|\mathbf{U})\right]\right) \tag{5.23}$$

utilisé pour définir l'ordre dans lequel les indépendances sont testées pour la recherche du squelette qui est résumée par l'algorithme 4. La recherche des v-structures, tout comme PC, est basée sur les triplets $X_i - X_k - X_j$ pour lesquels l'information mutuelle à trois variables est évaluée. Si, comme nous l'avons vu elle est négative, alors cela témoigne de la présence d'une v-structure et le triplet est orienté en conséquence. Les différentes étapes sont décrites par l'algorithme 5. Pour plus de détails sur l'implémentation de cette méthode, le lecteur peut se reporter à Affeldt et Isambert (2015) et Affeldt, Verny et al. (2016). Pour une comparaison de ses performances avec d'autres méthodes d'apprentissage par contraintes, il peut se reporter à Verny et al. (2017).

Ce chapitre a fait l'objet de l'introduction des BNs permettant de limiter la complexité du modèle de la loi jointe en tirant parti de ses indépendances. Ces dernières sont encodées au travers d'un DAG permettant une visualisation des relations entre variables mais aussi l'implémentation d'algorithmes efficaces pour l'apprentissage du modèle. Dans le prochain chapitre, nous allons voir que dans le cas continu, la structure de dépendance entre les variables est contenue dans la fonction de copule. Nous nous servirons plus tard de cette propriété pour l'apprentissage de la structure de réseaux bayésiens pour des variables continues.

Références

Affeldt, S. et Isambert, H. (2015). « Robust Reconstruction of Causal Graphical Models based on Conditional 2-point and 3-point Information. » In : *ACI@ UAI*, p. 1-29 (cf. p. 5, 90, 91, 93).

Affeldt, S., Verny, L. et Isambert, H. (2016). « 3off2 : A network reconstruction algorithm based on 2-point and 3-point information statistics ». In : *BMC bioinformatics*. T. 17. 2. BioMed Central, S12 (cf. p. 90, 93).

Bartlett, M. et Cussens, J. (2017). « Integer linear programming for the Bayesian network structure learning problem ». In : *Artificial Intelligence* 244, p. 258-271 (cf. p. 80).

Buntine, W. (1991). « Theory refinement on Bayesian networks ». In : *Uncertainty Proceedings 1991*. Elsevier, p. 52-60 (cf. p. 77).

Chickering, D. M. (1996). « Learning Bayesian networks is NP-complete ». In : *Learning from data*. Springer, p. 121-130 (cf. p. 75).

Colombo, D. et Maathuis, M. H. (2014). « Order-independent constraint-based causal structure learning ». In : *The Journal of Machine Learning Research* 15.1, p. 3741-3782 (cf. p. 90, 133, 162).

Cooper, G. F. et Herskovits, E. (1992). « A Bayesian method for the induction of probabilistic networks from data ». In : *Machine learning* 9.4, p. 309-347 (cf. p. 77).

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons (cf. p. 92).

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge university press (cf. p. 3, 75).

Geiger, D. et Heckerman, D. (1994). « Learning gaussian networks ». In : *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., p. 235-243 (cf. p. 5, 80, 135, 161).

Glover, F. et Laguna, M. (1998). « Tabu Search ». In : *Handbook of Combinatorial Optimization : Volume1–3*. Sous la dir. de D.-Z. Du et P. M. Pardalos. Boston, MA : Springer US, p. 2093-2229 (cf. p. 80).

Grünwald, P. D. et Grunwald, A. (2007). *The minimum description length principle*. MIT press (cf. p. 90, 149, 162, 163).

Heckerman, D., Geiger, D. et Chickering, D. M. (1995). « Learning Bayesian networks : The combination of knowledge and statistical data ». In : *Machine learning* 20.3, p. 197-243 (cf. p. 77, 79).

Kirkpatrick, S., Gelatt, C. D. et Vecchi, M. P. (1983). « Optimization by simulated annealing ». In : *science* 220.4598, p. 671-680 (cf. p. 80).

Koller, D. et Friedman, N. (2009). *Probabilistic graphical models : principles and techniques*. MIT press (cf. p. 3, 71, 72, 75, 76, 80, 131, 154).

Kontkanen, P. et Myllymäki, P. (2007). « A linear-time algorithm for computing the multinomial stochastic complexity ». In : *Information Processing Letters* 103.6, p. 227-233 (cf. p. 90, 91).

Neapolitan, R. E. (2004). *Learning bayesian networks*. T. 38. Pearson Prentice Hall Upper Saddle River, NJ (cf. p. 1, 3, 75).

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Elsevier (cf. p. 5, 70, 75).

Rissanen, J. (1983). « A universal prior for integers and estimation by minimum description length ». In : *The Annals of statistics*, p. 416-431 (cf. p. 90).

Rissanen, J. J. (1996). « Fisher information and stochastic complexity ». In : *IEEE transactions on information theory* 42.1, p. 40-47 (cf. p. 90, 91).

Roos, T., Silander, T., Kontkanen, P. et Myllymaki, P. (2008). « Bayesian network structure learning using factorized NML universal models ». In : *2008 Information Theory and Applications Workshop*. IEEE, p. 272-276 (cf. p. 90, 163).

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G. et Richardson, T. (2000). *Causation, prediction, and search*. MIT press (cf. p. 5, 85, 86, 161).

Trösser, F., Givry, S. de et Katsirelos, G. (2021). « Improved Acyclicity Reasoning for Bayesian Network Structure Learning with Constraint Programming ». In : *arXiv preprint arXiv :2106.12269* (cf. p. 80).

Verny, L., Sella, N., Affeldt, S., Singh, P. P. et Isambert, H. (2017). « Learning causal networks with latent variables from multivariate information in genomic data ». In : *PLoS computational biology* 13.10, e1005662 (cf. p. 93, 147).

Chapitre 6

Théorie des copules

Sommaire

La théorie des copules permet l'étude et la modélisation de dépendances entre variables aléatoires indépendamment de leur comportement individuel. Une fonction copule, ou tout simplement copule, est une fonction de répartition sur l'ensemble $[0,1]^n$ dont les marginales sont distribuées uniformément. Par le biais du théorème de Sklar, nous pouvons alors décomposer la distribution jointe d'un vecteur aléatoire $\mathbf{X}$ en l'ensemble de ses marginales univariées et d'une copule. L'ensemble des marginales contient le comportement isolé de chaque composante du vecteur aléatoire tandis que la copule contient leurs relations de dépendance. Lorsque la distribution jointe est continue, ce qui sera le cas dans le reste de cette thèse, la copule associée est unique.

Cette décomposition est intéressante pour l'estimation de modèles puisqu'elle permet de séparer l'apprentissage des marginales et l'apprentissage de la copule en deux sous-problèmes indépendants. Dans le cas de l'estimation paramétrique, elle permet également la construction de modèles multivariés à partir du choix d'une copule et d'un ensemble de marginales. Les modèles ainsi obtenus sont plus généraux que les modèles classiques qui sont pour la plupart contraints au cas où toutes les marginales sont distribuées selon une même loi ou bien au cas où la relation de dépendance entre variables est linéaire. Malheureusement, la plupart des modèles de copule paramétrique

sont bivariés et il n'en existe qu'un nombre limité qui soient définis pour une dimension n quelconque. De plus, comme nous le verrons, ceux-ci sont dérivés à partir des lois multivariées classiques et encodent donc la même relation de dépendance. Ces limitations peuvent être en partie levées par l'utilisation de la copule de Bernstein empirique qui permet l'estimation non-paramétrique de la copule à partir de polynômes de Bernstein. Cependant, comme pour toute distribution jointe, la manipulation de la copule devient difficile lorsque la dimension du problème est grande et ce d'autant plus pour un modèle non-paramétrique comme la copule de Bernstein.

Nous introduisons ici plusieurs notions de théorie des copules qui sont présentées en détails dans les ouvrages classiques du domaine comme Joe (1997), Nelsen (2007) ou Durante et al. (2016). Nous commençons par introduire la copule et le théorème de Sklar dans la première sous-section. Puis, nous montrons le rôle important qu'elle joue dans la construction de mesures de dépendance dont nous nous servons ensuite pour caractériser plusieurs copules paramétriques. Enfin, nous introduisons la copule de Bernstein qui permet une approximation de la copule par des polynômes de Bernstein.

6.1 Définitions et propriétés

Soit $\mathbf{X} = (X_1, \dots, X_n)^T$ un vecteur aléatoire de dimension n défini sur l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et soit $\mathbf{x} = (x_1, \dots, x_n)^T \in \overline{\mathbb{R}}^n$ une réalisation de $\mathbf{X}$. La définition de copule est facilitée en introduisant les notions de face inférieure et bord supérieur de $\mathbb{I}^n$:

Définition 6.1.1 (Face inférieure et bord supérieur de $\mathbb{I}^n$)**.** Soit un indice $j \in [\![1, n]\!]$, on appelle :

- Face inférieure j de $\mathbb{I}^n$ l'ensemble

$$\mathcal{F}_j^- = \{\mathbf{u} \in \mathbb{I}^n | \mathbf{u} = (u_1, \dots, u_{j-1}, 0, u_{j+1}, \dots, u_n)\}$$

- Bord supérieur j de $\mathbb{I}^n$ l'ensemble

$$\mathcal{B}_j^+ = \{\mathbf{u} \in \mathbb{I}^n | \mathbf{u} = (1, \dots, 1, u_j, 1, \dots, 1)\}$$

Nous notons $\mathcal{F}^- = \bigcup_{j=1}^n \mathcal{F}_j^-$ l'union des faces inférieures et $\mathcal{B}^+ = \bigcup_{j=1}^n \mathcal{B}_j^+$ l'union des bords supérieurs.

Une représentation graphique de ces ensembles dans le cas $n = 3$ est donnée sur la figure 6.1. Nous donnons à présent la définition de copule :

Définition 6.1.2. Une copule C est une fonction $C : \mathbb{I}^n \to \mathbb{I}$ vérifiant les propriétés suivantes :

1. elle est fondée (*grounded* en anglais) : si $\mathbf{u} \in \mathcal{F}^-$, alors $C(\mathbf{u}) = 0$,
2. elle est n-croissante (cf. définition 1.6.3),
3. pour tout $\mathbf{u} \in \mathcal{B}_j$, $C(\mathbf{u}) = u_j$

On note $\mathscr{C}_n$ l'ensemble des copules à n dimensions.

***Exemple* 6.1.1.** Soit la fonction $C_\theta : \mathbb{I}^2 \to \mathbb{I}$ définie par :

$$C_\theta(u, v) = uv(1 + \theta(1 - u)(1 - v)), \quad \theta \in [-1; 1] \tag{6.1}$$

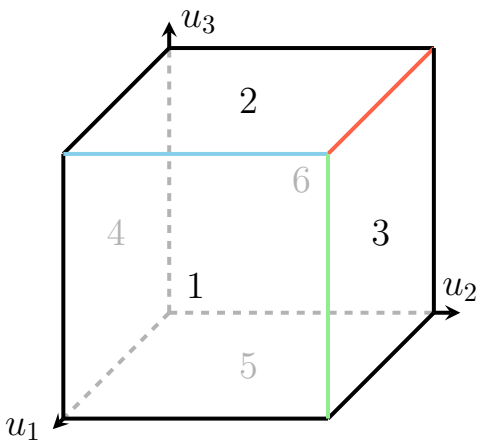


FIGURE 6.1 – Une copule à trois dimensions est définie sur le cube unité. Les faces 4, 5 et 6 correspondent respectivement aux faces inférieures $\mathcal{F}_2^-$, $\mathcal{F}_3^-$ et $\mathcal{F}_1^-$ sur lesquelles la copule est identiquement nulle. Les arrêtes rouge, bleue et verte correspondent respectivement aux bord supérieurs $\mathcal{B}_1^+$, $\mathcal{B}_2^+$ et $\mathcal{B}_3^+$ sur lesquels la copule correspond à l'identité selon la composante u_j.

Il est aisé de voir que $C(u,0) = C(0,v) = 0$ et que cette fonction est *fondée*. La dérivée croisée $\frac{\partial^2 C}{\partial u \partial v}$ existe et a pour expression :

$$c(u,v) = \frac{\partial^2 C}{\partial u \partial v} = 1 + \theta(1-2u)(1-2v) \tag{6.2}$$

On peut montrer que $\min_{u,v} c(u,v) = 1 - |\theta|$ et $\max_{u,v} c(u_v) = 1 + |\theta|$ et donc $0 \leq c(u,v) \leq 2$. Ainsi, d'après la propriété 1.6.1, le C-volume est positif. Enfin, $C(1,v) = v$, $C(u,1) = u$ et C est donc une copule. Celle-ci est appelée copule de Farlie-Gumbel-Morgenstern (FGM). Le graphe de la fonction pour $\theta = 1$ est représenté sur la figure 6.2a.

La copule vérifie les inégalités suivantes appelées bornes de Fréchet-Hoeffding :

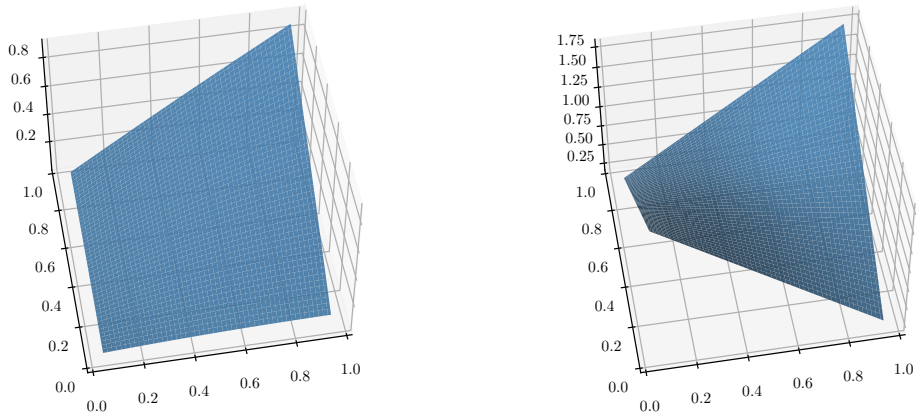
Théorème 6.1.1 (Bornes de Fréchet-Hoeffding)**.** Soit C une copule à n dimensions, alors $\forall \mathbf{u} \in \mathbb{I}^n$ nous avons les inégalités suivantes :

$$W_n(\mathbf{u}) = \max(1-d+u_1+\cdots+u_n, 0) \leq C(\mathbf{u}) \leq \min(u_1,\ldots,u_n) = M_n(\mathbf{u}) \tag{6.3}$$

La borne inférieure W_n n'est une copule que dans le cas particulier où $n = 2$ alors que la borne supérieure M_n l'est toujours et est appelée min-copule.

En plus de la définition fonctionnelle de la copule que nous venons de donner, une définition probabiliste existe. Dans ce cadre, la copule est une distribution sur un vecteur aléatoire $\mathbf{U}$ défini sur $\mathbb{I}^n$ et dont les marginales unidimensionnelles sont distribuées uniformément sur $\mathbb{I}$. Comme le montre le théorème suivant, ces deux définitions sont équivalentes :

Théorème 6.1.2. L'ensemble $\mathscr{C}_n$ coïncide avec l'ensemble des distributions continues restreintes à $\mathbb{I}^n$ et dont les marginales sont distribuées uniformément sur $\mathbb{I}$.



(a) Fonction de répartition (b) Fonction densité

FIGURE 6.2 – Représentation de la copule FGM et de sa densité.

Démonstration. Voir LEBRUN (2013). ■

Cette définition de la copule permet de définir les notions de copules marginales et conditionnelles selon les définitions 1.6.4 et 1.6.10. Si la copule est absolument continue, on peut en plus obtenir sa densité par dérivation :

Définition 6.1.3 (Copule densité)**.** Soit C une copule continue, sa densité [a] c est donnée par :

$$c(u_1, \ldots, u_n) = \frac{\partial^n C(u_1, \ldots, u_n)}{\partial u_1 \ldots \partial u_n} \tag{6.4}$$

a. Par abus de terminologie, la copule densité est aussi appelée copule.

***Exemple* 6.1.2.** Dans l'exemple précédent, nous avons utilisé la copule densité FGM (6.2) pour montrer qu'elle était n-croissante.

Le théorème de Sklar permet de relier une distribution jointe à ses marginales par le biais d'une copule :

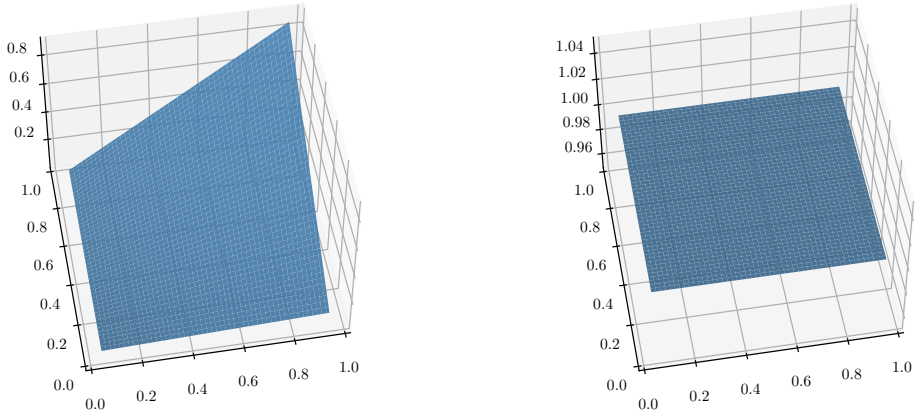
Théorème 6.1.3 (Sklar 1959)**.** Soit $\mathbf{X}$ un vecteur aléatoire, F sa distribution jointe et F_i ses marginales. Il existe une copule C telle que

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)). \tag{6.5}$$

Si les marginales F_i sont en plus continues, alors C est unique.
Réciproquement, si C est une copule de dimension n et $\{F_i\}$ est un ensemble de n distributions marginales, alors la fonction F définie selon (6.5) est une distribution jointe dont les marginales sont F_i.

Démonstration. Voir SCHWEIZER et al. (1974). ■

Dans le cas où C est absolument continue, ce théorème formalise ce que nous avons énoncé en introduction : la copule encode la dépendance entre les composantes de $\mathbf{X}$ tandis que les marginales encodent leur comportement individuel. En dérivant la relation 6.5, on obtient le corollaire suivant :



(a) Fonction de répartition (b) Fonction densité

FIGURE 6.3 – Représentation de la copule indépendante et de sa densité pour $n = 2$.

Corollaire 6.1.3.1. Soit F une distribution jointe absolument continue, F_i ses marginales et C sa copule. La densité c de la copule permet de relier la densité jointe f à ses marginales unidimensionnelles f_i au travers de la relation :

$$h(x_1, \dots, x_d) = c\left(F_1(x_1), \dots, F_d(x_d)\right) \prod_{i=1}^{d} f_i(x_i). \tag{6.6}$$

Le théorème de Sklar est intéressant pour la construction de distributions jointes à partir du choix d'un ensemble de marginales et d'une copule :

***Exemple* 6.1.3.** Soit F_1 et F_2 deux distributions exponentielles de paramètres λ_1 et λ_2 dont l'expression est :

$$F_i(x_i) = \left[1 - \exp(-\lambda_i x_i)\right] \mathbb{1}_{[0,+\infty[}(x_i) \tag{6.7}$$

et soit C la copule FGM de paramètre θ :

$$C(u, v) = uv(1 + \theta(1 - u)(1 - v)) \tag{6.8}$$

En utilisant le théorème de Sklar, on construit alors la distribution :

$$H(x_1, x_2) = C(F_1(x_1), F_2(x_2)) = (1 - e^{-\lambda_1 x_1})(1 - e^{-\lambda_2 x_2})(1 + \theta e^{-(\lambda_1 x_1 + \lambda_2 x_2)})$$

En plus de permettre la construction de nouvelles distributions, le théorème de Sklar permet également la construction de nouvelles copules en inversant la relation 6.5 :

$$C(u_1, \dots, u_n) = H(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)) \tag{6.9}$$

où $u_i = F_i(x_i)$ avec F_i continue. Dans le cas où les marginales F_i sont absolument continues, on peut dériver cette copule pour obtenir sa densité ou bien passer par la densité jointe en inversant le corollaire du théorème de Sklar :

$$c\left(u_1, \dots, u_n\right) = \frac{h\left(F^{-1}(u_1), \dots, F^{-1}(u_d)\right)}{\prod_{i=1}^{d} f_i\left(F^{-1}(u_i)\right)}. \tag{6.10}$$

En utilisant des distributions jointes connues, on peut alors en extraire leur copule (voir la sous-section 6.3). Rappelons que si une distribution marginale F_i n'est pas strictement croissante, elle est non-inversible et la fonction quantile (définition 1.5.1) est utilisée à la place.

***Exemple* 6.1.4.** Dans l'exemple 1.6.1, nous avions vu que le H-volume de la fonction sur $\mathbb{R}^n$ définie par :

$$H(\mathbf{u}) = \prod_{i=1}^{n} u_i$$

rejoignait la notion de volume au sens euclidien. Nous pouvons aisément montrer que la restriction de cette fonction à $\mathbb{I}^n$ définit une copule que l'on notera Π. Soit une distribution jointe F de marginales F_i et ayant pour copule Π. En utilisant le théorème de Sklar, nous avons la relation suivante :

$$F(x_1, \ldots, x_d) = \prod_{i=1}^{d} F_i(x_i). \tag{6.11}$$

qui correspond au cas où les variables sont indépendantes (1.44). Pour cette raison, la copule Π est appelée la copule indépendante. Elle est représentée dans le cas bidimensionnel sur la figure 6.3a. Remarquons pour finir que la copule FGM pour $\theta = 0$ rejoint la définition de la copule indépendante.

Une autre propriété importante de la copule est d'être invariante sous transformations croissantes des variables aléatoires :

Théorème 6.1.4. Soit $\mathbf{X}$ un vecteur aléatoire de dimension n ayant pour copule C et soit $\psi = (\psi_1, \ldots, \psi_n)$ un vecteur de n transformations strictement croissantes défini sur $\bigtimes_{i=1}^{n} \mathrm{Dom}\,(X_i)$. La copule du vecteur aléatoire $\psi(\mathbf{X})$ est également C.

Démonstration. Voir Nelsen (2007, p.25) pour une démonstration dans le cas $n = 2$. Cette démonstration se généralise facilement pour un n quelconque. ■

De plus, en utilisant cette dernière propriété avec $\psi_i = F_i$, nous avons que $H'(u_1, \ldots, u_d) = C(u_1, \ldots, u_d)$ ce qui permet de travailler directement avec la fonction copule et de regarder la structure de dépendance. Cependant, dans de nombreuses applications les F_i sont usuellement inconnues et les distributions empiriques sont utilisées à la place :

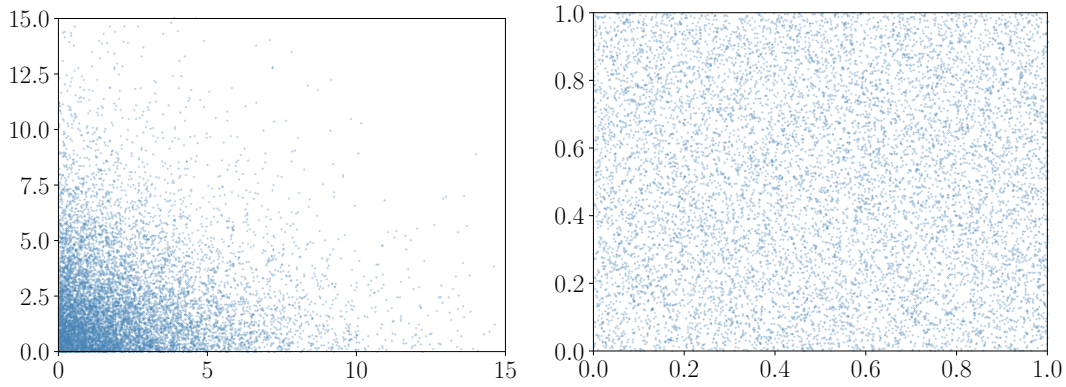
Définition 6.1.4 (Variables de rang)**.** Soit $\mathbf{X}$ un vecteur aléatoire et $\boldsymbol{d} = \{\boldsymbol{x}[j]\}_{1 \leq j \leq m}$ un échantillon de données contenant m réalisations de $\mathbf{X}$. On note $\hat{\boldsymbol{F}} = (\hat{F}_1, \ldots, \hat{F}_n)$ le vecteur des fonctions empiriques des composantes de $\mathbf{X}$. Les variables de rang $\mathbf{U}$ sont définies par :

$$\mathbf{U} = \hat{\boldsymbol{F}}(\mathbf{X}) = (\hat{F}_1(X_1), \ldots, \hat{F}_n(X_d)).$$

L'échantillon associé, appelé échantillon des rangs, est donné par $\mathcal{R} = \{\mathbf{u}[j]\}_{1 \leq j \leq m}$ où

$$\mathbf{u}[j] = \hat{\boldsymbol{F}}(\mathbf{x}[j]) = \left(\hat{F}_1(x_1[j]), \ldots, \hat{F}_n(x_n[j])\right).$$

La Figure 6.4 illustre un échantillon de données provenant d'un vecteur aléatoire (X_1, X_2) et les variables de rang (U_1, U_2) associées. Bien que les variables X_1 et X_2



(a) Échantillon (b) Échantillon des rangs

FIGURE 6.4 – Un échantillon de données distribué selon $X \sim \exp(0.5)$, $Y \sim \exp(0.5)$ et les variables de rang associées. Alors que les variables sont indépendantes, l'échantillon semble montrer une certaine dépendance. Cependant, la copule exhibe clairement l'indépendance des variables et l'apparente dépendance est en fait due au comportement individuel des variables.

semblent dépendantes d'après la représentation graphique de leur distribution, on peut voir avec l'échantillon de rang qu'elles ne le sont pas. Dans la sous-section 7.2, nous allons formaliser cette analyse graphique avec la définition d'un test d'indépendance basé sur la copule de Bernstein empirique.

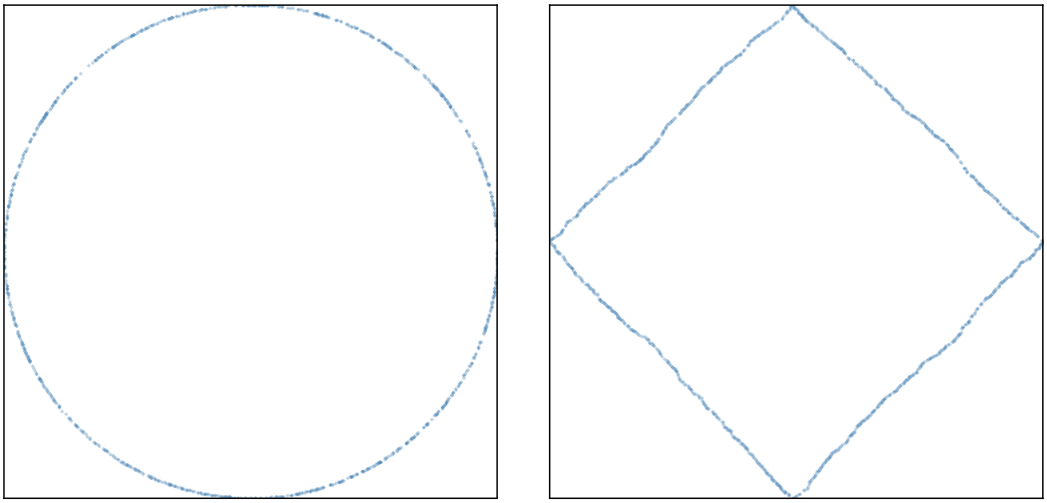
6.2 Mesures de dépendance

Une mesure de dépendance est un outil permettant de résumer la relation de dépendance entre deux variables aléatoires dans un scalaire. L'exemple le plus connu est la *corrélation linéaire* ou de *Pearson*. Pourtant nous allons voir que celle-ci ne possède pas la bonne propriété de ne dépendre que de la copule. SCARSINI (1984) a axiomatisé sous le nom de mesure de concordance un certain nombre de bonnes propriétés pour une mesure de dépendance et que vérifient le rho de Spearman et le tau de Kendall. Un autre type de mesure de dépendance importante, notamment en finance (MAI et al. 2014 ; RODRIGUEZ 2007), est la dépendance de queue dont les coefficients *Upper Tail Dependence* (UTD) et *Lower Tail Dependence* (LTD) sont un exemple. Enfin, nous discutons brièvement de l'information mutuelle et de son lien avec la copule qui sera plus tard étendu au cas de l'information mutuelle conditionnelle et de l'information mutuelle multivariée.

6.2.1 Corrélation linéaire

Définition 6.2.1 (Corrélation linéaire)**.** Soit (X_i, X_j) un vecteur aléatoire dont les variances sont finies et soient C sa copule et F_i et F_j ses marginales. La corrélation linéaire entre X_i et X_j est donnée par :

$$r(X_i, X_j) = \mathbb{E}\left[\left(\frac{X_i - \mathbb{E}[X_i]}{\sqrt{\mathbb{V}[X_i]}}\right)\left(\frac{X_j - \mathbb{E}[X_j]}{\sqrt{\mathbb{V}[X_j]}}\right)\right] \tag{6.12}$$



(a) Échantillon (b) Échantillon des rangs

FIGURE 6.5 – Un échantillon de données distribué uniformément sur le cercle unité et sa copule empirique. La corrélation linéaire de cet échantillon est nulle alors que les variables sont dépendantes.

Elle peut se réécrire en fonction de la copule et des marginales (LEHMANN 1966) :

$$r(X_i, X_j) = \frac{1}{\sqrt{\mathbb{V}[X_i]}\sqrt{\mathbb{V}[X_j]}} \int_0^1 \int_0^1 [C(u_i, u_j) - u_i v_j] \, \mathrm{d}F_i^{-1} \mathrm{d}F_j^{-1} \tag{6.13}$$

***Exemple* 6.2.1.** Nous dérivons la corrélation linéaire de la distribution H construite dans l'exemple 6.1.3 :

$$\begin{aligned}\mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] &= \int_0^{+\infty} \int_0^{+\infty} (x_1 - \frac{1}{\lambda_1})(x_2 - \frac{1}{\lambda_2}) h(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2 \\ &= \frac{\theta}{4}\end{aligned}$$

La corrélation linéaire a donc pour expression :

$$r(X_1, X_2) = \lambda_1 \lambda_2 \frac{\theta}{4} \tag{6.14}$$

Un exemple classique de deux variables X_1 et X_2 dépendantes pour lesquelles la corrélation linéaire est nulle est celui d'une distribution uniforme sur le cercle unité (figure 6.5). Dans ce cas, la copule est donnée par (NELSEN 2007, p.56) :

$$C(u, v) = \begin{cases} M_2(u, v), & |u - v| > \frac{1}{2} \\ W_2(u, v), & |u + v - 1| > \frac{1}{2} \\ \frac{u+v}{2} - \frac{1}{4}, & \text{sinon.} \end{cases} \tag{6.15}$$

En utilisant cette copule, on peut montrer que la corrélation linéaire est bien nulle. La copule n'étant pas la copule indépendante, cela est donc dû au comportement des marginales ce qui fait de la corrélation une mauvaise mesure de dépendance. En effet, puisque les relations de dépendances sont contenues dans la copule, une bonne mesure

de dépendance ne doit dépendre que de celle-ci. Notons toutefois que la dépendance entre les composantes d'une distribution gaussienne multivariée est paramétrée par une matrice de corrélation R dont les éléments R_{ij} correspondent à la corrélation linéaire entre X_i et X_j. Ainsi, dans ce cas particulier, une corrélation nulle témoigne d'une indépendance. Avec ce que nous venons de voir, cela montre donc que la relation de dépendance d'une distribution gaussienne est limitée. Pour finir, notons qu'un estimateur de la corrélation peut être obtenu en remplaçant l'espérance et la variance par leur équivalents empiriques (formule 1.14).

6.2.2 Mesures de concordance

Les mesures de concordance sont des fonctionnelles sur l'ensemble $\mathscr{C}_2$ vérifiant les propriétés suivantes :

Définition 6.2.2 (Mesure de concordance)**.** Une mesure de concordance est une application $\delta : \mathscr{C}_2 \to \mathbb{R}$ telle que :

1. δ est définie pour toute copule $C \in \mathscr{C}_2$,
2. $\forall C \in \mathscr{C}_2$, $\delta(C) = \delta(C^T)$, où C^T est la copule transposée : $C^T(u, v) = C(v, u)$.
3. Si $C \leq C'$ alors $\delta(C) \leq \delta(C')$,
4. $\delta(C) \in [-1, 1]$,
5. $\delta(\Pi) = 0$,
6. $\delta(C^{\sigma_1}) = \delta(C^{\sigma_2}) = -\delta(C)$, où $\delta(C^{\sigma_1})(u, v) = C(1-u, v)$ et $\delta(C^{\sigma_2})(u, v) = C(u, 1-v)$.
7. Si $C_m \xrightarrow[m\to+\infty]{} C$ uniformément, alors $\lim_{m\to+\infty} \delta(C_m) = \delta(C)$.

Des exemples de mesures de concordances sont le rho de Spearman et le tau de Kendall que nous allons présenter. Pour une démonstration que ces mesures vérifient bien les propriétés énoncées, nous renvoyons le lecteur au théorème 2.4.9 de Durante et al. (2016).

Définition 6.2.3 (Rho de Spearman)**.** Soient X_i et X_j deux variables aléatoires et soit C la copule associée à leur distribution jointe. Le rho de Spearman entre X_i et X_j est donné par :

$$\rho(X_i, X_j) = r\left(F_i(X_i), F_j(X_j)\right) \tag{6.16}$$

et peut être exprimé en fonction de la copule entre X_i et X_j

$$\rho(X_i, X_j) = 12 \int_{\mathbb{I}^2} C(u_i, u_j) \mathrm{d}u_i \mathrm{d}u_j - 3 = 12 \int_{\mathbb{I}^2} \left(C(u_i, u_j) - \Pi(u_i, u_j)\right) \mathrm{d}u_i \mathrm{d}u_j \tag{6.17}$$

À une constante près, le rho de Spearman est donc l'intégrale de la différence entre la copule C et la copule indépendante Π. En utilisant les bornes de Fréchet-Hoeffding, nous pouvons montrer que pour n'importe quelle copule, nous avons $-1 \leq \rho(X_i, X_j) \leq 1$. Observons que cette mesure de dépendance ne repose que sur la copule au contraire de la corrélation linéaire.

***Exemple* 6.2.2.** Le rho de Spearman pour deux variables X_1 et X_2 dont la copule est celle de Farlie-Gumbel-Morgenstern (6.1) est donné par :

$$\rho(X_1, X_2) = 12 \int_{\mathbb{I}^2} (C(u,v) - \Pi(u,v)) \, \mathrm{d}u\mathrm{d}v = 12\theta \left(\int_0^1 u(1-u)\mathrm{d}u \right)^2 = \frac{\theta}{3} \tag{6.18}$$

Ainsi, la copule FGM ne permet pas de modéliser des dépendances fortes puisque $-\frac{1}{3} \leq \rho(X_1, X_2) \leq \frac{1}{3}$.

Enfin, l'estimateur non-paramétrique usuel du rho de Spearman à partir d'un échantillon de données $\boldsymbol{d}$ de taille m, est donné par la corrélation entre les variables de rang (Genest et Favre 2007) :

$$\begin{aligned}
\hat{\rho}_m(X_i, X_j) &= \frac{\sum_{k=1}^{m} (u_i[k] - \bar{u}_i)(u_j[k] - \bar{u}_j)}{\sqrt{\sum_{k=1}^{m} (u_i[k] - \bar{u}_i)^2 \sum_{k=1}^{m} (u_j[k] - \bar{u}_j)^2}} \\
&= \frac{12}{m(m+1)(m+1)} \sum_{k=1}^{m} u_i[k] u_j[k] - 3\frac{m+1}{m-1} \in [-1, 1]
\end{aligned}$$

où $\bar{u}_i = \frac{1}{m}\sum_{k=1}^{m} u_i[k] = \frac{m+1}{2} = \frac{1}{m}\sum_{k=1}^{m} u_j[k] = \bar{u}_j$. On peut vérifier que cet estimateur converge asymptotiquement vers la quantité 6.17 en remplaçant dans cette relation la copule C par la copule empirique $\hat{C}_m$:

$$\begin{aligned}
12 \int_{\mathbb{I}^2} \hat{C}_m(u,v)\mathrm{d}u\mathrm{d}v - 3 &= \frac{12}{m} \sum_{k=1}^{m} \int_{\mathbb{I}^2} \mathbb{1}\{u_i[k] \leq u_i\} \mathbb{1}\{u_j[k] \leq v\} - 3\mathrm{d}u\mathrm{d}v \\
&= \frac{12}{m(m+1)(m+1)} \sum_{k=1}^{m} u_i[k] u_j[k] - 3 \\
&= \frac{m-1}{m+1} \hat{\rho}_m(X_i, X_j) \xrightarrow[m \to +\infty]{} \hat{\rho}_m(X_i, X_j)
\end{aligned}$$

Définition 6.2.4 (Tau de Kendall)**.** Soit $\mathbf{X} = (X_i, X_j)$ un vecteur aléatoire ayant pour copule C et pour marginales F_i et F_j. Soit $\mathbf{X}' = (X_i', X_j')$ une copie indépendante de $\mathbf{X}$. Le tau de Kendall entre X_i et X_j est donné par :

$$\tau(X_i, X_j) = \mathbb{P}\left((X_i - X_i') \left(X_j - X_j' \right) > 0 \right) - \mathbb{P}\left((X_i - X_i') \left(X_j - X_j' \right) < 0 \right)$$

et peut être exprimé en fonction de la copule C via

$$\tau(X_i, X_j) = 4 \int_{\mathbb{I}^2} C(u_i, u_j)\mathrm{d}C(u_i, u_j) - 1 = 1 - 4 \int_{\mathbb{I}^2} \frac{\partial C(u_i, u_j)}{\partial u_i} \frac{\partial C(u_i, u_j)}{\partial u_j} \mathrm{d}u_i \mathrm{d}u_j$$

Tout comme pour le rho de Spearman, on peut montrer à partir des bornes de Fréchet-Hoeffding que $-1 \leq \tau \leq 1$.

***Exemple* 6.2.3.** Reprenons l'exemple précédent et calculons cette fois-ci le tau

de Kendall pour X_1 et X_2 :

$$\begin{aligned}\tau(X_1, X_2) &= 4\int_{\mathbb{I}^2} C(u_1, u_2)c(u_1, u_2)\mathrm{d}u_1\mathrm{d}u_2 - 1 \\ &= 4\int_{\mathbb{I}^2} u_1u_2(1+\theta(1-u_1)(1-u_2))(1+\theta(1-2u_1)(1-2u_2)\mathrm{d}u_1\mathrm{d}u_2 - 1 \\ &= 4(\frac{\theta}{18}+\frac{1}{4}) - 1 = \frac{2}{9}\theta\end{aligned}$$

Ainsi, nous avons $-\frac{2}{9} \leq \tau(X_1, X_2) \leq \frac{2}{9}$ pour la copule FGM.

Enfin, l'estimateur non-paramétrique usuel du tau de Kendall à partir d'un échantillon de données $\boldsymbol{d}$ de taille m, est donné par (Genest et Favre 2007) :

$$\hat{\rho}_m(X_i, X_j) = \frac{P_m - Q_m}{\binom{m}{2}} = \frac{4}{m(m-1)}P_m - 1 \in [-1, 1] \tag{6.19}$$

où P_m et $Q_m = \binom{m}{2}$ sont respectivement le nombre de paires concordantes et discordantes. Deux paires $(x_i[k], x_j[k])$ et $(x_i[l], x_j[l])$ sont concordantes si $(u_i[k]-u_i[l])(u_j[k]-u_j[l]) \geq 0$ et discordantes dans le cas contraire. En utilisant la notation $S_{kl} = \mathbb{1}\{u_i[l] \leq u_i[k]\}\mathbb{1}\{u_j[l] \leq u_j[k]\}$, le nombre de paires concordantes dans un ensemble de données s'exprime comme :

$$P_m = \tfrac{1}{2}\sum_{k=1}^{m}\sum_{l\neq k}(S_{kl} + S_{lk}) = \sum_{k=1}^{m}\sum_{l\neq k} S_{kl} = \sum_{k=1}^{m}\sum_{l=1}^{m} S_{kl} - m \tag{6.20}$$

puisque $S_{kl} + S_{lk} = 1$ si et seulement si les paires $(u_i[k], u_j[k])$ et $(u_i[l], u_j[l])$ sont concordantes. Enfin, on remarque que :

$$\begin{aligned}\sum_{k=1}^{m}\sum_{l=1}^{m} S_{ij} &= \sum_{k=1}^{m}\sum_{l=1}^{m}\mathbb{1}\{u_i[l] \leq u_i[k]\}\mathbb{1}\{u_j[l] \leq u_j[k]\} \\ &= \int_{\mathbb{I}^2}\left(\sum_{l=1}^{m}\mathbb{1}\{u_i[l] \leq u_i\}\mathbb{1}\{u_j[l] \leq u_j\}\right)\left(\sum_{k=1}^{m}\delta(u_i = u_i[k])\delta(u_j = u_j[k])\right)\mathrm{d}u_i\mathrm{d}u_j \\ &= m^2\int_{\mathbb{I}^2}\hat{C}_m(u_i, u_j)\mathrm{d}\hat{C}_m(u_i, u_j)\end{aligned}$$

ce qui nous permet d'écrire l'estimateur du tau de Kendall en fonction de la copule empirique :

$$\hat{\rho}_m(X_i, X_j) = 4\frac{m}{m-1}\int_{\mathbb{I}^2}\hat{C}_m(u_i, u_j)\mathrm{d}\hat{C}_m(u_i, u_j) - \frac{m+3}{m-1} \xrightarrow[m\to+\infty]{} \rho(X_i, X_j) \tag{6.21}$$

6.2.3 Dépendance de queue

Définition 6.2.5 (Upper et Lower tail dependence)**.** Soient X_i et X_j deux variables aléatoires et soit C la copule associée à leur distribution jointe. Les coefficients *Upper* et *Lower tail dependence* (UTD et LTD) entre X_i et X_j sont donnés par (Joe 1993) :

$$\begin{aligned}\lambda_U(X_i, X_j) &= \lim_{t\to 1^-}\left(2 - \frac{1-C(t,t)}{1-t}\right) \\ \lambda_L(X_i, X_j) &= \lim_{t\to 0^+}\frac{C(t,t)}{t}\end{aligned}$$

si les limites existent.

***Exemple* 6.2.4.** Pour la copule FGM, les coefficients UTD et LTD sont nuls :

$$\lambda_U(X_i, X_j) = \lim_{t \to 1^-} \left(2 - \frac{1 - t^2(1 - \theta(1-t)^2)}{1-t} \right) = 0$$
$$\lambda_L(X_i, X_j) = \lim_{t \to 0^+} \frac{t^2(1 - \theta(1-t)^2)}{t} = 0$$

Nous allons voir que ces coefficients sont nuls pour la copule gaussienne mais que la copule de Student exhibe une dépendance de queue.

6.2.4 Information mutuelle

Définition 6.2.6 (Entropie de la copule)**.** Soit $\mathbf{X}$ un vecteur aléatoire à n dimensions et soit $C \in \mathscr{C}_n$ sa copule. L'entropie de la copule C est donnée par :

$$H_C(\mathbf{X}) = -\int_{\mathbb{I}^n} c(\mathbf{u}) \log c(\mathbf{u}) \mathrm{d}\mathbf{u} \tag{6.22}$$

Ma et al. (2011) ont montré que l'information mutuelle et la copule étaient reliées par la relation suivante :

$$I(X_i; X_j) = -H_C(X_i, X_j). \tag{6.23}$$

***Exemple* 6.2.5.** Pour la copule FGM, l'information mutuelle a pour expression :

$$I(X_1, X_2) = \frac{(\theta(4+\theta)+3)\log(1+\theta) + (\theta(4-\theta)-3)\log(1-\theta)}{8\theta} + \frac{\mathrm{Li}_2(\theta) - \mathrm{Li}_2(-\theta)}{4\theta} - \frac{5}{4}$$

avec $\mathrm{Li}_2(x) = -\int_0^x \frac{\log(1-t)}{t} \mathrm{d}t$ la fonction dilogarithme.

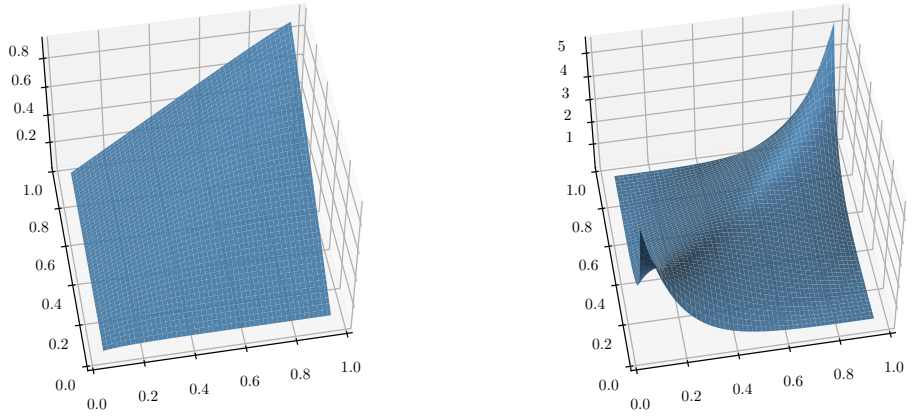
6.3 Copules paramétriques

Nous présentons ici plusieurs copules paramétriques qui seront notamment utilisées dans la construction de modèles de référence pour la comparaison des différentes méthodes d'apprentissage. Pour une présentation plus exhaustive des différentes copules paramétriques existantes, le lecteur peut se référer aux ouvrages de référence cités en introduction de cette section.

6.3.1 La copule Gaussienne

La copule gaussienne joue le même rôle central que la distribution gaussienne en théorie des probabilités et celle-ci est par exemple très utilisée en finance (Bouyé et al. 2000). Comme ce sera également le cas pour les copules de Student et de Dirichlet que nous allons voir, la copule gaussienne est définie en utilisant l'inversion du théorème de Sklar [1] (équation 6.9) pour la distribution normale standard multivariée Φ_R

1. D'autres méthodes existent pour la construction de copules et sont abordées dans le chapitre 3 de Nelsen (2007).



(a) Fonction de répartition (b) Fonction densité

FIGURE 6.6 – Visualisation d'une gaussienne à deux dimensions avec un paramètre de corrélation $\rho = 0.8$.

paramétrisée par une matrice de corrélation R. Elle a pour expression :

$$C_R^G(u_1, \ldots, u_n) = \Phi_R(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_n)) \tag{6.24}$$

avec ϕ la distribution normale standard unidimensionnelle de la distribution gaussienne. La fonction de répartition gaussienne n'ayant pas d'expression analytique, il en va de même pour la copule. Une illustration de copule gaussienne à deux dimensions est donnée sur la Figure 6.6. Notons que pour la définition de la copule gaussienne, nous avons utilisé la paramétrisation standard. Ceci est dû au fait que pour une distribution gaussienne multivariée $\Phi(x_1, \ldots, x_n)$ de moyenne $\boldsymbol{\mu}$ et matrice de covariance $\boldsymbol{\Sigma}$ celle-ci peut être ramenée à la distribution gaussienne multivariée standard en utilisant la transformation $\psi = (\psi_1, \ldots, \psi_n)$ tel que $\psi_i(x_i) = \frac{x_i - \mu_i}{\sigma_i^2}$. Les transformations ψ_i étant croissantes, d'après le théorème 6.1.4 elles n'affectent pas la copule ce qui nous permet d'utiliser la paramétrisation la plus simple. Pour finir, les mesures de concordance pour la copule gaussienne ont pour expression :

$$\rho(X_i, X_j) = \frac{6}{\pi} \arcsin(\frac{r_{ij}}{2}) \qquad \tau(X_i, X_j) = \frac{2}{\pi} \arcsin(r_{ij}) \tag{6.25}$$

tandis que les coefficients UTD et LTD sont nuls :

$$\lambda_L(X_i, X_j) = \lambda_U(X_i, X_j) = 0 \tag{6.26}$$

L'information mutuelle quant à elle s'écrit :

$$I(X_i; X_j) = -\tfrac{1}{2} \log(1 - r_{ij}) \tag{6.27}$$

.

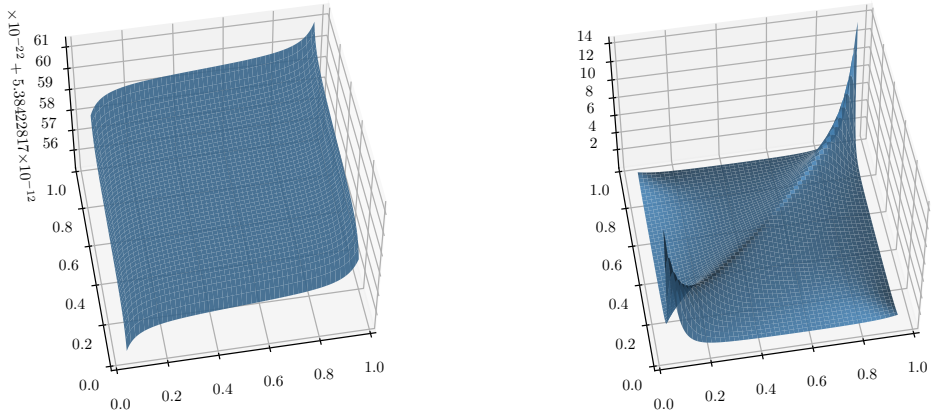
6.3.2 La copule de Student

La copule de Student, est définie à partir de la relation 6.9 et a pour expression :

$$C_{R,\nu}^S(u, v) = T_{R,\nu}(T_\nu^{-1}(u_1), \ldots, T_\nu^{-1}(u_n)).$$

Ses mesures de concordance sont données par :

$$\rho(X_i, X_j) = \frac{6}{\pi} \arcsin(\frac{r_{ij}}{2}) \qquad \tau(X_i, X_j) = \frac{2}{\pi} \arcsin(r_{ij}) \tag{6.28}$$



(a) Fonction de répartition

(b) Fonction densité

FIGURE 6.7 – Visualisation d'une copule de Student à deux dimensions avec un paramètre de corrélation $\rho = 0.8$ et un paramètre $\nu = 1$.

Comme nous pouvons le voir, le rho de Spearman et le tau de Kendall ont la même expression que pour la copule gaussienne. Ceci est dû au fait que les deux copules font partie de la famille des copules elliptiques pour lesquelles les deux expressions sont toujours valables (HULT et al. 2002 ; LINDSKOG et al. 2003). La principale différence entre les deux copules réside dans leur dépendance de queue :

$$\lambda_L(X_i, X_j) = \lambda_U(X_i, X_j) = 2t_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}}\right) \tag{6.29}$$

L'information mutuelle quant à elle s'écrit (ARELLANO-VALLE et al. 2013) :

$$I(X_i; X_j) = -\tfrac{1}{2}\log(1-r_{ij}) + \log\left(\frac{\Gamma(\frac{\nu}{2})\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu+1}{2})^2}\right) - (\nu+1)\psi(\tfrac{\nu+1}{2}) + \tfrac{\nu+2}{2}\psi(\tfrac{\nu+2}{2}) + \tfrac{\nu}{2}\psi(\tfrac{\nu}{2})$$

où $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$ est la fonction digamma. Une représentation de la copule de Student à deux dimensions et de sa densité est donnée sur la figure 6.7.

6.3.3 La copule de Dirichlet

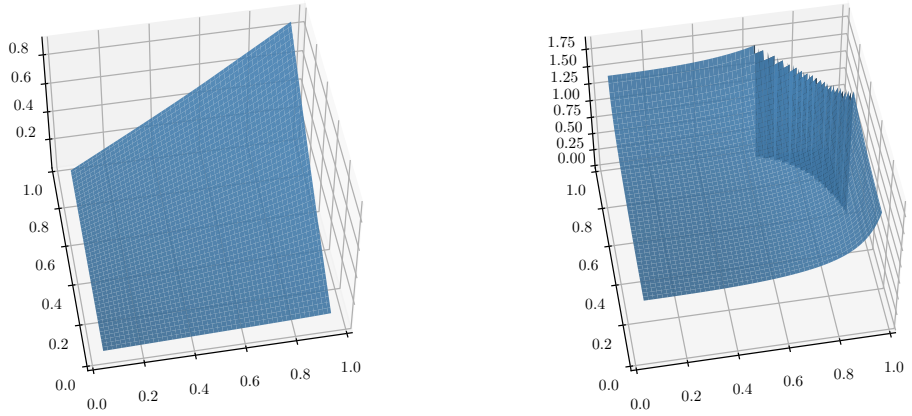
La particularité de la copule de Dirichlet est que son support est un sous ensemble de $\mathbb{I}^n$. Son expression est donnée par :

$$C^D_{\boldsymbol{\theta}}(\mathbf{u}) = \Delta_{\boldsymbol{\theta}}(I^{-1}_{u_1}(\theta_1, \theta_0 - \theta_1), \dots I^{-1}_{u_n}(\theta_n, \theta_0 - \theta_n))$$

où $\Delta_{\boldsymbol{\theta}}$ est la fonction de répartition d'une loi de Dirichlet de paramètre $\boldsymbol{\theta}$ et$I^{-1}_{u_i}(a, b)$ est l'inverse de la fonction bêta incomplète. Il n'existe pas d'expression analytique pour les mesures de dépendances que nous avons vues hormis pour l'information mutuelle qui est donnée par (EBRAHIMI et al. 2011) :

$$I(X_i; X_j) = H_G(\alpha_0 - \alpha_i) + H_G(\alpha_0 - \alpha_j) - H_G(\alpha_0 - \alpha_i - \alpha_j) - H_G(\alpha_0)$$

où $H_G(\alpha) = \log\Gamma(\alpha) - (\alpha - 1)\psi(\alpha) + \alpha$ est l'entropie de la distribution $\Gamma(\alpha, 1)$. Une représentation de la copule de Dirichlet à deux dimensions et de sa densité est donnée sur la figure 6.8.



(a) Fonction de répartition

(b) Fonction densité

FIGURE 6.8 – Visualisation d'une copule de Dirichlet à deux dimensions ayant pour paramètres $\boldsymbol{\theta} = (\frac{1}{3}, \frac{2}{3}, 1)$.

6.4 Copule de Bernstein empirique

Nous introduisons à présent la copule de Bernstein empirique (SANCETTA et al. 2004) dont la construction repose sur la copule empirique et l'opérateur d'approximation de Bernstein. Elle permet une estimation non-paramétrique de la copule à partir d'un échantillon et jouera, pour cette raison, un rôle central dans nos méthodes d'apprentissage. Nous présentons au préalable la copule empirique ainsi que les polynômes de Bernstein sur lesquels s'appuie l'opérateur d'approximation pour ensuite donner la définition de la copule de Bernstein empirique (EBC pour *Empirical Bernstein Copula*) et plusieurs de ses propriétés. Nous terminons en donnant la réécriture de sa fonction de répartition et de sa densité introduite par SEGERS et al. (2017) dans le cas particulier de la copule bêta et qui est ici étendue à la copule de Bernstein. Comme nous allons le voir, cette dernière permet des calculs numériques efficaces à partir de l'EBC.

6.4.1 La copule empirique

La copule étant une fonction de répartition, elle peut être estimée en calculant la fonction de répartition empirique à partir de l'échantillon des rangs (DEHEUVELS 1979) :

Définition 6.4.1 (Copule empirique)**.** Soit $\boldsymbol{d}$ un ensemble de m réalisations d'une loi multivariée. La copule empirique de l'échantillon a pour expression

$$\hat{C}_m(\mathbf{u}) = \frac{1}{m} \sum_{j=1}^{m} \prod_{i=1}^{n} \mathbb{1}\{U_i[j] \leq u_i\}.$$

où $U_i[j]$ sont les variables de rang.

Le théorème de Glivenko-Cantelli (BILLINGSLEY 2008, chapitre 20) nous assure que, presque sûrement, la copule empirique converge uniformément vers la copule C de la loi ayant généré l'échantillon :

$$\mathbb{P}\left(\lim_{m\to\infty} \|\hat{C}_m - C\|_\infty = 0\right) = 1. \tag{6.30}$$

Notons cependant que, malgré son nom, la copule empirique n'est pas une copule et ne peut donc pas être utilisée en tant que telle pour la construction de modèles multivariés. N'étant pas absolument continue, elle ne permet pas non plus d'estimer la densité de la copule limite si celle-ci existe. Enfin, elle ne peut pas être utilisée pour générer de nouveaux points en dehors de ceux de l'échantillon à partir duquel elle a été construite. Pour toutes ces raisons, et malgré le fait que la copule empirique permette une estimation non-paramétrique de la copule, nous allons dans la suite lui préférer la copule de Bernstein empirique qui pallie à ces défauts.

6.4.2 Polynômes et opérateur d'approximation de Bernstein

Les polynômes de Berstein ont été introduits en 1912 par Bernstein (1912) afin de donner une preuve constructive du théorème de Weierstrass affirmant que toute fonction f *continue* et définie sur un compact de $\mathbb{R}^n$ peut être approchée arbitrairement par un polynôme. Nous donnons ici leur définition :

Définition 6.4.2 (Polynôme de Bernstein)**.** Le polynôme de Bernstein de paramètres $(k, d) \in \mathbb{N}^2$ tel que $k \leq d$ s'écrit comme

$$b_{k,d}(u) = \binom{d}{k} u^k (1-u)^{d-k}. \tag{6.31}$$

L'ensemble des $d+1$ polynômes de Bernstein pour un degré d donné forment une base de l'espace vectoriel des polynômes de degré inférieur ou égal à d noté $\mathbb{R}_d[u]$. La figure 6.9 représente ces polynômes dans le cas $d = 5$. Les copules étant des fonctions continues et bornées définies sur $[0,1]^n$, le théorème de Weierstrass nous permet donc de conclure qu'elles sont approchables par des polynômes de Bernstein. À cet effet, nous allons introduire l'opérateur d'approximation de Bernstein mais devons avant cela donner les définitions de grille et cellule de $\mathbb{I}^n$:

Définition 6.4.3 (Grille et cellule de $\mathbb{I}^n$)**.**
Soit $\boldsymbol{K} = (K_1, \ldots, K_n) \in \mathbb{N}^n$ un multi-indice. On appelle :

- Grille de $\mathbb{I}^n$ de pas $\boldsymbol{K}$ l'ensemble de points

$$\mathcal{G}_{\boldsymbol{K}} = \left\{ \left(\tfrac{k_1}{K_1}, \ldots, \tfrac{k_n}{K_n} \right), 0 \leq k_i \leq K_i \right\} \tag{6.32}$$

  Si $\forall i \in [\![1, n]\!], K_i = K$, la grille est dite régulière et notée $\mathcal{G}_K$.
- Cellule $\boldsymbol{k} = (k_1, \ldots, k_n)$ de $\mathbb{I}^n$ l'ensemble

$$\mathcal{C}_{\boldsymbol{k}} = \left] \frac{k_1 - 1}{K_1}, \frac{k_1}{K_1} \right] \times \ldots \left] \frac{k_n - 1}{K_n}, \frac{k_n}{K_n} \right] \tag{6.33}$$

Définition 6.4.4 (Opérateur d'approximation de Bernstein)**.**
L'opérateur d'approximation de Bernstein, paramétré par $\boldsymbol{K}$ définissant la grille $\mathcal{G}_{\boldsymbol{K}}$, est l'opérateur linéaire défini par :

$$B_{\boldsymbol{K}}[\varphi](\mathbf{u}) = \sum_{k_1=0}^{K_1} \ldots \sum_{K_n=0}^{m_n} \varphi\left(\frac{k_1}{K_1}, \ldots, \frac{k_n}{K_n} \right) \prod_{j=1}^{n} b_{k_j, K_j}(u_j) \quad \forall \mathbf{u} \in [0,1]^n \tag{6.34}$$

On parle également de polynôme de Bernstein d'ordre $\boldsymbol{K}$ pour la fonction φ.

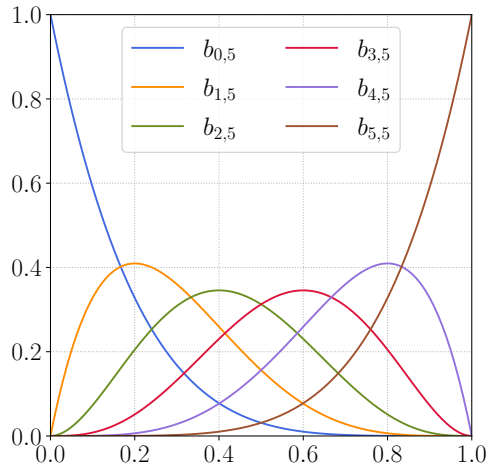


FIGURE 6.9 – Ensemble des polynômes de Bernstein pour $d = 5$.

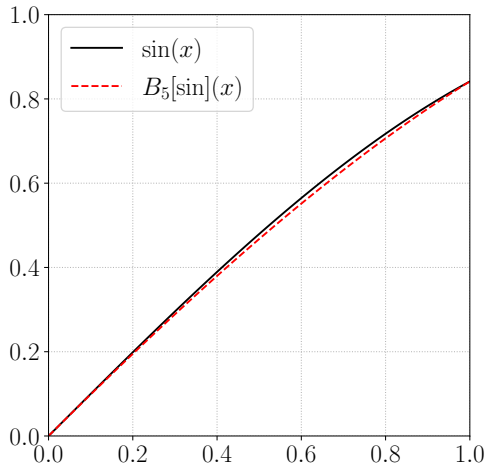


FIGURE 6.10 – Approximation de la fonction sinus par l'opérateur de Bernstein pour $d = 5$ sur $[0, 1]$.

La figure 6.10 représente l'approximation de la fonction sinus sur l'ensemble $[0, 1]$ par les polynômes de Bernstein de degré $K = 5$. Un résultat important, dont le théorème de Weierstrass est un corollaire, est que l'approximation de Bernstein d'une fonction φ converge uniformément vers la fonction φ :

Théorème 6.4.1. Soit $\varphi : [0, 1] \to \mathbb{R}$ une fonction continue et bornée. L'opérateur d'approximation de Bernstein converge uniformément vers φ :

$$\lim_{K \to +\infty} \|\varphi(x) - B_K[\varphi](x)\|_\infty = 0. \tag{6.35}$$

Démonstration. Voir LORENTZ (2012, p.5). ■

Il est montré dans le chapitre 8 de ANASTASSIOU et al. (2012) que ce résultat s'étend au cas multivarié. Une bonne propriété de l'opérateur d'approximation de Bernstein est d'être positif (DEVORE et al. 1993, p.2), c'est-à-dire que lorsque la fonction approchée φ est positive alors son approximation l'est aussi. Ceci est intéressant dans notre cas puisque nous cherchons à approcher des distributions de probabilité.

6.4.3 La copule de Bernstein

Les définitions et propriétés que nous venons de voir nous permettent à présent d'introduire la copule de Bernstein (SANCETTA et al. 2004) :

Définition 6.4.5 (Copule de Bernstein)**.** Une copule de Bernstein est une copule s'écrivant sous la forme $B_{\boldsymbol{K}}[\varphi]$.

Une question importante pour la suite est de savoir quelle condition doit vérifier φ pour que son approximation de Bernstein soit une copule. Il s'avère qu'elle doit être pour cela une copule discrète sur la grille $\mathcal{G}_{\boldsymbol{K}}$:

Définition 6.4.6. Une fonction $\varphi : \mathbb{I}^n \to \mathbb{I}$ est une *copule discrète* sur la grille $\mathcal{G}_{\boldsymbol{K}}$ si elle vérifie les trois conditions suivantes :

1. $\sum_{\ell_1=0}^{1} \cdots \sum_{\ell_n=0}^{1} (-1)^{\ell_1+\cdots+\ell_n} \varphi\left(\frac{k_1+\ell_1}{K_1}, \ldots, \frac{k_n+\ell_n}{K_n}\right) \geq 0, \quad \forall k_i \in [\![0, K_i - 1]\!]$
2. si $\mathbf{u} \in \mathcal{F}^-$, alors $\varphi(\mathbf{u}) = 0$
3. $\forall \mathbf{u} \in \mathcal{B}_j, \quad \varphi(\mathbf{u}) = u_j$

Nous remarquerons en particulier que toute copule est une copule discrète pour n'importe quelle grille $\mathcal{G}_{\boldsymbol{K}}$ sur $\mathbb{I}^n$.

Théorème 6.4.2. Si φ est une copule discrète sur la grille $\mathcal{G}_{\boldsymbol{K}}$, alors $C_{\boldsymbol{K}}^{B,\varphi} = B_{\boldsymbol{K}}[\varphi]$ est une copule de Bernstein.
Réciproquement, si $C_{\boldsymbol{K}}^{B,\varphi}$ est une copule de Bernstein, alors φ coïncide avec la fonction de répartition de la loi uniforme discrète sur chaque grille marginale $\mathcal{G}_{K_j} = \mathcal{B}_j \cap \mathcal{G}_{\boldsymbol{K}}$.

Une copule de Bernstein étant absolument continue, celle-ci possède une densité :

Théorème 6.4.3. Soit $C_{\boldsymbol{K}}^{B,\varphi}$ une copule de Bernstein, sa densité s'écrit :

$$c_{\boldsymbol{K}}^{B,\varphi}(u_1, \ldots, u_n) = \sum_{k_1=1}^{K_1} \cdots \sum_{k_n=1}^{K_n} p(k_1, \ldots, k_n) \prod_{j=1}^{n} K_j b_{k_j-1,K_j-1}(u_j), \quad \forall \mathbf{u} \in [0,1]^n \tag{6.36}$$

où $p(k_1, \ldots, k_n)$ est la loi de probabilité du vecteur aléatoire discret $(V_1, \ldots, V_n)$, tel que sa fonction de répartition coïncide avec φ sur la grille $\mathcal{G}_{\boldsymbol{K}}$ (et donc de marginales uniformes discrètes sur les grilles $\mathcal{G}_{K_i}$) :

$$\varphi\left(\frac{k_1}{K_1}, \ldots, \frac{k_n}{K_n}\right) = \mathbb{P}\left(V_1 \leq \frac{k_1}{K_1}, \ldots, V_n \leq \frac{k_n}{K_n}\right)$$
$$p(k_1, \ldots, k_n) = \mathbb{P}\left(V_1 = \frac{k_1}{K_1}, \ldots, V_n = \frac{k_n}{K_n}\right)$$

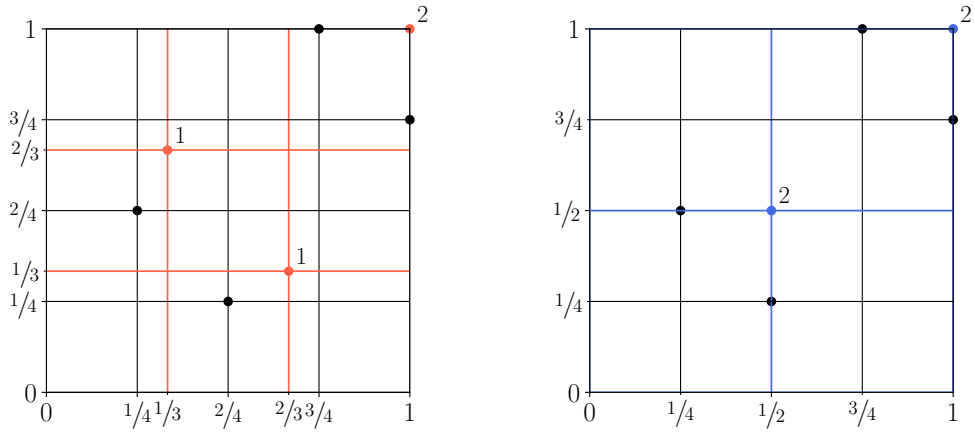
Démonstration. Voir Cottin et al. (2014) et Pfeifer et al. (2020). ■

6.4.4 La copule de Bernstein empirique

À partir d'ici, les grilles de $\mathbb{I}^n$ considérées seront des grilles régulières $\mathcal{G}_K$. La copule de Bernstein empirique correspond à la copule de Bernstein dont la fonction φ associée est la copule empirique. Cependant, cette dernière n'est une copule discrète que si le pas de la grille K divise la taille de l'échantillon m :

Théorème 6.4.4. Soit $\hat{C}_m$ la copule empirique d'un échantillon $\boldsymbol{d}$ supposé issu d'une loi à marginales continues. Celle-ci est une copule discrète par rapport à la grille régulière $\mathcal{G}_K$ si et seulement si K divise m.

***Exemple* 6.4.1.** Pour illustrer le théorème précédent, supposons que nous ayons l'échantillon de rang issu d'une loi à marginales continues suivant : $\boldsymbol{d} = \left\{(\frac{1}{4}, \frac{2}{4}), (\frac{2}{4}, \frac{1}{4}), (\frac{3}{4}, 1), (1, \frac{3}{4})\right\}$. D'après ce que nous venons de voir, la copule empirique de l'échantillon est une copule discrète pour la grille régulière et uniforme $\mathcal{G}_4$. En revanche, ce n'est pas le cas pour toute grille comme le montre la figure 6.11. En effet, si nous prenons la grille $\mathcal{G}_3$, dans ce cas la copule empi-



(a) K=3 (b) K=2

FIGURE 6.11 – Pour que la copule empirique soit une copule discrète, le pas de la grille doit être un multiple de la taille de l'échantillon.

rique n'est pas une loi uniforme sur les bords de la grille. Une manière graphique pour le voir est d'agréger les points de l'échantillon dans chaque cellule $\mathcal{C}_{\boldsymbol{k}}$ en un seul point auquel on associe un poids qui est égal au nombre de points agrégés. Pour que la copule empirique soit uniforme sur les bords de la grille, les poids des points résultants doivent être égaux. Nous voyons ici que ce n'est pas le cas puisque le point $(1, 1)$ a un poids de 2 tandis que les deux autres ont un poids de 1. En revanche, si nous prenons la grille $\mathcal{G}_2$, nous pouvons vérifier que tous les points résultants ont un même poids de 2. Ceci est attendu puisque 3 ne divise pas 4 alors que 2 le divise.

Le théorème précédent nous permet d'introduire la copule empirique de Bernstein :

Définition 6.4.7 (Copule de Bernstein empirique)**.** Sous les hypothèses du théorème 6.4.4, nous appelons copule de Bernstein empirique la copule $\hat{C}^B_{K,m} = B_K[\hat{C}_m]$.

Ainsi, étant donné un échantillon de taille m et d'après le théorème 6.4.4, seules certaines valeurs de K sont valables pour que l'EBC soit une copule. En pratique, le choix de $K \in [\![1, m]\!]$ est laissé à l'utilisateur et $K \lfloor \frac{m}{K} \rfloor$ points de l'échantillon sont gardés afin que la taille de l'échantillon soit un multiple de K. Remarquons cependant qu'il existe deux valeurs de K pour lesquelles la copule empirique est une copule discrète sans que nous n'ayons besoin d'ignorer certains points : $K = 1$ et $K = m$. Nous pouvons facilement voir avec le théorème 6.4.5 ci-dessous, que le cas $K = 1$ ne nous intéresse pas pour la modélisation de dépendance puisque l'EBC correspond dans cette situation à la copule indépendante. En revanche, le cas $K = m$ est intéressant et l'EBC correspond dans ce cas à la copule bêta (SEGERS et al. 2017). Toutefois, nous allons voir que la copule bêta ne nous donne pas de garantie de convergence pour la densité de la copule. L'expression de la densité de la copule de Bernstein est donnée par le théorème suivant :

Théorème 6.4.5 (Densité et fonction de répartition l'EBC)**.** Avec les mêmes notations que celles du théorème 6.4.3, la densité de l'EBC d'un échantillon $\boldsymbol{d}$

issu d'une loi à marginales continues a pour densité sur $\mathbb{I}^n$:

$$c_{\boldsymbol{K}}^{B,\varphi}(u_1,\dots,u_n) = \sum_{k_1=1}^{K_1}\cdots\sum_{k_n=1}^{K_n} p(k_1,\dots,k_n)\prod_{j=1}^{n}\beta_{(k_j,K-k_j+1)}(u_j) \tag{6.37}$$

et pour fonction de répartition :

$$C_{\boldsymbol{K}}^{B,\varphi}(u_1,\dots,u_n) = \sum_{k_1=1}^{K_1}\cdots\sum_{k_n=1}^{K_n} p(k_1,\dots,k_n)\prod_{j=1}^{n} I_{(k_j,K-k_j+1)}(u_j) \tag{6.38}$$

où $\beta_{(a,b)}$ et $I_{(a,b)}$ sont respectivement la densité et la fonction de répartition d'une loi bêta de paramètres (a,b).

Démonstration. Il suffit ensuite de remarquer que $b_{k_j-1,K-1}(u_j) = \frac{1}{K}\beta_{(k_j,K-k_j+1)}(u_j)$ et d'injecter ce résultat dans l'équation (6.4.3) pour obtenir la densité de l'EBC. Par intégration, on obtient alors sa fonction de répartition. ■

Nous nous intéressons à présent à la convergence de l'EBC et de sa densité. Pour ce qui est de la fonction de répartition, le théorème 6.4.1 nous assure que l'approximation de Bernstein d'une fonction continue sur $\mathbb{I}$ est consistante. Cependant, bien que la copule empirique soit définie sur $\mathbb{I}$, cette dernière n'est pas continue. Nous pouvons tout de même nous placer dans les conditions d'application du théorème en remarquant que toute copule discrète peut être prolongée en une copule (Nelsen 2007, Lemme 2.3.5.). En notant $\tilde{C}_m$ une prolongation de la copule empirique (en utilisant par exemple une interpolation multilinéaire), nous avons donc que $\hat{C}_{K,m}^B \to \tilde{C}_m$, quand $K \to +\infty$. Ensuite, par l'application du théorème de Glivenko-Cantelli, nous pouvons voir que l'EBC est un estimateur consistant de la copule, c'est-à-dire que :

$$\mathbb{P}\left(\lim_{K,m\to+\infty} \|\hat{C}_{K,m}^B - C\|_\infty\right) = 1.$$

Nous pouvons résumer cette discussion par le schéma suivant :

$$\hat{C}_{K,m}^B \xrightarrow{\text{Weierstrass}} \tilde{C}_m \longleftrightarrow \hat{C}_m \xrightarrow{\text{Glivenko-Cantelli}} C$$

Pour ce qui est de la densité de l'EBC, Sancetta et al. (2004) ont démontré le théorème suivant :

Théorème 6.4.6. Soit une copule C et sa densité c toutes deux lipschitziennes sur $\mathbb{I}^n$. Soit $\hat{c}_{K,m}^B$ la densité de l'EBC construite à partir d'un échantillon de C. Cet estimateur vérifie les propriétés suivantes :

1. Son biais, $\text{bias}(\hat{c}_{K,m}^B(\mathbf{u})) = \mathbb{E}\left[\hat{c}_{K,m}^B(\mathbf{u})\right] - c(\mathbf{u})$, est inférieur ou égal à K^{-1} à une constante $C \in \mathbb{R}_+^*$ près, ce que l'on note

$$\text{Bias}(\hat{c}_{K,m}^B) \lesssim K^{-1}$$

2. Soit $j \in [\![1,n]\!]$ et soit $\lambda_j = [u_j(1-u_j)]^{1/2}$.
   - Pour $u_j \in]0,1[$, nous avons

$$\mathbb{V}\left[\hat{c}_{K,m}^B(\mathbf{u})\right] \lesssim \left(m\prod_{j=1}^{n}\lambda_j\right)^{-1} K^{n/2}(1+K^{-1}), \tag{6.39}$$

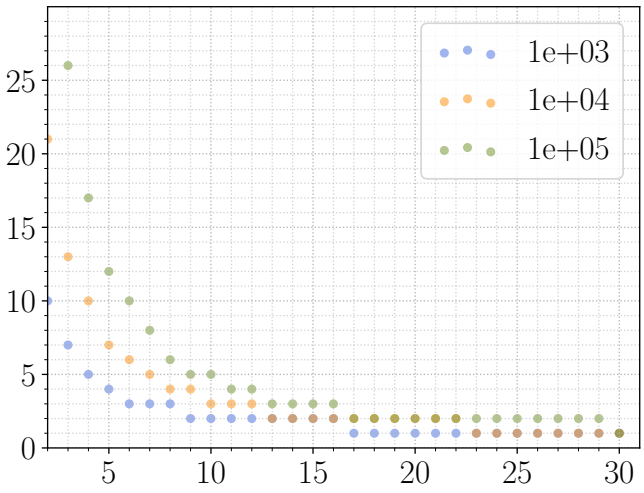


FIGURE 6.12 – Évolution de K_{MISE} en fonction de la dimension et pour différentes valeurs de la taille de l'échantillon.

- Pour $u_j \in \{0, 1\}$, nous avons

$$\mathbb{V}\left[\hat{c}^B_{K,m}(\mathbf{u})\right] = \frac{K^n}{m} c(\mathbf{u}) + O\left(\frac{K^{n-1}}{m}\right). \tag{6.40}$$

3. En utilisant comme métrique l'erreur quadratique moyenne intégrée (MISE pour *Mean Integrated Squared Error*), définie par

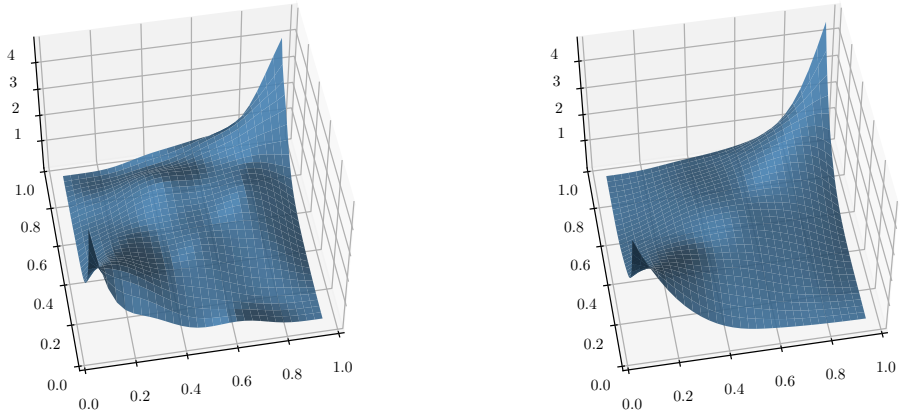
$$\mathbb{E}\left[\|\hat{c}^B_{K,m} - c\|_2^2\right] = \mathbb{E}\left[\int_{\mathbb{I}^n} \left(\hat{c}^B_{K,m}(\mathbf{u}) - c(\mathbf{u})\right)^2 \mathrm{d}\mathbf{u}\right], \tag{6.41}$$

nous avons $\hat{c}^B_{K,m}(\mathbf{u}) \to c(\mathbf{u})$

- Pour $u_j \in]0, 1[$, si $\frac{K^{n/2}}{m}$, quand $K, m \to \infty$,
- Pour $u_j \in \{0, 1\}$, si $\frac{K^n}{m} \to 0$ quand $K, m \to \infty$.

4. La valeur de K minimisant la MISE est :
   - $K \lesssim m^{2/(n+4)}$ si $\forall j \in [\![1, n]\!], u_j \in]0, 1[$,
   - $K \lesssim m^{1/(n+2)}$ si $\forall j \in [\![1, n]\!], u_j \in \{0, 1\}$.

Une première remarque par rapport à ce théorème est que si nous prenons $K = m$, c'est-à-dire la copule bêta, nous minimisons le biais de l'estimateur au prix d'une grande variance. Une autre remarque est que pour de grandes dimensions, la valeur optimale de K pour la MISE n'est pas utilisable pour la modélisation de dépendances puisque $K = \lfloor m^{2/(n+4)} \rfloor$ tend rapidement vers 1 lorsque n augmente. Par exemple, si la dimension du problème est $n = 20$, il faudra un échantillon de taille $m = 4096$ pour que la copule apprise avec ce critère soit différente de la copule indépendante. Pour éviter d'obtenir systématiquement la copule indépendante, nous prendrons plutôt $K_{\text{MISE}} = 1 + \lfloor m^{2/(n+4)} \rfloor$. Cependant, cela ne résout pas totalement le problème puisque K_{MISE} vaudra toujours 2 pour de grandes dimensions et la densité de l'EBC obtenue ne sera pas très expressive. Dans le prochain chapitre, nous allons introduire les réseaux bayésiens à base de copules qui permettent de découper la copule jointe de grande dimension en copules locales de moindres dimensions permettant alors l'apprentissage



(a) $m = 100$ (b) $m = 1000$

FIGURE 6.13 – Approximation de la densité d'une copule gaussienne de paramètre $\rho = 0.8$ par la densité de la copule de Bernstein empirique paramétrée par K_{MISE} et pour des échantillons de tailles $m = 100$ et $m = 1000$.

de copules avec K_{MISE}. La figure 6.13 illustre l'approximation de la densité d'une copule gaussienne par la densité de l'EBC pour des échantillons de taille $m = 100$ et $m = 1000$ en utilisant K_{MISE}.

6.4.5 Aspects numériques

Bien que l'expression 6.38 soit intéressante d'un point de vue théorique, elle l'est moins d'un côté pratique. En effet, l'application de cette formule pour évaluer l'EBC en un point $\mathbf{u} \in \mathbb{I}^n$ a un coût de mK^n : il faut parcourir les K^n cellules de la grille $\mathcal{G}_K$ et, pour chacune de ces cellules, tester l'appartenance de chaque point de l'échantillon. Or, il serait plus efficace d'organiser le parcours en sens inverse en déterminant pour chaque point la cellule à laquelle il appartient à partir de ses coordonnées. En suivant ce principe, SEGERS et al. (2017) ont introduit une écriture alternative de l'EBC et de sa densité :

Théorème 6.4.7 (Réécriture de l'EBC et de sa densité)**.** La copule de Bernstein empirique d'un échantillon $\boldsymbol{d}$ a pour fonction de répartition :

$$\hat{C}^B_{K,m}(\boldsymbol{u}) = \frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{n} I_{r_j[i],s_j[i]}(u_j) \tag{6.42}$$

et densité :

$$\hat{c}_B(\boldsymbol{u}) = \frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{n} \beta_{r_j[i],s_j[i]}(u_j) \tag{6.43}$$

où $r_j[i] = \lceil K u_j[i] \rceil$ et $s_j[i] = K - r_j[i] + 1$.

En utilisant cette nouvelle expression, l'évaluation de l'EBC ne demande que $m \times n$ opérations et ne dépend plus de K.

Un autre avantage numérique que procure cette réécriture réside dans l'échantillonnage de l'EBC. Pour cela, il suffit de remarquer que la formule (6.42) n'est autre qu'une mixture de lois bêta *indépendantes* dont les poids sont uniformes. Ainsi, il suffit de tirer d'abord un entier entre 1 et m de manière uniforme puis de tirer, de manière indépendante, une réalisation de chaque loi bêta associée à cet indice. Ceci nous donne donc

une méthode d'échantillonnage de complexité $O(n)$ beaucoup plus efficace par exemple que la méthode de rejet proposée dans Pfeifer et al. (2020).

Enfin, la réécriture de la densité de l'EBC permet quant à elle de faciliter le conditionnement. Pour le voir, il suffit d'utiliser le théorème de Bayes avec la formule (6.43) :

$$\hat{c}_B(u_l|u_1,\dots,_{l-1}) = \frac{\hat{c}_B(u_1,\dots,u_l)}{\hat{c}_B(u_1,\dots,u_{l-1})} = \frac{\sum_{i=1}^{m}\prod_{j=1}^{l}\beta_{(r_j[i],s_j[i])}(u_j)}{\sum_{i'=1}^{m}\prod_{j'=1}^{l-1}\beta_{(r_{j'}[i'],s_{j'}[i])}(u_{j'})}$$
$$= \sum_{i=1}^{m}\left(\frac{\prod_{j=1}^{l-1}\beta_{(r_j[i],s_j[i])}(u_j)}{\sum_{i'=1}^{m}\prod_{j'=1}^{l-1}\beta_{(r_{j'}[i'],s_{j'}[i'])}(u_{j'})}\right)\beta_{(r_l[i],s_l[i])}(u_l)$$

La densité conditionnelle est donc de nouveau une mixture de lois bêta indépendantes mais cette fois-ci avec des poids non-uniformes. D'après l'équation précédente, la complexité d'une évaluation est donc de $O(ml^2)$. Cette complexité peut encore être diminuée à $O(ml)$ en remarquant que certains calculs peuvent être réutilisés.

Dans le prochain chapitre, nous allons introduire le modèle des réseaux bayésiens à base de copules qui permet de décomposer la copule jointe en un ensemble de copules locales de moindres dimensions. Nous allons exploiter cette décomposition pour développer des algorithmes d'apprentissage efficaces utilisant la copule de Bernstein empirique.

Références

Anastassiou, G. A. et Gal, S. G. (2012). *Approximation theory : moduli of continuity and global smoothness preservation*. Springer Science & Business Media (cf. p. 113).

Arellano-Valle, R. B., Contreras-Reyes, J. E. et Genton, M. G. (2013). « Shannon Entropy and Mutual Information for Multivariate Skew-Elliptical Distributions ». In : *Scandinavian Journal of Statistics* 40.1, p. 42-62 (cf. p. 110).

Bernstein, S. (1912). « Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités ». In : *Comm. Kharkov Math. Soc.* 13.1, p. 1-2 (cf. p. 112).

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons (cf. p. 111).

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G. et Roncalli, T. (2000). « Copulas for finance-a reading guide and some applications ». In : *Available at SSRN 1032533* (cf. p. 108).

Cottin, C. et Pfeifer, D. (2014). « From Bernstein polynomials to Bernstein copulas ». In : *J. Appl. Funct. Anal* 9.3-4, p. 277-288 (cf. p. 114).

Deheuvels, P. (1979). « La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance ». In : *Bulletins de l'Académie Royale de Belgique* 65.1, p. 274-292 (cf. p. 111).

DeVore, R. A. et Lorentz, G. G. (1993). *Constructive approximation*. T. 303. Springer Science & Business Media (cf. p. 113).

Durante, F. et Sempi, C. (2016). *Principles of copula theory*. T. 474. CRC press Boca Raton, FL (cf. p. 98, 105).

Ebrahimi, N., Soofi, E. S. et Zhao, S. (2011). « Information measures of Dirichlet distribution with applications ». In : *Applied Stochastic Models in Business and Industry* 27.2, p. 131-150 (cf. p. 110).

Genest, C. et Favre, A.-C. (2007). « Everything you always wanted to know about copula modeling but were afraid to ask ». In : *Journal of hydrologic engineering* 12.4, p. 347-368 (cf. p. 106, 107).

Hult, H. et Lindskog, F. (2002). « Multivariate extremes, aggregation and dependence in elliptical distributions ». In : *Advances in Applied probability*, p. 587-608 (cf. p. 110).

Joe, H. (1993). « Parametric families of multivariate distributions with given margins ». In : *Journal of multivariate analysis* 46.2, p. 262-282 (cf. p. 107).

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press (cf. p. 2, 98).

Lebrun, R. (2013). « Contributions à la modélisation de la dépendance stochastique ». Thèse de doct. Université Paris-Diderot-Paris VII (cf. p. 100).

Lehmann, E. L. (1966). « Some concepts of dependence ». In : *The Annals of Mathematical Statistics*, p. 1137-1153 (cf. p. 104).

Lindskog, F., McNeil, A. et Schmock, U. (2003). « Kendall's tau for elliptical distributions ». In : *Credit Risk*. Springer, p. 149-156 (cf. p. 110).

Lorentz, G. G. (2012). *Bernstein polynomials*. American Mathematical Soc. (cf. p. 113).

Ma, J. et Sun, Z. (2011). « Mutual information is copula entropy ». In : *Tsinghua Science & Technology* 16.1, p. 51-54 (cf. p. 108, 142, 146).

Mai, J.-F. et Scherer, M. (2014). « How to Measure Dependence ? » In : *Financial Engineering with Copulas Explained*. London : Palgrave Macmillan UK, p. 35-48 (cf. p. 103).

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media (cf. p. 2, 98, 102, 104, 108, 116).

Pfeifer, D., Strassburger, D. et Philipps, J. (2020). « Modelling and simulation of dependence structures in nonlife insurance with Bernstein copulas ». In : *arXiv preprint arXiv :2010.15709* (cf. p. 114, 119).

Rodriguez, J. C. (2007). « Measuring financial contagion : A copula approach ». In : *Journal of empirical finance* 14.3, p. 401-423 (cf. p. 103).

Sancetta, A. et Satchell, S. (2004). « The Bernstein copula and its applications to modeling and approximations of multivariate distributions ». In : *Econometric theory* 20.3, p. 535-562 (cf. p. 3, 111, 113, 116).

Scarsini, M. (1984). « On measures of concordance. » In : *Stochastica* 8.3, p. 201-218 (cf. p. 103).

Schweizer, B. et Sklar, A. (1974). « Operations on distribution functions not derivable from operations on random variables ». eng. In : *Studia Mathematica* 52.1, p. 43-52 (cf. p. 100).

Segers, J., Sibuya, M. et Tsukahara, H. (2017). « The empirical beta copula ». In : *Journal of Multivariate Analysis* 155, p. 35-51 (cf. p. 111, 115, 118).

Partie III

Contributions à l'apprentissage des CBNs

Chapitre 7

CPC : un algorithme basé sur la distance de Hellinger

Sommaire

Dans le chapitre précédent, nous avons introduit la notion de copule qui, dans le cas d'une distribution multivariée, encode les dépendances entre variables aléatoires. Il paraît donc intéressant de vouloir faire le lien entre la théorie des copules et l'apprentissage de la structure d'un réseau bayésien. En particulier, nous avons vu qu'une bonne mesure de dépendance pouvait s'exprimer comme une fonctionnelle de la copule ou de sa densité. Pour cette raison, nous pouvons nous attendre à ce qu'une bonne statistique pour un test d'indépendance ne soit également fonction que de la copule. Dans ce cas, la copule de Bernstein empirique peut être en plus utilisée afin d'obtenir un test non-paramétrique.

En se basant sur ce principe et sur les travaux de Su et al. (2008a), Bouezmarni et al. (2010a) ont proposé un test d'indépendance conditionnelle non-paramétrique reposant sur la distance de Hellinger et la copule de Bernstein empirique. Ce test a ensuite été utilisé par Wan et al. (2014) afin d'implémenter un algorithme PC pour

la reconstruction de BNs continus dont les CPDs sont paramétrées par des modèles de mélange gaussiens (GMM pour *Gaussian Mixture Model* en anglais) (Reynolds 2009). Toutefois, cette paramétrisation pose deux problèmes : d'une part elle n'est pas cohérente avec celle utilisée lors de l'apprentissage de la structure et d'autre part elle ne tire pas parti du fait que la théorie des copules permette de séparer l'apprentissage des marginales de celui de la copule.

Les *Copula Bayesian Networks* (CBNs), introduit par Elidan (2010), sont une application des BNs à la théorie des copules : la copule jointe est factorisée en copules de moindres dimensions permettant de réduire la complexité du problème. Cette factorisation de la copule permet alors de paramétrer les CPDs en définissant une densité marginale et une copule densité *locale* pour chaque variable. La paramétrisation des CBNs permet donc de séparer la modélisation du comportement individuel de chaque variable de celle de leur dépendance. De plus, comme pour les BNs, la factorisation s'appuie sur les indépendances conditionnelles vérifiées par la copule et qui sont encodées au sein d'un DAG. La lecture des indépendances à partir de celui-ci étant la même que pour les BNs, les méthodes d'apprentissage de la structure peuvent donc être adaptées aux CBNs.

En utilisant la copule de Bernstein empirique pour paramétrer les copules locales, les CBNs permettent donc de résoudre les deux problèmes soulevés par l'utilisation de GMMs dans Wan et al. (2014). Du point de vue des BNs continus, ce modèle est également attrayant puisque l'utilisation de la copule de Bernstein permet de s'affranchir d'un choix de modèle (en général gaussien (Lauritzen et Wermuth 1989 ; Lauritzen 1992)) ou de discrétiser les variables. D'autres modèles de BNs continus abordent ce problème (Moral et al. 2001 ; Shenoy et West 2011 ; Langseth et al. 2012) mais leur apprentissage reste difficile (Romero et al. 2006) et sont donc en général limités à quelques variables. De même, il existe d'autres modèles graphiques à base de copules comme par exemple les *vine-copulas* (Bedford et al. 2002) ou plus généralement les *pair-copula constructions* (Czado 2010) dont l'objectif est la construction de copules multivariées à partir de copules bivariées. Pour cela, ils utilisent des graphes non-orientés et n'ont donc pas le même langage graphique que les BNs. Les *Non-parametric Bayesian belief networks* (Kurowicka et al. 2005 ; A. M. Hanea 2008 ; A. Hanea et al. 2015) utilisent, quant à eux, un DAG pour la structure de dépendance mais sont paramétrés par des *vine-copulas*, rendant leur échantillonnage difficile pour de grandes dimensions. Pour toutes ces raisons, nous leur préférerons le modèle des CBNs.

Dans ce chapitre, nous présentons le modèle des CBNs que nous paramétrons par la suite avec des copules de Bernstein empiriques. Cela va nous permettre de dériver un algorithme PC pour l'apprentissage de la structure d'un CBN cohérent avec sa paramétrisation. Nous présentons ensuite un protocole expérimental afin de tester nos algorithmes sur des données générées à partir de CBNs dont la structure est connue et de pouvoir donc mesurer leurs performances en terme structurel. Suite à cela, nous comparons notre algorithme CPC avec une méthode classique pour les BNs continus ainsi qu'avec la méthode d'apprentissage des CBNs proposée par Elidan (2010). Les travaux présentés dans ce chapitre ont fait l'objet de publications dans Lasserre et al. (2020) et Lasserre et al. (2021a).

7.1 Réseaux bayésiens de copules

Cette section reprend en grande partie l'article de Elidan (2010) et s'organise en deux parties : nous introduisons d'abord le modèle des CBNs et plusieurs de ses propriétés puis, dans un deuxième temps, nous présentons la méthode d'apprentissage proposée dans cet article et qui utilise le score BIC maximisé par une recherche TABU.

Cette méthode nous servira par la suite de référence pour comparer nos algorithmes de reconstruction des CBNs.

7.1.1 Définitions et propriétés

La définition des CBNs passe d'abord par la réécriture des densités conditionnelles d'un BN en fonction d'une copule densité. En utilisant la formule de Bayes et l'équation (6.6), on démontre facilement le lemme suivant :

Lemme 7.1.1. Soit $f(x|\mathbf{y})$ une densité conditionnelle avec $\mathbf{Y} = (Y_1, \ldots, Y_l)$ et soit $f(x)$ la densité marginale de X. Il existe une copule densité c telle que :

$$f(x|\mathbf{y}) = R_c(F(x), F(y_1), \ldots, F(y_l))f(x) \tag{7.1}$$

où R_c est le rapport :

$$R_c(u, v_1, \ldots, v_l) = \frac{c(F(x), F(y_1), \ldots, F(y_l))}{c(F(y_1), \ldots, F(y_l))} \tag{7.2}$$

et R_c vaut par définition 1 lorsque $\mathbf{y} = \emptyset$.
Réciproquement, étant donnée une copule densité c et un ensemble de distributions marginales, $R_c(F(x), F(y_1), \ldots, F(y_l))f(x)$ définit une densité conditionnelle.

En appliquant ce lemme aux densités conditionnelles de la règle de chaîne pour les BNs (équation 4.7), on démontre que la fonction copule se factorise sur le graphe :

Théorème 7.1.2 (Décomposition de la copule densité)**.** Soit G une structure de réseau bayésien pour une distribution $\mathbb{P}_{\mathbf{X}}$ ayant pour copule densité c. La copule densité se factorise sur le graphe G :

$$c(\mathbf{u}) = \prod_{i=1}^{n} R_{c_i}^G(u, \boldsymbol{v}) \tag{7.3}$$

où $\{c_i\}_{1 \leq i \leq n}$ est un ensemble de copules *locales* ne dépendant que de X_i et ses parents dans G.

Comme pour les BNs, l'inverse est aussi vrai :

Théorème 7.1.3. Soit $\mathcal{G}$ un DAG sur $\mathbf{X}$. De plus, soit $\{c_i(F(x_i), F(\mathrm{pa}_{i1}), \ldots, F(\mathrm{pa}_{ik_i}))\}$ un ensemble de copules densités strictement positives La fonction g définie comme :

$$g(x_1, \ldots, x_n) = \prod_{i=1}^{n} R_{c_i}(F(x_i), \{F(\mathrm{pa}_{il})\})f_i(x_i),$$

est une densité pour $\mathbf{X}$ vérifiant les indépendances $\mathcal{I}(\mathcal{G})$.

Il est important de noter qu'en dehors du cas où la structure est un arbre (Kirshner 2008), les densités f_i ne sont pas les marginales de g puisque le produit des facteurs R_{c_i} n'est pas une copule (Elidan 2010). Ceci est bien sûr un problème pour la construction de modèle, mais dans le cas de l'apprentissage de modèle qui nous intéresse, cette décomposition apporte une grande flexibilité puisque nous pouvons apprendre des modèles différents pour chaque copule locale et pour chaque densité f_i.

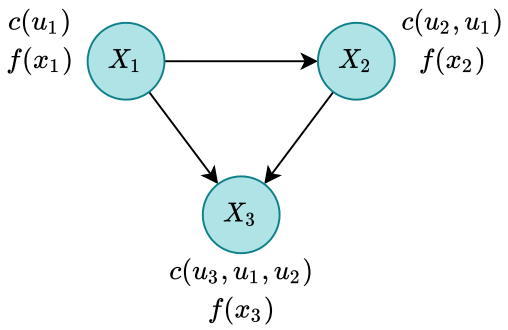


FIGURE 7.1 – Un CBN avec trois variables X_1, X_2 et X_3.

Nous pouvons à présent donner la définition d'un CBN :

Définition 7.1.1 (Copula Bayesian Network)**.** Un *Copula Bayesian Network* est une paire $\mathcal{C} = (G, \mathbb{P}_{\mathbf{X}})$ où G est une structure de réseau bayésien pour $\mathbb{P}_{\mathbf{X}}$. La densité de $\mathbf{X}$ se décompose alors sur G :

$$f(\mathbf{x}) = \prod_{i=1}^{n} R_{c_i}(F(x_i), \boldsymbol{F}(\mathbf{pa}_i)) f(x_i). \tag{7.4}$$

L'équation précédente est appelée règle de chaîne pour les *Copula Bayesian Networks*. La densité jointe f est spécifiée par les copules densités locales c_i et ses densités marginales f_i qui sont associées au nœud X_i dans le graphe.

Remarquons que lorsqu'un nœud n'a pas de parents dans G la copule densité qui lui est associée est la fonction valant identiquement 1 sur $\mathbb{I}^2$. Pour les variables avec au moins un parent, nous pouvons utiliser n'importe quel type de modèle paramétrique ou bien, comme nous le ferons dans la prochaine partie, des copules densités de Bernstein.

***Exemple* 7.1.1.** Soit $\mathcal{C}$ le CBN illustré sur la figure 7.1. Ce CBN encode la densité jointe des variables X_1, X_2 et X_3. Chaque nœud représente une variable aléatoire à laquelle est associée sa densité marginale f_i et une copule densité locale c_i. La structure du CBN encode la factorisation de la densité jointe :

$$\begin{aligned}
f(x_1, x_2, x_3) &= [R_{c_1}(F(x_1)) f(x_1)] \, [R_{c_2}(F(x_2), F(x_1)) f(x_2)] \\
&\qquad \times [R_{c_3}(F(x_3), F(x_1), F(x_2)) f(x_3)] \\
&= [1 \times f(x_1)] \left[\frac{c_2(F(x_2), F(x_1))}{c_2(F(x_1))} f(x_2) \right] \\
&\qquad \times \left[\frac{c_3(F(x_3), F(x_1), F(x_2))}{c_3(F(x_1), F(x_2))} f(x_3) \right] \\
&= c_2(F(x_2), F(x_1)) \frac{c_3(F(x_3), F(x_1), F(x_2))}{c_3(F(x_1), F(x_2))} f(x_1) f(x_2) f(x_3)
\end{aligned}$$

La simplification est due au fait que les marginales unidimensionnelles de la copule sont uniformes sur $\mathbb{I}$. La paramétrisation des marginales et des copules locales est laissée libre et pourrait être n'importe quel modèle.

Dans la suite, nous allons utiliser des CBNs couplés à la copule de Bernstein empirique pour obtenir un BN continu ne faisant pas d'hypothèse de modèle sur les densités conditionnelles et tirant parti de la liberté de modélisation que procure la théorie des

copules. En utilisant des tests d'indépendance conditionnelle fondés sur la copule, nous obtiendrons en plus un modèle cohérent à tous les niveaux de l'apprentissage.

7.1.2 Apprentissage de la structure

Les contributions de cette thèse portent sur l'implémentation de méthodes pour l'apprentissage des CBNs. Nous présentons ici la méthode proposée par Elidan (2010) qui servira par la suite de référence pour comparer nos algorithmes. Cette méthode est basée sur le score BIC adapté au cadre des CBNs et qui est maximisé selon la méthode de recherche locale *TABU list*. Pour rappel, le score BIC a pour expression :

$$\mathcal{S}_{\text{BIC}}(G) = f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}^G, G) - \frac{|\boldsymbol{\Theta}^G|}{2}\log m \tag{7.5}$$

où $|\boldsymbol{\Theta}^G|$ est le nombre de paramètres indépendants associés à la structure et qui dépend du modèle paramétrique utilisé. Elidan (2010) compare plusieurs modèles parmi lesquels nous ne retenons que la copule gaussienne C_R^G qui est celle obtenant les meilleurs résultats sur les cas d'applications présentés dans l'article. En utilisant la décomposition de la densité jointe (7.4), la log-vraisemblance s'écrit :

$$\begin{aligned} f(\boldsymbol{d}|\boldsymbol{\theta}, G) &= \sum_{j=1}^{m} \log f(x_1[j], \ldots, x_n[j]|\boldsymbol{\theta}, G) \\ &= \sum_{j=1}^{m} \log \prod_{i=1}^{n} R_{c_i}(F(x_i[j]), \{F(pa_{ik_i}[j])\}) f(x_i[j]|\boldsymbol{\theta}) \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} \log R_i(F(x_i[j]), \{F(pa_{ik_i}[j])\}) + \sum_{j=1}^{m}\sum_{i=1}^{n} \log f(x_i[j]) \end{aligned}$$

Elidan (2010) propose d'estimer les marginales de manière non-paramétrique en utilisant des fenêtres de Parzen (Parzen 1962) puis de les remplacer dans l'expression précédente. À la place, nous nous plaçons directement dans l'espace de la copule, c'est-à-dire en transformant les composantes de $\mathbf{X}$ selon $U_i = F_i(X_i)$ où F_i est la marginale de X_i. Dans ce cas, l'équation se simplifie en :

$$f(\boldsymbol{d}|\boldsymbol{\theta}, G) = \sum_{j=1}^{m}\sum_{i=1}^{n} \log R_{c_i}\left(u_i[j], \pi_{i1}[j], \ldots, \pi_{ik_i}[j]\right))$$

Rappelons que dans un cas applicatif où nous ne connaissons pas l'expression des marginales F_i, nous utilisons les variables de rang à la place ce qui, comme nous l'avons vu, revient à travailler avec la copule empirique. L'estimation de la matrice de corrélation R par maximum de vraisemblance pouvant être fastidieuse, une alternative est utilisée à la place. Elle repose sur la relation entre le tau de Kendall et la corrélation linéaire (équation (6.25)) pour les copules elliptiques. En utilisant l'estimateur non-paramétrique du tau de Kendall (équation (6.19)), les éléments de la matrice de corrélation pour un échantillon de données sont obtenus. La matrice de corrélation ainsi construite n'étant pas forcément définie semi-positive, l'utilisation de méthodes de régularisation (Rousseeuw et al. 1993) peut parfois être requise. D'après nos expériences, c'est en général le cas lorsque la taille de l'échantillon de données est petite. Toutefois, en utilisant à la place le rho de Spearman et sa relation avec la corrélation linéaire, il ne nous a pas été nécessaire d'avoir recours à ces méthodes. Notons que Elidan (2010) ne fait pas mention de l'utilisation de la décomposition du score BIC pour tirer pleinement parti de l'algorithme de recherche locale. Dans la prochaine partie, nous allons introduire plusieurs méthodes d'apprentissage des CBNs dont une amélioration utilisant cette propriété.

7.2 Test d'indépendance basé sur la distance de Hellinger

Nous présentons maintenant le test d'indépendance non-paramétrique BRT introduit par Bouezmarni et al. (2010b) et que nous allons utiliser pour implémenter un algorithme PC pour les CBNs. En sous-section 3.4.1, nous avions évoqué comment un test de qualité d'ajustement pouvait être mené dans le cas d'une variable aléatoire continue. Pour rappel, dans ce cas l'hypothèse nulle est donnée par $H_0 : F = F_0$ où F_0 est le modèle testé. Un détail important était le fait que l'on pouvait toujours se ramener au cas où F_0 est la distribution uniforme sur $\mathbb{I}$ en transformant les données selon $U = F(X)$. Dans le cas qui nous intéresse à présent, la distribution F_0 est une distribution à n dimensions et dont les marginales sont $F_{0i1\leq i\leq n}$. Nous pouvons alors appliquer la transformation $U_i = F_i(X_i)$ et donc obtenir la copule de $\mathbf{X}$. Ainsi, dans le cas multivarié le test de qualité d'ajustement revient à comparer la copule empirique à la copule C_0. Pour tester l'indépendance entre deux variables X et Y, la copule C_0 est la copule indépendante : $H_0 : C_0(u, v) = \Pi(u, v)$. Cependant, nous sommes intéressés par le cas plus général d'une indépendance conditionnelle $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ où $\mathbf{Z}$ est un ensemble conditionnant de taille l. Cette indépendance peut s'écrire en fonction des copules densités. En effet, nous avons :

$$
\begin{aligned}
f(x, y, \mathbf{z}) &= c_{XY\mathbf{Z}}(F(x), F(y), \boldsymbol{F}(\mathbf{z}))f(x)f(y)\prod_{i=1}^{l} f(z_i) \\
f(x, y, \mathbf{z}) &= f(\mathbf{z})f(x|\mathbf{z})f(y|\mathbf{z}) \\
&= \frac{c_{X\mathbf{Z}}(F(x), \boldsymbol{F}(\mathbf{z}))c_{Y\mathbf{Z}}(F(y), \boldsymbol{F}(\mathbf{z}))}{c_{\mathbf{Z}}(\boldsymbol{F}(\mathbf{z}))} f(x)f(y)\prod_{i=1}^{l} f(z_i)
\end{aligned}
$$

avec $\boldsymbol{F}(\mathbf{z}) = (F(z_1), \ldots, F(z_l))$. Soit en égalant les deux termes :

$$
c_{XY\mathbf{Z}}(F(x), F(y), \boldsymbol{F}(\mathbf{z})) = \frac{c_{X\mathbf{Z}}(F(x), \boldsymbol{F}(\mathbf{z}))c_{Y\mathbf{Z}}(F(y), \boldsymbol{F}(\mathbf{z}))}{c_{\mathbf{Z}}(\boldsymbol{F}(\mathbf{z}))} \tag{7.6}
$$

Pour comparer les deux modèles, une distance sur l'espace des distributions doit être choisie. Su et al. (2008b) utilisent la distance de Hellinger :

$$
h(c_{XY\mathbf{Z}}, \frac{c_{XY}c_{XZ}}{c_{\mathbf{Z}}}) = \int_{\mathbb{I}^{l+2}} \left(1 - \sqrt{\frac{c_{X,\mathbf{Z}}(u, \mathbf{w})c_{v,\mathbf{w}}(v, \mathbf{w})}{c_{X,Y,\mathbf{Z}}(u, v, \mathbf{w})c_{\mathbf{Z}}(\mathbf{w})}}\right)^2 \mathrm{d}C_{XY\mathbf{Z}}. \tag{7.7}
$$

Un estimateur non-paramétrique de cette quantité est alors obtenu en remplaçant les copules densités $c_{XY\mathbf{Z}}, c_{XY}, c_{XZ}$ et $c_{\mathbf{Z}}$ par leur copule densité de Bernstein et en remplaçant la copule $C_{XY\mathbf{Z}}$ par la copule empirique :

$$
\begin{aligned}
\hat{h} &= \int_{\mathbb{I}^{l+2}} \left(1 - \sqrt{\frac{\hat{c}^B_{X,\mathbf{Z}}(u, \mathbf{w})\hat{c}^B_{Y,\mathbf{Z}}(v, \mathbf{w})}{\hat{c}^B_{X,Y,\mathbf{Z}}(u, v, \mathbf{w})\hat{c}^B_{\mathbf{Z}}(\mathbf{w})}}\right)^2 \mathrm{d}\hat{C}_{XY\mathbf{Z}} \\
&= \frac{1}{m}\sum_{j=1}^{m} \left(1 - \sqrt{\frac{\hat{c}_{X,\mathbf{Z}}(x[m], \mathbf{z}[m])\hat{c}_{Y,\mathbf{Z}}(y[m], \mathbf{z}[m])}{\hat{c}_{X,Y,\mathbf{Z}}(x[m], y[m], \mathbf{z}[m])\hat{c}_{\mathbf{Z}}(\mathbf{z}[m])}}\right)^2
\end{aligned}
$$

Bouezmarni et al. (2012) ont montré que sous l'hypothèse nulle et sous certaines conditions sur le paramètre K, le test suivant était distribué asymptotiquement selon une loi normale standard :

$$
\mathrm{BRT} = \frac{K^{-(l+2)/2}}{\sigma}\left(4m\hat{h} - C_1K^{(l+2)/2} - B_1K - B_2K^{(l+1)/2} - B_3K^{1/2}\right)
$$

avec $C_1 = 2^{-(l+2)}\pi^{(l+2)/2}$, $\sigma = \sqrt{2}(\frac{\pi}{4})^{(l+2)/2)}$ et

$$B_1 = -\frac{\pi}{2} + \frac{1}{m}\sum_{j=1}^{m} \frac{(4\pi u[j](1-u[j]))^{-1/2}(4\pi v[j](1-v[j]))^{-1/2}}{\hat{c}_{XY}(u[j], v[j])}$$

$$B_2 = -2^{-l}\pi^{(l+1)/2} + \sum_{j=1}^{m} \frac{(4\pi u[j](1-u[j]))^{-1/2}\prod_{i=1}^{l}(4\pi w_i[j](1-w_i[j]))^{-1/2}}{\hat{c}_{XZ}(u[j], \boldsymbol{w}[j])}$$

$$B_3 = \pi^{-1/2}\frac{1}{m}\sum_{j=1}^{m} \frac{1}{\sqrt{u[j](1-u[j])}}$$

Ainsi, nous pouvons nous servir de ce test non-paramétrique afin de calculer des p-values et décider si oui ou non l'hypothèse d'indépendance conditionnelle $X \perp\!\!\!\perp Y \mid Z$ est rejetée. En particulier, nous allons implémenter un algorithme PC pour les CBNs à partir de ce test. Les étapes de l'algorithme sont les mêmes que celles présentées en 5.2.2.1 et seul le test change. Une fois la structure reconstruite, nous utilisons alors la formule 6.43 pour calculer les facteurs $R_{X_i|\mathbf{Pa}_i}$ et obtenir un modèle non-paramétrique. Comme pour l'algorithme utilisant le score BIC, nous travaillerons dans l'espace de la copule ce qui nous évitera d'avoir à estimer les marginales. Enfin, d'autres tests d'indépendances basés sur le schéma que nous venons de voir ont été proposés par Belalia et al. (2017). Dans le prochain chapitre, nous généralisons la dérivation de ces tests en utilisant la notion de f-divergence nous permettant d'avoir une méthode systématique pour l'obtention de tests d'indépendance non-paramétriques basés sur la copule de Bernstein empirique.

7.3 Protocole expérimental

Pour pouvoir mesurer les performances d'un algorithme d'apprentissage structurel, nous avons besoin d'échantillons dont nous connaissons la structure de dépendance. Ainsi, nous pouvons comparer la structure de référence avec celle apprise par l'algorithme. La distribution étant, la plupart du temps, inconnue dans des cas applicatifs, il en va de même pour sa structure de dépendance. Pour cette raison, nous devons recourir à des données synthétiques générées à partir d'une distribution que nous avons nous même construit et dont nous connaissons les indépendances. Cette construction repose sur le choix d'une structure de référence et d'une paramétrisation qui vont avoir un impact sur les résultats. Enfin, nous avons également besoin de choisir quelle métrique utiliser pour quantifier les différences entre la structure apprise et la structure de référence. Nous présentons ici en détails les choix que nous avons faits et qui sont à l'origine du protocole expérimental que nous allons utiliser par la suite pour comparer les différents algorithmes.

7.3.1 Structures de référence

Les données pour mesurer les performances des algorithmes ont été échantillonnées soit à partir de la structure du réseau ALARM (Beinlich et al. 1989) soit à partir de DAGs générés aléatoirement. Avec le réseau ALARM, nous avons une structure provenant d'une application concrète tandis qu'avec les structures aléatoires, nous pouvons obtenir des résultats plus généraux. Nous présentons maintenant les caractéristiques de ces structures et, pour les structures aléatoires, comment elles ont été générées.

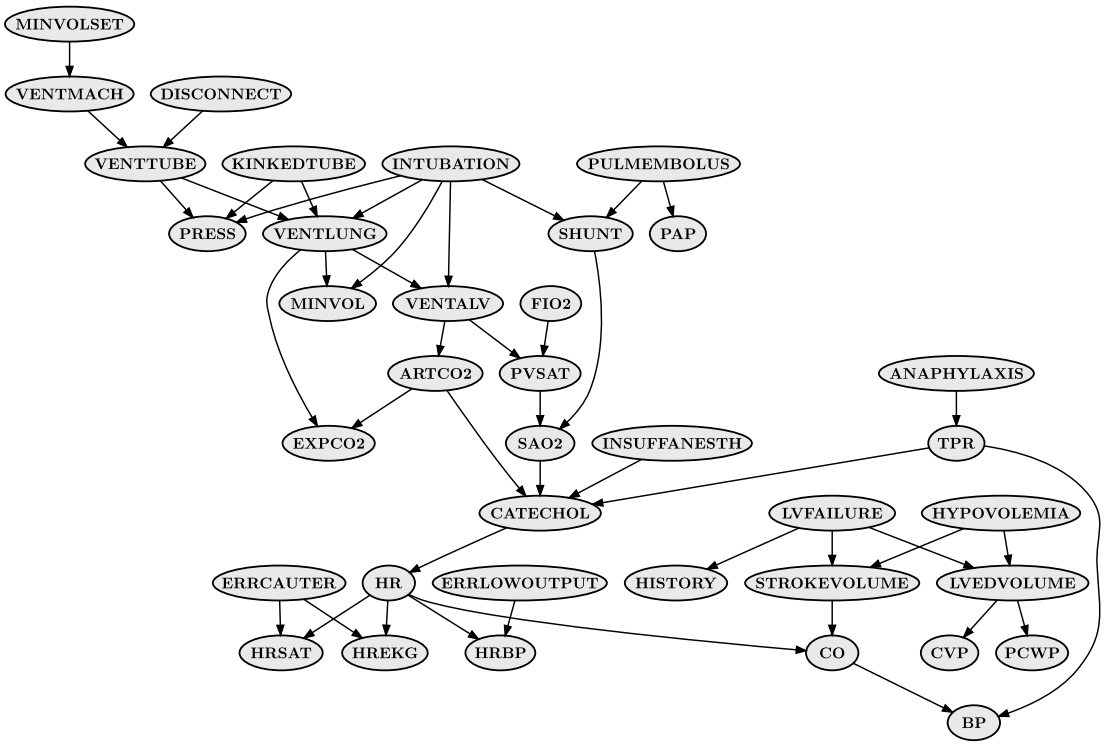


FIGURE 7.2 – Structure du réseau ALARM utilisée pour la construction de CBNs.

7.3.1.1 Le réseau ALARM

Le réseau ALARM, pour *A Logical Alarm Reduction Mechanism*, est un BN discret souvent utilisé pour tester les algorithmes d'apprentissage et servant à l'origine pour le diagnostic et le suivi médical de patients. La structure, représentée sur la figure 7.2, contient 37 nœuds, 46 arcs et 24 v-structures. Le détail des variables et des paramètres du BN ne nous intéressent pas pour la suite puisque nous voulons juste nous servir de la structure pour créer un CBN à partir duquel nous allons générer des données continues.

7.3.1.2 Structures aléatoires

Les structures aléatoires sont générées en suivant la procédure de IDE et al. (2002) qui, pour un nombre de nœuds et d'arcs fixé, propose de construire une *Monte-Carlo Markov Chain* (MCMC) convergeant vers une distribution uniforme sur l'ensemble des DAGs. Les structures aléatoires vont nous permettre d'observer l'évolution des performances des algorithmes en fonction du nombre de nœuds dans le graphe, c'est-à-dire de la dimension du problème. Il nous faut donc spécifier le nombre d'arcs voulu dans ces structures, sachant que l'utilisation de modèles graphiques n'a du sens que lorsque le modèle d'indépendance contient relativement peu d'arcs. En se basant sur le rapport des nombres de nœuds et d'arcs présents dans la structure ALARM, soit $\frac{46}{37} \approx 1.24$, nous avons décidé qu'étant donnée une structure de taille n, elle contiendrait $\lfloor 1.2 \times (n-1) \rfloor$ arcs. Un DAG aléatoire de taille 22 et généré selon les modalités que nous venons de décrire est représenté sur la figure 7.3.

7.3.2 Paramétrisation

Une fois le type de structure sélectionné, c'est-à-dire ALARM ou aléatoire, celle-ci doit être paramétrée selon la définition 7.1.1 pour obtenir un CBN. Comme nous tra-

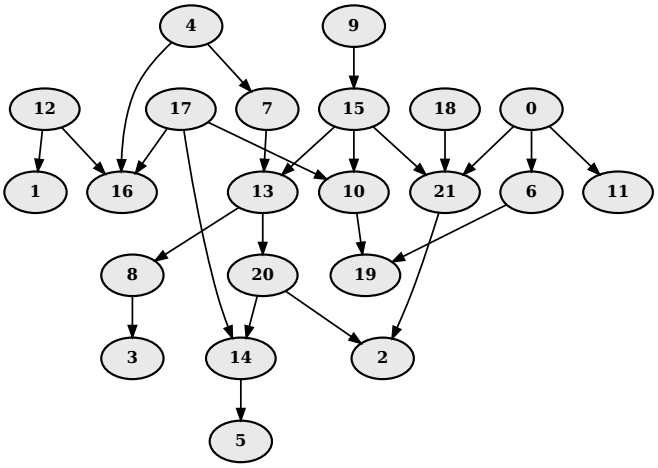
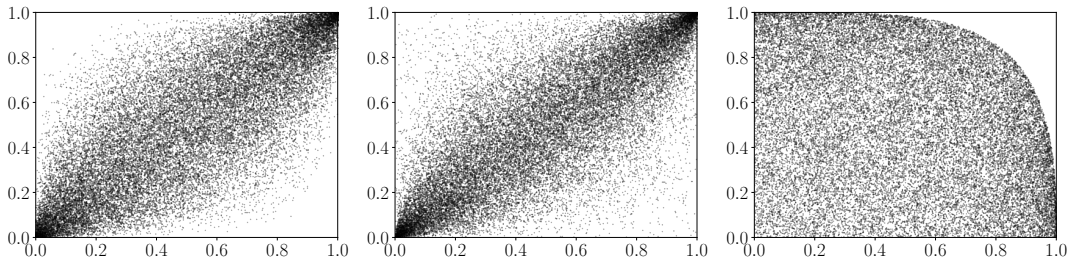


FIGURE 7.3 – Structure aléatoire de taille 22 générée aléatoirement et utilisée pour la construction de CBNs.

vaillons dans l'espace de la copule, nous n'avons besoin d'échantillonner que la copule et pour cette raison, nous n'aurons pas besoin de choisir un modèle pour les marginales f_{X_i} associées à chaque variable. Les copules locales $c_{X_i,\mathbf{Pa}_i}$ du CBN sont paramétrées avec les trois modèles de copules que nous avons présentés dans le chapitre précédent : gaussien, Student ou Dirichlet. Le modèle gaussien joue le rôle de modèle de référence puisque plusieurs des algorithmes que nous utilisons font l'hypothèse que les données sont distribuées selon une loi normale. La paramétrisation avec des copules de Student permet de rester proche du modèle gaussien mais avec la différence que ce modèle permet d'avoir des dépendances de queue. Ces deux modèles vont permettre d'évaluer les algorithmes non-paramétriques dans des situations où les algorithmes faisant l'hypothèse gaussienne sont *a priori* bien adaptés. Quant à la paramétrisation avec des copules de Dirichlet, celle-ci va servir à mettre les algorithmes en difficulté puisque la copule de Dirichlet est définie sur un support restreint (voir Figure (7.4)). De plus, elle va nous permettre de vérifier que les algorithmes non-paramétriques se généralisent bien à tout type de modèle contrairement aux algorithmes paramétriques. Pour ce qui est de la valeur des paramètres, nous les avons choisis de manière à induire de fortes corrélations entre les variables. Les copules gaussiennes et de Student sont paramétrées de sorte que leurs matrices de corrélation ont leurs coefficients hors-diagonale tous fixés à une même valeur $\rho = 0.8$. Pour que les copules de Student diffèrent légèrement des copules gaussiennes, nous avons fixé leur nombre de degrés de liberté à $\nu = 5$. Enfin, les copules de Dirichlet sont paramétrées avec $\boldsymbol{\alpha} = (\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}, 1)$ pour que leur support soit assez restreint. La figure (7.4) montre des échantillons obtenus à partir de copules bidimensionnelles paramétrées de cette manière.

7.3.3 Génération des données

Une fois les CBNs construits, nous pouvons nous en servir pour générer des données. Dans le chapitre 1, nous avons vu comment simuler une variable aléatoire selon n'importe quelle distribution à l'aide de la méthode de transformation inverse (cf. 1.5). La méthode dite de *forward sampling* (KOLLER et al. 2009, chapitre 12) consiste à utiliser cette technique successivement pour chaque variable et selon un ordre topologique relatif à la structure.



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

FIGURE 7.4 – Échantillons de copules densités gaussiennes, Student et Dirichlet. Le paramètre de corrélation de la copule gaussienne est fixé à $\rho = 0.8$, la copule de Student est prise avec $\nu = 5$ degrés de liberté et un paramètre de corrélation $\rho = 0.8$, les paramètres de la copule de Dirichlet sont fixés à $\boldsymbol{\alpha} = (1/3, 2/3, 1)$.

Définition 7.3.1 (Ordre topologique)**.** Soit $G = (V, A)$ un DAG et $(X, Y) \in V^2$ un couple de variables. Un ordre $\prec$ sur V est un ordre topologique relativement à G si, quand $X \to Y$ alors $X \prec Y$.

***Exemple* 7.3.1.** Considérons le DAG de la figure 4.4. Un ordre topologique possible est $A \prec_1 B \prec_1 C \prec_1 E \prec_1 D \prec_1 F$. Celui-ci n'est cependant pas unique et $B \prec_2 E \prec_2 F \prec_2 A \prec_2 C \prec_2 D$ est un autre ordre valable.

Le fait d'échantillonner les variables selon un ordre topologique nous assure que si la variable X_i à échantillonner a des parents $\mathbf{Pa}_i$, alors ceux-ci ont déjà été échantillonnés. En utilisant les valeurs $\mathbf{pa}_i[j]$ échantillonnées dans la densité conditionnelle associée à X_i, il suffit alors d'échantillonner selon $f_{X_i|\mathbf{pa}_i}$. En répétant cette opération jusqu'à ce que toutes les variables X_i de la structure aient été échantillonnées, on obtient alors une observation $\mathbf{x}[j]$ du vecteur $\mathbf{X}$. Il est simple d'obtenir un échantillon de taille quelconque en répétant la procédure le nombre de fois voulu. Les facteurs $R_{X_i|\mathbf{Pa}_i}$ associés à chaque nœud d'un CBN étant des densités conditionnelles, nous pouvons les échantillonner comme n'importe quelle densité classique et l'adaptation aux CBNs de la méthode du *forward sampling* est directe. Notons néanmoins que si la variable X_i n'a pas de parents, le facteur R_{X_i} vaut identiquement 1 ce qui revient à échantillonner selon une loi uniforme sur $\mathbb{I}$.

***Exemple* 7.3.2.** Reprenons le CBN de la figure 7.1 que l'on paramètre par des copules gaussiennes. Un ordre topologique (qui est unique dans ce cas) est $X_1 \prec X_2 \prec X_3$. Pour échantillonner X_1, nous faisons un tirage uniforme sur $\mathbb{I}$. À partir de la valeur $x_1[1]$ obtenue, nous pouvons à présent échantillonner X_2 selon la copule gaussienne c_{X_2,x_1} dont la corrélation est $\rho = 0.8$. Nous pouvons enfin échantillonner la variable X_3 dont la copule locale associée est c_{X_3,X_1,X_2} et dont la matrice de corrélation est donnée par :

$$\Sigma_{X_3,X_1,X_2} = \begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix} \tag{7.8}$$

7.3.4 Scores pour la structure

Pour mesurer les performances d'un algorithme d'apprentissage, nous générons des données selon la méthodologie que nous venons de décrire et reconstruisons une structure à partir de ces données. La structure apprise est ensuite comparée à la structure de référence ayant servi à générer les données. Reste alors à choisir une métrique permettant de résumer les différences entre les graphes au sein d'un scalaire. Les méthodes d'apprentissage utilisant des contraintes se déroulent en trois étapes : reconstruction du squelette, recherche des v-structures et propagation des contraintes. Pour cette raison, nous avons choisi de comparer les structures au niveau du squelette et du CPDAG. Nous utilisons pour cela le F-score et la distance de Hamming structurelle (SHD pour *Structural Hamming Distance*).

7.3.4.1 F-score

Bien que le F-score soit normalement utilisé pour mesurer les performances de classifieurs binaires, il peut être adapté à notre cas. En effet, dans le cas d'un graphe non-orienté nous sommes également dans un cas binaire puisqu'un lien est soit présent soit absent. On dit alors d'un lien que c'est un :

- **Vrai-positif** (VP), s'il est à la fois présent dans la structure de référence et dans la structure inférée.
- **Vrai-négatif** (VN), s'il est absent à la fois de la structure de référence et de la structure inférée.
- **Faux-positif** (FP), s'il est absent de la structure de référence mais présent dans la structure inférée.
- **Faux-négatif** (FN), s'il est présent dans la structure de référence mais absent de la structure inférée.

Cela nous permet de définir la précision (P) qui est la proportion de liens inférés qui sont effectivement dans la structure de référence et le rappel (R) (ou *recall* en anglais) qui est la proportion de liens présents dans la structure de référence qui ont été inférés par l'algorithme. Ces deux quantités ont pour expression :

$$\mathrm{P} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}} \quad \text{et} \quad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{7.9}$$

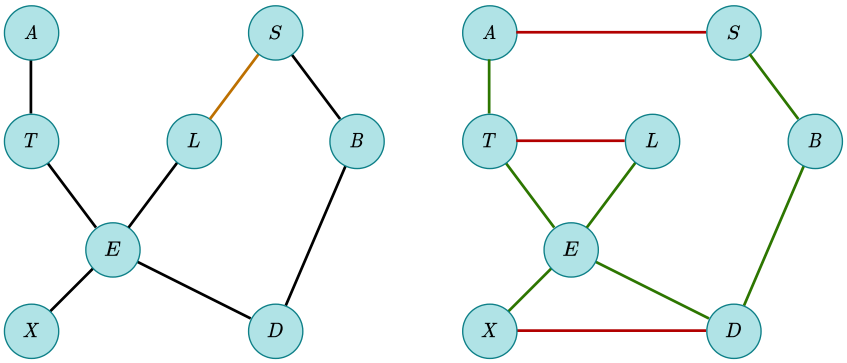
Enfin, le F-score est défini comme la moyenne hyperbolique de la précision et du rappel :

$$\mathrm{F} = \left(\frac{\mathrm{P}^{-1} + \mathrm{R}^{-1}}{2} \right)^{-1} = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}. \tag{7.10}$$

Si le squelette de référence a été parfaitement reconstruit par l'algorithme d'apprentissage, la valeur du F-score est de 1. Donc, plus le F-score est proche de cette valeur plus l'algorithme est performant.

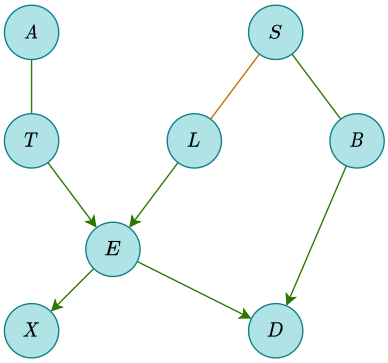
7.3.4.2 Distance de Hamming structurelle

Dans le cas d'un graphe orienté, nous ne pouvons pas utiliser le F-score puisqu'un arc peut être absent, orienté dans un sens ou bien orienté dans l'autre. Nous ne sommes plus dans un cadre binaire et devons recourir à une autre métrique. Rappelons que plusieurs DAGs différents peuvent encoder les mêmes indépendances et l'orientation de certains arcs n'a d'influence que sur la paramétrisation. Pour cette raison, nous utilisons la distance de Hamming structurelle (SHD pour *Structural Hamming Distance*) (Colombo et al. 2014) qui se calcule sur le CPDAG plutôt que sur le DAG. Elle

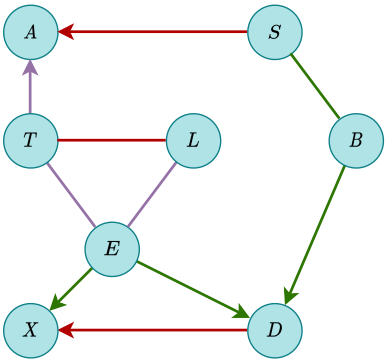


(a) Squelette de référence

(b) Squelette appris



(c) CPDAG de référence

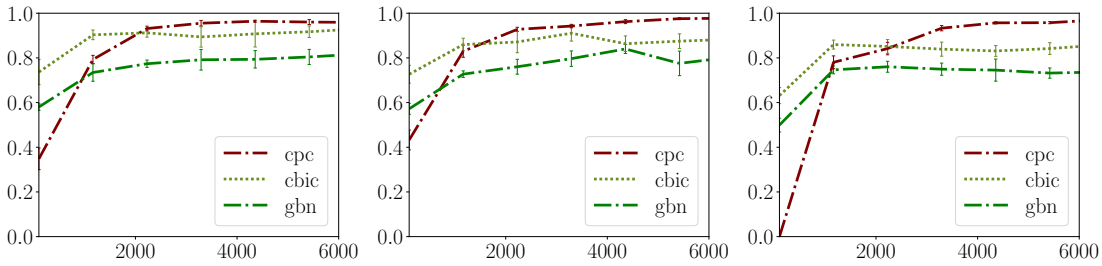


(d) CPDAG appris

FIGURE 7.5 – Comparaison entre le squelette/CPDAG appris et le squelette/CPDAG de référence. Le squelette appris a un F-score de $\frac{49}{75}$ et le CPDAG appris à une SHD de 7. Pour les squelettes, les liens en vert correspondent aux vrais-positifs, les liens en rouge aux faux-positifs et les liens en orange aux faux-négatifs. Pour les CPDAG, les liens en vert sont bien orientés, les liens en violet sont mal orientés, les liens en rouge doivent être supprimés et les liens en orange doivent être ajoutés.

décompte le nombre d'opérations nécessaires pour passer d'une structure à une autre. Ces opérations sont l'insertion, la suppression ou bien le changement d'orientation d'un lien/arc. Si le CPDAG de référence est parfaitement reconstruit, alors la valeur de la SHD est nulle. Ainsi, plus la SHD est proche de cette valeur plus l'algorithme est performant. L'exemple suivant donne le calcul du F-score et de la SHD pour les squelettes et CPDAGs donnés sur la figure 7.5.

***Exemple* 7.3.3.** Supposons que le squelette et le CPDAG de références soient donnés par les graphes des figures 7.5a et 7.5c et que l'algorithme d'apprentissage reconstruise un DAG ayant pour squelette et CPDAG les graphes des figures 7.5b et 7.5d. Pour le squelette, nous pouvons voir qu'il existe 7 vrai-positifs (en vert), 3 faux-positifs (en rouge), et 1 faux-négatif (en orange). La précision de l'algorithme est donc de $P = \frac{7}{7+3} = \frac{7}{10}$ et son rappel est de $R = \frac{7}{7+1} = \frac{7}{8}$. Enfin, le F-score a une valeur de $F = 2\frac{P \times R}{P+R} = \frac{49}{75}$. Pour ce qui est du CPDAG, il existe 4 liens/arcs qui ont la bonne orientation (en vert), 3 liens/arcs à supprimer (en rouge), 3 liens/arcs ayant une mauvaise orientation (en violet) et 1 lien/arc à ajouter (en orange). Le nombre total d'opérations pour passer d'un CPDAG à l'autre est donc de SHD = 7.



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

FIGURE 7.6 – Évolution du F-score pour les méthodes CBIC, CPC et GBN en fonction de la taille de l'échantillon d'apprentissage. La moyenne des résultats est calculée sur 5 réinitialisations avec différents échantillons générés à partir de la structure du réseau ALARM.

7.3.5 Code source et plugin otagrum

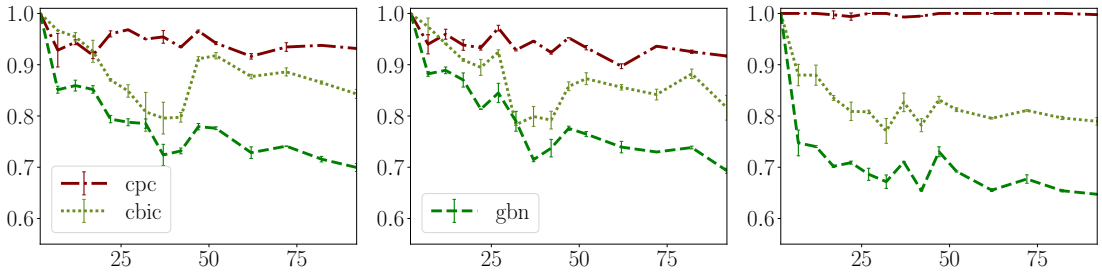
L'ensemble des méthodes pour la reconstruction de CBNs que nous présentons dans cette thèse sont utilisables via le *plugin* pour la bibliothèque OpenTURNS appelé *otagrum*. Ce *plugin* fait le pont entre les bibliothèques OpenTURNS (BAUDIN et al. 2016) et aGrUM (DUCAMP et al. 2020) pour implémenter les CBNs ainsi que l'ensemble des algorithmes d'apprentissage. La première permet la manipulation de modèles continus multivariés, et en particulier de copules, tandis que la deuxième permet la manipulation de modèles graphiques. Dans la même philosophie qu'OpenTURNS et aGrUM, le code source d'otagrum est écrit dans le langage C++ et ses classes et méthodes sont ensuite interfacées en langage python à l'aide de l'outil SWIG. Le code source est directement accessible et téléchargeable sur le répertoire gitHub openturns/otagrum. Les détails de l'installation du *plugin* ainsi que sa documentation sont disponibles en ligne. De même, le code source pour les *scripts* permettant la génération des données synthétiques, l'apprentissage des structures à partir des algorithmes et la production des courbes d'apprentissage est disponible sur le répertoire gitHub MLasserre/otagrum-experiments.

7.4 Résultats numériques

Nous présentons maintenant les résultats des expériences numériques qui ont été menées en suivant le protocole expérimental que nous venons de décrire. Les algorithmes comparés ici sont l'algorithme CPC, l'algorithme de ELIDAN (2010) – que nous appelons CBIC dans la suite – et un algorithme d'apprentissage utilisant un score bayésien gaussien (GEIGER et HECKERMAN 1994) pour la reconstruction de GBNs (cf. 4.5.2) – appelé GBN dans la suite. Comme pour le score BIC, le score bayésien gaussien est maximisé sur l'ensemble des DAGs en utilisant une méthode de recherche locale. Afin d'éviter les maxima locaux, cette recherche est, pour CBIC et GBN, réinitialisée 5 fois tout en utilisant une liste *TABU* de taille 5. Enfin, pour éviter les graphes trop denses, la recherche est contrainte aux graphes dont le nombre maximum de parents d'un nœud est 4. Quant à la méthode CPC, nous fixons la significativité des tests d'indépendance à $p = 0.05$.

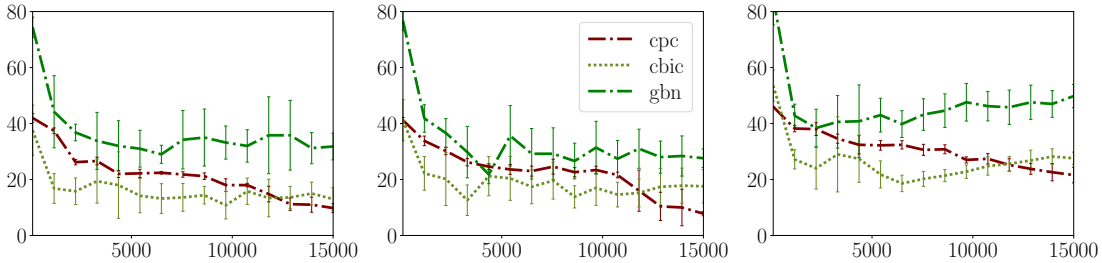
7.4.1 Performances pour la reconstruction du squelette

La figure 7.6 donne l'évolution du F-score des structures reconstruites par chacune des méthodes en fonction de la taille de l'échantillon d'apprentissage lorsque la struc-



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

FIGURE 7.7 – Évolution du F-score pour les méthodes CBIC, CPC et GBN en fonction de la dimension des graphes aléatoires. Pour chaque dimension, les résultats sont moyennés sur 2 graphes aléatoires différents et sur 5 échantillons de taille $m = 10\,000$.



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

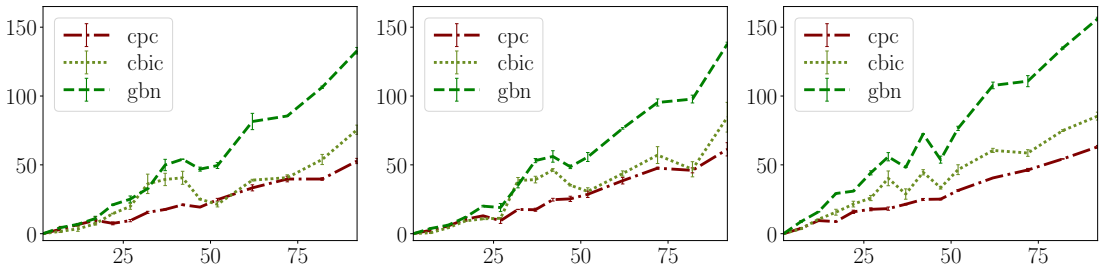
FIGURE 7.8 – Évolution de la SHD pour les méthodes CBIC, CPC et GBN en fonction de la taille de l'échantillon d'apprentissage. La moyenne des résultats est calculée sur 5 réinitialisations avec différents échantillons générés à partir de la structure du réseau ALARM.

ture de référence est celle du réseau ALARM. Pour une taille donnée, ces résultats sont moyennés sur 5 échantillons différents. Nous observons sur ces figures que l'algorithme CPC a de meilleures performances que CBIC et GBN pour n'importe quelle paramétrisation des CBNs, dès lors que la taille de l'échantillon est suffisante ($m \geq 1000$). Bien que les méthodes CBIC et GBN utilisent toutes deux des scores gaussiens, la méthode CBIC est celle qui obtient les meilleurs résultats pour n'importe quelle taille d'échantillon. Enfin, malgré le fait que la paramétrisation de Dirichlet soit éloignée du modèle gaussien, la reconstruction du squelette ne semble pas en être affectée.

La figure 7.7 donne l'évolution du F-score en fonction de la taille des structures aléatoires utilisées pour générer les données. Pour chaque dimension, les résultats sont moyennés sur 2 structures aléatoires différentes et, pour chacune de ces structures, sur 5 échantillons différents de taille $m = 10\,000$. La dimension des graphes aléatoires va de $n = 2$ à $n = 92$. Nous observons une nouvelle fois que l'algorithme CPC est celui qui a les meilleurs résultats pour n'importe quel type de données et que l'algorithme CBIC est supérieur à l'algorithme GBN pour n'importe quelle dimension. Dans le cas de données issues de copules de Dirichlet, nous observons que les performances de CBIC et GBN décroissent plus rapidement que dans les cas gaussien et de Student contrairement à CPC qui montre même de meilleurs résultats dans ce cas là.

7.4.2 Performances pour la reconstruction du CPDAG

De même que pour le F-score, les figures 7.8 et 7.9 montrent respectivement l'évolution de la SHD en fonction de la taille de l'échantillon pour la structure du réseau ALARM et l'évolution de la SHD en fonction de la dimension pour les structures aléa-



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

FIGURE 7.9 – Évolution de la SHD pour les méthodes CBIC, CPC et GBN en fonction de la dimension des graphes aléatoires. Pour chaque dimension, les résultats sont moyennés sur 2 graphes aléatoires différents et sur 5 échantillons de taille $m = 10\,000$.

toires. Ces courbes sont également moyennées sur 5 échantillons dans le premier cas et 2×5 échantillons dans le deuxième. De nouveau, nous observons que les performances des algorithmes CBIC et CPC sont supérieures à celles de l'algorithme GBN pour n'importe quelle taille d'échantillon dans le cas de la structure ALARM et pour n'importe quelle dimension dans le cas des structures aléatoires. Bien que pour de grandes tailles d'échantillons l'algorithme CPC reste le meilleur des trois algorithmes pour la structure ALARM, l'algorithme CBIC a, cette fois-ci, de meilleures performances lorsque cette taille est inférieure à $m \sim 12\,000$. En revanche, dans le cas des structures aléatoires, les performances de l'algorithme CPC restent supérieures pour quasiment toutes les dimensions.

Ainsi, notre algorithme est globalement meilleur que l'algorithme proposé par ELIDAN (2010) et les algorithmes basés sur les copules montrent de meilleurs résultats que l'algorithme classique pour les GBNs. Dans le prochain chapitre, nous introduisons un autre algorithme d'apprentissage pour les CBNs appelé CMIIC. Nous allons voir que celui-ci surclasse tous les algorithmes que nous avons introduits dans ce chapitre.

Références

BAUDIN, M., DUTFOY, A., IOOSS, B. et POPELIN, A.-L. (2016). « OpenTURNS : An Industrial Software for Uncertainty Quantification in Simulation ». In : *Handbook of Uncertainty Quantification*. Sous la dir. de R. GHANEM, D. HIGDON et H. OWHADI. Cham : Springer International Publishing, p. 1-38 (cf. p. 135).

BEDFORD, T. et COOKE, R. M. (2002). « Vines–a new graphical model for dependent random variables ». In : *The Annals of Statistics* 30.4, p. 1031-1068 (cf. p. 3, 124).

BEINLICH, I. A., SUERMONDT, H. J., CHAVEZ, R. M. et COOPER, G. F. (1989). « The ALARM monitoring system : A case study with two probabilistic inference techniques for belief networks ». In : *AIME 89*. Springer, p. 247-256 (cf. p. 129).

BELALIA, M., BOUEZMARNI, T., LEMYRE, F. et TAAMOUTI, A. (2017). « Testing independence based on Bernstein empirical copula and copula density ». In : *Journal of Nonparametric Statistics* 29.2, p. 346-380 (cf. p. 129, 145, 147, 163).

BOUEZMARNI, T., ROMBOUTS, J. V. et TAAMOUTI, A. (2010a). « Asymptotic properties of the Bernstein density copula estimator for α-mixing data ». In : *Journal of Multivariate Analysis* 101.1, p. 1-10 (cf. p. 123).

BOUEZMARNI, T., ROMBOUTS, J. V. et TAAMOUTI, A. (2010b). « Asymptotic properties of the Bernstein density copula estimator for α-mixing data ». In : *Journal of Multivariate Analysis* 101.1, p. 1-10 (cf. p. 128, 141, 142).

Bouezmarni, T., Rombouts, J. V. et Taamouti, A. (2012). « Nonparametric copula-based test for conditional independence with applications to Granger causality ». In : *Journal of Business & Economic Statistics* 30.2, p. 275-287 (cf. p. 128, 161, 163).

Colombo, D. et Maathuis, M. H. (2014). « Order-independent constraint-based causal structure learning ». In : *The Journal of Machine Learning Research* 15.1, p. 3741-3782 (cf. p. 90, 133, 162).

Czado, C. (2010). « Pair-copula constructions of multivariate copulas ». In : *Copula theory and its applications*. Springer, p. 93-109 (cf. p. 3, 124).

Ducamp, G., Gonzales, C. et 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*. T. 138. Proceedings of Machine Learning Research. Skørping, Denmark, p. 609-612 (cf. p. 135).

Elidan, G. (2010). « Copula bayesian networks ». In : *Advances in neural information processing systems*, p. 559-567 (cf. p. 3, 5, 6, 124, 125, 127, 135, 137, 141, 142, 149, 161).

Geiger, D. et Heckerman, D. (1994). « Learning gaussian networks ». In : *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., p. 235-243 (cf. p. 5, 80, 135, 161).

Hanea, A., Napoles, O. M. et Ababei, D. (2015). « Non-parametric Bayesian networks : Improving theory and reviewing applications ». In : *Reliability Engineering & System Safety* 144, p. 265-284 (cf. p. 124).

Hanea, A. M. (2008). « Algorithms for non-parametric Bayesian belief nets ». In : (cf. p. 124).

Ide, J. S. et Cozman, F. G. (2002). « Random generation of Bayesian networks ». In : *Brazilian symposium on artificial intelligence*. Springer, p. 366-376 (cf. p. 130).

Kirshner, S. (2008). « Learning with tree-averaged densities and distributions ». In : *Advances in Neural Information Processing Systems*, p. 761-768 (cf. p. 125).

Koller, D. et Friedman, N. (2009). *Probabilistic graphical models : principles and techniques*. MIT press (cf. p. 3, 71, 72, 75, 76, 80, 131, 154).

Kurowicka, D. et Cooke, R. (2005). « Distribution-free continuous Bayesian belief ». In : *Modern statistical and mathematical methods in reliability* 10, p. 309 (cf. p. 124).

Langseth, H., Nielsen, T. D., Rumi, R. et Salmerón, A. (2012). « Mixtures of truncated basis functions ». In : *International Journal of Approximate Reasoning* 53.2, p. 212-227 (cf. p. 2, 124).

Lasserre, M., Lebrun, R. et Wuillemin, P.-H. (2020). « Constraint-Based Learning for Non-Parametric Continuous Bayesian Networks ». In : *FLAIRS 33 - 33rd Florida Artificial Intelligence Research Society Conference*. Miami, United States : AAAI, p. 581-586 (cf. p. 5, 124).

Lasserre, M., Lebrun, R. et Wuillemin, P.-H. (2021a). « Constraint-based learning for non-parametric continuous bayesian networks ». In : *Annals of Mathematics and Artificial Intelligence*, p. 1-18 (cf. p. 6, 124).

Lauritzen, S. L. (1992). « Propagation of probabilities, means, and variances in mixed graphical association models ». In : *Journal of the American Statistical Association* 87.420, p. 1098-1108 (cf. p. 124).

Lauritzen, S. L. et Wermuth, N. (1989). « Graphical models for associations between variables, some of which are qualitative and some quantitative ». In : *The annals of Statistics*, p. 31-57 (cf. p. 2, 124).

Moral, S., Rumí, R. et Salmerón, A. (2001). « Mixtures of truncated exponentials in hybrid Bayesian networks ». In : *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer, p. 156-167 (cf. p. 2, 124).

Parzen, E. (1962). « On estimation of a probability density function and mode ». In : *The annals of mathematical statistics* 33.3, p. 1065-1076 (cf. p. 127).

Reynolds, D. A. (2009). « Gaussian mixture models. » In : *Encyclopedia of biometrics* 741, p. 659-663 (cf. p. 124).

Romero, V., Rumí, R. et Salmerón, A. (2006). « Learning hybrid Bayesian networks using mixtures of truncated exponentials ». In : *International Journal of Approximate Reasoning* 42.1-2, p. 54-68 (cf. p. 2, 124).

Rousseeuw, P. J. et Molenberghs, G. (1993). « Transformation of non positive semidefinite correlation matrices ». In : *Communications in Statistics–Theory and Methods* 22.4, p. 965-984 (cf. p. 127).

Shenoy, P. P. et West, J. C. (2011). « Inference in hybrid Bayesian networks using mixtures of polynomials ». In : *International Journal of Approximate Reasoning* 52.5, p. 641-657 (cf. p. 2, 124).

Su, L. et White, H. (2008a). « A Nonparametric Hellinger Metric Test for Conditional Independence ». In : *Econometric Theory* 24.4, p. 829-864 (cf. p. 123).

Su, L. et White, H. (2008b). « A nonparametric Hellinger metric test for conditional independence ». In : *Econometric Theory*, p. 829-864 (cf. p. 128).

Wan, J. et Zabaras, N. (2014). « A probabilistic graphical model based stochastic input model construction ». In : *J. Comput. Physics* 272, p. 664-685 (cf. p. 123, 124).

Chapitre 8

CMIIC : un algorithme basé sur l'information mutuelle

Sommaire

Dans le chapitre précédent, nous avons donné un cadre pour l'apprentissage de réseaux bayésiens continus non-paramétriques en combinant le modèle des CBNs (Elidan 2010) avec celui de la copule de Bernstein empirique. En utilisant le test d'indépendance non-paramétrique proposé par Bouezmarni et al. (2010b), nous avons alors proposé un algorithme PC pour l'apprentissage de la structure d'un CBN cohérent avec sa paramétrisation. Ce test d'indépendance est dérivé selon l'approche classique des statistiques et, comme l'ensemble des tests d'hypothèse que nous avons vus jusqu'ici, dépend donc d'une fonction mesurant l'écart entre la distribution sous l'hypothèse H_0 et la distribution sous l'hypothèse H_1. Il se trouve que ces fonctions sont réunies sous la définition de *f-divergence*, introduite indépendamment par Csiszár (1964) [1] et Ali et al. (1966), vérifiant un certain nombre de propriétés. En utilisant cette définition et la copule de Bernstein empirique, nous proposons ici un cadre général permettant de dériver un ensemble de tests d'indépendance non-paramétriques. L'entropie relative

1. Le lecteur non magyarophone pourra se référer à Csiszár (1967).

(cf. définition 2.1.3), qui fait partie de cette classe de fonctions, joue un rôle particulier pour l'inférence paramétrique basée sur le maximum de vraisemblance mais aussi dans les méthodes de sélection de modèle dans les contextes bayésien et NML. Pour cette raison, nous présentons le test d'indépendance qui lui est associé et dont la statistique de test est l'information conditionnelle. Pour en obtenir un estimateur, nous généralisons le lien existant entre l'entropie de la copule et l'information mutuelle (Ma et al. 2011) à l'information conditionnelle. Cet estimateur est ensuite utilisé d'une part pour accélérer la recherche locale menée par l'algorithme de reconstruction des CBNs proposé par Elidan (2010) et d'autre part pour implémenter un algorithme MIIC non-paramétrique. Les travaux présentés dans ce chapitre ont fait l'objet d'une publication dans Lasserre et al. (2021b).

8.1 Cadre pour la dérivation de tests d'indépendance non-paramétriques

Nous généralisons ici la dérivation du test non-paramétrique de Bouezmarni et al. (2010b) basé sur la distance de Hellinger et la copule de Bernstein empirique que nous avons vues dans le chapitre précédent. Cette généralisation va donner lieu à un cadre permettant la dérivation systématique de plusieurs tests d'indépendance non-paramétriques. Pour cela, nous commençons par remarquer que l'ensemble des statistiques de test pour la sélection de modèle que nous avons vues jusqu'à présent sont construites à partir de différentes fonctions résumant les différences entre la distribution sous l'hypothèse H_0 et sous l'hypothèse H_1, au sein d'un scalaire. Il s'avère que ces fonctions sont différentes instantiations d'une même quantité appelée f-divergence :

Définition 8.1.1 (f-divergence). Soit $f : \mathbb{R}_*^+ \to \mathbb{R}$ une fonction convexe telle que $f(1) = 0$ et soient $\mathbb{P}$ et $\mathbb{Q}$ deux distributions de probabilité définies sur l'espace probabilisable $(\Omega, \mathcal{A})$. Si $\mathbb{P} \ll \mathbb{Q}$, la f-divergence entre les deux distributions est donnée par

$$D_f\left(\mathbb{P}||\mathbb{Q}\right) = \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\right]. \tag{8.1}$$

La fonction f est appelée fonction génératrice et l'on note $\mathcal{F}$ l'ensemble de ces fonctions. Soit p et q les densités de $\mathbb{P}$ et $\mathbb{Q}$ par rapport à une mesure μ, la f-divergence s'écrit alors

$$D_f\left(\mathbb{P}||\mathbb{Q}\right) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x)\mathrm{d}\mu(x). \tag{8.2}$$

Nom de la divergence	Formule	Fonction génératrice
Entropie relative	$\int_\Omega p(x) \log \frac{p(x)}{q(x)} \mathrm{d}\mu$	$u \log u$
Divergence χ^2	$\int_\Omega \frac{(q(x)-p(x))^2}{p(x)} \mathrm{d}\mu$	$(u-1)^2$
Variation totale	$\frac{1}{2}\int_\Omega \lvert f(x) - g(x)\rvert \, \mathrm{d}\mu$	$\frac{1}{2}\lvert u-1\rvert$
Distance de Hellinger	$\int_\Omega \left(1 - \sqrt{\frac{p(x)}{q(x)}}\right)^2 q(x)\mathrm{d}\mu$	$(1-\sqrt{u})^2$

Table 8.1 – Ensemble des f-divergence rencontrées.

Le tableau 8.1 résume les principales f-divergences que nous avons rencontrées jus-

qu'ici ainsi que la fonction convexe qui leur est associée [2]. Par exemple, le test χ^2 que nous avons présenté au chapitre 3 dérive d'une f-divergence du même nom et dont la fonction convexe associée est $f(u) = (u-1)^2$. La f-divergence vérifie plusieurs propriétés liées à sa fonction génératrice :

Proposition 8.1.1 (Propriétés liées à la fonction génératrice)**.** Soient $\mathbb{P} \ll \mathbb{Q}$ deux distributions définies sur un même espace mesurable. Pour l'ensemble de ces distributions, la f-divergence vérifie les propriétés suivantes :

- **Linéarité** : Soit f une fonction génératrice telle que $f = \lambda_1 f_1 + \lambda_2 f_2$, alors $D_f(\mathbb{P}||\mathbb{Q}) = \lambda_1 D_{f_1}(\mathbb{P}||\mathbb{Q}) + \lambda_2 D_{f_2}(\mathbb{P}||\mathbb{Q})$.
- **Annulation pour une fonction génératrice linéaire** : La f-divergence s'annule si et seulement si sa fonction génératrice est définie comme $f(x) = c(x-1)$, avec $c \in \mathbb{R}$.
- **Invariance par ajout d'un terme linéaire** : Soit $c \in \mathbb{R}$ et soit $f^*(x) = f(x) + c(x-1)$, alors $D_{f^*}(\mathbb{P}||\mathbb{Q}) = D_f(\mathbb{P}||\mathbb{Q})$.

Démonstration. La linéarité vis-à-vis de la fonction génératrice découle de celle de l'espérance. Soit $f(u) = c(u-1)$, alors $D_f(\mathbb{P}||\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[f(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}})\right] = c\left(\mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right] - 1\right) = 0$, ce qui démontre la deuxièmre propriété. L'invariance par ajout d'un terme linéaire découle directement des deux autres propriétés. ■

L'invariance de la f-divergence par l'ajout d'un terme linéaire à la fonction génératrice nous permet de définir des classes d'équivalences sur $\mathcal{F}$ telles que f_1 et f_2 appartiennent à une même classe si $f_1(x) = f_2(x) + c(x-1)$. En reprenant l'exemple de la divergence χ^2, une autre fonction génératrice pourrait être $f(u) = u^2 - 1 = (u-1)^2 + 2(u-1)$. Les formules et les calculs peuvent parfois se trouver simplifiés par le bon choix d'un représentant de la classe. Pour une fonction génératrice donnée, plusieurs des bonnes propriétés que nous avons vues pour l'entropie relative se généralisent à la f-divergence :

Proposition 8.1.2 (Propriétés de base)**.** Soient $\mathbb{P}_{\mathbf{X}} \ll \mathbb{Q}_{\mathbf{X}}$ deux distributions de probabilités définies sur un même espace mesurable et soit f une fonction génératrice. La f-divergence vérifie les propriétés suivantes :

- **Positivité** : $D_f(\mathbb{P}||\mathbb{Q}) \geq 0$ et $D_f(\mathbb{P}||\mathbb{Q}) = 0$ si et seulement si $\mathbb{P} = \mathbb{Q}$.
- **Invariance par changement de variable** : Soit ϕ une fonction bijective différentiable et soit $\mathbf{Y} = \phi(\mathbf{X})$. Dans ce cas, $D_f(\mathbb{P}_{\mathbf{Y}}||\mathbb{Q}_{\mathbf{Y}}) = D_f(\mathbb{P}_{\mathbf{X}}||\mathbb{Q}_{\mathbf{X}})$.

Démonstration. Ces propriétés se démontrent de la même manière que pour l'entropie relative. ■

Enfin, il existe également une notion de f-divergence conditionnelle généralisant la définition 2.1.4 :

Définition 8.1.2 (f-divergence conditionnelle)**.** Soit $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ un vecteur aléatoire et soient $\mathbb{P}_{\mathbf{X}} \ll \mathbb{Q}_{\mathbf{X}}$ deux distributions jointes définies sur un même espace probabilisable. La f-divergence conditionnelle est définie par

$$D_f(\mathbb{P}_{\mathbf{X}_2|\mathbf{X}_1}||\mathbb{Q}_{\mathbf{X}_2|\mathbf{X}_1}|\mathbb{P}_{\mathbf{X}_1}) = \mathbb{E}_{\mathbb{P}_{\mathbf{X}_1}}\left[D_f(\mathbb{P}_{\mathbf{X}_2|\mathbf{X}_1}||\mathbb{Q}_{\mathbf{X}_2|\mathbf{X}_1})\right] \tag{8.3}$$

2. Pour une liste plus exhaustive, voir Sason (2018).

La règle de chaine (cf. proposition 2.1.4) en revanche ne s'étend pas à toutes les f-divergences et reste un cas particulier de l'entropie relative. Nous pouvons toutefois simplifier la f-divergence quand les deux distributions comparées possèdent une marginale en commun :

Proposition 8.1.3. Soit $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ un vecteur aléatoire et soient $\mathbf{X}_1$ et $\mathbf{X}_2$ deux sous-vecteurs disjoints. Si $\mathbb{P}_{\mathbf{X}} \ll \mathbb{Q}_{\mathbf{X}}$ sont deux distributions jointes de ce vecteur telles que $\mathbb{P}_{\mathbf{X}_1} = \mathbb{Q}_{\mathbf{X}_1}$, alors

$$D_f(\mathbb{P}_{\mathbf{X}} || \mathbb{Q}_{\mathbf{X}}) = D_f(\mathbb{P}_{\mathbf{X}_2|\mathbf{X}_1} || \mathbb{Q}_{\mathbf{X}_2|\mathbf{X}_1} | \mathbb{P}_{\mathbf{X}_1}) \tag{8.4}$$

Maintenant que nous avons donné les principales propriétés et définitions liées à la f-divergence, revenons à l'équation (8.2) exprimant la f-divergence en fonction du rapport des densités des deux distributions. Cette formule est intéressante dans le cadre d'un test d'indépendance puisque cela veut dire que la statistique de test n'est fonction que des densités des copules. Pour le voir, rappelons que dans le cas du test d'une indépendance conditionnelle $X_1 \perp X_2 | \mathbf{X}_3$, les distributions que nous cherchons à comparer sont :

$$\begin{cases} M_0 : \mathbf{X} \sim \mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_0} = \mathbb{P}_{\mathbf{X}_3|\boldsymbol{\theta}_{\mathbf{X}_3}} \otimes \mathbb{P}_{X_1,X_2|\mathbf{X}_3,\boldsymbol{\theta}_{X_1X_2|\mathbf{X}_3}} \\ M_1 : \mathbf{X} \sim \mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_1} = \mathbb{P}_{\mathbf{X}_3|\boldsymbol{\theta}_{\mathbf{X}_3}} \otimes \mathbb{P}_{X_1|\mathbf{X}_3,\boldsymbol{\theta}_{X_1|\mathbf{X}_3}} \otimes \mathbb{P}_{X_2|\mathbf{X}_3,\boldsymbol{\theta}_{X_2|\mathbf{X}_3}} \end{cases} \tag{8.5}$$

Pour n'importe quelle fonction génératrice, nous avons ainsi :

$$\begin{aligned}
&D_f\left(\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_0} || \mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_1}\right) \\
&\quad = D_f\left(\mathbb{P}_{X_1,X_2|\mathbf{X}_3,\boldsymbol{\theta}_{X_1X_2|\mathbf{X}_3}} || \mathbb{P}_{X_1|\mathbf{X}_3,\boldsymbol{\theta}_{X_1|\mathbf{X}_3}} \otimes \mathbb{P}_{X_2|\mathbf{X}_3,\boldsymbol{\theta}_{X_2|\mathbf{X}_3}} | \mathbb{P}_{\mathbf{X}_3|\boldsymbol{\theta}_{\mathbf{X}_3}}\right) \\
&\quad = \mathbb{E}_{\mathbb{P}_{\mathbf{X}_3}}\left[D_f(\mathbb{P}_{X_1,X_2|\mathbf{X}_3,\boldsymbol{\theta}_{X_1X_2|\mathbf{X}_3}} || \mathbb{P}_{X_1|\mathbf{X}_3,\boldsymbol{\theta}_{X_1|\mathbf{X}_3}} \otimes \mathbb{P}_{X_2|\mathbf{X}_3,\boldsymbol{\theta}_{X_2|\mathbf{X}_3}})\right] \\
&\quad = \mathbb{E}_{\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_1}}\left[f\left(\frac{\mathrm{d}\mathbb{P}_{X_1,X_2|\mathbf{X}_3,\boldsymbol{\theta}_{X_1X_2|\mathbf{X}_3}}}{\mathrm{d}(\mathbb{P}_{X_1|\mathbf{X}_3,\boldsymbol{\theta}_{X_1|\mathbf{X}_3}} \otimes \mathbb{P}_{X_2|\mathbf{X}_3,\boldsymbol{\theta}_{X_2|\mathbf{X}_3}})}\right)\right] \\
&\quad = \int_{\Omega_{\mathbf{X}}} \frac{p(x_1,x_2|\mathbf{x}_3,\boldsymbol{\theta}_{X_1X_2|\mathbf{X}_3})}{p(x_1|\mathbf{x}_3,\boldsymbol{\theta}_{X_1|\mathbf{X}_3}) \otimes p(x_2|\mathbf{x}_3,\boldsymbol{\theta}_{X_2|\mathbf{X}_3})} \mathrm{d}\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_1} \\
&\quad = \int_{\Omega_{\mathbf{X}}} f\left(\frac{R_{X_1X_2|\mathbf{X}_3}(P_{X_1}(x_1|\boldsymbol{\theta}_{X_1}),P_{X_2}(x_2|\boldsymbol{\theta}_{X_2})|\boldsymbol{P}_{\mathbf{X}_3}(\mathbf{x}_3|\boldsymbol{\theta}_{X_3}))}{R_{X_1|\mathbf{X}_3}(P_{X_1}(x_1|\boldsymbol{\theta}_{X_1})|\boldsymbol{P}_{\mathbf{X}_3}(\mathbf{x}_3|\boldsymbol{\theta}_{X_3}))R_{X_2|\mathbf{X}_3}(P_{X_2}(x_2|\boldsymbol{\theta}_{X_2})|\boldsymbol{P}_{\mathbf{X}_3}(\mathbf{x}_3|\boldsymbol{\theta}_{X_3}))}\right) \mathrm{d}\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_1} \\
&\quad = \int_{\mathbb{I}^{l+2}} f\left(\frac{R_{X_1X_2|\mathbf{X}_3}(u_1,u_2|\mathbf{u}_3)}{R_{X_1|\mathbf{X}_3}(u_1|\mathbf{u}_3)R_{X_2|\mathbf{X}_3}(u_2|\mathbf{u}_3)}\right) \frac{R_{X_1|\mathbf{X}_3}(u_1|\mathbf{u}_3)R_{X_2|\mathbf{X}_3}(u_2|\mathbf{u}_3)}{R_{X_1,X_2|\mathbf{X}_3}(u_1,u_2\mathbf{u}_3)} \mathrm{d}C_{\mathbf{X}|\boldsymbol{\theta},M_0}(\mathbf{u})
\end{aligned} \tag{8.6}$$

où nous avons utilisé la proposition 8.1.3 pour la première égalité, le lemme 7.1.1 pour la cinquième et le changement de variable $U_i = P_{X_i}(x_i|\boldsymbol{\theta}_{X_i})$ pour la dernière. À partir de ce résultat, nous pouvons obtenir un estimateur des densités de copule soit en utilisant une famille de copules paramétriques, dont les paramètres peuvent par exemple être inférés par la méthode du maximum de vraisemblance ou par tout autre estimateur que nous avons vu dans le chapitre 3, ou en utilisant une estimation non-paramétrique de la densité de la copule et donc dans notre cas avec la copule de Bernstein empirique. En intégrant par rapport à la copule empirique, nous obtenons l'estimateur suivant pour la statistique de test :

$$\begin{aligned}
t_f(\boldsymbol{d}) &= \int_{\mathbb{I}^{l+2}} f\left(\frac{\hat{R}_{X_1X_2|\mathbf{X}_3}(u_1,u_2|\mathbf{u}_3)}{\hat{R}_{X_1|\mathbf{X}_3}(u_1|\mathbf{u}_3)\hat{R}_{X_2|\mathbf{X}_3}(u_2|\mathbf{u}_3)}\right) \frac{\hat{R}_{X_1|\mathbf{X}_3}(u_1|\mathbf{u}_3)\hat{R}_{X_2|\mathbf{X}_3}(u_2|\mathbf{u}_3)}{\hat{R}_{X_1X_2|\mathbf{X}_3}(u_1,u_2|\mathbf{u}_3)} \mathrm{d}\hat{C}_m(\mathbf{u}) \\
&= \sum_{i=1}^{m} f\left(\frac{\hat{R}_{X_1X_2|\mathbf{X}_3}(u_1[i],u_2[i]|\mathbf{u}_3[i])}{\hat{R}_{X_1|\mathbf{X}_3}(u_1[i]|\mathbf{u}_3[i])\hat{R}_{X_2|\mathbf{X}_3}(u_2[i]|\mathbf{u}_3[i])}\right) \frac{\hat{R}_{X_1|\mathbf{X}_3}(u_1[i]|\mathbf{u}_3[i])\hat{R}_{X_2|\mathbf{X}_3}(u_2[i]|\mathbf{u}_3[i])}{\hat{R}_{X_1X_2|\mathbf{X}_3}(u_1[i],u_2[i]|\mathbf{u}_3[i])}.
\end{aligned}$$

Remarquons pour finir que dans le cas d'un test d'indépendance marginale, l'équation 8.6 nous donne

$$D_f\left(\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_0}||\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_1}\right) = \int_{\mathbb{I}^2} f\left(c_{X_1X_2}(u_1,u_2)\right)\mathrm{d}u_1\mathrm{d}u_2, \tag{8.7}$$

ce qui nous permet de dériver en plus des mesures de dépendances.

8.2 Test d'indépendance basé sur la divergence relative

Maintenant que nous avons proposé un cadre pour la dérivation systématique de test d'indépendance conditionnelle étant donnée une f-divergence, nous allons nous en servir pour obtenir un test à partir de l'entropie relative. Ce test est une généralisation de celui proposé par Belalia et al. 2017, lequel se limite à une indépendance marginale.

8.2.1 Le choix de l'entropie relative

L'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\theta}}$ peut être réinterprété comme étant celui du minimum de l'entropie relative entre la distribution $\tilde{\mathbb{P}}$ ayant généré l'échantillon d'observations $\boldsymbol{d}$ et la distribution paramétrique $\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta}}$ [3]. En effet, nous pouvons réécrire la vraisemblance comme :

$$f(\boldsymbol{d}|\boldsymbol{\theta}) = \exp\left[\sum_{i=1}^{m}\log f(\mathbf{x}[i]|\boldsymbol{\theta})\right] = \exp\left[m\int_{\Omega_{\mathbf{X}}}\log f(\mathbf{x}|\boldsymbol{\theta})\mathrm{d}\hat{\mathbb{P}}_m\right] = \exp\left[-mH(\hat{\mathbb{P}}_m||\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta}})\right] \tag{8.8}$$

et dans la limite d'un échantillon de taille infinie, nous avons :

$$-\frac{1}{m}\log f(\boldsymbol{d}|\boldsymbol{\theta}) \xrightarrow[m\to+\infty]{} H(\tilde{\mathbb{P}}_{\mathbf{X}}||\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta}}) = H(\tilde{\mathbb{P}}_{\mathbf{X}}) + D(\tilde{\mathbb{P}}_{\mathbf{X}}||\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta}}).$$

où nous avons utilisé la proposition 2.1.6. Maximiser la vraisemblance revient alors à minimiser l'entropie relative entre $\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta}}$ et $\tilde{\mathbb{P}}$ puisque $H(\tilde{\mathbb{P}}_{\mathbf{X}})$ ne dépend pas de $\boldsymbol{\theta}$. Par cohérence avec l'apprentissage des paramètres, nous voulons utiliser la même f-divergence pour l'apprentissage de la structure et donc utiliser l'entropie relative pour dériver une statistique de test. Un autre argument en sa faveur est que, comme nous l'avons remarqué plus haut, l'entropie relative vérifie la règle de chaine contrairement aux autres f-divergences. Nous allons tirer parti de cette propriété pour décomposer la statistique de test en une somme de termes ce qui nous permettra d'éviter certains calculs redondants.

8.2.2 L'information conditionnelle comme statistique de test

Reprenons à présent l'équation 8.6 pour l'entropie relative, c'est-à-dire avec la fonction génératrice $f(u) = u\log(u)$:

$$\begin{aligned}
D\left(\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_0}||\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_1}\right) &= \int_{\mathbb{I}^{l+2}}\log\left(\frac{R_{X_1X_2|\mathbf{X}_3}(u_1,u_2|\mathbf{u}_3)}{R_{X_1|\mathbf{X}_3}(u_1|\mathbf{u}_3)R_{X_2|\mathbf{X}_3}(u_2|\mathbf{u}_3)}\right)\mathrm{d}C_{\mathbf{X}|\boldsymbol{\theta},M_0}(\mathbf{u}) \\
&= \int_{\mathbb{I}^{l+2}}\log R_{X_1X_2|\mathbf{X}_3}(u_1,u_2|\mathbf{u}_3)\mathrm{d}C_{\mathbf{X}|\boldsymbol{\theta},M_0}(\mathbf{u}) \\
&\quad - \int_{\mathbb{I}^{l+2}}\log R_{X_1|\mathbf{X}_3}(u_1|\mathbf{u}_3)\mathrm{d}C_{\mathbf{X}|\boldsymbol{\theta},M_0}(\mathbf{u}) \\
&\quad - \int_{\mathbb{I}^{l+2}}\log R_{X_2|\mathbf{X}_3}(u_2|\mathbf{u}_3)\mathrm{d}C_{\mathbf{X}|\boldsymbol{\theta},M_0}(\mathbf{u}) \\
&= -H_c(X_1,X_2,\mathbf{X}_3) + H_c(X_1,\mathbf{X}_3) + H_c(X_2,\mathbf{X}_3) - H_c(\mathbf{X}_3)
\end{aligned}$$

3. Cette réinterprétation peut mener à une vision géométrique de l'estimation paramétrique (Kulhavỳ 1996).

L'entropie relative nous permet donc d'écrire la statistique de test comme une somme d'entropies de copule. Remarquons que par définition de l'information mutuelle (définition 2.2.2), nous avons également :

$$D\left(\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_0}||\mathbb{P}_{\mathbf{X}|\boldsymbol{\theta},M_1}\right) = D\left(\mathbb{P}_{X_1,X_2|\mathbf{X}_3,\boldsymbol{\theta}_{X_1X_2|\mathbf{X}_3}}||\mathbb{P}_{X_1|\mathbf{X}_3,\boldsymbol{\theta}_{X_1|\mathbf{X}_3}} \otimes \mathbb{P}_{X_2|\mathbf{X}_3,\boldsymbol{\theta}_{X_2|\mathbf{X}_3}}|\mathbb{P}_{\mathbf{X}_3|\boldsymbol{\theta}_{\mathbf{X}_3}}\right) = I(X_1;X_2|\mathbf{X}_3). \tag{8.9}$$

et la statistique de test associée à l'entropie relative est donc l'information conditionnelle. En égalisant les deux derniers résultats, nous pouvons alors relier l'information conditionnelle à l'entropie de la copule et généraliser le résultat de Ma et al. (2011) :

Théorème 8.2.1. Soit $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ un vecteur aléatoire avec $\mathbf{X}_1$,$\mathbf{X}_2$ et $\mathbf{X}_3$ trois sous-vecteurs disjoints. L'information entre $\mathbf{X}_1$ et $\mathbf{X}_2$ conditionnellement à $\mathbf{X}_3$ est liée à l'entropie de la copule par :

$$I(\mathbf{X}_1;\mathbf{X}_2|\boldsymbol{U}) = H_c(\mathbf{X}_1,\boldsymbol{U}) + H_c(\mathbf{X}_2,\boldsymbol{U}) - H_c(\mathbf{X}_1,\mathbf{X}_2,\boldsymbol{U}) - H_c(\boldsymbol{U}) \tag{8.10}$$

Notons que, bien que nous ayons remplacé X_1 et X_2 par des vecteurs aléatoires dans l'énoncé du théorème, sa démonstration ne change pas. La relation 8.10 est donc similaire à la relation entre l'information conditionnelle et l'entropie (équation 2.22) à cela près que l'entropie de la copule est par définition nulle pour l'ensemble vide et également nulle pour une seule variable puisque dans ce cas là la copule densité vaut identiquement 1 sur $\mathbb{I}$ et son logarithme est alors nul. Nous retrouvons bien dans ce cas la relation démontrée dans Ma et al. (2011) :

$$I(X_1;X_2|\emptyset) = H_c(X_1,\emptyset) + H_c(X_2,\emptyset) - H_c(X_1,X_2,\emptyset) - H_c(\emptyset) = -H_c(X_1,X_2).$$

L'estimation de la statistique de test sur un échantillon $\boldsymbol{d}$ de taille m, donnée par l'équation 8.9, nous donne donc un estimateur pour l'information conditionnelle :

$$t_f(\boldsymbol{d}) = \hat{I}(X_1;X_2|\mathbf{X}_3) = \hat{H}_c(X_1,\mathbf{X}_3) + \hat{H}_c(X_2,\mathbf{X}_3) - \hat{H}_c(X_1,X_2,\mathbf{X}_3) - \hat{H}_c(\mathbf{X}_3) \tag{8.11}$$

qui s'écrit comme la somme des estimateurs des entropies de copules $\hat{H}_c$ ayant pour expression

$$\hat{H}_c(\mathbf{X}) = -\sum_{j=1}^{m} \log \hat{c}(\mathbf{x}[j]) = -\int_{\mathbb{I}^n} \log \hat{c}(\mathbf{x}) \mathrm{d}\hat{\mathbb{C}}_m = H_c(\hat{\mathbb{C}}_m||\mathbb{C}_{\mathbf{X}|\boldsymbol{\theta}}). \tag{8.12}$$

Rappelons que $\hat{c}$ est un estimateur de la copule densité calculé soit à partir d'un modèle paramétrique, soit à partir de la copule de Bernstein empirique. Certains estimateurs d'entropie de copule sont communs à différents tests d'indépendance conditionnelle. Par exemple, tous les tests d'indépendances conditionnellement à $\mathbf{X}_3$ contiendront le terme $\hat{H}_c(\mathbf{X}_3)$. Ainsi, ces estimateurs peuvent être stockés et réutilisés de manière à accélérer nos algorithmes pour l'apprentissage de la structure d'un CBN puisque ceux-ci réalisent un grand nombre de tests. Comme nous l'avons dit plus haut, cette propriété est spécifique à l'entropie relative qui permet cette décomposition.

8.3 Implémentation de l'algorithme CMIIC

Jusqu'à présent, nous nous sommes focalisés sur comment dériver une statistique de test et avons ignoré la question de la définition d'une région de rejet. Pour cela,

il est nécessaire de déterminer la distribution de la statistique de test afin de pouvoir calculer des *p-values* et décider alors du rejet ou non de l'hypothèse nulle. Il s'avère que même dans la limite des grands échantillons, cette tâche est souvent compliquée. Dans le cas de l'entropie relative, Belalia et al. (2017) ont proposé une approximation de la statistique de test pour une indépendance marginale ($\mathbf{U} = \emptyset$) et qui, dans la limite des grands échantillons, possède une distribution normale standard. Cette approximation n'est cependant pas suffisante pour l'implémentation d'un algorithme PC comme dans le chapitre précédent puisque celui-ci nécessite l'usage de tests d'indépendance *conditionnelle*. Quand bien-même, Verny et al. (2017) ont montré que dans le cas de variables aléatoires discrètes, l'algorithme PC était moins performant que l'algorithme MIIC et une version continue de cet algorithme serait donc préférable. Justement, ce dernier fait usage de l'information conditionnelle pour décider d'une indépendance et plus précisément de l'estimateur que nous avons dérivé dans la section précédente. Ainsi, cela va nous permettre de dériver une version continue et non-paramétrique de l'algorithme pour l'apprentissage de la structure d'un CBN.

L'algorithme MIIC, comme nous l'avons vu (cf. sous-section 5.2.2), utilise l'approche NML afin de dériver un test d'indépendance. Plus tôt, nous avons fait l'hypothèse que les données étaient discrètes et distribuées selon une loi catégorielle. Ici, nous faisons l'hypothèse que la distribution ayant généré les données appartient à un modèle paramétrique continu quelconque (gaussien, exponentiel, Dirichlet, etc.). En utilisant l'écriture 8.8, le rapport des maxima de vraisemblance s'écrit comme :

$$\frac{f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, M_1)}{f(\boldsymbol{d}|\hat{\boldsymbol{\theta}}, M_0)} = \exp\left[-m\left(H\left(\hat{\mathbb{P}}_m||\mathbb{P}_{\mathbf{X}|\hat{\boldsymbol{\theta}},M_1}\right) - H\left(\hat{\mathbb{P}}_m||\mathbb{P}_{\mathbf{X}|\hat{\boldsymbol{\theta}},M_0}\right)\right)\right].$$

Les deux entropies croisées se décomposant comme :

$$\begin{cases} H(\hat{\mathbb{P}}_m||\mathbb{P}_{\mathbf{X}|\hat{\boldsymbol{\theta}},M_1}) = H\left(\hat{\mathbb{P}}||\mathbb{P}_{\mathbf{U}}\right) + H\left(\hat{\mathbb{P}}||\mathbb{P}_{X_1|\mathbf{U}}\right) + H\left(\hat{\mathbb{P}}||\mathbb{P}_{X_2|\mathbf{U}}\right) \\ H(\hat{\mathbb{P}}_m||\mathbb{P}_{\mathbf{X}|\hat{\boldsymbol{\theta}},M_0}) = H\left(\hat{\mathbb{P}}||\mathbb{P}_{\mathbf{U}}\right) + H\left(\hat{\mathbb{P}}||\mathbb{P}_{X_1,X_2|\mathbf{U}}\right) \end{cases}, \tag{8.13}$$

le rapport des densités NML a alors pour expression :

$$\frac{f_{\mathrm{NML}}(\boldsymbol{d}|M_1)}{f_{\mathrm{NML}}(\boldsymbol{d}|M_0)} = \exp\left[-m\left(H\left(\hat{\mathbb{P}}_m||\mathbb{P}_{X_1|\mathbf{U}}\right) + H\left(\hat{\mathbb{P}}_m||\mathbb{P}_{X_2|\mathbf{U}}\right) - H\left(\hat{\mathbb{P}}_m||\mathbb{P}_{X_1,X_2|\mathbf{U}}\right)\right) + q_{X_1;X_2|\mathbf{U}}\right] \tag{8.14}$$

avec, comme précédemment, $q_{X_1;X_2|\mathbf{U}} = \log\left(Z(M_0)/Z(M_1)\right)$. Cette dernière expression peut s'exprimer en fonction seulement de copules, faisant apparaitre l'estimateur de l'information conditionnelle que nous avons introduit plus tôt. Pour le voir, remarquons qu'étant donnés deux vecteurs aléatoires $\mathbf{X} = (X_1, \ldots, X_d)$ et $\mathbf{Y}$, nous avons

$$\begin{aligned}
H(\hat{\mathbb{P}}_m||\mathbb{P}_{\mathbf{X}|\mathbf{Y}}) &= -\frac{1}{m}\sum_{i=1}^{m} \log f(\mathbf{x}[i]|\mathbf{y}[i], \boldsymbol{\theta}) \\
&= -\frac{1}{m}\sum_{i=1}^{m} \log\left(R_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}[i], \mathbf{y}[i]) \prod_{j=1}^{d} f(x_j[i])\right) \\
&= -\frac{1}{m}\sum_{i=1}^{m} \left[\log c_{\mathbf{X},\mathbf{Y}}(\boldsymbol{F}(\mathbf{x}[i]), \boldsymbol{F}(\mathbf{y}[i])) - \log c_{\mathbf{Y}}(\boldsymbol{F}(\mathbf{y}[i])) - \sum_{j=1}^{d} \log f(x_j[i])\right] \\
&= H_c(\hat{\mathbb{C}}_m||\mathbb{C}_{\mathbf{X},\mathbf{Y}}) - H_c(\hat{\mathbb{C}}_m||\mathbb{C}_{\mathbf{Y}}) + \sum_{j=1}^{d} H(\hat{\mathbb{P}}_{j,m}||\mathbb{P}_{X_j}) \\
&= \hat{H}_c(\mathbf{X}, \mathbf{Y}) - \hat{H}_c(\mathbf{Y}) + \sum_{j=1}^{d} H(\hat{\mathbb{P}}_{j,m}||\mathbb{P}_{X_j}).
\end{aligned}$$

En utilisant cette relation dans le rapport des densités NML, nous avons donc :

$$\begin{aligned}\frac{f_{\text{NML}}(\boldsymbol{d}|M_1)}{f_{\text{NML}}(\boldsymbol{d}|M_0)} &= \exp\left[-m\left(\hat{H}_c\left(X_1,\mathbf{U}\right)-\hat{H}_c\left(\mathbf{U}\right)+\hat{H}_c\left(X_2,\mathbf{U}\right)-\hat{H}_c\left(X_1,X_2,\mathbf{U}\right)\right)+q_{X_1;X_2|\mathbf{U}}\right]\\ &= \exp\left[-m\hat{I}(X_1;X_2|\mathbf{U})+q_{X_1;X_2|\mathbf{U}}\right].\end{aligned}$$

De la même manière que dans le cas discret, l'indépendance conditionnelle est acceptée si le rapport est supérieur à 1, c'est-à-dire si $\hat{I}'(X_1;X_2|\mathbf{U}) = \hat{I}(X_1;X_2|\mathbf{U}) - \frac{q_{X_1;X_2|\mathbf{U}}}{m} < 0$. Reste alors à estimer l'information à trois points afin de déterminer l'ordre dans lequel les variables sont traitées ainsi que de la présence ou non de v-structures. Il nous suffit pour cela d'utiliser l'équation 2.27, nous permettant de relier l'information multivariée conditionnelle à l'entropie de la copule :

Théorème 8.3.1. Soit $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{U})$ un vecteur aléatoire avec $\mathbf{X}_1$,$\mathbf{X}_2$, $\mathbf{X}_3$ et $\mathbf{U}$ quatre sous-vecteurs disjoints. L'information entre $\mathbf{X}_1$, $\mathbf{X}_2$ et $\mathbf{X}_3$ conditionnellement à $\mathbf{U}$ est liée à l'entropie de la copule par :

$$\begin{aligned}I(\mathbf{X}_1;\mathbf{X}_2;\mathbf{X}_3|\boldsymbol{U}) &= I(\mathbf{X}_1;\mathbf{X}_2|\mathbf{U}) - I(\mathbf{X}_1;\mathbf{X}_2|\mathbf{X}_3,\mathbf{U})\\ &= H_c(\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3,\boldsymbol{U}) + H_c(\mathbf{X}_1,\boldsymbol{U}) + H_c(\mathbf{X}_2,\boldsymbol{U}) + H_c(\mathbf{X}_3,\mathbf{U})\\ &\quad - H_c(\mathbf{X}_1,\mathbf{X}_2,\boldsymbol{U}) - H_c(\mathbf{X}_1,\mathbf{X}_3,\boldsymbol{U}) - H_c(\mathbf{X}_2,\mathbf{X}_3,\boldsymbol{U}) - H_c(\boldsymbol{U}).\end{aligned} \tag{8.15}$$

L'estimateur pour l'information multivariée conditionnelle est alors simplement défini en remplaçant les informations conditionnelles dans l'équation précédente par leurs estimateurs, c'est-à-dire :

$$\hat{I}(X_1;X_2;X_3|\mathbf{U}) = \hat{I}(X_1;X_2|\mathbf{U}) - \hat{I}(X_1;X_2|\mathbf{U}\cup X_3). \tag{8.16}$$

Une fois de plus, l'estimateur peut s'exprimer en fonction uniquement d'estimateurs d'entropies de copules qui peuvent être stockés et réutilisés entre différents tests pour accélérer les calculs. Comme dans le cas discret nous avons alors :

$$\frac{f_{\text{NML}}(\boldsymbol{d}|\hat{\boldsymbol{\theta}},M_1^+)}{f_{\text{NML}}(\boldsymbol{d}|\hat{\boldsymbol{\theta}},M_0^+)} = \frac{f_{\text{NML}}(\boldsymbol{d}|\hat{\boldsymbol{\theta}},M_1)}{f_{\text{NML}}(\boldsymbol{d}|\hat{\boldsymbol{\theta}},M_0)}\exp\left[m\hat{I}(X_1;X_2;X_3|\mathbf{U})+q_{X_1;X_2;X_3|\mathbf{U}}\right] \tag{8.17}$$

avec $q_{X_1;X_2;X_3|\mathbf{U}} = q_{X_1;X_2|\mathbf{U}} - q_{X_1;X_2|\mathbf{U}\cup X_3}$. En résumé, l'algorithme reste inchangé si l'on utilise l'estimateur de l'information conditionnelle - et l'estimateur de l'information multivariée conditionnelle qui en découle - à la place des estimateurs que nous avons vus dans le cas discret. La différence réside dans le calcul des estimateurs qui repose sur la théorie des copules et nous permet donc d'utiliser la densité de la copule de Bernstein empirique $\hat{c}_m^B$ pour obtenir une version non-paramétrique de l'algorithme MIIC. Une version paramétrique peut également être obtenue en utilisant n'importe quel modèle de copule densité paramétrique pour lequel les paramètres sont estimés, par exemple, avec la méthode du MLE.

Pour finir, nous devons discuter des termes correctifs $q_{X_i;X_j|\mathbf{U}}$ et $q_{X_i;X_j;X_k|\mathbf{U}}$. Ceux-ci proviennent de la constante de normalisation des densités NML qui, dans le cas continu, est définie par l'intégrale

$$Z(M) = \int_{\boldsymbol{d}\in\mathcal{D}} f(\boldsymbol{d}|\hat{\boldsymbol{\theta}},M)\mathrm{d}\boldsymbol{d}. \tag{8.18}$$

Comme nous l'avons commenté plus tôt, pour certains modèles, la constante n'est finie que si le domaine $\mathcal{D}$ est restreint. Dans le cas continu, nous intégrons sur $\mathbb{R}$ et l'intégrale diverge donc le plus souvent. Pour résoudre ce problème, une solution possible est

d'utiliser le *lucky* NML (LNML) (Grünwald et al. 2007) qui rajoute une fonction de *luckiness* jouant un rôle très similaire à celui de la densité *a priori* dans le cadre de l'approche bayésienne. La méthode LNML sera abordée un peu plus en détails en fin de thèse lorsque nous discuterons des perspectives à nos travaux. En effet, comme les calculs associés sont difficiles et que cette méthode n'est de toute manière adaptée qu'au cas paramétrique, nous avons fait à la place le choix plus simple de fixer ces corrections à une constante α telle que $I'(X_i; X_j|\mathbf{U}) = I(X_i; X_j|\mathbf{U}) - \alpha$ et $I'(X_i; X_j; X_k|\mathbf{U}) = I(X_i; X_j; X_k|\mathbf{U}) + \alpha$. Bien que l'utilisation d'un même paramètre pour l'information conditionnelle et l'information multivariée conditionnelle soit discutable, nous avons fait ce choix afin d'éviter l'ajout de plusieurs hyper-paramètres. Celui-ci 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. Enfin, il peut être déterminé en utilisant, par exemple, la méthode de *cross-validation*.

8.4 Accélération de l'algorithme utilisant le score BIC

Dans le chapitre précédent, nous avons présenté la méthode pour l'apprentissage de la structure d'un CBN introduite par Elidan (2010). Celle-ci utilise le score BIC qui est maximisé sur l'espace des *DAG* via une recherche TABU. Pour rappel, son expression pour une structure $\mathcal{G}$ est donnée par :

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

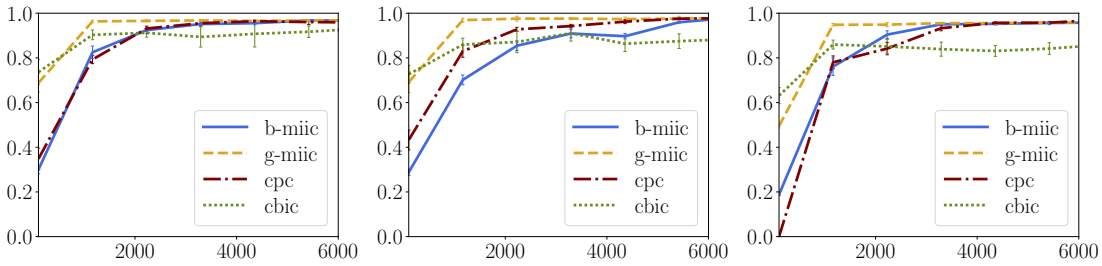
où ℓ est la log-vraisemblance, $\hat{\theta}$ sont les estimateurs des paramètres du maximum de vraisemblance (MLE) et $|\Theta_{\mathcal{G}}|$ est le nombre de paramètres libres associés à la structure du graphe. L'inconvénient de cette technique est que le score doit être calculé sur l'ensemble du graphe à chaque fois qu'une modification locale sur le graphe est effectuée par la recherche TABU. Comme nous en avons discuté en 5.2.1.2 pour le cas discret, la solution est d'utiliser la décomposition de la vraisemblance en une somme d'informations mutuelles permettant alors de ne recalculer le score que sur les familles ayant été modifiées. Il s'avère que nous pouvons faire la même chose ici en remarquant que :

$$\begin{aligned}
\ell(\boldsymbol{d} : \hat{\boldsymbol{\theta}}, G) &= \sum_{i=1}^{m}\sum_{j=1}^{n} \log R_j \left(u_j[i], \pi_{j1}[i], \ldots, \pi_{jk_j}[i]\right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} \left[\log c_{X_j,\mathbf{Pa}_j}\left(u_j[i], \pi_{j1}[i], \ldots, \pi_{jk_j}[i]\right) - \log c_{\mathbf{Pa}_j}\left(\pi_{j1}[i], \ldots, \pi_{jk_j}[i]\right)\right] \\
&= -m\sum_{j=1}^{n}\left[\hat{H}_c\left(X_j, \mathbf{Pa}_j\right) - \hat{H}_c\left(\mathbf{Pa}_j\right)\right] = m\sum_{i=1}^{n}\hat{I}(X_i; \mathbf{Pa}_i),
\end{aligned}$$

où nous avons utilisé la relation (8.10) pour la dernière égalité. Nous pouvons alors utiliser les formules de la proposition 5.2.1 pour calculer les variations de score associées à une modification locale et accélérer l'algorithme proposé par Elidan (2010). Une nouvelle fois, nous avons fait apparaître l'estimateur de l'information conditionnelle associé à la statistique de test dérivée à partir de l'entropie relative. Le reste de l'algorithme reste inchangé et les copules densités sont estimées à partir de modèles gaussiens.

8.5 Comparaison des algorithmes d'apprentissage

Nous reprenons à présent la comparaison des algorithmes d'apprentissage pour les CBNs avec le protocole expérimental que nous avons introduit lors du chapitre précédent. Au F-score et à la SHD, nous ajoutons ici le temps d'apprentissage des algorithmes

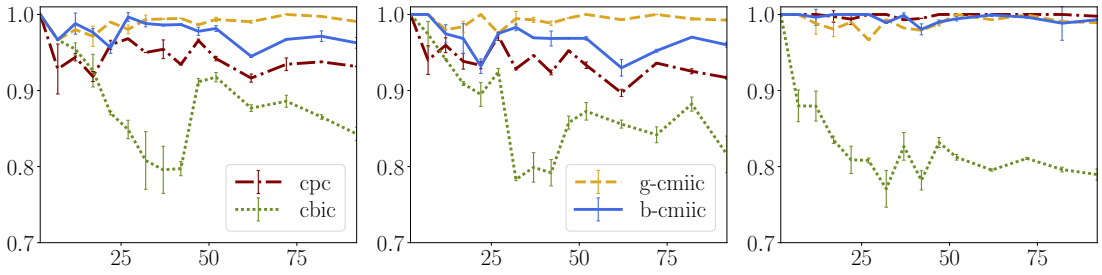


(a) Gaussien (b) Student (c) Dirichlet

FIGURE 8.1 – Évolution du F-score pour les méthodes CBIC, CPC, G-CMIIC et B-CMIIC en fonction de la taille de l'échantillon d'apprentissage. Les résultats sont moyennés sur 5 échantillons différents générés à partir de la structure du réseau ALARM.

comme critère de performance. Avec les méthodes CPC et CBIC, nous comparons deux versions différentes de CMIIC. La première utilise la copule de Bernstein empirique pour obtenir une estimation de l'information mutuelle tandis que la deuxième utilise pour cela la copule gaussienne, dont les paramètres sont estimés en suivant la même procédure que pour CBIC. Nous nous référons à ces deux versions en les appelant respectivement B-CMIIC et G-CMIIC. Dans les deux cas, le paramètre de normalisation est fixé à $\alpha = 0,01$. Comme nous allons le voir, cette valeur s'avère être expérimentalement un bon compromis entre le nombre de données nécessaires pour l'apprentissage et la validité des tests d'indépendances. Pour ce qui est de l'algorithme CBIC, nous utilisons la version accélérée tirant parti de la décomposition du score BIC. Enfin, l'algorithme GBN ayant des performances systématiquement inférieures à CPC et CBIC, nous ne reproduisons pas les courbes qui lui sont associées.

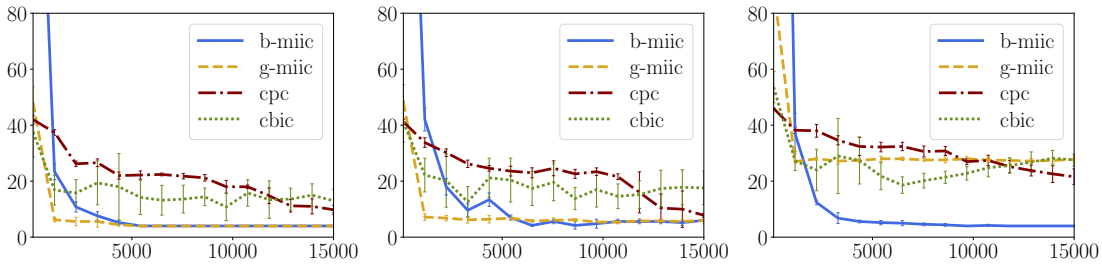
8.5.1 Performances pour la reconstruction du squelette



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

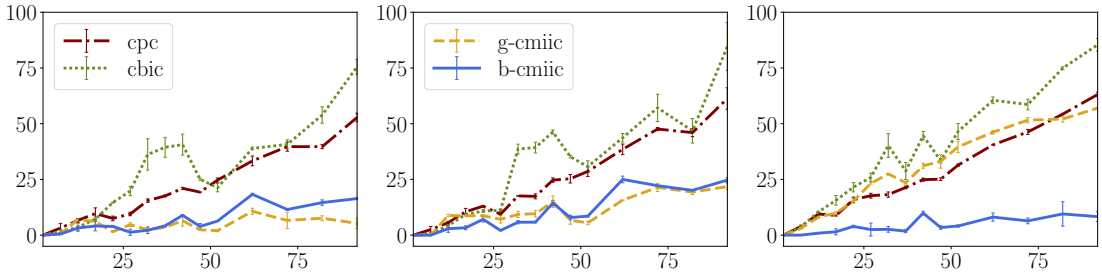
FIGURE 8.2 – Évolution du F-score pour les méthodes CBIC, CPC, G-CMIIC et B-CMIIC en fonction de la dimension des graphes aléatoires. Pour chaque dimension, les résultats sont moyennés sur 2 graphes aléatoires différents et sur 5 échantillons de taille $m = 10\,000$.

La figure 8.1 donne l'évolution du F-score des structures reconstruites par chacune des méthodes en fonction de la taille de l'échantillon d'apprentissage lorsque la structure de référence est celle du réseau ALARM. Pour une taille donnée, ces résultats sont moyennés sur 5 échantillons différents. Nous observons que l'algorithme G-CMIIC est celui qui converge le plus rapidement des quatre algorithmes. Néanmoins, B-CMIIC et CPC convergent approximativement vers la même valeur lorsque la taille de l'échantillon est suffisante. Pour ce qui est de la méthode CBIC, celle-ci a de moins bonnes performances que les trois autres peu importe le modèle utilisé pour paramétrer les



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

FIGURE 8.3 – Évolution de la SHD pour les méthodes CBIC, CPC, G-CMIIC et B-CMIIC en fonction de la taille de l'échantillon d'apprentissage. Les résultats sont moyennés sur 5 échantillons différents générés à partir de la structure du réseau ALARM.



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

FIGURE 8.4 – Évolution de la SHD pour les méthodes CBIC, CPC, G-CMIIC et B-CMIIC en fonction de la dimension des graphes aléatoires. Pour chaque dimension, les résultats sont moyennés sur 2 graphes aléatoires différents et sur 5 échantillons de taille $m = 10\,000$.

CBNs.

La figure 8.2 donne l'évolution du F-score en fonction de la taille des structures aléatoires utilisées pour générer les données. Pour chaque dimension, les résultats sont moyennés sur 2 structures aléatoires différentes et, pour chacune de ces structures, sur 5 échantillons différents de taille $m = 10\,000$. Pour des données provenant de copules gaussiennes et de Student, nous pouvons voir que G-CMIIC a des performances légèrement supérieures à CPC et B-CMIIC. Étonnamment, G-CMIIC conserve de bons résultats même lorsque les données proviennent de copules de Dirichlet et que la dimension augmente. Dans ce dernier cas, les méthodes CPC et B-CMIIC ont des performances comparables à G-CMIIC. Enfin, tout comme pour la structure ALARM, les performances de l'algorithme CBIC sont inférieures à celles des autres algorithmes.

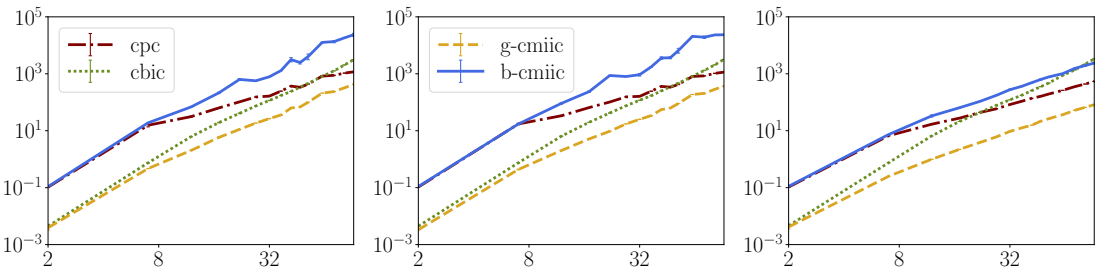
8.5.2 Performances pour la reconstruction du CPDAG

Les figures 8.3 et 8.4 montrent respectivement l'évolution de la SHD en fonction de la taille de l'échantillon pour la structure du réseau ALARM et l'évolution de la SHD en fonction de la dimension pour les structures aléatoires. Ces courbes sont également moyennées sur 5 échantillons dans le premier cas et 2×5 échantillons dans le deuxième. Ici aussi, la méthode G-CMIIC récupère presque parfaitement le CPDAG de référence lorsque le modèle générant les données est gaussien ou de Student. De plus, elle a besoin de moins de données que les autres algorithmes pour converger. Toutefois, ses performances se dégradent dans le cas où les données sont générées à partir de copules de Dirichlet. À son tour, l'algorithme B-CMIIC possède des performances comparables à celles de G-CMIIC lorsque la taille des échantillons est suffisante. En revanche, ses

performances ne dépendent pas du modèle génératif, ce qui permet d'illustrer l'avantage d'une méthode non-paramétrique par rapport à une méthode paramétrique. Nous pouvons observer que même dans le cas de données générées avec des copules de Dirichlet et pour une taille de graphe de $n = 92$, l'algorithme B-CMIIC retrouve presque parfaitement le CPDAG de référence. Dans le cas de la structure ALARM, la méthode CPC semble converger vers la même valeur de SHD que la méthode B-CMIIC mais nécessite pour cela beaucoup plus de données que cette dernière. Dans le cas des structures aléatoires, ses performances diminuent plus rapidement que celles de l'algorithme B-CMIIC lorsque la dimension augmente. Quant à la méthode CBIC, ses performances sont une nouvelle fois inférieures à celles des autres algorithmes et se dégradent rapidement lorsque la dimension augmente.

8.5.3 Complexité temporelle

Les temps d'apprentissage (en secondes) ont été calculés pour les quatre méthodes en fonction de la dimension des graphes aléatoires et pour des échantillons de taille $m = 10\,000$. Les résultats pour des tailles de graphe allant de $n = 2$ à $n = 92$, sont présentés sur la figure 8.5. En contrepartie de ses bonnes performances pour la reconstruction du graphe, nous observons que l'algorithme B-CMIIC est celui dont le temps d'apprentissage est le plus long. Cela n'est pas surprenant puisque celui-ci utilise un modèle de copule non-paramétrique qui est plus complexe et donc plus difficile à estimer qu'un modèle paramétrique. En effet, nous pouvons voir que l'algorithme G-MIIC est justement le plus rapide alors que l'architecture de l'algorithme est exactement la même que celle de B-CMIIC. À ce titre, G-CMIIC devrait être préféré lorsque l'on sait que l'hypothèse gaussienne est valide. Au contraire, lorsqu'aucune information sur la distribution sous-jacente n'est disponible, l'algorithme B-CMIIC devrait être préféré en raison de la généralité de ses performances. Pour ce qui est des algorithmes CBIC et CPC, ceux-ci ont globalement des temps d'apprentissage compris entre ceux de G-CMIIC et B-CMIIC. Pour de petites tailles de graphes, l'algorithme CBIC est plus rapide que l'algorithme CPC alors que c'est l'inverse pour de grande tailles. Cela s'explique du fait que la taille de l'espace des DAGs augmente de façon exponentielle par rapport au nombre de nœuds (cf. 4.3) et la recherche locale menée par CBIC devient de plus en plus complexe.



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

FIGURE 8.5 – Temps d'apprentissage en secondes pour les méthodes CBIC, CPC, G-CMIIC et B-CMIIC en fonction de la dimension des graphes aléatoires. Pour chaque dimension, les résultats sont moyennés sur 2 graphes aléatoires différents et sur 5 échantillons de taille $m = 10\,000$.

8.6 Application « *wine quality* »

Nous terminons ce chapitre en testant et en comparant nos algorithmes sur des données d'application. Pour cela, nous avons choisi d'utiliser l'échantillon de données *wine quality*, publié dans Cortez et al. (2009) et disponible en libre accès sur le UCI Machine Learning Repository.

8.6.1 Description des données

Cet échantillon provient de l'analyse physico-chimique et sensorielle de *vinho verde*, un vin produit dans le nord-ouest du Portugal. Pour chaque vin est mesuré : son acidité fixe (*fixed acidity*), son acidité volatile (*volatile acidity*), sa concentration en acide citrique (*citric acid*), sa concentration en sucre résiduel (*residual sugar*), sa concentration en sel (*chlorid*), sa concentration en dioxyde de soufre libre (*free dioxide sulfur*), sa concentration totale en dioxyde de soufre (*total dioxide sulfur*), sa densité (*density*), son pH (*pH*), sa concentration en sulfate de potassium (*sulfates*) et enfin son degré d'alcool (*alcohol*). Le dioxyde de soufre (SO_2), ajouté au vin avant fermentation, est utilisé comme antioxydant et comme conservateur. Le SO_2 n'ayant pas réagi lors de la fermentation correspond à ce que l'on appelle SO_2 libre tandis que la concentration avant fermentation correspond à la concentration totale de SO_2. De même, le sucre résiduel correspond à la quantité de sucre n'ayant pas été transformée en alcool par les levures lors de la fermentation. Les acidités fixe et volatile, quant à elles, correspondent respectivement à la concentration en acide tartrique et en acide acétique du vin. À ces mesures s'ajoute l'évaluation de la qualité du vin par au moins trois testeurs lui donnant une note entre 0 et 10 et ces notes sont agrégées en prenant leur médiane. Au total, l'échantillon est donc composé de 12 variables aléatoires dont 11 qui sont continues et 1 (la qualité) qui est discrète. Malgré le fait que nos algorithmes aient été conçus *a priori* pour des variables continues, nous allons voir que nous obtenons tout de même des résultats probants dans le cas hybride. Pour finir, l'échantillon contient à la fois des variétés de vins rouges et de vins blancs que nous traitons séparément puisqu'elles sont différentes au niveau de leur analyse physico-chimique comme le montre le tableau 8.2. Ce dernier, que nous avons repris de Cortez et al. (2009), résume plusieurs des statistiques de ces deux sous-échantillons composés de 1559 exemples pour les vins rouges et de 4898 exemples pour les vins blancs.

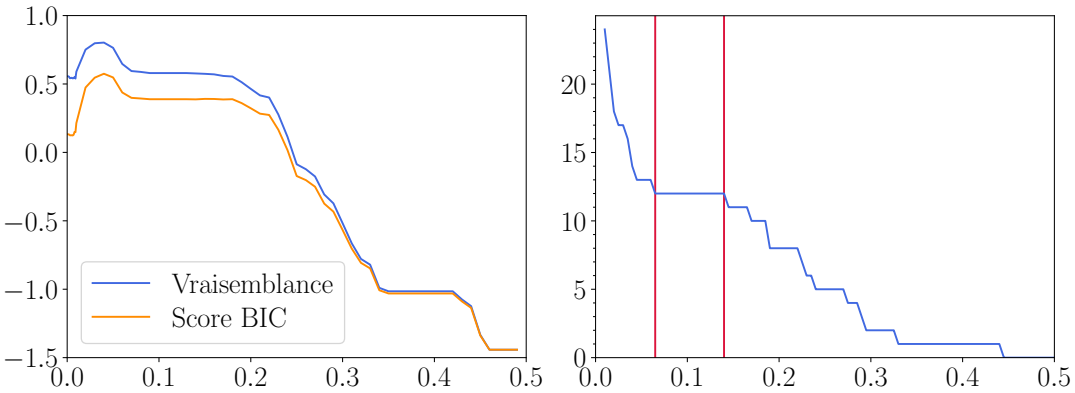
L'objectif de l'étude menée par Cortez et al. (2009) est, d'une part, de mieux comprendre le lien entre les variables physico-chimiques et la qualité des vins et, d'autre part, de permettre l'estimation de la qualité d'un vin à partir de l'observation du reste des variables. Ce dernier problème étant un problème de classification, il ne rentre pas dans le cadre de cette thèse puisque, comme nous l'avons dit en introduction, nous nous concentrons ici sur l'apprentissage du modèle. Toutefois, l'apprentissage de la structure peut nous permettre de mieux comprendre le lien entre chacune des variables et en particulier de savoir quelles variables sont pertinentes pour déterminer la qualité du vin. C'est donc dans ce cadre que nous allons appliquer nos algorithmes à ces données.

8.6.2 Apprentissage de la structure

Nous utilisons maintenant les trois algorithmes basés sur les CBNs (CBIC, CPC et CMIIC) pour apprendre la structure de dépendance entre les variables. Pour limiter le nombre de graphes que nous présentons et parce que l'algorithme CMIIC est celui qui a montré les meilleures performances lors des expériences précédentes, nous n'utilisons que ce dernier sur l'échantillon des vins blancs. Les méthodes CBIC et CPC ne sont donc appliquées qu'à l'échantillon des vins rouges.

	Vins rouges			Vins blancs		
Variables (unités)	Min	Max	Mean	Min	Max	Mean
Acidité fixe (g (acide tartrique)/dm^3)	4.60	15.9	8.32	3.80	14.2	6.85
Acidité volatile (g (acide acétique)/dm^3)	0.12	1.58	0.528	0.08	1.10	0.278
Acide citrique (g/dm^3)	0.00	1.00	0.271	0.00	1.66	0.334
Sucre résiduel (g/dm^3)	0.900	15.5	2.54	0.600	65.8	6.39
Chlorides (g (chlorure de sodium)/dm^3)	0.120	0.611	0.0875	0.0458	0.611	0.346
Dioxide de sulfure libre (mg/dm^3)	1.00	72.0	15.9	2.00	289	35.3
Dioxide de sulfure total (mg/dm^3)	6.00	289	46.5	9.00	440	138
Densité (g/dm^3)	0.990	1.00	0.997	0.987	1.04	0.994
pH	2.74	4.01	3.31	2.72	3.82	3.19
Sulfates (g (sulfate de potassium)/dm^3)	0.330	2.00	0.658	0.220	1.08	0.490
Alcool (%vol)	8.40	14.9	10.4	8.00	14.2	10.5

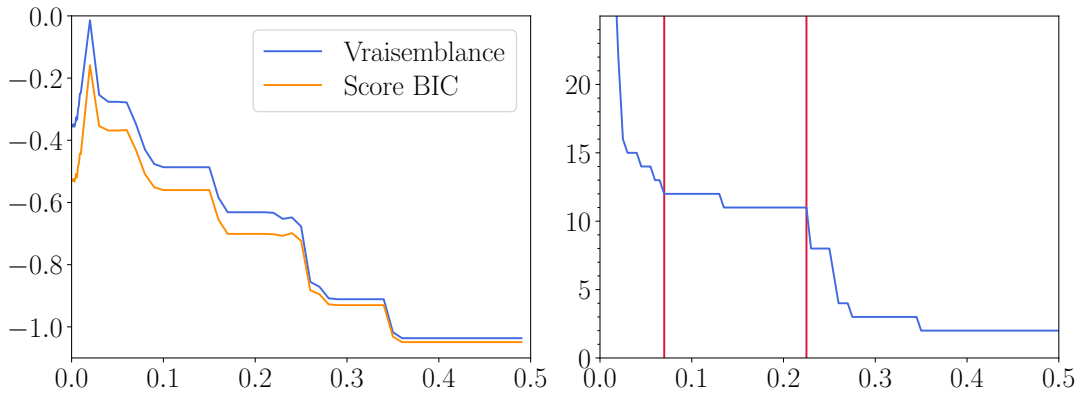
TABLE 8.2 – Analyse des données physico-chimique des vins rouges et blancs.



(a) Vraisemblance (b) Nombre d'arcs

FIGURE 8.6 – La figure de gauche représente l'évolution de la vraisemblance et du score BIC (naïf) en fonction de α. La vraisemblance est calculée en utilisant le modèle appris avec B-CMIIC à partir de l'échantillon des vins **rouges** et en utilisant une *cross-validation* de 10 blocs. La figure de droite représente l'évolution du nombre d'arcs dans le CBN appris en fonction de α.

L'algorithme CBIC ayant tendance à reconstruire des graphes denses, nous limitons le nombre de parents qu'un nœud peut avoir à 1. Nous espérons ainsi minimiser le nombre de faux-positifs et capturer les dépendances les plus importantes (KOLLER et al. 2009, p. 808). Le graphe obtenu avec cette méthode est représenté sur la figure 8.8. Pour l'algorithme CPC, nous fixons la valeur du seuil de significativité des tests d'indépendance à la valeur traditionnelle de $p = 0.05$. Le graphe obtenu est représenté sur la figure 8.9. Quant à l'algorithme CMIIC, nous utilisons sa version non-paramétrique et sélectionnons la valeur du paramètre α en maximisant la vraisemblance du modèle obtenu. Pour plus de généralité, nous utilisons la méthode de *cross-validation* avec $k = 10$ blocs (*folds* en anglais) afin de calculer la vraisemblance sur des données différentes de celles utilisées pour l'apprentissage. Brièvement, cette méthode consiste à diviser l'échantillon en k sous-échantillons, puis de réaliser k apprentissages différents en laissant à chaque fois un des sous-échantillons de côté sur lequel est ensuite calculée la vraisemblance du modèle. En répétant cette opération pour plusieurs valeurs de α, nous obtenons ainsi k courbes de vraisemblance. La figure 8.6a représente la courbe moyenne pour le cas du vin rouge et nous pouvons voir expérimentalement que son maximum est atteint en $\alpha^* = 0.04$. Le graphe obtenu pour cette valeur de α est représenté sur la figure 8.10. Contrairement à ce que l'on observe habituellement, la



(a) Vraisemblance (b) Nombre d'arcs

FIGURE 8.7 – La figure de gauche représente l'évolution de la vraisemblance et du score BIC (naïf) en fonction de α. La vraisemblance est calculée en utilisant le modèle appris avec B-CMIIC à partir de l'échantillon des vins **blancs** et en utilisant une *cross-validation* de 10 blocs. La figure de droite représente l'évolution du nombre d'arcs dans le CBN appris en fonction de α.

structure sélectionnée en utilisant la vraisemblance est creuse. En effet, celle-ci tend à préférer les structures complexes et doit alors être corrigée par un terme de complexité pour contrebalancer cet effet. Ici, l'utilisation de la copule de Bernstein empirique avec K_{MISE} comporte naturellement une pénalisation des structures denses puisque, comme nous l'avons vu dans le chapitre 6, plus la dimension de l'échantillon est grande plus K_{MISE} est faible et moins le modèle est expressif. Toutefois, remarquons que pour une dimension donnée, la valeur de K_{MISE} augmente avec la taille de l'échantillon. Ainsi, pour de plus grands échantillons, il se peut que la correction due au K_{MISE} ne soit pas suffisante. C'est justement ce que l'on observe si on trace la même courbe pour l'échantillon des vins blancs contenant plus d'exemples. Le résultat est donné sur la figure la figure 8.7a et la courbe moyenne atteint son maximum en $\alpha^* = 0.02$. Le graphe associé, qui est représenté sur la figure 8.11, contient un grand nombre d'arcs et l'utilisation d'une pénalité semble donc nécessaire. Cependant, le score BIC et ses dérivés (AIC, DIC, etc.) ne sont définis que pour des modèles paramétriques. Une idée naïve serait de remplacer le nombre de paramètres libres (3.65) par le nombre d'arcs que contient le modèle. Comme nous pouvons le voir sur les figures 8.6a et 8.7a, cette pénalité n'est pas suffisante puisque la valeur de α pour laquelle le maximum est atteint n'est pas affectée. Ceci n'est pas surprenant puisque rien n'indique *a priori* que le nombre d'arcs soit dans la même plage de valeurs que la vraisemblance. Une meilleure idée serait plutôt d'utiliser une fonction r du nombre d'arcs n_a pour obtenir une métrique similaire au score BIC :

$$d(G_\alpha) = \frac{1}{m} \log f(\boldsymbol{d}|G_\alpha) - \frac{r(n_a^{G_\alpha})}{2} \frac{\log m}{m}. \tag{8.19}$$

Cette solution reste encore à explorer et nous avons ici choisi d'utiliser une heuristique en observant l'évolution du nombre d'arcs dans la structure apprise en fonction de α. En effet, nous pouvons voir sur les figures 8.6b et 8.7b que ces courbes comportent un plateau pour certaines valeurs de α que nous avons délimité par deux lignes verticales rouges. La fonction r serait donc constante pour ces valeurs et nous avons choisi de prendre la plus petite valeur de α définissant ce plateau. Dans le cas du vin blanc, nous avons $\alpha^* = 0.06$ dont le graphe associé est représenté sur la figure 8.12.

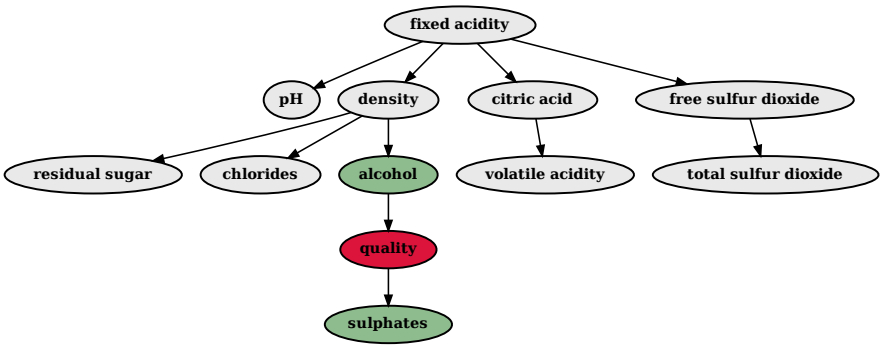


FIGURE 8.8 – Structure reconstruite avec l'algorithme CBIC à partir de l'échantillon des vins rouges. Le nombre de parents que peut avoir un nœud est limité à 1. La variable colorée en rouge est la variable cible tandis que l'ensemble des variables colorées en vert correspond à sa couverture de Markov.

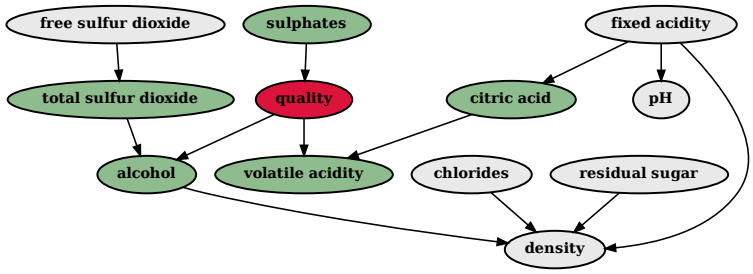


FIGURE 8.9 – Structure reconstruite avec l'algorithme CPC à partir de l'échantillon des vins rouges. Le seuil de significativité des tests d'indépendances est fixé à $p = 0.005$. La variable colorée en rouge est la variable cible tandis que l'ensemble des variables colorées en vert correspond à sa couverture de Markov.

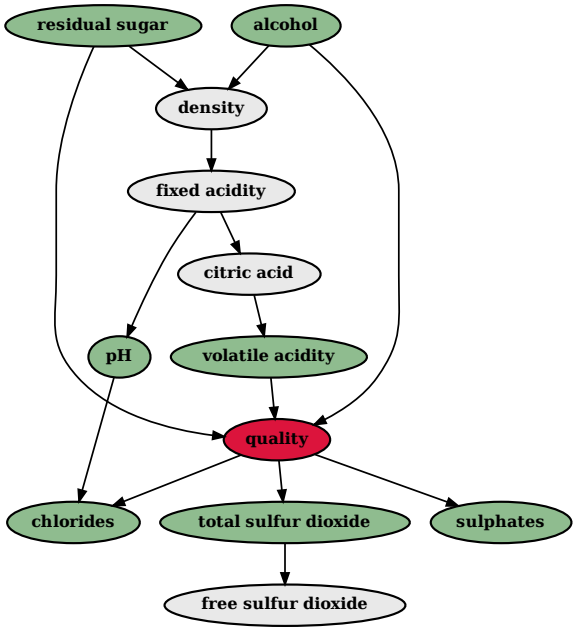


FIGURE 8.10 – Structure reconstruite avec l'algorithme B-CMIIC à partir de l'échantillon des vins rouges. Le paramètre α a été fixé à 0.04 en utilisant la méthode de *cross-validation*. La variable colorée en rouge est la variable cible tandis que l'ensemble des variables colorées en vert correspond à sa couverture de Markov.

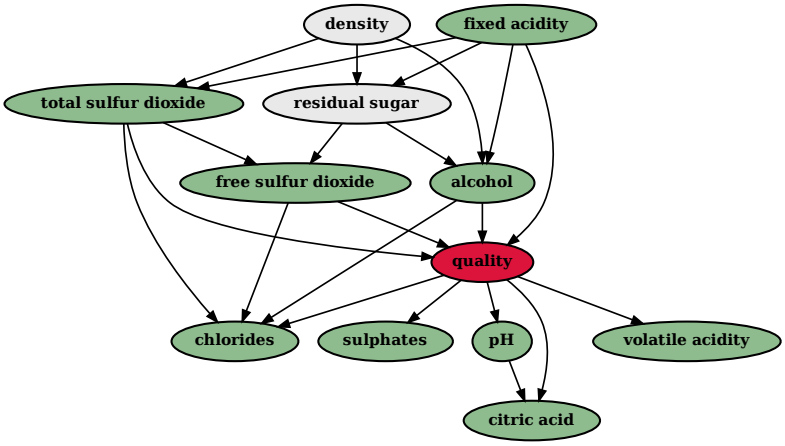


FIGURE 8.11 – Structure reconstruite avec l'algorithme B-CMIIC pour l'échantillon des vins blancs. Le paramètre α est fixé à 0.02. La variable colorée en rouge est la variable cible tandis que l'ensemble des variables colorées en vert correspond à sa couverture de Markov.

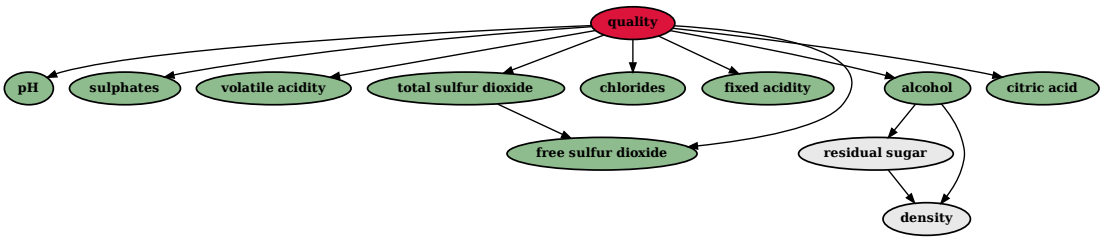


FIGURE 8.12 – Structure reconstruite avec l'algorithme B-CMIIC pour l'échantillon des vins blancs. Le paramètre α est fixé à 0.06. La variable colorée en rouge est la variable cible tandis que l'ensemble des variables colorées en vert correspond à sa couverture de Markov.

8.6.3 Sélection de variables

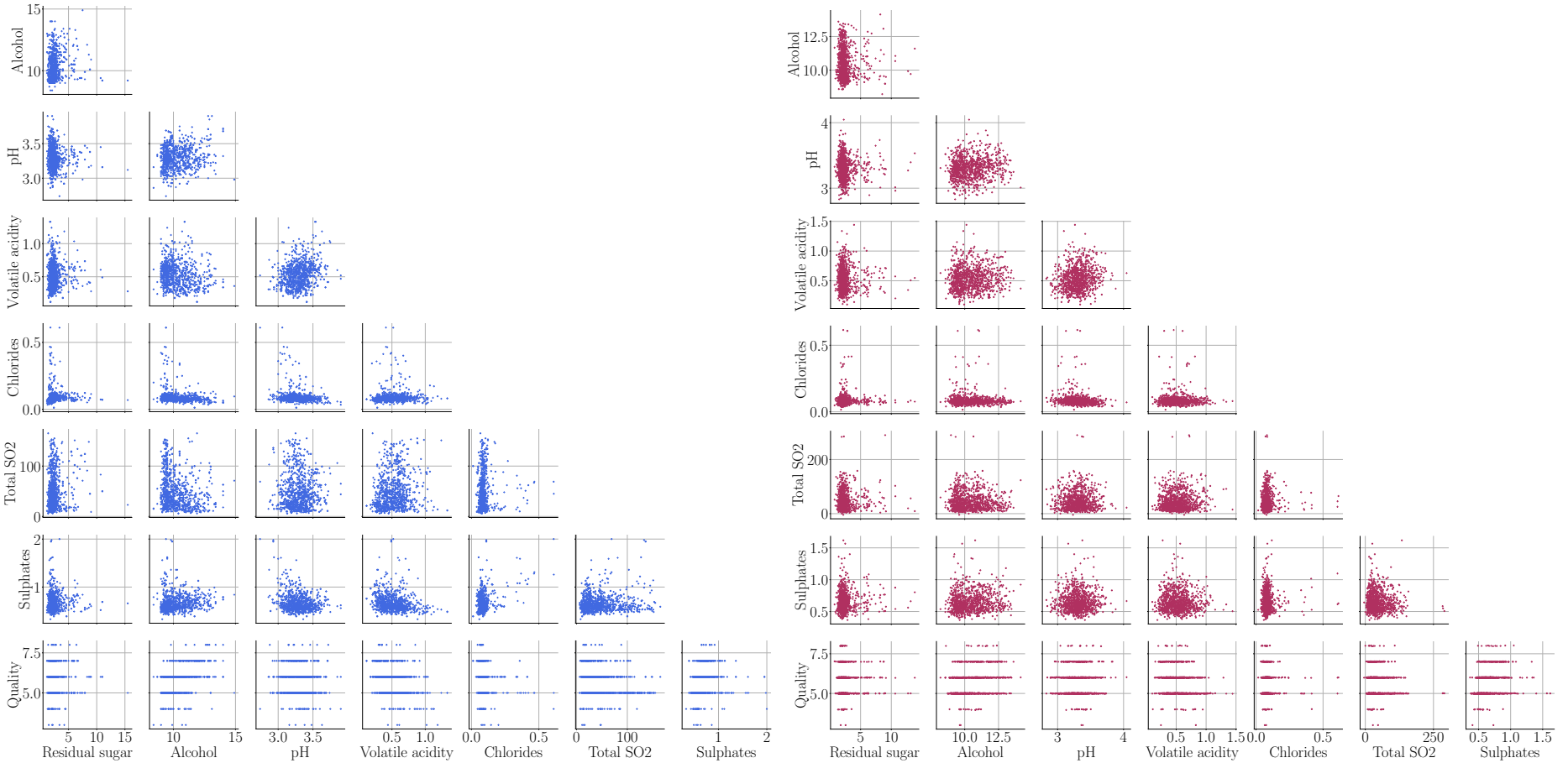
N'ayant pas ici de structure de référence, une comparaison exhaustive des différentes structures demanderait l'avis d'experts du domaine. Nous pouvons néanmoins extraire les variables importantes pour la prédiction de la qualité du vin à partir de ce que l'on appelle sa *couverture de Markov*. Celle-ci est définie comme l'ensemble des nœuds qui d-séparent *quality* du reste du graphe. Ainsi, si toutes les variables de la couverture de Markov sont observées, l'observation des autres variables n'apporte aucune information supplémentaire pour sa prédiction. D'après la définition 4.3.4, elle est donc composée des parents et des enfants de *quality* ainsi que des autres parents de ses enfants. Pour chacun des graphes que nous avons reconstruits, la variable cible est colorée en rouge tandis que les nœuds appartenant à sa couverture de Markov sont colorés en vert. Dans le cas du vin rouge, les variables *alcohol* et *sulphates* font partie des couvertures de Markov de *quality* peu importe l'algorithme utilisé. Les méthodes CPC et CMIIC s'accordent en plus sur la présence de *volatile acidity* et *total sulfur dioxide* dans celle-ci. Dans le cas du vin blanc, seuls *density* et *residual sugar* n'en font pas partie.

8.6.4 Apprentissage du modèle

Pour terminer, nous faisons une étude qualitative du modèle construit à partir de la structure obtenue avec l'algorithme CMIIC dans le cas du vin rouge et en utilisant un *kernel smoothing* gaussien pour les marginales. Pour cela, nous comparons un échan-

tillon provenant de ce modèle avec l'échantillon de référence. Afin de limiter le nombre de variables, nous nous restreignons à la variable *quality* et aux variables formant sa couverture de Markov. Pour pouvoir représenter ces échantillons, nous utilisons des nuages de points pour chaque paire de variables. Le résultat est donné sur la figure 8.13. Nous pouvons voir que la répartition des nuages obtenus à partir du modèle est assez proche de celle des données de référence mais que certains d'entre eux semblent moins diffus, en particulier ceux faisant intervenir la variable *total sulfur dioxide*. De plus, nous remarquerons que même la répartition de la variable *quality*, qui est discrète, est bien reproduite par le modèle.

Ainsi, nous avons vu avec ce cas d'application que nos algorithmes permettaient de bien représenter une loi jointe de 12 variables sans jamais la reconstruire et en utilisant seulement des densités conditionnelles. Ce modèle s'accompagne également d'une structure graphique représentant les dépendances entre ces variables nous permettant de valider le modèle obtenu avec des experts du domaine. Comme nous l'avons mentionné à plusieurs reprises, l'objectif de Cortez et al. (2009) est de permettre la prédiction de la qualité du vin en fonction des autres variables. Dans le prochain chapitre, nous concluons cette thèse en donnant plusieurs pistes de recherche afin de réaliser cette tâche dans le cadre des CBNs.



(a) Données de référence

(b) Données générées

FIGURE 8.13 – Distributions des données de références et de données générées pour chaque paire de variables dans le cas du vin rouge. Le modèle ayant généré les données a été construit à partir de la structure apprise avec B-CMIIC et ses marginales ont été estimées avec un *kernel smoothing* gaussien.

Références

Ali, S. M. et Silvey, S. D. (1966). « A general class of coefficients of divergence of one distribution from another ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 28.1, p. 131-142 (cf. p. 141).

Belalia, M., Bouezmarni, T., Lemyre, F. et Taamouti, A. (2017). « Testing independence based on Bernstein empirical copula and copula density ». In : *Journal of Nonparametric Statistics* 29.2, p. 346-380 (cf. p. 129, 145, 147, 163).

Bouezmarni, T., Rombouts, J. V. et Taamouti, A. (2010b). « Asymptotic properties of the Bernstein density copula estimator for α-mixing data ». In : *Journal of Multivariate Analysis* 101.1, p. 1-10 (cf. p. 128, 141, 142).

Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T. et Reis, J. (2009). « Using data mining for wine quality assessment ». In : *International Conference on Discovery Science*. Springer, p. 66-79 (cf. p. 153, 158).

Csiszár, I. (1964). « Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten ». In : *Magyer Tud. Akad. Mat. Kutato Int. Koezl.* 8, p. 85-108 (cf. p. 141).

Csiszár, I. (1967). « Information-type measures of difference of probability distributions and indirect observation ». In : *studia scientiarum Mathematicarum Hungarica* 2, p. 229-318 (cf. p. 141).

Elidan, G. (2010). « Copula bayesian networks ». In : *Advances in neural information processing systems*, p. 559-567 (cf. p. 3, 5, 6, 124, 125, 127, 135, 137, 141, 142, 149, 161).

Grünwald, P. D. et Grunwald, A. (2007). *The minimum description length principle*. MIT press (cf. p. 90, 149, 162, 163).

Koller, D. et Friedman, N. (2009). *Probabilistic graphical models : principles and techniques*. MIT press (cf. p. 3, 71, 72, 75, 76, 80, 131, 154).

Kulhavỳ, R. (1996). *Recursive nonlinear estimation : a geometric approach*. T. 216. Springer (cf. p. 36, 145).

Lasserre, M., Lebrun, R. et 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*. T. 35. 13, p. 12139-12146 (cf. p. 6, 142).

Ma, J. et Sun, Z. (2011). « Mutual information is copula entropy ». In : *Tsinghua Science & Technology* 16.1, p. 51-54 (cf. p. 108, 142, 146).

Sason, I. (2018). « On f-divergences : Integral representations, local behavior, and inequalities ». In : *Entropy* 20.5, p. 383 (cf. p. 143).

Verny, L., Sella, N., Affeldt, S., Singh, P. P. et Isambert, H. (2017). « Learning causal networks with latent variables from multivariate information in genomic data ». In : *PLoS computational biology* 13.10, e1005662 (cf. p. 93, 147).

Conclusion

Dans cette thèse, nous nous sommes intéressés à l'implémentation de méthodes pour l'apprentissage de distributions continues en grandes dimensions. Pour cela, nous avons utilisé le modèle des CBNs faisant le lien entre les réseaux bayésiens et la théorie des copules. Les réseaux bayésiens nous permettent de tirer parti des indépendances conditionnelles de la distribution jointe pour réduire la complexité des algorithmes tandis que la fonction copule qui contient toute l'information sur la dépendance entre les variables est au centre des tests d'indépendances utilisés pour l'apprentissage de la structure du réseau bayésien. De plus, la combinaison des deux modèles permet une plus grande liberté de modélisation en utilisant conjointement les factorisations de la distribution induites par les réseaux bayésiens et par le théorème de Sklar. Toutefois, pour éviter de faire une hypothèse de modèle, nous avons utilisé la copule de Bernstein empirique à la fois pour obtenir des tests d'indépendance non-paramétriques mais aussi pour paramétrer les CBNs, obtenant ainsi une cohérence entre apprentissage de la structure et apprentissage des paramètres. Enfin, les CBNs ayant le même langage graphique que les réseaux bayésiens, nous avons pu adapter les algorithmes classiques pour apprendre leur structure. Dans ce contexte, nous avons proposé deux algorithmes d'apprentissage :

– *Une méthode d'apprentissage par contraintes appelée CPC* et basée sur l'algorithme classique PC (Spirtes et al. 2000). Pour le test d'indépendance, nous avons utilisé celui de Bouezmarni et al. (2012) reposant sur la distance de Hellinger. Nous avons ensuite comparé cet algorithme à celui proposé par (Elidan 2010) ainsi qu'à un algorithme de recherche locale utilisant un score bayésien gaussien (Geiger et Heckerman 1994). Nous avons pu voir que l'utilisation de la copule de Bernstein empirique permettait bien à notre algorithme de se généraliser à des données non-gaussiennes.

– *L'algorithme CMIIC reposant sur l'extension du lien entre la copule et l'information multivariée conditionnelle.* À l'occasion de ce chapitre, nous avons également présenté les f-divergences et explicité leur lien avec la copule dans le cadre d'un test d'indépendance conditionnelle. Cela nous a permis de généraliser les deux tests d'indépendance que nous avons présentés et d'entrevoir la dérivation d'autres tests étant donnée une fonction génératrice. L'algorithme CMIIC s'est ensuite avéré avoir les meilleures performances en terme de scores structurels comparativement à l'ensemble des méthodes présentées. En particulier, nous avons pu voir que même pour des problèmes de dimension $n \approx 100$, les graphes reconstruits étaient proches de la structure de référence. Enfin, nous avons également testé nos algorithmes sur un cas d'application provenant de l'analyse physico-chimique et sensorielle de vins. À cette occasion, nous avons comparé les différentes structures obtenues et, par le biais de la couverture de Markov, nous avons pu sélectionner les variables les plus importantes pour déterminer la qualité du vin.

Perspectives

Suite aux travaux que nous avons présentés dans cette thèse, nous proposons plusieurs pistes de recherches qui nous semblent pertinentes pour améliorer nos algorithmes d'apprentissage ainsi que pour permettre des inférences dans les CBNs.

Amélioration de l'algorithme CPC

Nous voyons deux points d'amélioration pour l'algorithme CPC que nous avons introduit dans le chapitre 7. Le premier et le plus facile à mettre en œuvre serait d'utiliser les extensions permettant d'éviter le choix d'un ordre sur les variables qui sont discutées dans Colombo et al. (2014). En effet, les modifications à apporter à notre algorithme pour cela ne seraient que de l'ordre de l'implémentation et ne demanderaient donc aucun travail théorique. Le deuxième point d'amélioration concerne le test d'indépendance conditionnelle utilisé pour l'apprentissage du squelette. Nous avons vu dans le chapitre 8 que la distance de Hellinger comme l'entropie relative faisaient partie d'une classe de métriques appelées f-divergences permettant de quantifier la différence entre deux distributions. Nous avons également montré comment à partir de la copule empirique de Bernstein et étant donnée une divergence, nous pouvions en dériver un test d'indépendance conditionnelle non-paramétrique. Par une étude comparative de différentes f-divergences, nous pourrions alors améliorer les résultats obtenus ou du moins laisser le choix à l'utilisateur. Cependant, cela demande de dériver la distribution de la statistique de test associée ce qui n'est pas toujours aisé même dans le cas asymptotique.

Amélioration de l'algorithme CMIIC

Lors de la présentation de notre algorithme CMIIC, nous avons choisi d'introduire un paramètre α pour prendre en compte le biais d'échantillon fini dans le calcul de l'information mutuelle. La raison pour laquelle nous avons choisi une correction constante plutôt que d'utiliser la complexité comme dans le cas discret, vient du fait que d'une part le NML suppose un modèle paramétrique et que d'autre part cette dernière diverge pour la plupart des modèles continus. Ainsi, malgré de très bons résultats, l'algorithme CMIIC reste perfectible puisque nous nous attendons à ce qu'en réalité la correction soit dépendante non seulement de la taille de l'échantillon mais aussi de la taille de l'ensemble conditionnant. Pour cela, nous envisageons deux solutions possibles en fonction de si le modèle de copule utilisé est paramétrique ou non.

Dans le cas paramétrique, nous pourrions utiliser l'extension de l'approche NML appelée *Lucky Normalized Maximum Likelihood* (LNML). Plusieurs versions différentes existent (Grünwald et al. 2007, section 11.3) et nous considérons ici la version LNML-2. La densité LNML se définit relativement à une fonction de *luckiness* $a : \boldsymbol{\Theta} \to \mathbb{R}$ et a pour expression :

$$f^a_{\text{LNML}}(\boldsymbol{d}|M) = \frac{f(\boldsymbol{d}|\tilde{\boldsymbol{\theta}}, M)e^{-a(\tilde{\boldsymbol{\theta}}|M)}}{C(M,a)}, \quad \text{avec } C(M,a) = \int_{\boldsymbol{d}\in\mathcal{D}} f(\boldsymbol{d}|\tilde{\boldsymbol{\theta}}, M)e^{-a(\tilde{\boldsymbol{\theta}}|M)}\mathrm{d}\boldsymbol{d}$$

avec $\tilde{\theta} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} [-\log f(\boldsymbol{d}|\boldsymbol{\theta}, M) + a(\boldsymbol{\theta}|M)]$. La fonction de *luckiness* est arbitraire et se rapproche en cela de la densité *a priori* dans l'approche bayésienne. Pour une discussion sur les similarités et les différences entre ces deux fonctions, le lecteur pourra se référer à Grünwald et al. (2007, section 17.2.1). Remarquons qui si nous prenons une *luckiness* uniforme, $a(\boldsymbol{\theta}|M) = 0$, nous retrouvons le NML classique. Dans le cas gaussien, nous pourrions utiliser les résultats de Miyaguchi (2017) donnant l'expression analytique de la complexité pour plusieurs fonctions de *luckiness*. En étendant les

travaux de Roos et al. (2008) au LNML, nous pourrions alors factoriser la complexité sur la structure d'un réseau bayésien et ainsi obtenir une pénalité adaptée. Lorsque les calculs ne peuvent pas être menés sous forme analytique, les développements asymptotiques présentés dans Grünwald et al. (2007, section 11.3.1) peuvent être utilisés afin d'obtenir une approximation.

Dans le cadre non-paramétrique, nous proposons de dériver la distribution de l'estimateur et nous en servir pour calculer des *p-values* comme nous l'avons fait avec le test non-paramétrique basé sur la distance de Hellinger. Remarquons qu'en procédant ainsi, nous ne serions donc plus dans une approche MDL de la sélection de modèle mais plutôt dans une approche classique. Belalia et al. (2017), dans le même esprit que Bouezmarni et al. (2012), ont proposé une correction de l'estimateur de l'information mutuelle afin que celui-ci soit distribué asymptotiquement selon une loi normale standard. Nous pourrions alors nous servir de ces travaux comme point de départ pour dériver une correction pour l'information conditionnelle et l'information multivariée, nécessaires à l'implémentation de l'algorithme MIIC. Notons toutefois que nous avons menés de premières expériences sur la statistique de test proposée par Belalia et al. (2017) mais ceux-ci se sont avérés peu concluants puisqu'elle ne semble pas être asymptotiquement distribuée selon une loi normale standard. Enfin, bien que le NML soit restreint au cas paramétrique, l'approche MDL peut être utilisée dans un cadre non-paramétrique et Grünwald et al. (2007, chapitre 13) donnent plusieurs pistes en ce sens que nous pourrions explorer.

Extension au cas hybride

Dans la pratique, la majorité des cas d'applications portent sur des vecteurs aléatoires dont chaque composante peut être soit discrète soit continue. Pour cette raison, nos algorithmes ont besoin d'être étendus à ce cas appelé cas hybride. Toutefois, nous avons vu avec les données du vin que nos algorithmes pouvaient déjà fonctionner avec des variables continues ou discrètes puisque la qualité du vin prenait des valeurs discrètes tandis que les autres variables étaient continues. Ces résultats étant encore préliminaires, nous avons besoin de mener une étude plus approfondie sur le comportement des algorithmes dans ce cas là. Insistons tout de même que d'un point de vue théorique, le théorème de Sklar peut toujours s'appliquer lorsque les variables sont discrètes mais la copule associée n'est alors plus unique. En revanche, la copule discrète définie sur les points de la mesure discrète est unique et il suffit alors d'utiliser n'importe quel prolongement continu de cette dernière pour pouvoir adapter nos algorithmes.

Inférence dans les réseaux bayésiens continus

Comme nous l'avons vu en introduction, l'un des objectifs des réseaux bayésiens est de permettre de raisonner dans l'incertain. Actuellement, nos algorithmes se concentrent sur l'apprentissage du modèle pour faire de la découverte de connaissances ou bien de l'échantillonnage. Une suite logique de nos travaux est donc de proposer des algorithmes d'inférence pour les CBNs afin de pouvoir ensuite faire de la prédiction ou de la classification. Étant donné un vecteur aléatoire $\mathbf{X}$ de densité jointe $f_{\mathbf{X}}$ et deux sous-vecteurs $\mathbf{Y}$ et $\mathbf{Z}$ qui en sont issus, une inférence consiste à obtenir la densité conditionnelle $f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}_o) = \frac{f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z}_o)}{f_{\mathbf{Z}}(\mathbf{z}_o)}$ où $\mathbf{z}_o$ est une observation *fixée* de $\mathbf{Z}$. Pour cela, il suffit de calculer le numérateur puisqu'on peut obtenir le dénominateur en marginalisant ensuite sur $\mathbf{Y}$:

$$f_{\mathbf{Z}}(\mathbf{z}_o) = \int_{\Omega_{\mathbf{Y}}} f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}, \mathbf{z}_o) \mathrm{d}\mathbf{y}.$$

Pour obtenir le numérateur, nous devons évaluer $f_{\mathbf{X}}$ en $\mathbf{z}_o$ puis marginaliser sur les composantes de $\mathbf{X}$ qui ne sont ni dans $\mathbf{Y}$ ni dans $\mathbf{Z}$ et qui forment le sous-vecteur que l'on note $\mathbf{W}$:

$$f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z}_o) = \int_{\Omega_{\mathbf{W}}} f_{\mathbf{W},\mathbf{Y},\mathbf{Z}}(\mathbf{w},\mathbf{y},\mathbf{z}_o)\mathrm{d}\mathbf{w}. \tag{8.20}$$

Cette tâche, comme l'apprentissage, est un problème NP-difficile (Cooper 1990) dont les réseaux bayésiens permettent de réduire la complexité et de mettre en place des algorithmes efficaces afin de pouvoir tout de même mener les calculs. Dans le cas discret, ces intégrales se réduisent à des sommes qui peuvent être factorisées grâce aux indépendances du réseau bayésien. La méthode générique utilise ce qu'on appelle des arbres de jonction (Madsen et al. 1999) obtenus à partir de la structure du réseau bayésien et qui représentent graphiquement les calculs à mener pour obtenir les densités conditionnelles. De plus, certains résultats partiels sont stockés au sein de cette structure ce qui permet de recycler les calculs lorsque plusieurs inférences sont réalisées. Ces algorithmes requièrent toutefois que la classe de modèle utilisée pour les densités conditionnelles soit stable par marginalisation et par produit (Shenoy 1997) ce qui n'est pas a priori notre cas. Remarquons toutefois que pour mener une intégration numérique, nous opérons l'analogue à une discrétisation ce qui pourrait permettre d'utiliser des réseaux bayésiens discrets et leurs algorithmes d'inférence pour calculer ces intégrales.

En effet, soit g une fonction définie sur $E \subseteq \mathbb{R}$ et supposons que nous voulions calculer l'intégrale suivante :

$$I[g] = \int_a^b g(x)\omega(x)\mathrm{d}x$$

où $[a,b] \subseteq \mathbb{R}$ et ω est une fonction de poids. Les méthodes d'intégration par quadrature de Gauss (Kincaid et al. 2009) consistent à estimer cette intégrale par la somme suivante :

$$\hat{I}_p[g] = \sum_{i=1}^{p} \alpha_i g(x[i]).$$

où les p valeurs distinctes $x[i]$ sont appelées les points de la quadrature tandis que les α_i sont appelés coefficients de la quadrature. La position des points est choisie de manière à minimiser l'erreur du résultat et correspond aux zéros d'une famille de polynômes orthonormés déterminée par la fonction de poids utilisée. Par exemple, lorsque cette fonction est $\omega(x) = 1$, ces polynômes sont ceux de Legendre. De même, les coefficients sont déterminés à partir de formules dépendant de la fonction de poids. La méthode s'étend facilement au cas d'une intégrale multiple à n dimensions :

$$\hat{I}_{\boldsymbol{p}}[g] = \sum_{i_1=1}^{p_1} \cdots \sum_{i_n=1}^{p_n} \alpha_{1,i_1} \ldots \alpha_{n,i_n} g(x_1[i_1],\ldots,x_n[i_n]),$$

avec $\boldsymbol{p} = (p_1,\ldots,p_n)$. En appliquant cette formule pour le calcul de la densité conditionnelle, nous obtenons alors la somme suivante :

$$\hat{f}^{\boldsymbol{p}}_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z}_o) = \sum_{i_1=1}^{p_1} \cdots \sum_{i_n=1}^{p_n} \alpha_{1,i_1} \ldots \alpha_{n,i_n} f_{\mathbf{W},\mathbf{Y},\mathbf{Z}}(\mathbf{w}[\boldsymbol{i}],\mathbf{y},\mathbf{z}_o)$$

où $\boldsymbol{i} = (i_1,\ldots,i_k)$ avec k la dimension du vecteur $\mathbf{W}$. En utilisant la décomposition de la loi jointe sur la structure du graphe, cette somme peut être factorisée de manière à réduire le nombre d'opération.

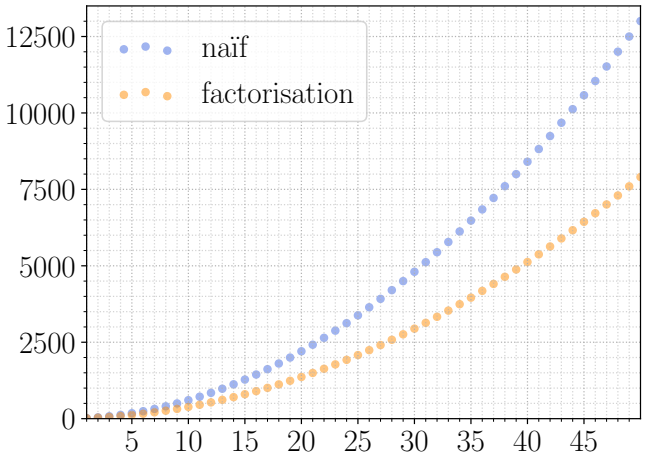


FIGURE 8.14 – Nombre d'opérations nécessaires pour le calcul de la densité marginale de X_3 en fonction du nombre de points de quadrature et pour une structure de BN $X_1 \to X_2 \to X_3$.

Pour illustrer notre propos, prenons le cas d'un réseau bayésien dont la structure est une chaîne $X_1 \to X_2 \to X_3$ et pour lequel nous voulons obtenir la densité marginale sur X_3. Pour cela, nous devons mener le calcul :

$$f(x_3) = \int_{\Omega_{X_1}} \int_{\Omega_{X_2}} f(x_1, x_2, x_3) \mathrm{d}x_1 \mathrm{d}x_2 = \int_{\Omega_{X_1}} \int_{\Omega_{X_2}} f(x_1) f(x_2|x_1) f(x_3|x_2) \mathrm{d}x_1 \mathrm{d}x_2$$

En utilisant une quadrature de Gauss, nous avons alors :

$$\hat{f}_{\boldsymbol{p}}(x_3) = \sum_{i_1=0}^{p_1} \sum_{i_2=0}^{p_2} \alpha_{1,i_1} \alpha_{2,i_2} f(x_1[i_1]) f(x_2[i_2]|x_1[i_1]) f(x_3|x_2[i_2])$$

Tel quel, ce calcul demande $5p_1p_2 + 5(p_1 + p_2) + 4$ opérations. Or, certains termes peuvent être factorisés et la somme peut être réécrite :

$$\hat{f}_{\boldsymbol{p}}(x_3) = \sum_{i_2=0}^{p_2} \alpha_{2,i_2} f(x_3|x_2[i_2]) \sum_{i_1=0}^{p_1} \alpha_{1,i_1} f(x_1[i_1]) f(x_2[i_2]|x_1[i_1]),$$

ce qui réduit le nombre d'opérations à $3p_1p_2 + 3p_1 + 5p_2 + 4$. Pour une meilleur visualisation, nous avons tracé le nombre d'opérations en fonction de $p_1 = p_2 = p$ sur la figure 8.14. Les arbres de jonctions permettent justement d'organiser ces calculs efficacement pour n'importe quelle structure et nous pensons donc que par une instanciation, dont les détails restent encore à préciser, des tables de probabilités conditionnelles d'un réseau bayésien discret avec les coefficients et les valeurs de la fonction aux points de quadrature, nous pourrions donc aboutir à des inférences continues.

Dans le cas où cette approche ne serait pas possible, nous pourrions à la place utiliser des méthodes approchées qui estiment les densités conditionnelles en échantillonnant la distribution. Le problème reste NP-difficile mais, comme nous l'avons vu avec la méthode du *forward sampling*, l'exploitation des indépendances permet là aussi de réduire la complexité. Par conséquent, nous pourrions utiliser des méthodes plus sophistiquées, comme les méthodes de Monte-Carlo par chaîne de Markov, afin d'obtenir de meilleures estimations.

Références

Belalia, M., Bouezmarni, T., Lemyre, F. et Taamouti, A. (2017). « Testing independence based on Bernstein empirical copula and copula density ». In : *Journal of Nonparametric Statistics* 29.2, p. 346-380 (cf. p. 129, 145, 147, 163).

Bouezmarni, T., Rombouts, J. V. et Taamouti, A. (2012). « Nonparametric copula-based test for conditional independence with applications to Granger causality ». In : *Journal of Business & Economic Statistics* 30.2, p. 275-287 (cf. p. 128, 161, 163).

Colombo, D. et Maathuis, M. H. (2014). « Order-independent constraint-based causal structure learning ». In : *The Journal of Machine Learning Research* 15.1, p. 3741-3782 (cf. p. 90, 133, 162).

Cooper, G. F. (1990). « The computational complexity of probabilistic inference using Bayesian belief networks ». In : *Artificial intelligence* 42.2-3, p. 393-405 (cf. p. 164).

Elidan, G. (2010). « Copula bayesian networks ». In : *Advances in neural information processing systems*, p. 559-567 (cf. p. 3, 5, 6, 124, 125, 127, 135, 137, 141, 142, 149, 161).

Geiger, D. et Heckerman, D. (1994). « Learning gaussian networks ». In : *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., p. 235-243 (cf. p. 5, 80, 135, 161).

Grünwald, P. D. et Grunwald, A. (2007). *The minimum description length principle*. MIT press (cf. p. 90, 149, 162, 163).

Kincaid, D., Kincaid, D. R. et Cheney, E. W. (2009). *Numerical analysis : mathematics of scientific computing*. T. 2. American Mathematical Soc. (cf. p. 164).

Madsen, A. L. et Jensen, F. V. (1999). « Lazy propagation : a junction tree inference algorithm based on lazy evaluation ». In : *Artificial Intelligence* 113.1-2, p. 203-245 (cf. p. 164).

Miyaguchi, K. (2017). « Normalized Maximum Likelihood with Luckiness for Multivariate Normal Distributions ». In : *arXiv preprint arXiv :1708.01861* (cf. p. 162).

Roos, T., Silander, T., Kontkanen, P. et Myllymaki, P. (2008). « Bayesian network structure learning using factorized NML universal models ». In : *2008 Information Theory and Applications Workshop*. IEEE, p. 272-276 (cf. p. 90, 163).

Shenoy, P. P. (1997). « Binary join trees for computing marginals in the Shenoy-Shafer architecture ». In : *International Journal of approximate reasoning* 17.2-3, p. 239-263 (cf. p. 2, 164).

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G. et Richardson, T. (2000). *Causation, prediction, and search*. MIT press (cf. p. 5, 85, 86, 161).

Bibliographie

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Khalek, S. A., Abdelalim, A., Aben, R., Abi, B., Abolins, M., Abouzeid, O. et al. (2012). « Combined search for the Standard Model Higgs boson in p p collisions at s= 7 TeV with the ATLAS detector ». In : *Physical Review D* 86.3, p. 032003 (cf. p. 49).

Affeldt, S. et Isambert, H. (2015). « Robust Reconstruction of Causal Graphical Models based on Conditional 2-point and 3-point Information. » In : *ACI@ UAI*, p. 1-29 (cf. p. 5, 90, 91, 93).

Affeldt, S., Verny, L. et Isambert, H. (2016). « 3off2 : A network reconstruction algorithm based on 2-point and 3-point information statistics ». In : *BMC bioinformatics*. T. 17. 2. BioMed Central, S12 (cf. p. 90, 93).

Ali, S. M. et Silvey, S. D. (1966). « A general class of coefficients of divergence of one distribution from another ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 28.1, p. 131-142 (cf. p. 141).

Anastassiou, G. A. et Gal, S. G. (2012). *Approximation theory : moduli of continuity and global smoothness preservation*. Springer Science & Business Media (cf. p. 113).

Arellano-Valle, R. B., Contreras-Reyes, J. E. et Genton, M. G. (2013). « Shannon Entropy and Mutual Information for Multivariate Skew-Elliptical Distributions ». In : *Scandinavian Journal of Statistics* 40.1, p. 42-62 (cf. p. 110).

Bartlett, M. et Cussens, J. (2017). « Integer linear programming for the Bayesian network structure learning problem ». In : *Artificial Intelligence* 244, p. 258-271 (cf. p. 80).

Baudin, M., Dutfoy, A., Iooss, B. et Popelin, A.-L. (2016). « OpenTURNS : An Industrial Software for Uncertainty Quantification in Simulation ». In : *Handbook of Uncertainty Quantification*. Sous la dir. de R. Ghanem, D. Higdon et H. Owhadi. Cham : Springer International Publishing, p. 1-38 (cf. p. 135).

Bedford, T. et Cooke, R. M. (2002). « Vines–a new graphical model for dependent random variables ». In : *The Annals of Statistics* 30.4, p. 1031-1068 (cf. p. 3, 124).

Beinlich, I. A., Suermondt, H. J., Chavez, R. M. et Cooper, G. F. (1989). « The ALARM monitoring system : A case study with two probabilistic inference techniques for belief networks ». In : *AIME 89*. Springer, p. 247-256 (cf. p. 129).

Belalia, M., Bouezmarni, T., Lemyre, F. et Taamouti, A. (2017). « Testing independence based on Bernstein empirical copula and copula density ». In : *Journal of Nonparametric Statistics* 29.2, p. 346-380 (cf. p. 129, 145, 147, 163).

Benhamou, E. et Melot, V. (2018). « Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation ». In : *arXiv preprint arXiv :1808.09171* (cf. p. 53).

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media (cf. p. 57).

Berger, J. O. et Sellke, T. (1987). « Testing a point null hypothesis : The irreconcilability of p values and evidence ». In : *Journal of the American statistical Association* 82.397, p. 112-122 (cf. p. 49).

Bernstein, S. (1912). « Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités ». In : *Comm. Kharkov Math. Soc.* 13.1, p. 1-2 (cf. p. 112).

Bhatti, M. I. et Do, H. Q. (2019). « Recent development in copula and its applications to the energy, forestry and environmental sciences ». In : *International Journal of Hydrogen Energy* 44.36, p. 19453-19473 (cf. p. 2).

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons (cf. p. 111).

Bollobás, B. (2013). *Modern graph theory*. T. 184. Springer Science & Business Media (cf. p. 65).

Bondy, J. A. et Murty, U. S. R. (1976). *Graph theory with applications*. T. 290. Macmillan London (cf. p. 65).

Bouezmarni, T., Rombouts, J. V. et Taamouti, A. (2010a). « Asymptotic properties of the Bernstein density copula estimator for α-mixing data ». In : *Journal of Multivariate Analysis* 101.1, p. 1-10 (cf. p. 123).

Bouezmarni, T., Rombouts, J. V. et Taamouti, A. (2010b). « Asymptotic properties of the Bernstein density copula estimator for α-mixing data ». In : *Journal of Multivariate Analysis* 101.1, p. 1-10 (cf. p. 128, 141, 142).

Bouezmarni, T., Rombouts, J. V. et Taamouti, A. (2012). « Nonparametric copula-based test for conditional independence with applications to Granger causality ». In : *Journal of Business & Economic Statistics* 30.2, p. 275-287 (cf. p. 128, 161, 163).

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G. et Roncalli, T. (2000). « Copulas for finance-a reading guide and some applications ». In : *Available at SSRN 1032533* (cf. p. 108).

Box, G. E. P. (1958). « A note on the generation of random normal deviates ». In : *Ann. Math. Statist.* 29, p. 610-611 (cf. p. 22).

Bretto, A., Faisant, A. et Hennecart, F. (2012). *Éléments de théorie des graphes*. Springer (cf. p. 65).

Buntine, W. (1991). « Theory refinement on Bayesian networks ». In : *Uncertainty Proceedings 1991*. Elsevier, p. 52-60 (cf. p. 77).

Butler, R. W. (2007). *Saddlepoint approximations with applications*. T. 22. Cambridge University Press (cf. p. 56).

Candelpergher, B. (2013). *Théorie des probabilités* (cf. p. 11, 13, 14).

Chickering, D. M. (1996). « Learning Bayesian networks is NP-complete ». In : *Learning from data*. Springer, p. 121-130 (cf. p. 75).

Chickering, D. M. et Heckerman, D. (1997). « Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables ». In : *Machine learning* 29.2, p. 181-212 (cf. p. 58).

Colombo, D. et Maathuis, M. H. (2014). « Order-independent constraint-based causal structure learning ». In : *The Journal of Machine Learning Research* 15.1, p. 3741-3782 (cf. p. 90, 133, 162).

Cooper, G. F. (1990). « The computational complexity of probabilistic inference using Bayesian belief networks ». In : *Artificial intelligence* 42.2-3, p. 393-405 (cf. p. 164).

Cooper, G. F. et Herskovits, E. (1992). « A Bayesian method for the induction of probabilistic networks from data ». In : *Machine learning* 9.4, p. 309-347 (cf. p. 77).

Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T. et Reis, J. (2009). « Using data mining for wine quality assessment ». In : *International Conference on Discovery Science*. Springer, p. 66-79 (cf. p. 153, 158).

Cottin, C. et Pfeifer, D. (2014). « From Bernstein polynomials to Bernstein copulas ». In : *J. Appl. Funct. Anal* 9.3-4, p. 277-288 (cf. p. 114).

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons (cf. p. 92).

Cover, T. M. et Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons (cf. p. 31).

Csiszár, I. (1964). « Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten ». In : *Magyer Tud. Akad. Mat. Kutato Int. Koezl.* 8, p. 85-108 (cf. p. 141).

Csiszár, I. (1967). « Information-type measures of difference of probability distributions and indirect observation ». In : *studia scientiarum Mathematicarum Hungarica* 2, p. 229-318 (cf. p. 141).

Czado, C. (2010). « Pair-copula constructions of multivariate copulas ». In : *Copula theory and its applications*. Springer, p. 93-109 (cf. p. 3, 124).

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge university press (cf. p. 3, 75).

DeGroot, M. H. (2005). *Optimal statistical decisions*. T. 82. John Wiley & Sons (cf. p. 44).

Deheuvels, P. (1979). « La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance ». In : *Bulletins de l'Académie Royale de Belgique* 65.1, p. 274-292 (cf. p. 111).

Dempster, A. P. (1968). « A generalization of Bayesian inference ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 30.2, p. 205-232 (cf. p. 1).

DeVore, R. A. et Lorentz, G. G. (1993). *Constructive approximation*. T. 303. Springer Science & Business Media (cf. p. 113).

Diestel, R. (2005). « Graph theory 3rd ed ». In : *Graduate texts in mathematics* 173 (cf. p. 65).

Dubois, D. et Prade, H. (1988). « Possibility Theory - An Approach to Computerized Processing of Uncertainty ». In : (cf. p. 1).

Ducamp, G., Gonzales, C. et 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*. T. 138. Proceedings of Machine Learning Research. Skørping, Denmark, p. 609-612 (cf. p. 135).

Durante, F. et Sempi, C. (2016). *Principles of copula theory*. T. 474. CRC press Boca Raton, FL (cf. p. 98, 105).

Ebrahimi, N., Soofi, E. S. et Zhao, S. (2011). « Information measures of Dirichlet distribution with applications ». In : *Applied Stochastic Models in Business and Industry* 27.2, p. 131-150 (cf. p. 110).

Efron, B. et Hinkley, D. V. (1978). « Assessing the accuracy of the maximum likelihood estimator : Observed versus expected Fisher information ». In : *Biometrika* 65.3, p. 457-483 (cf. p. 57).

Elidan, G. (2010). « Copula bayesian networks ». In : *Advances in neural information processing systems*, p. 559-567 (cf. p. 3, 5, 6, 124, 125, 127, 135, 137, 141, 142, 149, 161).

Fréchet, M. (1951). « Sur les tableaux de corrélation dont les marges sont données ». In : *Ann. Univ. Lyon, 3ˆ e serie, Sciences, Sect. A* 14, p. 53-77 (cf. p. 2).

Freedman, D. (1997). « Some issues in the foundation of statistics ». In : *Topics in the Foundation of Statistics*. Springer, p. 19-39 (cf. p. 42).

Geiger, D. et Heckerman, D. (1994). « Learning gaussian networks ». In : *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., p. 235-243 (cf. p. 5, 80, 135, 161).

Geiger, D. et Pearl, J. (1990). « On the logic of causal models ». In : *Machine Intelligence and Pattern Recognition*. T. 9. Elsevier, p. 3-14 (cf. p. 71).

Genest, C. et Favre, A.-C. (2007). « Everything you always wanted to know about copula modeling but were afraid to ask ». In : *Journal of hydrologic engineering* 12.4, p. 347-368 (cf. p. 106, 107).

Genest, C., Gendron, M. et Bourdeau-Brien, M. (2009). « The advent of copulas in finance ». In : *The European journal of finance* 15.7-8, p. 609-618 (cf. p. 2).

Glover, F. et Laguna, M. (1998). « Tabu Search ». In : *Handbook of Combinatorial Optimization : Volume1–3*. Sous la dir. de D.-Z. Du et P. M. Pardalos. Boston, MA : Springer US, p. 2093-2229 (cf. p. 80).

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media (cf. p. 31, 34).

Grimmett, G. et Stirzaker, D. (2020). *Probability and random processes*. Oxford university press (cf. p. 11).

Grünwald, P. D. et Grunwald, A. (2007). *The minimum description length principle*. MIT press (cf. p. 90, 149, 162, 163).

Hanea, A., Napoles, O. M. et Ababei, D. (2015). « Non-parametric Bayesian networks : Improving theory and reviewing applications ». In : *Reliability Engineering & System Safety* 144, p. 265-284 (cf. p. 124).

Hanea, A. M. (2008). « Algorithms for non-parametric Bayesian belief nets ». In : (cf. p. 124).

Heckerman, D., Geiger, D. et Chickering, D. M. (1995). « Learning Bayesian networks : The combination of knowledge and statistical data ». In : *Machine learning* 20.3, p. 197-243 (cf. p. 77, 79).

Hogg, R. V., McKean, J. et Craig, A. T. (2005). *Introduction to mathematical statistics*. Pearson Education (cf. p. 49).

Hult, H. et Lindskog, F. (2002). « Multivariate extremes, aggregation and dependence in elliptical distributions ». In : *Advances in Applied probability*, p. 587-608 (cf. p. 110).

Ide, J. S. et Cozman, F. G. (2002). « Random generation of Bayesian networks ». In : *Brazilian symposium on artificial intelligence*. Springer, p. 366-376 (cf. p. 130).

Jaynes, E. (1963). *Brandeis Summer Institute Lectures in Theoretical Physics : Statistical Physics* (cf. p. 32).

Joe, H. (1993). « Parametric families of multivariate distributions with given margins ». In : *Journal of multivariate analysis* 46.2, p. 262-282 (cf. p. 107).

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press (cf. p. 2, 98).

Kallenberg, O. et Kallenberg, O. (1997). *Foundations of modern probability*. T. 2. Springer (cf. p. 27).

Khinchin, A. Y. (1957). « Mathematical foundations of information theory ». In : (cf. p. 32).

Kincaid, D., Kincaid, D. R. et Cheney, E. W. (2009). *Numerical analysis : mathematics of scientific computing*. T. 2. American Mathematical Soc. (cf. p. 164).

Kirkpatrick, S., Gelatt, C. D. et Vecchi, M. P. (1983). « Optimization by simulated annealing ». In : *science* 220.4598, p. 671-680 (cf. p. 80).

Kirshner, S. (2008). « Learning with tree-averaged densities and distributions ». In : *Advances in Neural Information Processing Systems*, p. 761-768 (cf. p. 125).

Koller, D. et Friedman, N. (2009). *Probabilistic graphical models : principles and techniques*. MIT press (cf. p. 3, 71, 72, 75, 76, 80, 131, 154).

Kolmogorov, A. N. et Bharucha-Reid, A. T. (2018). *Foundations of the theory of probability : Second English Edition*. Courier Dover Publications (cf. p. 11).

Kontkanen, P. et Myllymäki, P. (2007). « A linear-time algorithm for computing the multinomial stochastic complexity ». In : *Information Processing Letters* 103.6, p. 227-233 (cf. p. 90, 91).

Kulhavỳ, R. (1996). *Recursive nonlinear estimation : a geometric approach*. T. 216. Springer (cf. p. 36, 145).

Kurowicka, D. et Cooke, R. (2005). « Distribution-free continuous Bayesian belief ». In : *Modern statistical and mathematical methods in reliability* 10, p. 309 (cf. p. 124).

Langseth, H., Nielsen, T. D., Rumi, R. et Salmerón, A. (2012). « Mixtures of truncated basis functions ». In : *International Journal of Approximate Reasoning* 53.2, p. 212-227 (cf. p. 2, 124).

Lasserre, M., Lebrun, R. et Wuillemin, P.-H. (2020). « Constraint-Based Learning for Non-Parametric Continuous Bayesian Networks ». In : *FLAIRS 33 - 33rd Florida Artificial Intelligence Research Society Conference*. Miami, United States : AAAI, p. 581-586 (cf. p. 5, 124).

Lasserre, M., Lebrun, R. et Wuillemin, P.-H. (2021a). « Constraint-based learning for non-parametric continuous bayesian networks ». In : *Annals of Mathematics and Artificial Intelligence*, p. 1-18 (cf. p. 6, 124).

Lasserre, M., Lebrun, R. et 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*. T. 35. 13, p. 12139-12146 (cf. p. 6, 142).

Lauritzen, S. L. (1992). « Propagation of probabilities, means, and variances in mixed graphical association models ». In : *Journal of the American Statistical Association* 87.420, p. 1098-1108 (cf. p. 124).

Lauritzen, S. L. et Wermuth, N. (1989). « Graphical models for associations between variables, some of which are qualitative and some quantitative ». In : *The annals of Statistics*, p. 31-57 (cf. p. 2, 124).

Lebrun, R. (2013). « Contributions à la modélisation de la dépendance stochastique ». Thèse de doct. Université Paris-Diderot-Paris VII (cf. p. 100).

Lehmann, E. L. et Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media (cf. p. 49, 50).

Lehmann, E. L. (1966). « Some concepts of dependence ». In : *The Annals of Mathematical Statistics*, p. 1137-1153 (cf. p. 104).

Lindskog, F., McNeil, A. et Schmock, U. (2003). « Kendall's tau for elliptical distributions ». In : *Credit Risk*. Springer, p. 149-156 (cf. p. 110).

Lorentz, G. G. (2012). *Bernstein polynomials*. American Mathematical Soc. (cf. p. 113).

Ma, J. et Sun, Z. (2011). « Mutual information is copula entropy ». In : *Tsinghua Science & Technology* 16.1, p. 51-54 (cf. p. 108, 142, 146).

Madsen, A. L. et Jensen, F. V. (1999). « Lazy propagation : a junction tree inference algorithm based on lazy evaluation ». In : *Artificial Intelligence* 113.1-2, p. 203-245 (cf. p. 164).

Mai, J.-F. et Scherer, M. (2014). « How to Measure Dependence ? » In : *Financial Engineering with Copulas Explained*. London : Palgrave Macmillan UK, p. 35-48 (cf. p. 103).

Massey Jr, F. J. (1951). « The Kolmogorov-Smirnov test for goodness of fit ». In : *Journal of the American statistical Association* 46.253, p. 68-78 (cf. p. 53).

Matsumoto, M. et Nishimura, T. (1998). « Mersenne twister : a 623-dimensionally equidistributed uniform pseudo-random number generator ». In : *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 8.1, p. 3-30 (cf. p. 22).

McGill, W. J. (1954). « Multivariate information transmission ». In : *Psychometrika* 19.2, p. 97-116 (cf. p. 38).

Mittelhammer, R. C. (2013). *Mathematical Statistics for Economics and Business*. Springer Science & Business Media (cf. p. 49, 52).

Miyaguchi, K. (2017). « Normalized Maximum Likelihood with Luckiness for Multivariate Normal Distributions ». In : *arXiv preprint arXiv :1708.01861* (cf. p. 162).

Moral, S., Rumií, R. et Salmerón, A. (2001). « Mixtures of truncated exponentials in hybrid Bayesian networks ». In : *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer, p. 156-167 (cf. p. 2, 124).

Neapolitan, R. E. (2004). *Learning bayesian networks*. T. 38. Pearson Prentice Hall Upper Saddle River, NJ (cf. p. 1, 3, 75).

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media (cf. p. 2, 98, 102, 104, 108, 116).

Ouvrard, J.-Y. (2004). « Probabilités : Tome II ». In : *Master-Agrégation, Cassini* (cf. p. 11).

Parzen, E. (1962). « On estimation of a probability density function and mode ». In : *The annals of mathematical statistics* 33.3, p. 1065-1076 (cf. p. 127).

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan kaufmann (cf. p. 1).

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Elsevier (cf. p. 5, 70, 75).

Pearl, J. et Paz, A. (1985). *Graphoids : A graph-based logic for reasoning about relevance relations*. University of California (Los Angeles). Computer Science Department (cf. p. 68).

Pfeifer, D., Strassburger, D. et Philipps, J. (2020). « Modelling and simulation of dependence structures in nonlife insurance with Bernstein copulas ». In : *arXiv preprint arXiv :2010.15709* (cf. p. 114, 119).

Rao, C. (2002). « Karl Pearson chi-square test the dawn of statistical inference ». In : *Goodness-of-fit tests and model validity*. Springer, p. 9-24 (cf. p. 52).

Reynolds, D. A. (2009). « Gaussian mixture models. » In : *Encyclopedia of biometrics* 741, p. 659-663 (cf. p. 124).

Rissanen, J. (1983). « A universal prior for integers and estimation by minimum description length ». In : *The Annals of statistics*, p. 416-431 (cf. p. 90).

Rissanen, J. J. (1996). « Fisher information and stochastic complexity ». In : *IEEE transactions on information theory* 42.1, p. 40-47 (cf. p. 90, 91).

Robert, C. (2007). *The Bayesian choice : from decision-theoretic foundations to computational implementation*. Springer Science & Business Media (cf. p. 45, 46, 49, 58).

Robinson, R. W. (1977). « Counting unlabeled acyclic digraphs ». In : *Combinatorial mathematics V*. Springer, p. 28-43 (cf. p. 67).

Rodriguez, J. C. (2007). « Measuring financial contagion : A copula approach ». In : *Journal of empirical finance* 14.3, p. 401-423 (cf. p. 103).

Romero, V., Rumí, R. et Salmerón, A. (2006). « Learning hybrid Bayesian networks using mixtures of truncated exponentials ». In : *International Journal of Approximate Reasoning* 42.1-2, p. 54-68 (cf. p. 2, 124).

Roos, T., Silander, T., Kontkanen, P. et Myllymaki, P. (2008). « Bayesian network structure learning using factorized NML universal models ». In : *2008 Information Theory and Applications Workshop*. IEEE, p. 272-276 (cf. p. 90, 163).

Rosenthal, J. S. (2006). *First Look At Rigorous Probability Theory, A*. World Scientific Publishing Company (cf. p. 11).

Rousseeuw, P. J. et Molenberghs, G. (1993). « Transformation of non positive semidefinite correlation matrices ». In : *Communications in Statistics–Theory and Methods* 22.4, p. 965-984 (cf. p. 127).

Salmon, W. C. (2017). *The foundations of scientific inference*. University of Pittsburgh Press (cf. p. 42).

Salvadori, G. et De Michele, C. (2007). « On the use of copulas in hydrology : theory and practice ». In : *Journal of Hydrologic Engineering* 12.4, p. 369-380 (cf. p. 2).

Sancetta, A. et Satchell, S. (2004). « The Bernstein copula and its applications to modeling and approximations of multivariate distributions ». In : *Econometric theory* 20.3, p. 535-562 (cf. p. 3, 111, 113, 116).

Sason, I. (2018). « On f-divergences : Integral representations, local behavior, and inequalities ». In : *Entropy* 20.5, p. 383 (cf. p. 143).

Scarsini, M. (1984). « On measures of concordance. » In : *Stochastica* 8.3, p. 201-218 (cf. p. 103).

Schervish, M. J. (2012). *Theory of statistics*. Springer Science & Business Media (cf. p. 28).

Schwarz, G. (1978). « Estimating the dimension of a model ». In : *The annals of statistics* 6.2, p. 461-464 (cf. p. 57).

Schweizer, B. et Sklar, A. (1974). « Operations on distribution functions not derivable from operations on random variables ». eng. In : *Studia Mathematica* 52.1, p. 43-52 (cf. p. 100).

Segers, J., Sibuya, M. et Tsukahara, H. (2017). « The empirical beta copula ». In : *Journal of Multivariate Analysis* 155, p. 35-51 (cf. p. 111, 115, 118).

Shachter, R. D. et Kenley, C. R. (1989). « Gaussian influence diagrams ». In : *Management science* 35.5, p. 527-550 (cf. p. 72).

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton university press (cf. p. 1).

Shenoy, P. P. (1997). « Binary join trees for computing marginals in the Shenoy-Shafer architecture ». In : *International Journal of approximate reasoning* 17.2-3, p. 239-263 (cf. p. 2, 164).

Shenoy, P. P. et West, J. C. (2011). « Inference in hybrid Bayesian networks using mixtures of polynomials ». In : *International Journal of Approximate Reasoning* 52.5, p. 641-657 (cf. p. 2, 124).

Sklar, A. (1959). « Fonctions de répartition à n dimensions et leurs marges ». In : *Publ. Inst. Statist. Univ. Paris* 8, p. 229-231 (cf. p. 2).

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G. et Richardson, T. (2000). *Causation, prediction, and search*. MIT press (cf. p. 5, 85, 86, 161).

Su, L. et White, H. (2008a). « A Nonparametric Hellinger Metric Test for Conditional Independence ». In : *Econometric Theory* 24.4, p. 829-864 (cf. p. 123).

Su, L. et White, H. (2008b). « A nonparametric Hellinger metric test for conditional independence ». In : *Econometric Theory*, p. 829-864 (cf. p. 128).

Te Sun, H. (1980). « Multiple mutual informations and multiple interactions in frequency data ». In : *Info. Control* 46.26-45, p. 4 (cf. p. 38).

Trösser, F., Givry, S. de et Katsirelos, G. (2021). « Improved Acyclicity Reasoning for Bayesian Network Structure Learning with Constraint Programming ». In : *arXiv preprint arXiv :2106.12269* (cf. p. 80).

Verma, T. et Pearl, J. (1988). *Influence diagrams and d-separation*. UCLA, Computer Science Department (cf. p. 68).

Verma, T. et Pearl, J. (1990). *Equivalence and synthesis of causal models*. UCLA, Computer Science Department (cf. p. 72).

Verny, L., Sella, N., Affeldt, S., Singh, P. P. et Isambert, H. (2017). « Learning causal networks with latent variables from multivariate information in genomic data ». In : *PLoS computational biology* 13.10, e1005662 (cf. p. 93, 147).

Walley, P. (1990). « Statistical Reasoning with Imprecise Probabilities ». In : (cf. p. 1).

Wan, J. et Zabaras, N. (2014). « A probabilistic graphical model based stochastic input model construction ». In : *J. Comput. Physics* 272, p. 664-685 (cf. p. 123, 124).

Williams, D. (1991). *Probability with martingales*. Cambridge university press (cf. p. 11).

Zadeh, L. A. (1996). « Fuzzy sets ». In : *Fuzzy sets, fuzzy logic, and fuzzy systems : selected papers by Lotfi A Zadeh*. World Scientific, p. 394-432 (cf. p. 1).

Résumé : 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. Cependant, les distributions de grandes dimensions sont difficiles à manipuler et peuvent conduire à des calculs coûteux en temps et en ressources.
Les réseaux bayésiens de copules (CBNs) tirent parti à la fois des réseaux bayésiens (BNs) et de la théorie des copules pour représenter de manière compacte de telles distributions multivariées. Les réseaux bayésiens s'appuient sur les indépendances conditionnelles afin de réduire la complexité du problème, tandis que les fonctions copules permettent de modéliser les relations de dépendance entre les variables aléatoires.
L'objectif de cette thèse est de donner un cadre commun aux deux domaines et de proposer de nouveaux algorithmes d'apprentissage pour les réseaux bayésiens de copules. Pour ce faire, nous utilisons le fait que les CBNs possèdent le même langage graphique que les BNs ce qui nous permet d'adapter leurs méthodes d'apprentissage à ce modèle. De plus, en utilisant la copule empirique de Bernstein à la fois pour concevoir des tests d'indépendance conditionnelle et pour estimer les copules, nous évitons de faire des hypothèses paramétriques, ce qui donne une plus grande généralité à nos méthodes.

Mots-clés : *Réseaux bayésiens, théorie des copules, apprentissage non-paramétrique*

Abstract : Modeling multivariate continuous distributions is a task of central interest in statistics and machine learning with many applications in science and engineering. However, high-dimensional distributions are difficult to handle and can lead to intractable computations.
The Copula Bayesian Networks (CBNs) take advantage of both Bayesian networks (BNs) and copula theory to compactly represent such multivariate distributions. Bayesian networks rely on conditional independences in order to reduce the complexity of the problem, while copula functions allow to model the dependence relation between random variables.
The goal of this thesis is to give a common framework to both domains and to propose new learning algorithms for copula Bayesian networks. To do so, we use the fact that CBNs have the same graphical language as BNs which allows us to adapt their learning methods to this model. Moreover, using the empirical Bernstein copula both to design conditional independence tests and to estimate copulas from data, we avoid making parametric assumptions, which gives greater generality to our methods.

Keywords : *Bayesian networks, copula theory, non-parametric learning*