# Utilisation de copules paramétriques en présence de données observationnelles : cadre théorique et modélisations.

Charles Fontaine

# THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'Université de Montpellier

Spécialité : **Biostatistique**

Préparée au sein de l'unité de recherche EA 2415 :
Laboratoire de Biostatistique, d'Epidémiologie et de Santé Publique
Et de l'école doctorale n°166 : Information, Structures, Systèmes (I2S)

Présentée et soutenue publiquement par :

**Charles Fontaine**

Le 19 Septembre 2016

# Utilisation de copules paramétriques en présence de données observationnelles: cadre théorique et modélisations.

Directeur de thèse : Jean-Pierre Daurès
Co-directeur de thèse : Paul Landais

Jury composé de :

| | | | |
|---|---|---|---|
| Salim BOUZEBDA | Professeur | Université de technologie de Compiègne | Rapporteur |
| Gérard DURU | Professeur | Université Claude-Bernard-Lyon-1 | Rapporteur |
| Gary COLLINS | Professeur | University of Oxford | Examinateur |
| Yannick LE MANACH | PU–PH | McMaster University | Examinateur |
| Jean-Pierre DAURÈS | PU–PH | Université de Montpellier | Directeur |
| Paul LANDAIS | PU–PH | Université de Montpellier | Co-directeur |

Institut Universitaire de Recherche Clinique (IURC)
641, avenue de doyen Gaston Giraud
34093 Montpellier Cedex 5

*À ma chère tante Jeannette,*
*qui consacra sa vie*
*à transmettre les connaissances.*

*«Les statistiques sont aux [professionnels]*
*ce que les lampadaires sont aux ivrognes :*
*elles sont plus utiles pour*
*s'appuyer que pour s'éclairer.»*

*Jacques Parizeau (1930-2015)*
*Premier ministre du Québec*

# Remerciements

A priori, il est impératif que je remercie le Professeur Jean-Pierre Daurès. Je lui suis reconnaissant qu'il ait accepté de travailler avec moi pour quelques années, de me suivre dans mon projet de thèse sur l'univers des fonctions de répartitions multivariées et de m'orienter dans des sphères concrètes de la recherche médicale où l'application des copules permet d'innover. Je le remercie également pour les échanges constructifs desquels émanaient des idées ingénieuses. Par ailleurs, je dois remercier le Professeur Paul Landais pour ses judicieux conseils, les échanges toujours cordiaux puis la finesse et la qualité des corrections qu'il a pu effectuer sur mes divers travaux. Enfin, je tiens à remercier ces deux professeurs pour l'influence qu'ils ont eu sur moi à vouloir continuer mon cursus académique dans la sphère médicale et ainsi rafiner la qualité de mes futurs travaux de recherche.

Ensuite, je dois également remercier le Professeur Yannick Le Manach pour les idées de projets que nous avons partagées. Je le remercie aussi pour la qualité de son accueil lors de mes séjours à Hamilton et la convivialité qui règnait lorsque nous travaillions, ou simplement que nous buvions une bière ensemble ! Je remercie également le Professeur Taoufik Bouezmarni qui m'a fait découvrir lors de ma maîtrise en mathématiques, le domaine des copules, qui m'a donné l'occasion de débuter une thèse avec lui et qui a continué à échanger des idées statistiques avec moi malgré ces impératifs administratifs ayant justifié la fin de notre collaboration.

Par ailleurs, je remercie le Professeur Gary Collins pour la qualité des corrections effectuées sur notre article commun présenté au chapitre 3 de cette thèse, et d'avoir accepté d'être examinateur dans ce jury. De plus, je remercie sincèrement le Professeur Salim Bouzebda et le Professeur Gérard Duru d'avoir accepté de rapporter cette thèse. Je les remercie également pour les commentaires et les questions pertinentes, qui ont été constructives dans l'élaboration de la version finale de ce manuscrit. Je suis honoré que vous sègiez sur ce jury !

Cette thèse n'aurait pas aboutie sans l'existence d'amitiés de qualité au cours de ces dernières années. Je commence par remercier Paméla, pour ces taquineries et blagues que nous nous sommes faits et qui ont su mettre de l'ambiance au bureau ; et pour les encouragements à travailler ces quelques fins de semaine, pour bien avancer. J'ai bien aimé lui apprendre divers mots de la langue de Molière tels que les noms des nuages... Tu demeurera mon amie à qui j'ai le plus de plaisir à apprendre à patiner !

Je remercie Audrey pour sa bonne humeur contagieuse. Cette bonne humeur s'est transmet dès son arrivée au bureau et se prolonge dans l'expression de ses chefs-d'oeuvre pâtissiers ! Notre escapade en terre shakespearienne demeurera pour moi un bon souvenir ! J'espère que notre amitié continuera à grandir lorsque l'on suivra le D.U. de recherche clinique ensemble !

Je remercie Long pour sa collaboration sur certains projets et je remercie mon ami québécois Félix pour ces agréables discussions, souvent de nature mathématique, toujours autour d'une bière de qualité supérieure. Par ailleurs, je remercie tous les autres amis que j'aurais pu oublier et toutes ces jolies dames de Montpellier qui ont pu agrémenter ces années de thèse.

Enfin, mais pas la moindre, je remercie Nathalie pour les moments de rigolade que nous avons eus pendant ces pauses au bureau. Elle est, à l'instar de moi, une personne taquine en qui identifié une amitié qui, j'ose l'espérer, durera des lustres. Nathalie, lorsque je saurai danser, je t'inviterai pour un tango ou toute autre danse compliquée, mais d'ici là, on doit aller courir un marathon !

Pour terminer, je dois remercier profondément ma famille : ma mère, ma soeur et Michèle pour leurs encouragements sincères dans mes projets d'études, et mon père pour son soutien financier et pour avoir cru en moi malgré le caractère «non-linéaire» de mon cheminement académique.

# Table des matières

# Avant-propos

Depuis la publication des découvertes de l'illustre Ronald Fischer quant à la convergence de données de masse vers une loi spécifique, le domaine de la recherche clinique a été transformé d'une série de protocoles expérimentaux émanants d'hypothèses d'origines incongrues en une science robuste se basant sur les outils que peuvent lui apporter les statistiques inférentielles et, plus particulièrement, la biostatistique. Ainsi, tout protocole expérimental se base sur l'expérience (i.e. données de masse provenant des études antérieures) et on en tire des conclusions à partir de l'observation faite au niveau des patients. On qualifie donc ces données recueillies au niveau des patients de données observationnelles. Le fait de mesurer des variables au niveau d'individus qui ont des différences majeures malgré ce qui pourrait sembler une certaine homogénéité face à une maladie (e.g. niveaux de perception de la douleur différents, degrés d'assiduité aux thérapies différents, coûts de traitements différents) entrainent des particularités importantes sur ces dites données. Alors, il faut être en mesure d'élaborer des stratégies d'analyse statistique qui ne soient pas sensibles à de telles particularités.

On prend, pour commencer, l'exemple de données observationnelles émanant de données qualititatives et/ou dichotomiques. On pourrait penser à la mesure de l'efficacité d'un traitement (efficace=1, non-efficace=0) sachant que des covariables l'affectent (e.g. exposition ou non à un produit toxique ). Alors, étant donné que les variables ont pour support l'ensemble naturel positif $\mathbb{N}^+$, l'analyse de ces données, discrètes, ne peut pas se baser sur le théorème central limite et il faut sortir du cadre inférentiel statistique standard. Certes, il y a la régression logistique qui permet de traiter ce types de données ; mais l'imposition de la linéarité entre les prédicteurs mise en place par ce modèle est contraignante et, parfois, inappropriée.

Un autre exemple type de la particularité que peuvent présenter les données observationnelles se situe au niveau des données comparatives entre deux bras thérapeutiques et, en particulier lorsque ces derniers ne sont pas équilibrés. Bien sûr, il est possible d'ajuster les deux bras par diverses techniques d'analyse du score de propension (e.g. appariement, discrimination, etc.), mais il arrive que la démarche expérimentale présente encore des complications spécifiques à ces données une fois que l'équilibre entre les bras est effectué. On prend l'exemple de l'analyse des coûts et de l'efficacité d'une thérapie. En se basant sur l'approche de Willan (2002), il est possible d'obtenir une modélisation explicite des quantités d'intérêt (ICER et INB) mais cette approche impose ici aussi la linéarité de la régression aux covariables d'intérêt. Par ailleurs, étant donné que les temps de survie (et particulièrement lorsqu'ils sont ajustés à la qualité de vie des patients)

ont une distribution qui tend à être asymétrique, un modèle basé sur la régression linéaire est inadéquat.

Le dernier exemple que l'on considère ici pour illustrer la nature singulière des données expérimentales est la présence sur les variables étudiées du phénomène de censure. Nonobstant le type de censure présente sur les variables étudiées, il faut avant tout considérer la perte d'information que génère ce cas particulier du phénomène des données manquantes. On prend comme exemple le cas d'une relation linéaire simple entre deux variables, potentiellement toutes deux censurées. Alors, la droite interpolant cette relation est, au mieux, décalée vers le bas et, dans tous les cas, elle est biaisée. Donc, l'utilisation des méthodes classiques d'estimation de la régression en présence de censure à droite peut être une astuce adéquate si la relation entre les variables est réellement linéaire. Autrement, on se retrouve encore dans la situation de l'imposition d'un modèle supposant la linéarité à des données inadéquates.

L'objectif de ce travail de thèse est avant tout de proposer des procédures de modélisation théoriques pour certains champs d'utilisation des données observationnelles ; procédures qui doivent être flexibles face à la structure et à la dispersion des données constituant les variables d'intérêt. Ainsi, l'utilisation des copules paramétriques sera au centre de toutes les modélisation proposées étant donné qu'elles permettent d'établir une relation simple entre les fonctions de répartition marginales des variables d'intérêt à partir d'un (ou de plusieurs) paramètre(s) qui se déduit grâce à une relation fonctionnelle propre à chaque copule avec une mesure robuste entre les marges : la concordance entre ces dernières (évaluée, par exemple, par le tau de Kendall, le rho de Spearman, le bêta de Blomqvist ou le coefficient de Gini). Ainsi, les résultats de ce travail ont l'avantage de sortir du cadre paramétrique standard des statistiques appliquées aux données observationnelles, sans toutefois avoir l'inconvénient du cadre non-paramétrique des statistiques qui tend à introduire un surapprentissage au niveau des données.

Ce mémoire de thèse présente trois chapitres de contenu nouveau, chacun ayant donné lieu à au moins un article scientifique. Ces trois chapitres présentent une modélisation dans un contexte particulier concernant notamment les données observationnelles et on note qu'ils sont indépendants entre-eux. Ainsi, la notation utilisée est propre à chaque chapitre et un symbole particulier retrouvé dans un chapitre n'aura nécessairement pas la même signification dans le chapitre suivant.

L'organisation de ce mémoire est la suivante : au premier chapitre, on rappelle certains concepts de dépendance et la théorie générale des copules afin d'introduire le lecteur aux outils utilisés dans les chapitres suivants. Au deuxième chapitre, on présente une modélisation basée sur les copules dans le cas où l'on utilise les données observationnelles en tant que données médico-économiques avec un ajustement sur la qualité de vie. Spécifiquement, il s'agit d'une réécriture, basée sur la fonction de densité jointe en termes de copules, de l'analyse coût-efficacité. Au troisième chapitre, on présente le cas le plus fréquent quant aux données observationnelles : les données discrètes. Ainsi, on présente une méthodologie pour déterminer le score de propension en évitant l'utilisation de la régression logistique, mais uniquement en se basant sur l'utilisation

de la fonction copule liant la variable traitement aux covariables d'intérêt. Enfin, au quatrième chapitre, on s'intéresse au cas des données de survie en présence de données censurées. On y propose une régression basée sur une copule semi-paramétrique.

# Chapitre 1

# Miscellanées : fonction copule et mesures de dépendance

Ce chapitre se veut un bref rappel des principaux éléments de la théorie des copules et des concepts de dépendance qui y sont liés. Il s'agit principalement d'une prémisse à la lecture de cette thèse pour le néophyte du concept de la fonction copule. On y présente les principales caractéristiques de ce concept, on y détaille les propriétés fondamentales puis on donne une esquisse de certaines démonstrations aidant à comprendre le concept.

L'intérêt inhérent à l'étude des copules, leur application en statistiques descriptives et les concepts de dépendance qui y sont liés est un phénomène plutôt moderne en statistique, la preuve étant que ce terme est apparu pour la première fois dans l'*Encyclopedia of Statistical Sciences* en 1997.[1] Pourtant, des pans entiers de plusieurs sciences centrales telles que l'économie, la physique moderne et la recherche clinique tentent depuis longtemps d'établir des relations entre au moins deux variables qui sont bien souvent dépendantes. L'idée centrale est donc d'établir une loi conjointe entre plusieurs variables aléatoires afin d'ultimement utiliser cette loi dans l'analyse d'un phénomène d'intérêt.

Soient les variables aléatoires $X_1, ..., X_d$. En établissant l'hypothèse de l'indépendance entre les $d$ variables aléatoires, il est facile d'obtenir la loi conjointe via le produit des fonctions marginales, pour $d \geqslant 2$ :

$$\begin{aligned} \mathbb{P}(X_1 \leqslant x_1, ..., X_d \leqslant x_d) &= F(x_1, ..., x_d) \\ &= F_1(x_1) \times ... \times F_d(x_d). \end{aligned}$$

Cependant, une telle hypothèse d'indépendance est-elle toujours adéquate ? Peut-on, à titre d'exemple, affirmer que le coût cumulé d'un traitement pour un individu est indépendant de son temps de survie ? Ainsi, il faut nécessairement prendre en compte la dépendance entre ces variables, la force de cette dépendance et sa structure, puis l'intégrer à travers la modélisation de la loi désirée. C'est à cette fin que s'inscrit la prise en compte de la fonction copule. Effectivement, le terme *copula* a été introduit par Abe Sklar, mathématicien américain, en 1959 dans son article *Fonctions de répartition à n dimensions et leurs marges*,[2] et provient du latin *cōpŭlo* signifiant "lier ensemble, attacher".[3] Il s'agit donc d'un outil statistique qui non seulement structure une

forme et un niveau de dépendance entre deux ou plusieurs variable aléatoires, mais permet une réécriture non-biaisée de la fonction de répartition jointe nonobstant le fait que les variables soient indépendantes ou non.

C'est ainsi que le théorème de Sklar[2] (qui sera étudié plus loin) stipule que pour tout $d$-ensemble de fonctions de répartition $F_j(x_j) = \mathbb{P}(X_j \leqslant x_j)$, $j = 1, ..., d$, il existe au moins une fonction bijective $C : [0,1]^d \to [0,1]$ ayant des marges uniformes telle que, $\forall (x_1, ..., x_d) \in \mathbb{R}^d, d \geqslant 2$, on ait la fonction de distribution

$$F(x_1, ..., x_d) = C\left[F_1(x_1), ..., F_d(x_d)\right].$$

Par ailleurs, si les marges sont continues, alors $C$ est unique.

Afin de bien préparer le lecteur aux prochains chapitres sans toutefois lui allourdir la tâche en lui présentant des aspects théoriques pour lesquels il n'y a aucune mention dans l'ensemble de ce travail, ce chapitre est divisé comme suit : on commence par présenter les aspects propres à la fonction copule tout en se focalisant principalement sur les copules de nature paramétrique et en présentant quelques exemples de cet outil statistique. Puis, on présente les principaux aspects liés à la mesure de la dépendance entre deux fonctions de répartition marginales, soient une présentation des aspects théoriques liés au tau de Kendall et aux mesures d'association.

## 1.1 Fonctions copules

On définit ici la pierre angulaire de la théorie des fonctions de distribution multivariées : la fonction copule. On commence par introduire une notation de base avant définir la dite fonction et d'énoncer le théorème de Sklar.

On notera dans cette section le mot "copule" tant pour les fonctions multivariées discrètes ( nommées également sous-copules) que pour les fonctions multivariées continues. On rappelle que les marges de la copule sont uniformes. Alors, pour $V$, une variable aléatoire suivant une loi uniforme sur l'intervalle $[0,1]$,

$$\mathbb{P}\left(V \leqslant v\right) = \begin{cases} 0 & \text{si } u < 0; \\ u & \text{si } 0 \leqslant u \leqslant 1; \\ 1 & \text{si } 1 < u. \end{cases}$$

**Définition 1.1.1.** *Une copule, notée $C$, bivariée est une fonction dont le domaine est*

- *en cas de fonctions marginales continues :* $[0,1] \times [0,1]$ *;*
- *en cas de fonctions marginales discrètes :* $S_1 \times S_2$ *où* $S_1$ *et* $S_2$ *sont des sous-ensembles de* $[0,1]$ *contenant* $0$ *et* $1$*.*

De cette définition du domaine de la fonction copule, on peut définir la fonction copule de par ses propriétés principales.

**Définition 1.1.2.** *Une copule bivariée est une fonction $C$ avec les propriétés :*

— *$C$ est caractérisée «grounded», i.e.*

$$C(u,0) = C(0,v) = 0 \text{ pour tout } u, v \in [0,1];$$

— *$C$ est 2-croissante, i.e.*

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geqslant 0 \begin{cases} \forall u_1,u_2,v_1,v_2 \in [0,1] \text{ si les marges sont continues,} \\ \forall u_1,u_2 \in S_1, \forall v_1,v_2 \in S_2 \text{ sinon;} \end{cases}$$

— *Les marges sont uniformes, i.e.*

$$C(u,1) = u \text{ et } C(1,v) = v \begin{cases} \forall u,v \in [0,1] \text{ si les marges sont continues,} \\ \forall u \in S_1, \forall v \in S_2 \text{ sinon;} \end{cases}$$

En généralisant la 2-croissance des fonctions copules bivariées au cas multivarié, et grâce au lemme suivant, on est en mesure de montrer la régularité au sens de Lipschitz de la copule multivariée.

**Lemme 1.1.1.** *Soient $\alpha_i$ et $\beta_i$, $i \in \mathbb{N}^+$ des suites de nombres complexes de module inférieurs à 1 ($|\alpha_i| \leqslant 1, |\beta_i| \leqslant 1 \forall i \in \mathbb{N}^+$ ). Alors, $\forall n \geqslant 1$,*

$$\left| \prod_{i=1}^{n} \alpha_i - \prod_{i=1}^{n} \beta_i \right| \leqslant \sum_{i=1}^{n} |\alpha_i - \beta_i| .$$

**Théorème 1.1.1.** *Soit $C$ une copule multivariée de dimension $d \geqslant 2$. $C$ est lipschitzienne : pour tout $(u_1,...,u_d)$ et $(v_1,...,v_d)$ qui appartiennent au domaine de $C$, on a*

$$|C(u_1,...,u_d) - C(v_1,...,v_d)| \leqslant \sum_{i=1}^{d} |u_i - v_i| .$$

*Ainsi, $C$ est uniformément continue sur son domaine.*

**Démonstration :**

Soient les variables aléatoires suivant la loi uniforme sur l'intervalle $[0,1]$ : $\boldsymbol{U} = (U_1,...,U_d)$ ; et $C$ une copule de dimension $d \geqslant 2$ définie par $\mathbb{P}(U_1 \leqslant u_1,...,U_d \leqslant u_d)$. On a, lorsque $\boldsymbol{u} = (u_1,...,u_d), \boldsymbol{v} = (v_1,...,v_d) \in \mathbb{R}^d$,

$$\begin{aligned} |C(u_1,...,u_d) - C(v_1,...,v_d)| = \left|\mathbb{E}\left[\mathbb{1}_{(\boldsymbol{U} \leqslant \boldsymbol{u})}\right] - \mathbb{E}\left[\mathbb{1}_{(\boldsymbol{U} \leqslant \boldsymbol{v})}\right]\right| &\leqslant \mathbb{E}\left[\left|\prod_{i=1}^{n} \mathbb{1}_{(U_i \leqslant u_i)} - \prod_{i=1}^{n} \mathbb{1}_{(U_i \leqslant v_i)}\right|\right] \\ &\leqslant \mathbb{E}\left[\sum_{i=1}^{n} \left|\mathbb{1}_{(U_i \leqslant u_i)} - \mathbb{1}_{(U_i \leqslant v_i)}\right|\right] \\ &= \sum_{i=1}^{n} |u_i - v_i| \end{aligned}$$

La validité de ce développement réside en l'utilisation de l'inégalité de Jensen pour la première inégalité et du lemme précédent pour la seconde.

□

On peut ainsi qualifier la fonction copule de dimension $d$ de fonction de répartiton multivariée définie sur $[0,1]^d$ et dont les marges sont uniformes. Dans le cas multivarié, il est simple d'avoir l'intuition que le caractère «*grounded*» et 2-croissant de la copule implique qu'elle soit bornée de part et d'autre, pour $u$ et $v$ fixés. On qualifie de telles limites *bornes de Fréchet*. En se projetant au cas multivarié, on peut établir formellement l'existence des bornes de Fréchet telles que qualifées au théorème suivant.

**Théorème 1.1.2.** *Pour toute copule $C$ de dimension $d \geqslant 2$ et pour tout $\boldsymbol{u} = (u_1, ..., u_d) \in [0,1]^d$, la copule $C(\boldsymbol{u})$ est bornée telle que*

$$C^-(\boldsymbol{u}) \leqslant C(\boldsymbol{u}) \leqslant C^+(\boldsymbol{u})$$

*où les bornes de Fréchet sont telles que* $C^-(\boldsymbol{u}) = \max\left(0, \sum_{i=1}^{d} u_i - d + 1\right)$ *et que* $C^+ = \min_{\{1 \leqslant i \leqslant d\}} u_i$.

**Démonstration :**

Pour la partie gauche de cette inégalité, soit $\boldsymbol{U} = (U_1, ..., U_d)$ des variables aléatoires qui suivent une loi uniforme sur $[0,1]^d$, où $d \geqslant 2$, et $C$ la copule associée à la loi jointe de ces variables. Soit $\boldsymbol{u} = (u_1, ..., u_d) \in [0,1]^d$. On a donc :

$$\begin{aligned} C(\boldsymbol{u}) &= \mathbb{P}\left(\boldsymbol{U} \leqslant \boldsymbol{u}\right) \\ &= \mathbb{P}\left(U_1 \leqslant u_1, ..., U_d \leqslant u_d\right) \\ &= 1 - \mathbb{P}\left(\bigcup_{i=1}^{d} \{U_i > u_i\}\right) \\ &\geqslant 1 - \sum_{i=1}^{d} \mathbb{P}\left(U_i \geqslant u_i\right) \\ &= 1 - d + \sum_{i=1}^{d} u_i. \end{aligned}$$

Or, étant une fonction de répartition, $C$ est toujours supérieure ou égale à 0. Alors, $C(\boldsymbol{u}) \geqslant \max\left(0, \sum_{i=1}^{d} u_i - d + 1\right)$. La partie droite de cette inégalité se montre sous ce raisonnement :

$$C(\boldsymbol{u}) = \mathbb{P}(\boldsymbol{U} \leqslant u) \leqslant \mathbb{P}(U_i \leqslant u_i) = u_i.$$

Ainsi, $C(\boldsymbol{u}) \leqslant \min_{\{1 \leqslant i \leqslant d\}} u_i$.

□

### 1.1.1 Théorème de Sklar

Maintenant que l'on a défini une copule par ses attributs (e.g. propriété lipschitzienne et bornes), il est temps de présenter la copule en tant qu'une représentation de la fonction jointe de répartition de $d$ variables. Pour ce faire, on commence par introduire l'inversion de la fonction de répartition.

**Proposition 1.1.1.** *Soit $F$ une fonction de répartition définie sur $\mathbb{R}$, croissante et continue à droite. Soit $F^{-1}$ son inverse généralisé croissant et défini à gauche par $F^{-1}(t) = \inf\{x|F(x) = t\}$. Soit $U$ une variable aléatoire de loi uniforme sur $[0,1]$. Alors, pour tout $x \in \mathbb{R}$ et $s \in (0,1]$,*

$$F(x) \geqslant s \Leftrightarrow x \geqslant F^{-1}(s).$$

*Par ailleurs, $F$ est la fonction de répartition de la variable aléatoire $F^{-1}(U)$.*

**Démonstration :**

On prend un epsilon arbitraire dans l'intervalle ouvert $(0,s)$. Alors,$\forall \epsilon \in (0,s)$, on retrouve les équivalences

$$F^{-1}(s-\epsilon) \leqslant x \Leftrightarrow s - \epsilon \leqslant F(x) \Leftrightarrow s \leqslant F(x) \Leftrightarrow F^{-1}(s) \leqslant x.$$

De ces relations, on a donc $\forall x \in \mathbb{R}$,

$$\mathbb{P}\left(F^{-1}(U) \leqslant x\right) = \mathbb{P}\left(U \leqslant F(x)\right) = F(x).$$

□

Maintenant, on est en mesure d'énoncer le théorème de Sklar.

**Théorème 1.1.3.** ***[Théorème de Sklar] :***

*Soit $H$ la fonction de répartition jointe entre les variables $X_1, ..., X_d$ avec les distributions marginales $F_1, ..., F_d$ pour $d \geqslant 2$. Alors, il existe une copule $C$ telle que, pour tout $x_1, ..., x_d \in \mathbb{R}^d$,*

$$H(x_1, x_2, ..., x_d) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)).$$

*Si $F_1(x_1), F_2(x_2), ..., F_d(x_d)$ sont continues, alors $C$ est unique. Autrement, $C$ est non-unique sur $[0,1]^d$, mais est unique sur le support de ses marges, soit $Ran(F_1) \times Ran(F_2) \times ... \times Ran(F_d)$.*

En se concentrant sur la construction d'une copule via le corollaire suivant, il sera possible de démontrer le théorème de Sklar.

**Corollaire 1.** *Soit $H$ étant la fonction de répartition jointe de dimension $d$ entre les fonctions de distribution marginales $F_1, ..., F_d, d \geqslant 2$, et $C$ la copule unissant ces dernières. Par ailleurs, soit $F_1^{-1}, ..., F_d^{-1}$ étant l'inverse généralisé des fonctions de distribution marginales. Alors, pour $\boldsymbol{u} = (u_1, ..., u_d) \in [0,1]^d$ ,*

$$\begin{aligned} C(\boldsymbol{u}) &= C(F_1(x_1), ..., F_d(x_d)) \\ &= H(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)). \end{aligned}$$

**Démonstration :**

Soient $\boldsymbol{U} = (U_1, ..., U_d)$ des variables aléatoires suivant une loi uniforme sur l'intervalle $[0,1]$ et $C$ une copule de dimension $d \geqslant 2$ telle qu'elle soit la fonction de répartition de ces variables. Soit $F_i, i \in \{1, ..., d\}$ une fonction de répartition prenant valeur dans $\mathbb{R}$. En posant

$\boldsymbol{X} = \left(F_1^{-1}(U_1), ..., F_d^{-1}(U_d)\right)$ pour $\boldsymbol{x} = (x_1, ..., x_d) \in \mathbb{R}^d$, on écrit la fonction de répartition $H$ de $X$ telle que

$$\begin{aligned} H(\boldsymbol{x}) &= H(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)) \\ &= \mathbb{P}\left(F_1^{-1}(U_1) \leqslant x_1, ..., F_d^{-1}(U_d) \leqslant x_d\right) \\ &= \mathbb{P}\left(U_1 \leqslant F_1(x_1), ..., F_d(x_d)\right) \\ &= C\left(F_1(x_1), ..., F_d(x_d)\right) \\ &= C(\boldsymbol{u}). \end{aligned}$$

□

**Démonstration du théorème 1.1.3 :**

Le lien entre la fonction de répartition jointe et la fonction copule se prouve via l'application directe du corollaire 1. Pour prouver l'unicité de la fonction copule en cas de marges continues, on pose $V$ une variable aléatoire suivant une loi uniforme sur l'intervalle unitaire et indépendante de $X_i, i = 1, ..., d$. On pose, par ailleurs, $U_i = F_i(X_i), i = 1, ..., d$ si $F_i$ est continue. Comme les variables $X_i$ et $F_i^{-1}(U_i)$ ont la même fonction de répartition, elles ont la même loi et conséquemment, $F(X_i)$ et $F(F^{-1}(U_i))$ ont la même loi, et $F(F^{-1}(U_i))$ est donc une variable aléatoire valant $U_i$ pour $X_i$ finie presque sûrement. Ainsi, $U_i$ est de loi uniforme sur $[0, 1]$ et, $\forall x_i \in \mathbb{R}$, on a presque sûrement l'égalité $\{C_i \leqslant x_i\} = \{U_i \leqslant F_i(x_i)\}$. Soit la copule $C$ qui est également la fonction de répartition de $\boldsymbol{U} = (U_1, ..., U_d)$. Alors,

$$\mathbb{P}\left(X_1 \leqslant x_1, ..., X_d \leqslant x_d\right) = \mathbb{P}\left(U_1 \leqslant F_1(x_1), ..., U_d \leqslant F_d(x_d)\right) = C(F_1(x_1, ..., F_d(x_d)).$$

Enfin, la continuité de la fonction copule et l'image contenant l'ouvert $(0, 1)$ assurent l'unicité de la copule en cas de fonctions marginales continues.

□

De ces démonstrations et en se basant sur le corollaire 1, on est en mesure de construire une copule. Ainsi, on prend l'exemple suivant pour illustrer la démarche associée.

**Exemple :**

On prend la distribution logistique bivariée d'Ali et al.[4] telle que la fonction de répartition jointe soit donnée par

$$H(x_1, x_2) = \left[1 + e^{-x_1} + e^{-x_2} + (1 - \theta)e^{-x_1 - x_2}\right]^{-1}$$

où $\theta$ représente un paramètre de dépendance. On voit que lorsque $\theta = 0$, il y a indépendance entre $X_1$ et $X_2$ ; puis que lorsque $\theta = 1$, on a la distribution logistique bivariée de Gumbel.[5] En intégrant la fonction jointe de répartition par rapport, dans un cas, à $x_2$, et dans l'autre cas, à $x_1$, on a respectivement les distributions marginales telles que $F_1(x_1) = (1 + e^{-x_1})^{-1}$ et $F_2(x_2) = (1 + e^{-x_2})^{-1}$ et, leurs inverses sont de la forme $F^{-1}(u) = -\log\left(\frac{1}{u} - 1\right)$. Ainsi, du

corollaire 1, on obtient la copule d'Ali-Mikhail-Haq par la construction :

$$\begin{aligned} C(u,v) &= H(F_1^{-1}(u), F_2^{-1}(v)) \\ &= \left[1 + e^{\ln(\frac{1}{u}-1)} + e^{\ln(\frac{1}{v}-1)} + (1-\theta)e^{\ln(\frac{1}{u}-1)+\ln(\frac{1}{v}-1)}\right]^{-1} \\ &= \frac{uv}{1-\theta(1-u)(1-v)}. \end{aligned}$$

### 1.1.2 Exemples de copules paramétriques

On présente ici deux exemples de copules pour les deux principales familles (au sens large) de copules paramétriques, soit la famille des copules elliptiques pour laquelle on présente la copule gaussienne et celle de Student ; puis la famille des copules archimédiennes pour laquelle on présente la copule de Clayton et celle de Gumbel.

On qualifie de distribution elliptique toute loi jointe du couple $\boldsymbol{X} = (X_1, X_2)$ pour laquelle il est possible de réécrire ce couple par la décomposition[6] $\mathbf{X} = \mu + \mathbf{R\Sigma^{1/2}U}$ où $\mu$ représente la moyenne, R une variable aléatoire à valeur positive et indépendante de $U$, et $U$ une variable aléatoire uniformément distribuée sur le disque contenu sur $\mathbb{R}^2$ ayant pour rayon l'intervalle $[0, 1]$. Alors, on peut représenter la fonction de densité jointe de tels couples par des courbes de niveaux prenant généralement la forme d'une ellipse.

Pour ce qui en est de la notion de famille de copules archimédiennes, elle a été introduite par Genest et Mackay[7] et son principe est que pour une copule de générateur $\phi$, la transformation $\omega(\phi)$ appliquée aux marges crée l'indépendance entre les composantes. On remarque que chaque copule a son propre générateur.

**Définition 1.1.3.** *Soit $\phi$ le générateur d'une copule archimédienne tel qu'il soit une fonction de classe $C^2$ de sorte que $\phi(1) = 0, \phi'(u) \leqslant 0$ et $\phi''(u) > 0$ . On qualifie une copule d'archimédienne si, pour $d \geqslant 2$, si $\sum_{i=1}^{d} \phi(u_i) \leqslant \phi(0)$,*

$$C(u_1, ..., u_d) = \prod_{i=1}^{d} \phi^{-1}\left(\phi(u_i) + ... + \phi(u_d)\right)$$

*et si $\sum_{i=1}^{d} \phi(u_i) > \phi(0)$, $C(u_1, ..., u_d) = 0$.*

Pour concrétiser l'idée d'une fonction nommée *générateur*, on prend l'exemple de la copule de Frank pour laquelle le générateur est $-\ln\left[\frac{e^{-\theta u}-1}{e^{-\theta}-1}\right]$ où $\theta > 0$ est le paramètre de dépendance de cette copule.

#### 1.1.2.1 Copule gaussienne

Le premier type de copules paramétriques que l'on présente ici est la copule gaussienne (ou normale). Par son nom, on remarque bien qu'elle est issue de la distribution normale multivariée. Ainsi, il s'agit d'une copule non-adaptée aux valeurs extrêmes étant donné qu'elle ne présente

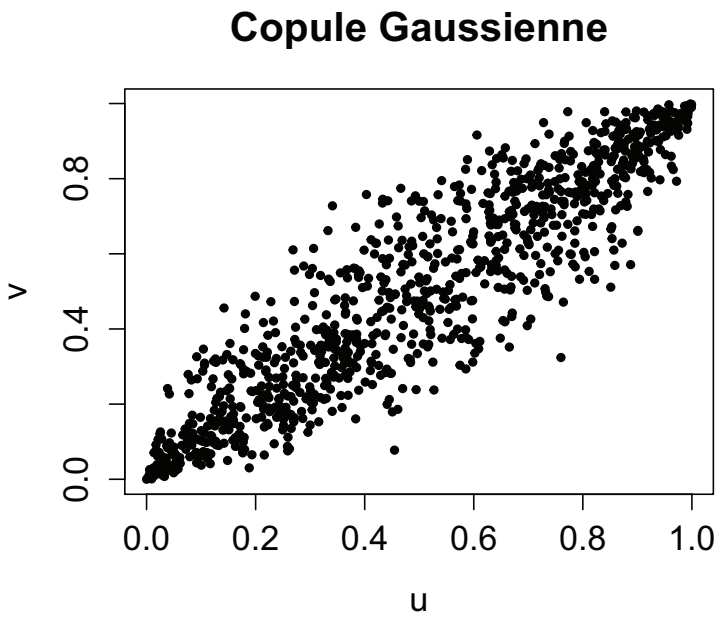


FIGURE 1.1 – Exemple de dispersion de 2000 observations issues d'une copule gaussienne avec une force de dépendance liée à un tau de Kendall de 0,5.

pas de dépendance de queue. On note que la copule gaussienne a pour coefficient de dépendance le coefficient de corrélation linéaire standard, soit le $\rho$ de Pearson.

**Définition 1.1.4.** *Soient $\rho \in [-1,1]$ le coefficient de corrélation de Pearson, $\Phi^{-1}$ l'inverse d'une fonction de répartition gaussienne centrée réduite et* $\Phi_\rho(u,v) = \int_{-\infty}^{v}\int_{-\infty}^{u} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{\frac{-(s^2+t^2-2\rho st)}{2(1-\rho^2)}\right\} dsdt$*, la loi multinormale de dimension 2. Alors, pour $(u,v) \in [0,1]^2$,*

$$C(u,v) = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right).$$

En dérivant la copule gaussienne, on obtient sa densité de copule telle que

$$\begin{aligned} c(u,v) &= \frac{\partial^2}{\partial u \partial v} C(u,v) \\ &= \frac{1}{\sqrt{1-\rho^2}} \exp\left\{\frac{2\rho\Phi^{-1}(u)\Phi^{-1}(v) - \rho^2(\Phi^{-1}(u)^2 + \Phi^{-1}(v)^2)}{2(1-\rho^2)}\right\}. \end{aligned}$$

À la figure 1.1, on aperçoit la forme de la dépendance entre les marges inhérente à une copule gaussienne pour un paramètre de dépendance moyen ($\rho \approx 0.92$). On y remarque en effet l'absence d'une dépendance notoire entre les extrêmes des deux distributions, et la symétrie caractéristique de la loi normale dans le nuage de dispersion des couples $(u,v)$.

#### 1.1.2.2 Copule de Student

La copule de Student est issue, comme son som l'indique, de la distribution multivariée de Student. Sa construction est dans le continuum de la copule gaussienne mais, contrairement à cette dernière, elle réussit à bien capter les dépendances aux extrêmes, tant positives que négatives, de par ses queues lourdes.

**Définition 1.1.5.** *Soient $\rho \in [-1,1]$ le coefficient de corrélation de Pearson, $T^{-1}$ l'inverse d'une fonction de répartition de Student centrée réduire univariée,*
$T_{\rho,\kappa}(u,v) = \int_{-\infty}^{v}\int_{-\infty}^{u} \frac{1}{2\pi\sqrt{1-\rho^2}}\left[1+\frac{s^2+t^2-2\rho st}{\kappa(1-\rho^2)}\right]^{-(\kappa+2)/2} dsdt$, *la distribution de la loi de Student de dimension 2 où $\kappa \geqslant 0$ représente le nombre de degrés de liberté. Alors, pour $(u,v) \in [0,1]^2$,*

$$C(u,v) = T_{\rho,\kappa}\left(T^{-1}(u), T^{-1}(v)\right).$$

Pour obtenir la densité de la copule de Student, il suffit d'utiliser la définition de la densité de copule :

$$\begin{aligned} c(u,v) &= \frac{\partial^2}{\partial u \partial v} T_{\rho,\kappa}\left(T^{-1}(u), T^{-1}(v)\right) \\ &= \frac{f(u,v)}{f(u)\times f(v)} \\ &= \frac{\kappa}{2\sqrt{1-\rho^2}} \frac{\Gamma(\kappa/2)^2}{\Gamma((\kappa+1)/2)^2} \frac{\left[1+\frac{u^2+v^2-2\rho uv}{\kappa(1-\rho^2)}\right]^{-(\kappa+2)/2}}{\left[\left(1+\frac{u^2}{\kappa}\right)\left(1+\frac{v^2}{\kappa}\right)\right]^{-(\kappa+2)/2}} \end{aligned}$$

où $\Gamma$ représente la fonction gamma, $f(u,v)$ la densité jointe d'une loi de Student et $f(u), f(v)$ les densités marginales.

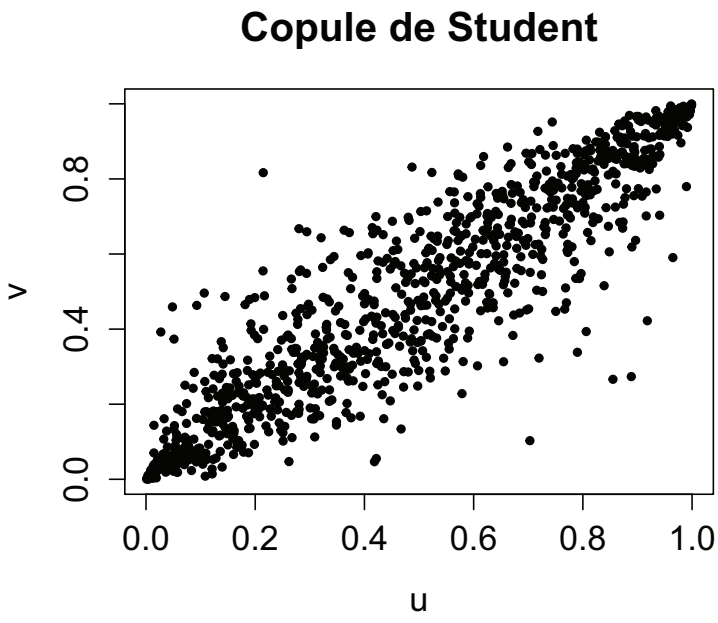


FIGURE 1.2 – Exemple de dispersion de 2000 observations issues d'une copule de Student avec une force de dépendance liée à un tau de Kendall de 0,5.

À la figure 1.2, on peut voir que la loi jointe des observations au sens de la copule de Student est une loi aux queues de dépendance lourdes et modélisant une probabilité d'événements extrêmes joints supérieure à celle modélisée par la copule gaussienne.

#### 1.1.2.3 Copule de Clayton

La copule de Clayton modélise la relation entre deux fonctions de répartition avec une très forte dépendance à la partie inférieure de la distribution, et un nuage de points qui tend à croître lorsque l'on se déplace vers l'extrémité supérieur. Ainsi, il s'agit d'une copule modélisant une relation totalement asymétrique.

**Définition 1.1.6.** *On définit la copule de Clayton par la fonction, pour un paramètre de dépendance* $\theta \in [-1, \infty) \backslash \{0\}$*,*

$$C(u_1, ..., u_d) = \left(u_1^{-\theta} + ... + u_d^{-\theta} - d + 1\right)^{-1/\theta}$$

*où* $d \geqslant 2$*. On note que le générateur de cette copule est* $\phi(z) = z^{-\theta} - 1$*.*

Ainsi, dans le cas bivarié, la copule de Clayton a la forme $C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$ et, on en déduit sa densité de copule bivariée

$$\begin{aligned} c(u, v) &= \frac{\partial^2}{\partial u \partial v} C(u, v) \\ &= (\theta + 1)(uv)^{-(\theta+1)}(u^{-\theta} + v^{-\theta} - 1)^{-\left(\frac{2\theta+1}{\theta}\right)}. \end{aligned}$$

On remarque que pour toute copule archimédienne, tel que montré par Savu et Trede,[8] la densité de copule est fonction du générateur de cette dernière. Ainsi, pour toute copule archimédienne, la densité

$$c(u_1, ..., u_d) = (\phi^{-1})^{(d)}(\phi(u_1), ..., \phi(u_d)) \prod_{i=1}^{d} \phi'(u_i)$$

où $\phi^{-1}$ est l'inverse du générateur, $\phi'$ la dérivée du gérénateur et $(\phi^{-1})^{(d)}$ la d-ième dérivée du générateur, est valide.

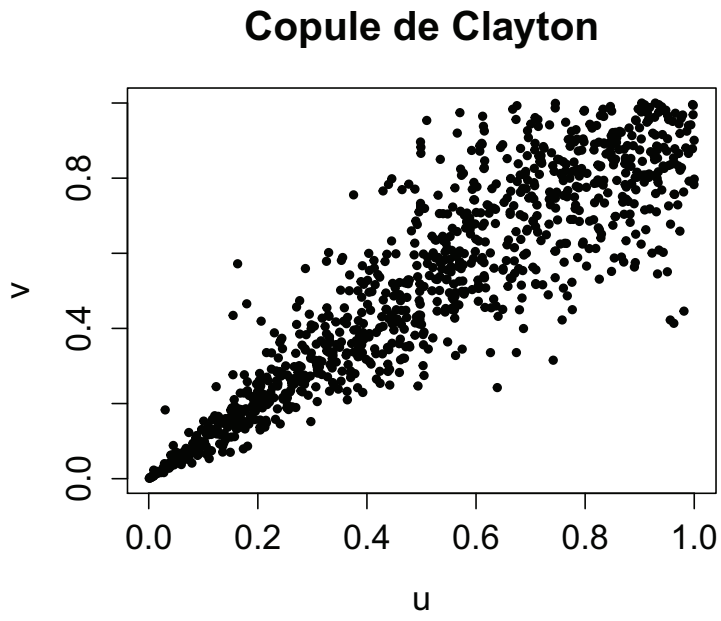


FIGURE 1.3 – Exemple de dispersion de 2000 observations issues d'une copule de Clayton avec une force de dépendance liée à un tau de Kendall de 0,5.

À la figure 1.3, une relation de dépendance moyenne est représentée ($\theta = 6$). On y voit clairement l'asymétrie dans la relation de dépendance entre les variables aléatoires.

#### 1.1.2.4 Copule de Gumbel

La copule de Gumbel modélise une dépendance de l'extrémité supérieure de la distribution. On note son asymétrie provoquée simplement par une plus grande dépendance au niveau de l'extrémité supérieure de la distribution versus une dépendance presque nulle à la partie inférieure.

**Définition 1.1.7.** *On définit la copule de Gumbel par la fonction, pour un paramètre de dépendance $\theta \geqslant 1$,*

$$C(u_1, ..., u_d) = \exp\left\{ -\left[ \sum_{i=1}^{d} (-\ln(u_i))^{\theta} \right]^{1/\theta} \right\}$$

*où $d \geqslant 2$ et pour laquelle le générateur est $\phi(z) = (-\ln(z))^{\theta}$.*

Donc, dans le cas bivarié, la copule de Gumbel a la forme $C(u,v) = \exp\{-\left[(-\ln(u))^\theta + (-\ln(v))\theta\right]^{1/\theta}\}$. Pour trouver la densité bivariée, il faut utiliser le générateur de cette copule. Ainsi,

$$\begin{aligned} c(u,v) &= (\phi^{-1})^{(d)}(\phi(u_1),...,\phi(u_d))\prod_{i=1}^{d}\phi'(u_i) \\ &= -\frac{\phi'(C(u,v))\phi'(u)\phi'(v)}{\phi'(C(u,v))^3} \\ &= C(u,v)\left[\phi(u)+\phi(v)\right]^{\frac{1}{\theta}-2}\left[\theta - 1 + (\phi(u)+\phi(v))^{1/\theta}\right]\frac{(-\ln(u))^{\theta-1}(-\ln(v))^{\theta-1}}{uv}. \end{aligned}$$

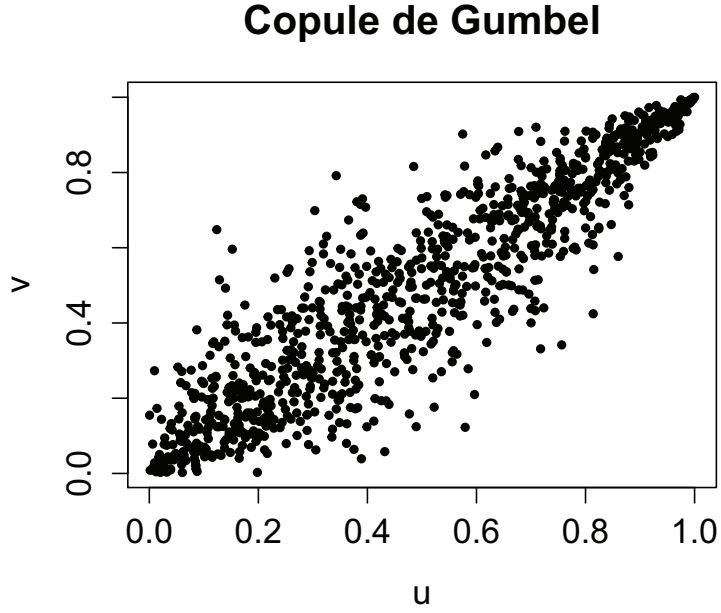


FIGURE 1.4 – Exemple de dispersion de 2000 observations issues d'une copule de Gumbel avec une force de dépendance liée à un tau de Kendall de 0,5.

À la figure 1.4, pour une force de dépendance standard ($\theta = 4$), on remarque la force de la dépendance modélisée par la queue supérieure de la distribution, contrairement à la dépendance à la partie inférieure ; ainsi que l'asymétrie de la copule. On remarque, par ailleurs, qu'il s'agit de la seule copule archimédienne vérifiant la propriété de max-stabilité, c'est-à-dire que $C(u_1^n,...,u_d^n) = C^n(u_1,...,u_d)$ pour $d \geqslant 2$ et $n \geqslant 1$.

## 1.2 Mesures de dépendance

L'idée d'introduire les principales mesures de dépendance ici, ainsi que les principales mesures d'association relève de l'estimation qui sera faite dans les chapitres ultérieurs quant à l'inférence du paramètre de dépendance d'une copule paramétrique. Effectivement, pour coupler deux ou

plusieurs fonctions de répartition multivariées qui sont composées de lois marginales identiquement distribuées, il faut mesurer la dépendance entre les marges et, cela se fait à partir de la mesure d'un ordre partiel entre les couples de données composant les observations.

Étant donné que les mesures de concordance se calculent sur des couples de données et que pour obtenir le paramètre de dépendance d'une copule de dimension $d \geqslant 3$, il est plus consistant et computationnellement plus simple d'utiliser une méthode basée sur la vraisemblance des données qu'une méthode basée sur l'inversion d'une mesure de concordance (voir le chapitre 3), on présente ici les mesures de dépendance pour des distributions bivariées uniquement.

**Définition 1.2.1.** *Soient $(x_i, y_i)$ et $(x_j, y_j)$, deux couples d'observations d'un vecteur de variables aléatoires $(X, Y)$. Alors,*

- *si $(x_i - x_j)(y_i - y_j) > 0$, les couples $(x_i, y_i)$ et $(x_j, y_j)$ sont dits concordants ;*
- *si $(x_i - x_j)(y_i - y_j) < 0$, les couples $(x_i, y_i)$ et $(x_j, y_j)$ sont dits discordants.*

On note qu'il est également possible d'établir une comparaison de la concordance entre deux fonctions de répartitions bivariées de mêmes marges sans calculer de relation d'ordre sur les couples d'observations.

**Définition 1.2.2.** *Soient $G$ et $H$, deux fonctions de répartition bivariées appartenant à la même classe de fonctions de répartitions bivariées de marges $F_1$ et $F_2$. alors $G$ a une plus grande concordance que $F$ si, $\forall x_1, x_2 \in \mathbb{R}$,*

$$F(x_1, x_2) \leqslant G(x_1, x_2).$$

### 1.2.1 Tau de Kendall

Étant la mesure de dépendance la plus largement utilisée dans la littérature sur l'estimation de copules paramétriques, le tau de Kendall[9] a l'avantage, a contrario au coefficient de corrélation de Pearson, de ne pas reposer sur une hypothèse de linéarité entre les variables en étant plutôt une mesure d'association.

**Définition 1.2.3.** *Soient $(X_1, Y_1)$ et $(X_2, Y_2)$ deux vecteurs aléatoires indépendants et identiquements distribués de même fonction de répartition jointe $H$, qui sont en fait deux copies indépendantes des variables aléatoires continues $X$ et $Y$. Alors, on définit le tau de Kendall ($\tau$) comme étant la probabilité de concordance moins la probabilité de discordance de ces vecteurs aléatoires tel que*

$$\tau = \tau_{(X,Y)} = \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right].$$

Pour démontrer le rôle que joue la fonction copule dans la mesure d'association qu'est le tau de Kendall, on montre que ce dernier dépend la fonction de répartition jointe $H_{X,Y}$ seulement via sa copule.

**Théorème 1.2.1.** *Soient $\tau$ le tau de Kendall tel que présenté à la définition 1.2.3 et $(X_1, Y_1)$, $(X_2, Y_2)$ deux copies indépendantes de $(X, Y)$, des variables aléatoires continues, ayant pour fonction de répartition jointe $H$, et des fonctions de répartition marginales communes $F$ (pour $X_1$ et $X_2$) et $G$ (pour $Y_1$ et $Y_2$). Soit $C$, la copule associée à la fonction de répartition jointe $H(x, y) = C(F(x), G(y))$. Alors,*

$$\tau_{X,Y} = 4 \int\int_{\{[0,1]\times[0,1]\}} C(u, v) dC(u, v) - 1.$$

**Démonstration :**

Étant donné que les variables aléatoires sont continues, on peut émettre l'égalité

$$\begin{aligned}
\tau_{X,Y} &= \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right] \\
&= \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - (1 - \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right]) \\
&= 2\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - 1.
\end{aligned}$$

Toutefois, il est possible de réécrire cette probabilité de condordance telle que

$$\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] = \mathbb{P}(X_1 > X_2, Y_1 > Y_2) + \mathbb{P}(X_1 < X_2, Y_1 < Y_2).$$

On commence par évaluer la probabilité $\mathbb{P}(X_1 < X_2, Y_1 < Y_2)$ par sa définition :

$$\begin{aligned}
\mathbb{P}(X_1 < X_2, Y_1 < Y_2) &= \int\int_{\mathbb{R}^2} \mathbb{P}(X_1 < x, Y_1 < y) dH(x, y) \\
&= \int\int_{\mathbb{R}^2} C(F(x), G(y)) dC(F(x), G(y)) \\
&= \int\int_{\{[0,1]\times[0,1]\}} C(u, v) dC(u, v)
\end{aligned}$$

en utilisant les transformations de probabilité $u = F(x)$ et $v = G(y)$. Similairement,

$$\begin{aligned}
\mathbb{P}(X_1 > X_2, Y_1 > Y_2) &= \int\int_{\mathbb{R}^2} \mathbb{P}(X_1 > x, Y_1 > y) dH(x, y) \\
&= \int\int_{\mathbb{R}^2} \left[1 - F(x) - G(y) + C(F(x), G(y))\right] dH(x, y) \\
&= \int\int_{\mathbb{R}^2} \left[1 - F(x) - G(y) + C(F(x), G(y))\right] dC(F(x), G(y)) \\
&= \int\int_{\{[0,1]\times[0,1]\}} \left[1 - u - v + C(u, v)\right] dC(u, v).
\end{aligned}$$

Par contre, compte tenu que $C$ est la distribution jointe du couple $(U, V)$ de variables aléatoires de loi uniforme définies sur $(0, 1)$, $\mathbb{E}[U] = \mathbb{E}[V] = 0,5$. Alors,

$$\mathbb{P}(X_1 > X_2, Y_1 > Y_2) = \int\int_{\{[0,1]\times[0,1]\}} C(u, v) dC(u, v).$$

C'est ainsi que l'on réécrit le tau de Kendall :

$$\begin{aligned}\tau_{X,Y} &= 2\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - 1 \\ &= 2\left[\int\int_{\{[0,1]\times[0,1]\}} C(u,v)dC(u,v) + \int\int_{\{[0,1]\times[0,1]\}} C(u,v)dC(u,v)\right] - 1 \\ &= 4\int\int_{\mathbb{R}^2} C(u,v)dC(u,v) - 1.\end{aligned}$$

□

### 1.2.2 Rho de Spearman

Similairement au tau de Kendall, le rho de Spearman, $\tilde{\rho}$, est une mesure d'association basée sur des mesures de concordance et de discordance. Là où il y a un changement par rapport au tau de Kendall est que $\tilde{\rho}$ prend en compte les vecteurs aléatoires indépendants et identiquement distribués, $(X_1, Y_1), (X_2, Y_2)$ et $(X_3, Y_3)$.

**Définition 1.2.4.** *Soient $(X_1, Y_1), (X_2, Y_2)$ et $(X_3, Y_3)$ trois copies indépendantes des variables aléatoires continues $X$ et $Y$, et ayant la même fonction de répartition jointe $H$ et les mêmes fonctions de répartition marginales $F$ et $G$. Alors, on définit le rho de Spearman ($\tilde{\rho}$) comme étant proportionnel à la probabilité de concordance moins la probabilité de discordance de ces vecteurs aléatoires tel que*

$$\tilde{\rho} = \tilde{\rho}_{(X,Y)} = 3\left(\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right] - \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) < 0\right]\right).$$

Il est également possible de démontrer que la mesure du rho de Spearman de deux variables aléatoires $X$ et $Y$ est intimement liée à la fonction de répartition jointe $H$ de ces variables.

**Théorème 1.2.2.** *Soient $\tilde{\rho}$ le rho de Spearman tel que présenté à la définition 1.2.4 et $(X_1, Y_1)$, $(X_2, Y_2)$, $(X_3, Y_3)$ trois copies indépendantes de $(X, Y)$, des variables aléatoires continues, ayant pour fonction de répartition jointe $H$, et des fonctions de répartition marginales communes $F$ et $G$ pour $X$ et $Y$ respectivement. Soit $C$, la copule associée à la fonction de répartition jointes $H(x, y) = C(F(x), G(y))$. Alors,*

$$\tilde{\rho} = \tilde{\rho}_{X,Y} = 12\int\int_{\{[0,1]\times[0,1]\}} C(u,v)dudv - 3.$$

**Démonstration :**

On commence par réécrire $\tilde{\rho}$ étant donné la continuité des variables aléatoires. Alors, on a l'égalité

$$\begin{aligned}\tilde{\rho}_{X,Y} &= 3\left(\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right] - \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) < 0\right]\right) \\ &= 3\left(\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right] - \left(1 - \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right]\right)\right) \\ &= 6\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right] - 3.\end{aligned}$$

On peut réécrire cette probabilité telle que

$$\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] = \mathbb{P}(X_1 > X_2, Y_1 > Y_2) + \mathbb{P}(X_1 < X_2, Y_1 < Y_2).$$

On note que l'on a la fonction de répartition jointe qui peut être réécrite en terme de sa copule $H_{X_1,Y_1}(x,y) = H_{X_2,Y_2}(x,y) = H_{X_3,Y_3}(x,y) = C(F(x),G(y))$. Par contre, étant donné l'indépendance entre les copies $(X_1,Y_1),(X_2,Y_2),(X_3,Y_3)$, on note $\Pi$ la copule équivalente à la fonction de répartition jointe entre les variables $X_2$ et $Y_3$ telle que $\bar{H}_{X_2,Y_3}(x,y) = \Pi(F(x),G(y))$. On remarque alors que $\Pi$ est la copule d'indépendance telle que $\Pi(u,v) = uv$. Ainsi, on peut évaluer $\mathbb{P}(X_1 < X_2, Y_1 < Y_2)$ de façon analogue à la démonstration du théorème 1.2.1 :

$$\begin{aligned}
\mathbb{P}(X_1 < X_2, Y_1 < Y_3) &= \int\int_{\mathbb{R}^2} \mathbb{P}(X_1 < x, Y_1 < y) dH_{X_2,Y_3}(x,y) \\
&= \int\int_{\mathbb{R}^2} C(F(x),G(y)) d\Pi(F(x),G(y)) \\
&= \int\int_{\{[0,1]\times[0,1]\}} C(u,v) d\Pi(u,v) \\
&= \int\int_{\{[0,1]\times[0,1]\}} C(u,v) dudv
\end{aligned}$$

en utilisant les transformations de probabilité $u = F(x)$ et $v = G(y)$. Par ailleurs,

$$\begin{aligned}
\mathbb{P}(X_1 > X_2, Y_1 > Y_2) &= \int\int_{\mathbb{R}^2} \mathbb{P}(X_1 > x, Y_1 > y) dH_{X_2,Y_3}(x,y) \\
&= \int\int_{\mathbb{R}^2} \left[1 - F(x) - G(y) + C(F(x),G(y))\right] dH_{X_2,Y_3}(x,y) \\
&= \int\int_{\mathbb{R}^2} \left[1 - F(x) - G(y) + C(F(x),G(y))\right] d\Pi(F(x),G(y)) \\
&= \int\int_{\{[0,1]\times[0,1]\}} \left[1 - u - v + C(u,v)\right] dudv \\
&= \int\int_{\{[0,1]\times[0,1]\}} C(u,v) dudv
\end{aligned}$$

où on obtient le passage à la dernière ligne en raison d'arguments similaires à ceux de la démonstration du théorème 1.2.1. C'est ainsi que l'on réécrit le rho de Spearman :

$$\begin{aligned}
\tilde{\rho}_{X,Y} &= 6\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right] - 3 \\
&= 6\left[\int\int_{\{[0,1]\times[0,1]\}} C(u,v) dudv + \int\int_{\{[0,1]\times[0,1]\}} C(u,v) dudv\right] - 3 \\
&= 12\int\int_{\mathbb{R}^2} C(u,v) dudv - 3.
\end{aligned}$$

□

Pour conclure ces rappels des deux principales mesures d'association utilisées dans l'inférence des copules paramétriques, on va établir une relation entre $\tau$ et $\tilde{\rho}$. Cette relation, établie à travers le théorème suivant, vient de Durbin et Stuart,[10] et la démonstration est adaptée de Kruskal[11] et de Nelsen.[12]

**Théorème 1.2.3.** *Soient $X$ et $Y$, deux variables aléatoires de lois continues, $\tau = \tau_{X,Y}$ et $\tilde{\rho} = \tilde{\rho}_{X,Y}$ respectivement les mesures d'association du tau de Kendall et du rho de Spearman*

*mesurées sur ces deux variables aléatoires. Alors,*

$$\frac{1+\tilde{\rho}}{2} \geqslant \left[\frac{1+\tau}{2}\right]^2$$

*et*

$$\frac{1-\tilde{\rho}}{2} \geqslant \left[\frac{1-\tau}{2}\right]^2.$$

**Démonstration :**

On commence par rappeler, tel qu'utilisé à la démonstration du théorème 1.2.1, que l'on peut réécrire $\tau$ tel que $\tau = 2\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - 1$ pour $(X_1, Y_1), (X_2, Y_2)$, deux copies indépendantes de $(X, Y)$. Cela dit, on a $(X_1, Y_1), (X_2, Y_2)$ et $(X_3, Y_3)$ trois copies indépendantes de $(X, Y)$ et ayant la même fonction de répartition $H$. Si on note $p$ la probabilité que deux de ces copies soient concordantes avec la troisième, on a

$$\begin{aligned}
p &= \mathbb{P}\left[(X_2, Y_2) \text{ et } (X_3, Y_3) \text{ sont concordants avec } (X_1, Y_1)\right] \\
&= \int\int_{\mathbb{R}^2} \mathbb{P}\left[(X_2, Y_2) \text{ et } (X_3, Y_3) \text{ sont concordants avec } (x, y)\right] dH(x, y) \\
&= \int\int_{\mathbb{R}^2} \mathbb{P}\left[(X_2 - x)(Y_2 - y) > 0\right] \mathbb{P}\left[(X_3 - x)(Y_3 - y) > 0\right] dH(x, y) \\
&= \int\int_{\mathbb{R}^2} \left(\mathbb{P}\left[(X_2 - x)(Y_2 - y) > 0\right]\right)^2 dH(x, y) \\
&\geqslant \left[\int\int_{\mathbb{R}^2} \left(\mathbb{P}\left[(X_2 - x)(Y_2 - y) > 0\right]\right) dH(x, y)\right]^2 \\
&= \left[\int\int_{\mathbb{R}^2} \left(\mathbb{P}\left[(X_2 > x)(Y_2 > y)\right] + \mathbb{P}\left[(X_2 < x)(Y_2 < y)\right]\right) dH(x, y)\right]^2 \\
&= \left[\mathbb{P}\left[(X_2 - X_1)(Y_2 - Y_1) > 0\right]\right]^2 \\
&= \left(\frac{1+\tau}{2}\right)^2
\end{aligned}$$

où la dernière égalité se justifie avec le rappel de la réécriture du tau de Kendall présenté au début de cette démonstration, et l'inégalité se justifie par $\mathbb{E}[Z^2] \geqslant (\mathbb{E}[Z])^2$ où $Z = \mathbb{P}\left[(X_2 - X_1)(Y_2 - Y_1) > 0 | (X_1, Y_1)\right]$. Soit $\tilde{\rho} = 6\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right] - 3$. En permutant les indices de $X$ et de $Y$, on obtient la forme symétrique de $\tilde{\rho}$ telle que

$$\begin{aligned}
\tilde{\rho} &= \{\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right] + \mathbb{P}\left[(X_1 - X_3)(Y_1 - Y_2) > 0\right] \\
&\quad + \mathbb{P}\left[(X_2 - X_1)(Y_2 - Y_3) > 0\right] + \mathbb{P}\left[(X_2 - X_3)(Y_2 - Y_1) > 0\right] \\
&\quad + \mathbb{P}\left[(X_3 - X_1)(Y_3 - Y_2) > 0\right] + \mathbb{P}\left[(X_3 - X_2)(Y_3 - Y_1) > 0\right]\} - 3.
\end{aligned}$$

Soit $p_{ijk} = \mathbb{P}\left[Y_i < Y_j < Y_k | X_1 < X_2 < X_3\right]$. Étant donné que $\tilde{\rho}$ ne varie pas lorsque l'on permute les indices des variables aléatoires $X$ et $Y$, on peut émettre l'hypothèse que $X_1 < X_2 < X_3$. Alors, $\tilde{\rho} = 2\mathbb{P}\left[Y_1 < Y_3\right] - 1. = 2(p_{123} + p_{132} + p_{213}) - 1$. Si on permute les indices de la définition de $p$,

on remarque alors

$$\begin{aligned} p &= \frac{1}{3}\{\mathbb{P}\left[(X_2, Y_2) \text{ et } (X_3, Y_3) \text{ sont concordants avec } (X_1, Y_1)\right] \\ &\quad +\mathbb{P}\left[(X_1, Y_1) \text{ et } (X_3, Y_3) \text{ sont concordants avec } (X_2, Y_2)\right] \\ &\quad +\mathbb{P}\left[(X_1, Y_1) \text{ et } (X_2, Y_2) \text{ sont concordants avec } (X_3, Y_3)\right]\} \\ &= \frac{1}{3}\{(p_{123} + p_{132}) + p_{123} + (p_{123} + p_{213})\} \\ &= p_{123} + \frac{1}{3}p_{132} + \frac{1}{3}p_{213}. \end{aligned}$$

On conclut cette démonstration en remarquant que

$$\frac{1+\tilde{\rho}}{2} = p_{123} + p_{132} + p_{213} \geqslant p = p_{123} + \frac{1}{3}p_{132} + \frac{1}{3}p_{213} \geqslant \left(\frac{1+\tau}{2}\right)^2.$$

Pour la deuxième inégalité à démontrer, on réécrit cette démonstration en montrant la probabilité de discordance au lieu de la probabilité de concordance.

□

## 1.3 Discussion

Avec rappel des quelques définitions présentées dans ce chapitre ainsi qu'avec la présentation des quelques aspects techniques des théorèmes et de leur démonstration pour la fonction copule et pour les principales mesures d'association utilisées pour inférer une copule paramétrique, le lecteur dispose de l'essentiel des outils nécessaires pour pouvoir lire et comprendre les modélisations présentées dans les chapitres suivants. Au prochain chapitre, on va proposer une approche alternative de la modélisation de la relation entre le coût et l'efficacité d'une nouvelle thérapie versus les alternatives pré-existantes sur le marché. Pour ce faire, les notions de dépendance seront particulièrement utiles. Ainsi, le choix d'une famille de copules réprésentant la bonne «forme» de dépendance et la mesure de la force de cette dernière via le tau de Kendall seront des sujets qui y seront approfondis. Ensuite, au chapitre 3, on traitera de données discrètes pour effectuer la mesure du score de propension. Alors, il faudra adapter la théorie présentée ici à ce type de données. Enfin, le chapitre 4 traitera du phénomène de censure sur les observations. Donc, on y verra comment les fonctions marginales ainsi que le paramètre de dépendance de la copule peuvent être estimés avec consistance dans un tel cas.

# Chapitre 2

# Copules et données médico-économiques : cas de l'analyse coût-efficacité

Le concept de fonction copule et la théorie qui en est issue sont des notions essentielles pour comprendre la relation de dépendance entre plusieurs covariables et leurs lois marginales. Dans un schème de données provenant d'essais thérapeutiques, la présence de censure sur les variables d'intérêt est systématique et peut mener à une interprétation biaisée de la relation de dépendance entre les fonctions de distribution marginales et, qui plus est, en une inférence biaisée de la fonction de distribution jointe. Un cas particulier de ces schémas de données provenant d'essais thérapeutiques est l'*analyse coût-efficacité* (et sa variante qu'est l'*analyse coût-utilité*), qui a montré sa nécessité dans plusieurs études médico-économiques ; études où le phénomène de censure intervient. Ce chapitre discute d'une modélisation basée sur la fonction copule pour la fonction de répartition jointe, d'une méthode d'estimation du coût (tant ponctuel que cumulatif) et de la survie ajustée sur la qualitée de vie (QALY) lors d'une analyse coût-efficacité en présence de censure. Cette méthode n'est basée sur aucune supposition de linéarité sur les variables estimées contrairement aux approches classiques ; mais sur une estimation ponctuelle réalisée à partir des distributions marginales des covariables, et du lien de dépendance entre ces dernières. Ainsi, ce chapitre fera un état des lieux des concepts primordiaux dans l'analyse coût-efficacité tout en présentant les diverses approches rencontrées dans la littérature, puis présentera la nouvelle approche proposée dans ce travail de thèse. Ensuite une méthode d'analyse de sous-groupes pour les cohortes randomisées et non-randomisées est présentée ; méthode qui innove en tenant compte des sous-groupes à l'intérieur des bras prédéterminés dans les essais thérapeutiques pour les calculs de variance et de covariance, calculs nécessaires pour la détermination d'intervalles de confiance précis. Enfin, une analyse comparative sur des données cliniques de soins d'acupuncture en traitement primaire pour les céphalées de tension est présentée.

| **Méthodologie médico-économique** | **Réponse obtenue** *(Outcome)* | **Résultat** |
|---|---|---|
| **Analyse Coût-Bénéfices** (CBA) | | • Retour sur l'investissement (ROI)<br>• Bénéfices Net (BNAI) |
| **Analyse Coût-Efficacité** (CEA) | • Valeurs cliniques<br>• Annés de vie gagnées | • Ratio Coût-Efficacité Incrémental (ICER) par année de vie gagnée<br>• Bénéfice Net Incrémental (INB) |
| **Analyse Coût-Utilité** (CUA) | • Années de vie gagnées ajustées par la qualité de vie (QALY) | • ICER par QALY<br>• INB avec QALY |

TABLEAU 2.1 – Distinctions entre les différents types d'analyses médico-économiques.

## 2.1 Introduction

En raison de la croissance au cours du temps de la variété de traitements possibles pour un problème de santé spécifique et, corollairement, à la croissance de l'ensemble des coûts monétaires directs (e.g. coût spécifique d'une hospitalisation) et indirects (coût d'opportunité) pour ce problème spécifique, les sciences médico-économiques tendent à se concentrer sur le concept de coût-efficacité des nouvelles thérapies versus celles préexistantes. Cette pratique mène donc à une étude statistique complète depuis que la pratique commune dans les laboratoires occidentaux est de collecter les données individuelles des coûts sur les patients dans les études randomisées. Ainsi, il est d'usage de calculer avec ces données le bénéfice incrémental net de l'utilisation de la nouvelle thérapeutique versus celle d'usage commun.

Durant les dernières décennies, l'analyse coût-efficacité (CEA) pour les nouvelles thérapies est devenu un sujet contemporain de travail pour les biostatisticiens. Cette analyse est utilisée à l'intérieur de deux schémas particuliers : la CEA qui se base sur la modélisation et la CEA qui se base sur un apprentissage par essais et erreurs. La différence majeure entre les deux approches est que dans le cas de l'analyse basée sur l'apprentissage par essais et erreurs, les données sont situées au niveau individuel des patients dans une étude spécifique, et cela peut conduire à un phénomène de surapprentissage quant aux conclusions statistiques ; ce qui mène à des biais et des erreurs d'interprétation lorsque les résultats sont ramenés au niveau de populations. En contraste, la CEA basée sur la modélisation est basée sur des données plus facilement généralisables car une structure d'aide à la décision peut être imposée au travail statistique (e.g. imposition de fonctions marginales paramétriques aux covariables).

Au tableau 2.1, il est présenté les différences majeures entre les types d'analyses médico-économiques actuellement rencontrées dans la littérature. Dans le cadre de ce chapitre, nous nous limiterons à l'analyse coût-utilité que nous présenterons en fait comme une spécification de l'analyse coût-efficacité. Ainsi, on utilise la sémantique coût-efficacité (ou CEA) car tous les résultats présentés sont généralisables sans travail supplémentaire de la part du lecteur pour les années de vies gagnées sans ajustement sur la qualité de vie.

## 2.2 Quantités d'intérêt et approches dans la littérature

Il sera question ici de présenter les deux principales quantités d'intérêt : le ratio incrémental coût-efficacité (**ICER**) et le bénéfice incrémental net (**INB**). Pour y arriver, on note qu'on présente ces deux quantités en utilisant la variable exprimant les temps de survie $T$ et, par la suite, on introduira la variable des temps de survie ajustés à la qualité de vie $T_{adj}$ et utilisera cette dernière pour les calculs d'intérêt au lieu de $T$.

Soit l'intervalle $(0, \tau]$ étant la durée d'intérêt de l'étude médico-économique ; $\tau$ étant le dernier moment de suivi du patient. Cet intervalle est divisible en $K$ sous-intervalles $[\alpha_k, \alpha_{k+1})$ de longueur $\kappa_k$ où $0 = \alpha_1 < \alpha_2 < ... < \alpha_{K+1} = \tau$ où la borne supérieure de l'intervalle représente le moment de saisie des coûts instantanés de traitement dans le dit intervalle, coûts dénotés $c_k, k \in \{1, 2, ..., K\}$. Ainsi, on obtient la variable aléatoire exprimant les coûts cumulatifs d'une thérapeutique sur la durée de l'étude telle que

$$C = \sum_{k=0}^{K} c_k \kappa_k.$$

Soit $j$ étant l'indice notant le bras thérapeutique de l'étude ($j = 1$ pour le bras traité et $j = 0$ pour le bras contrôle). Alors, on définit le ratio incrémental coût-efficacité tel que

$$\text{ICER} = \frac{\mathbb{E}[C_{j=1}] - \mathbb{E}[C_{j=0}]}{\mathbb{E}[T_{j=1}] - \mathbb{E}[T_{j=0}]}.$$

où $\mathbb{E}(\bullet)$ est l'espérance mathématique. Cette quantité est donc un indicateur des coûts monétaires de l'utilisation d'une nouvelle thérapeutique quant aux temps de survie. Par ailleurs, on définit le bénéfice incrémental net comme la différence

$$\text{INB}(\lambda) = \lambda \left(\mathbb{E}[T_{j=1}] - \mathbb{E}[T_{j=0}]\right) - \left(\mathbb{E}[C_{j=1}] - \mathbb{E}[C_{j=0}]\right)$$

où $\lambda$ représente une quantité prédéterminée, arbitraire, de la volonté de payer pour une unité supplémentaire d'efficacité (de survie).

Dans la littérature, plusieurs article proposent des voies pour estimer ces quantités. Pour commencer, Willan et Lin[13] proposent une approche basée sur la moyenne échantillonnale. Soient $T_{ji}$ et $C_{ji}$ étant respectivement les mesures de l'efficacité et des coûts cumulés pour le patient $i, i = 1, 2, ..., n_j$ assigné à la thérapeutique $j$. Alors, on obtient les termes d'espérance

$$\mathbb{E}\begin{bmatrix} T_j \\ C_j \end{bmatrix} = \begin{bmatrix} \mu_j \\ \nu_j \end{bmatrix}$$

où $\mu = \mu_{j=1} - \mu_{j=0}$ et $\nu = \nu_{j=1} - \nu_{j=0}$, et de variance $\mathbb{V}(T_j) = \sigma_j^2$, $\mathbb{V}(C_j) = \omega_j^2$, $cov(C_j, T_j) = \rho_j \sigma_j \omega_j$. Si on se place dans un schéma de minimisation, $\mu$ est considéré comme étant nul et, alors, $\nu$ est considéré le paramètre d'intérêt. Alors, en cas de données complètes, ils proposent d'estimer le bénéfice incrémental net comme une simple relation linéaire de différence impliquant les estimateurs $\hat{\mu}$, $\hat{\nu_{j=1}}$ et $\hat{\nu_{j=0}}$. Par ailleurs, Willan et O'Brien[14] proposent d'estimer l'intervalle

de confiance de l'ICER et sa variance grâce à l'application du théorème de Fieller. En présence de données censurées, ils proposent d'estimer la fonction de survie $\hat{S}_j(t)$ sur chaque bras thérapeutique en utilisant l'estimateur de Kaplan-Meier,[15] et ensuite d'estimer $\hat{\mu}_j$ en intégrant la fonction de survie jusqu'au temps $\tau$ (dernier moment d'observation de l'individu).

Par la suite, les principales méthodes développées ont considéré l'ajustement des temps de survie à la qualité de vie. C'est à partir de 2003 avec le travail de Willan et al.[16] qu'on a commencé à considérer la variable *QALY* (Quality Adjusted Life Years) ; variable que l'on va noter ici $T_{adj}$. Soit $Q_{jki}$ étant la qualité de vie observée sur l'intervalle temporel $k, k \in \{1, 2, ..., K\}$ pour le patient $i$ contraint au traitement $j$. Il s'agit en fait de temps de survie contractés par le facteur de la qualité de vie. Par ailleurs, on note $\mathbb{E}[q_{ji}] = \mu_j$.

On détermine la valeur de $Q_{jki}$ ainsi : soit l'individu $i$ assigné au bras $j$ avec une qualité de vie mesurée aux moments $t_{ji1}, t_{ji2}, ..., t_{jim_{ji}}$ avec des scores de qualité de vie $q_{ji1}, q_{ji2}, ..., q_{jim_{ij}}$ qui sont en fait les valeurs de l'utilité. Alors, $Q_{jki} = \int_{a_k}^{a_{k+1}} q(t)dt$ est la somme pondérée du temps passé dans les différents états de qualité de vie où

$$q(t) = \begin{cases} q_{ji1} & \text{si } 0 \leqslant t < t_{ji1}; \\ q_{jih} + \dfrac{(q_{ji,h+1} - q_{jih})(t - t_{jih})}{t_{ji,h+1} - t_{jih}} & \text{si } t_{jih} \leqslant t < t_{ji,h+1}; \\ q_{jim_{ji}} & \text{si } t_{jim_{ji}} \leqslant t < X_{ji}; \\ 0 & \text{si } t \geqslant X_{ji}, \end{cases}$$

et où $X_{ji} = min(T_{ji}, \eta_{ji}), \delta_{ji} = \mathbb{1}_{\{T_{ji} < \eta_{ji}\}}$ où $\eta$ représente la variable censure. De plus, soit $Y_{jki} = \mathbb{1}(X_{ji} \geqslant a_k$ et $[X_{ji} \geqslant a_{k+1}$ ou $\delta_{ji} = 1])$, qui indique si l'individu $i$ sur le bras traité $j$ est en vie au moment $a_k$ et n'est pas censuré sur l'intervalle $[a_k, a_{k+1})$ ; et soit $Y_{jk} = \sum_{i=1}^{n_j} Y_{jki}$. Si on note $\bar{Q}_{jk} = \sum_{i=1}^{n_j} (Y_{jki} Q_{jki})/Y_{jk}$, alors, avec une expression connue de la variance,[16] on obtient l'estimation de $\mu_j$ ajusté à la qualité de vie prenant la forme

$$\hat{\mu}_j = \sum_{k=1}^{K} \hat{S}_j(a_k) \bar{Q}_{jk}.$$

Plus récemment, Willan et al.[17] ont proposé de réaliser l'analyse coût-efficacité entière en utilisant les méthodes de régression linéaire. Soit $C_i$ le coût cumulatif pour l'individu $i$, alors $\mathbb{E}(C_{ji}) = \beta_{C_j}^T Z_{C_{ji}}$, $i = 1, 2, ..., n_j$ où $Z_{C_{ji}}$ est un vecteur de covariables affectant le coût et $\beta_{C_j}$ est le vecteur des coefficients de régression (à estimer). Ainsi, en utilisant une pondération basée sur l'inverse de la probabilité de censure (IPCW), ils proposent une méthode pour estimer la seconde composante de $\beta_{C_j}$ qui est la différence moyenne des coûts ajustés aux autres covariables entre les deux bras, $\hat{\Delta}_c$, et sa variance associée. Une méthodologie similaire est effectuée pour la variable $T_{adj}$. Donc, ils proposent d'estimer l'ICER ajusté sur la qualité de vie par $\hat{\Delta}_c / \hat{\Delta}_{t_{adj}}$ et d'utiliser le théorème de Fieller afin de déterminer l'intervalle de confiance au niveau $100(1 - \alpha)\%$. En ce qui a trait à l'INB, ils proposent comme estimation d'utiliser $b_\lambda = \lambda \hat{\Delta}_{t_{adj}} - \hat{\Delta}_c$ avec, comme

expression de la variance $\hat{\sigma}^2_\lambda = \lambda^2\hat{\sigma}^2_{\Delta_e} + \hat{\sigma}^2_{\Delta_c} - 2\lambda\hat{\sigma}_{\Delta_c\Delta_e}$. Alors, si $b_\lambda/\hat{\sigma}_\lambda$ est plus grand que le quantile $z_{1-\alpha}$, au niveau $\alpha$, la nouvelle thérapie est considérée *coût-efficace*.

De cette approche basée sur la régression linéaire, plusieurs variantes paramétriques et semi-paramétriques[18] sont nées. Le problème principal de ces estimateurs pour l'analyse coût-efficacité basé sur la régression linéaire est que même si l'estimateur IPCW est consistent, il n'est pas efficace dans la mesure où s'il y a présence d'un individu censuré avant ou au temps $\alpha_{K+1}$, ce dernier ne contribue pas à la sommation qui constitue cet estimateur et de l'information statistiquement significative est dès lors perdue. Par ailleurs, en l'absence de censure, cette approche est équivalente à la résolution des moindres carrés ordinaires pour laquelle l'hypothèse de linéarité peut mener à de sérieux biais en cas de non-linéarité. Pour ces raisons, une méthodologie basée sur les copules paramétriques d'analyse coût-efficacité et de modélisation des fonctions de densité jointes entre les coûts cumulatifs et la variable QALY, pour chaque bras thérapeutique, et donc basée seulement sur la dépendance entre les covariables et l'information a priori sur les distributions des variables d'intérêt, est présentée ici.

## 2.3 Modèle

### 2.3.1 Détermination de QALY en termes de temps et de qualité de vie

Dans l'éventualité où le temps de survie ajusté sur la qualité de vie est déjà mesuré, il est évidemment indiqué de procéder directement à l'estimation des paramètres de sa distribution. Toutefois, cette situation est plutôt rare : il est d'usage que les cliniciens n'aient en leur possession que les variables *temps de survie* et *utilité (i.e. qualité de vie)*. Tel que montré dans la section précédente du même chapitre, la méthode d'ajustement classique du temps de survie sur la qualité de vie est donnée par

$$T_{adj}(\omega) = \int_0^{T(\omega)} Q(v(t))dt$$

où $Q(v(t)) \in [0,1]$ est la qualité de vie sur l'intervalle temporel d'intérêt et $v(t)$ est l'état de santé du sujet au temps $t$. Or, étant donné que la fonction $H(t) = \int_0^t Q(v(y))dy$ est monotone croissante, il est possible de réécrire la fonction de distribution cumulative de $T_{adj}$ en une composition de fonctions telle que

$$F_{adj}(y) = F \circ H^{-1}(y)$$

où $H^{-1}(\cdot)$ est la fonction de l'inverse généralisé de $H(t)$ et l'on peut également réécrire la fonction de densité de probabilité de la façon suivante :

$$f_{adj}(y) = f[H^{-1}(y)]\frac{1}{Q[v(H^{-1}(y))]}$$

où $f_{adj}$ est la fonction de densité de $T_{adj}$ et $f$ est celle de $T$. Donc, pour un individu $i$ étant soumis au traitement $j$, la variable aléatoire représentant $QALY$, $E_{adj_{ji}}$, est telle que

$$E_{adj_{ji}} = inf_i\left[T_{adj_{ji}}, \eta_{adj_{ji}}\right]$$

où $\eta_{adj_{ji}}$ représente le phénomène de censure ajustée sur la qualité de vie des patients $i$ dans le bras thérapeutique $j$ et $T_{adj_{ji}}$ représente le temps de survie soumis au même ajustement. Par ailleurs, notons $C_{ji}(t)$ le coût cumulatif de l'invididu $i$ ayant la thérapie $j$ jusqu'à l'instant $t$. Ainsi, on obtient les relations de dépendance suivantes :

#### 2.3.1.1 Relations de dépendance entre les variables d'intérêt dans le cadre de l'analyse coût-utilité

- **Les variables aléatoires $T_{adj_{ji}}$ et $\eta_{ji}$ sont dépendantes.**

*Démonstration :*

Pour simplifier la notation, supposons ce qui suit comme représentant un individu $i$ assigné à un bras thérapeutique $j$. De plus, soit $T_{adj}(\omega) = \int_0^{T(\omega)} Q(t)dt$. Ayant les temps observés $X(\omega) = \inf(\eta(\omega), T(\omega))$, si $\eta(\omega) \leqslant T(\omega)$, on obtient

$$\begin{aligned} T_{adj}(\omega) &= \int_0^{\eta(\omega)} Q(t)dt + \int_{\eta(\omega)}^{T(\omega)} Q(t)dt \\ &= \eta_{adj}(\omega) + \int_{\eta(\omega)}^{T(\omega)} Q(t)dt \\ &= \eta_{adj}(\omega) + f(\eta(\omega)) \end{aligned}$$

où $f$ est une fonction de $\eta(\omega)$. Alors, $T_{adj}(\omega)$ est dépendant de $\eta(\omega)$.

□

- **Les variables aléatoires $C_{ji}$ et $\eta_{ji}$ sont dépendantes.**

*Démonstration :*

Pour alléger la notation, on suppose ce qui suit comme étant pour un individu $i$ ayant la thérapeutique $j$. On a :

$$C(\omega) = \begin{cases} \sum_0^{T(\omega)} c_k \kappa_k & \text{si } T(\omega) \leqslant \eta(\omega); \\ \sum_0^{\eta(\omega)} c_k \kappa_k & \text{si } \eta(\omega) \leqslant T(\omega). \end{cases}$$

Alors, avec $E(\omega) = \inf(\eta(\omega), T(\omega))$,

$$C(E(\omega)) = \mathbb{1}_{[T(\omega) \leqslant \eta(\omega)]} C(T(\omega)) + \mathbb{1}_{[\eta(\omega) \leqslant T(\omega)]} C(\eta(\omega)).$$

Ainsi, en acceptant que la variable $T(\omega)$ soit indépendante de $\eta(\omega)$, on remarque que les variables $\eta(\omega)$ et $C(\omega)$ sont dépendantes.

□

- **Les variables aléatoires $T_{adj_{ji}}$ et $\eta_{adj_{ji}}$ sont indépendantes.**

*Démonstration :*

Soient

$$\begin{aligned} T_{adj}(\omega) &= \int_0^{T(\omega)} Q(t)dt \\ &= H[T(\omega)] \end{aligned}$$

et

$$\begin{aligned} \eta_{adj}(\omega) &= \int_0^{\eta(\omega)} Q(t)dt \\ &= H[\eta(\omega)] \end{aligned}$$

où $H$ est une fonction borélienne inversible. Étant donné que $T(\omega)$ et $\eta(\omega)$ sont des temps indépendants, il est trivial de remarquer que $H[T(\omega)]$ et $H[\eta(\omega)]$ le sont également. Alors, $T_{adj_{ji}}$ et $\eta_{adj_{ji}}$ sont indépendants.

□

### 2.3.2 Estimation des paramètres inhérents aux distributions

Pour commencer, même dans l'éventualité où les distributions réelles des coûts cumulés et de QALY sont inconnus, il est possible de faire l'inférence des deux principaux paramètres empiriques : la moyenne et la variance. En fait, on va considérer ici chaque bras de l'essai, pour chacune des deux distributions d'intérêt, comme étant une variable aléatoire distincte avec sa propre moyenne et sa propre variance, mais avec la même loi de distribution que sa réciproque dans le bras opposé. Par ailleurs, on va assumer l'idée de l'existence d'une censure non-administrative, ce qui conduit à optimiser le processus d'estimation.

Soit $\mathbf{Z_{ji}^{C}}$ étant le vecteur de dimension $d$ de covariables qui affectent les coûts sur le bras $j$, $j = 0, 1$, pour la population groupée et soit $\mathbf{Z_{ji}^{T_a dj}}$ étant son équivalence pour QALY. Alors, tel que proposé par Thompson et Nixon,[19] et Stamey et al.,[20] la fonction exprimant la moyenne pour les coûts sur le bras $j$ est exprimée par

$$\mu_j^C = \alpha_0 + \alpha_1 z_{1j}^C + ... + \alpha_d z_{dj}^C$$

et, pour QALY, par

$$\mu_j^{T_a dj} = \beta_0 + \beta_1 z_{1j}^{T_a dj} + ... + \beta_d z_{dj}^{T_a dj}.$$

De ce fait, ces constructions sont des modèles de régression linéaire avec censure sur les covariables. En utilisant la méthode de Lin,[21] on peut estimer le vecteur des coefficients de régression $\alpha_C$ par une somme sur les $k$ périodes d'intérêt telle que $\hat{\alpha}_C = \sum_{k=1}^{K} \hat{\alpha}_{c_k}$, où en utilisant la méthode de pondération de l'inverse de la probabilité de censure (IPCW) telle que, pour un individu $i$ appartement au bras $j$, on ait

$$\hat{\alpha}_{c_k} = \left( \sum_{i=1}^{n} \frac{\delta_{jki}^{\star}}{\hat{G}(X_{jki}^{\star})} Z_j^C (Z_j^C)^t \right)^{-1} \sum_{i=1}^{n} \frac{\delta_{jki}^{\star} c_{jki}}{\hat{G}(X_{jki}^{\star})} Z_j^C$$

où $X^{\star}_{jki} = \min(X_{ji}, a_{k+1})$, $\hat{G}(\bullet)$ est l'estimateur de Kaplan-Meier de $G(\bullet)$, $\delta^{\star}_{jki} = \delta_{ji} + (1 - \delta_{ji})\mathbb{1}(X_{ji} \geqslant a_{k+1})$ et, tel que décrit précédemment, $X_{ji}$ est le temps minimal entre soit le temps compris dans l'intervalle temporel de la randomisation au décès de l'individu, ou soit celui compris dans l'intervalle temporel de la randomisation au moment de censure ; et $\delta_{ji} = \mathbb{1}(T_{ji} \leqslant \eta_{ji})$. Le même cheminement est utilisé pour trouver $\beta$, le vecteur des coefficients de régression pour QALY. Alors, de cette inférence sur les coefficients, il est possible de déterminer la moyenne ajustée sur la survie.

On note qu'il est possible d'utiliser toute autre technique de pondération avec un biais inférieur ou égal à celui de Lin.[21] Ainsi, en ce qui a trait à la variance des variables aléatoires d'intérêt, on propose d'utiliser le résultat de Buckley et James[22] (voir également Miller et Halpern[23]), qui est une généralisation des techniques IPCW basée sur l'espérance conditionnelle des variables aléatoires et tendant donc à avoir un biais minimal. Ainsi, la variance approximative pour la distribution des coûts cumulés dans un bras thérapeutique spécifique ($j$ étant fixé) est

$$\hat{\sigma}^2_C = \frac{1}{\sum_{l=1}^{n} \delta_l - 2} \sum_{i=1}^{n} \delta_i \left( \hat{e}^0_i - \frac{1}{\sum_{l=1}^{n} \delta_l} \sum_{j=1}^{n} \delta_j \hat{e}^0_j \right)^2$$

où $\hat{e}^0_i$ est un terme d'erreur tel que $\hat{e}^0_i = C_i - \mathbf{Z^C_j}\hat{\beta}_j$. Une approche similaire est réalisée pour la distribution de QALY.

### 2.3.3 Détermination des distributions paramétriques

Pour la modélisation des coûts, trois lois de probabilité communes sont fréquemment utilisées : la loi Gamma, la loi normale et la loi lognormale. La paramétrisation de ces dernières est facilement réalisable sachant la moyenne et la variance de la distribution. Soit $\mu_C$ étant la moyenne et $\sigma^2_C$ étant la variance des coûts, et $j$ le bras thérapeutique. Alors, le choix de la distribution paramétrique va être l'une des options suivantes :

1. $C_j \sim Normal(\mu_{C_j}, \sigma^2_{C_j})$;
2. $C_j \sim Gamma(\mu_{C_j}, \rho_{C_j})$
3. $C_j \sim Lognormal(\nu_{C_j}, \tau^2_{C_j})$;

où $\nu_C$ et $\tau^2_C$ sont la moyenne et la variance des log-coûts, i.e. $\nu_C = 2log(\mu_C) - \frac{1}{2}log(\sigma^2_C + \mu^2_C)$ et $\tau^2_C = log(\sigma^2_C + \mu^2_C) - 2log(\mu_C)$. Par ailleurs, $\rho_C$ est le paramètre de forme de la distribution Gamma, qui est tel que $\rho_C = \mu^2_C/\sigma^2_C$. Ainsi, une fois que chaque modélisation est effectuée, la sélection de la distribution paramétrique qui a la meilleure adéquation aux données s'effectue sur la base du critère de la déviance. En fait, la meilleure adéquation aux données est celle qui a la plus petite déviance, soit le négatif de deux fois la log-vraisemblance.

Dans le cas de la variable QALY, le choix est limité à une seule option compte tenu de l'aspect symétrique que prennent les données de cette distribution :

1. $T_{adj_j} \sim Normal(\mu_{T_{adj_j}}, \sigma^2_{T_{adj_j}})$.

### 2.3.4 Inférence sur le tau de Kendall

Dans le but ultérieur de calculer le paramètre de dépendance pour chaque copule testée dans la procédure de sélection de la copule, on doit obtenir un paramètre de dépendance global indépendant des types de familles de fonctions multivariées : le tau de Kendall. L'idée est que de réaliser l'inférence sur le tau de Kendall une seule fois au lieu d'une fois sur chaque famille de copule à tester mène à une seule procédure d'inférence. Bien que l'estimateur du maximum de vraisemblance puisse être utilisé pour obtenir le paramètre de la copule, dans ce travail, on a considéré uniquement la méthode d'inversion du tau de Kendall dans le but de diminuer le travail computationnel nécessaire. En fait, pour chaque famille de copules paramétriques, il existe une relation directe entre le paramètre de la copule et le tau de Kendall. On considère le couple de variables aléatoires $(C, T_{adj})$ pour un bras thérapeutique donné $j$, $j = 0, 1$. Par ailleurs, on considère $(C^{\{1\}}, T_{adj}^{\{1\}})$ et $(C^{\{2\}}, T_{adj}^{\{2\}})$, deux copies indépendantes du couple $(C, T_{adj})$. Alors, ces copies sont dites concordantes si $(C^{\{1\}} - C^{\{2\}})(T_{adj}^{\{1\}} - T_{adj}^{\{2\}}) > 0$ et discordantes autrement. Le tau de Kendall, qui est donc une mesure de concordance, est défini par Kendall[9] tel que

$$\begin{aligned}
\tau_K &= \mathbb{P}[(C^{\{1\}} - C^{\{2\}})(T_{adj}^{\{1\}} - T_{adj}^{\{2\}}) > 0] - \mathbb{P}[(C^{\{1\}} - C^{\{2\}})(T_{adj}^{\{1\}} - T_{adj}^{\{2\}}) < 0] \\
&= 2 \cdot \mathbb{P}[(C^{\{1\}} - C^{\{2\}})(T_{adj}^{\{1\}} - T_{adj}^{\{2\}}) > 0] - 1 \\
&= \mathbb{E}[(2 \cdot \mathbb{1}[C^{\{1\}} - C^{\{2\}} > 0] - 1)(2 \cdot \mathbb{1}[T_{adj}^{\{1\}} - T_{adj}^{\{2\}} > 0] - 1)] \\
&= \mathbb{E}[a_{12} b_{12}]
\end{aligned}$$

où $\mathbb{E}$ est l'espérance, $\mathbb{1}$ est la fonction indicatrice, $a_{12} = 2 \cdot \mathbb{1}[C^{\{1\}} - C^{\{2\}} > 0] - 1$ et $b_{12} = 2 \cdot \mathbb{1}[T_{adj}^{\{1\}} - T_{adj}^{\{2\}} > 0] - 1$. Dans un cadre plus général, on a les couples $(C^{\{1\}}, T_{adj}^{\{1\}}), (C^{\{2\}}, T_{adj}^{\{2\}}), ..., (C^{\{n\}}, T_{adj}^{\{n\}})$ où toutes les valeurs de $C^{\{r\}}, T_{adj}^{\{r\}}$, $r = 1, ..., n$ sont uniques. Ainsi, on peut écrire $a_{rs} = 2 \cdot \mathbb{1}[C^{\{r\}} - C^{\{s\}} > 0] - 1$ et $b_{rs} = 2 \cdot \mathbb{1}[T_{adj}^{\{r\}} - T_{adj}^{\{s\}} > 0] - 1$ où $r$ et $s$ sont les indices de ces réplications indépendantes. En l'absence de censure, l'estimation de $\tau$ est donnée par

$$\hat{\tau}_K = \binom{n}{2}^{-1} \sum_{1 \leqslant r < s \leqslant n} a_{rs} b_{rs}$$

où $n$ est la taille de l'échantillon. En fait, il s'agit simplement des $n(n-1)/2$ couples d'observations bivariées qui peuvent être construits multipliés par la soustraction du nombre de paires discordantes au nombre de paires concordantes. Sous la contrainte de la censure, l'approche d'Oakes[24] propose d'ajouter une indicatrice de complétude des observations $L_{rs} = \mathbb{1}[\min(C^{\{r\}}, C^{\{s\}}) < \min(\eta_C^{\{r\}}, \eta_C^{\{s\}}), \min(T_{adj}^{\{r\}}, T_{adj}^{\{s\}}) < \min(\eta_{T_{adj}}^{\{r\}}, \eta_{T_{adj}}^{\{s\}})]$ à cette équation telle que

$$\tilde{\tau}_K = \binom{n}{2}^{-1} \sum_{1 \leqslant r < s \leqslant n} L_{rs} a_{rs} b_{rs}$$

où $\eta^{\{r\}}$ et $\eta^{\{s\}}$ représentent les variables censurées sous chaque copie indépendante. Le problème de cet estimateur est son manque de consistance dans la situation où les données sont hautement corrélées. De cette façon, il est recommandé dans ce travail d'utiliser l'estimateur renormalisé

d'Oakes[25] pour lequel la consistance a été démontré peu importe le niveau de corrélation entre les variables aléatoires. Ainsi, l'estimateur

$$\hat{\tau}_K = \frac{\sum_{\{1 \leqslant r < s \leqslant n\}} L_{rs} a_{rs} b_{rs}}{\sum_{\{1 \leqslant r < s \leqslant n\}} L_{rs}}$$

est simplement le ratio du nombre de paires discordantes non-censurées soustraites au nombre de paires concordantes non-censurées par rapport au nombre total de paires non-censurées.

#### 2.3.4.1 Sélection bayésienne de la copule

Pour sélectionner la meilleure copule quant à son adéquation aux données, quelques possibilités sont présentées dans la littérature. Pour les distributions de données complètes, Genest et Rivest[26] proposent une procédure basée sur la transformée de l'intégrale de probabilité. D'autre part, en se basant sur un test d'adéquation aux données, Lakhal-Chaieb[27] propose une procédure pour les données censurées lorsque les distributions sont estimées à l'aide d'une fonction de survie (e.g. estimation de Kaplan-Meier). Cependant, lorsque disponible, une connaissance a priori des lois de probabilité que suivent les fonctions marginales à l'intérieur de la copule est une information non-négligeable et se doit d'être prise en considération pour l'inférence sur la famille de copules lorsque cette dernière est inconnue, afin de minimiser le risque d'erreur. Dans leur article, Dos Santos Silva et al.[28] proposent une méthodologie de sélection basée sur les critères d'information (e.g. AIC, BIC, ...) On note $F_{T_{adj}}(y)$ et $F_C(\chi)$ les fonctions de répartition pour QALY et les coûts cumulatifs sur un bras thérapeutique donné. Alors, on a :

$$f(y, \chi|\Phi) = c(F_{T_{adj}}(y|\Phi_{T_{adj}}), F_C(\chi|\Phi_C)|\Phi_\theta) f_{T_{adj}}(y|\Phi_{T_{adj}}) f_C(\chi|\Phi_C)$$

où $\Phi_C$ dénote un vecteur de paramètres pour la distribution des coûts, $\Phi_{T_{adj}}$ est le vecteur de paramètres pour la distribution de QALY, $\Phi_\theta$ est le paramètre de dépendance qui est, en fait, une fonctionnelle du tau de Kendall et $\Phi = \Phi_C \cup \Phi_{T_{adj}} \cup \Phi_\theta$ est l'union de tous ces vecteurs de paramètres. Aussi, $f$ et $F$ dénotent respectivement la fonction de masse et la fonction de répartitions. On remarque qu'une écriture similaire peut être effectuée pour une modélisation multivariée. Tel qu'illustré par Genest et al.,[29] on peut trouver le paramètre de copule de n'importe quelle copule paramétrique en ayant uniquement la mesure du tau de Kendall,[9] même si les données sont multivariées.[30]

Soit $\boldsymbol{x}$ un échantillon bivarié de taille $n$ provenant de cette densité. Soit $M_k$ une famille de copules donnée, $k = 1..m$, où $m$ est le nombre de modèles (familles) que l'on désire comparer afin de déterminer celui ayant la meilleure adéquation aux données. Alors, la fonction de vraisemblance est donnée par

$$L(\boldsymbol{x}|\Phi, M_k) = \prod_{j=1}^{n} c(F_{T_{adj}}(y_j|\Phi_{T_{adj}}, M_k), F_C(\chi_j|\Phi_C, M_k)|\Phi_\theta, M_k) f_{T_{adj}}(y_j|\Phi_{T_{adj}}, M_k) f_C(\chi_j|\Phi_C, M_k).$$

Ainsi, en utilisant la déviance, qui est en fait $D(\Phi_k) = -2ll(\boldsymbol{x}|\Phi, M_k)$ où $ll$ représente la fonction de log-vraisemblance, Dos Santos Silva et al.[28] proposent d'utiliser le critère d'information basé

sur la déviance (DIC) qui est

$$DIC(M_k) = 2E[D(\Phi_k)|\boldsymbol{x}, M_k] - D(E[\Phi_k|\boldsymbol{x}, M_k]),$$

où ils proposent l'utilisation des approximations de Monte Carlo pour les quantités $E[D(\Phi_k)|\boldsymbol{x}, M_k]$ et $E[\Phi_k|\boldsymbol{x}, M_k]$ qui sont respectivement $L^{-1}\sum_{l=1}^{L} D(\Phi_k^l)$ et $L^{-1}\sum_{l=1}^{L}\Phi_k^l$. Ici, on suppose que $\left\{\Phi_k^{(1)}, ..., \Phi_k^{(L)}\right\}$ est un échantillon provenant de la distribution a posteriori $f(\Phi_k|\boldsymbol{x}, M_k)$. Alors, on choisit le modèle de copule ayant le plus petit DIC.

### 2.3.5 Ratio incrémental coût-efficacité

À partir des densités jointes estimées, $f(y_j, \chi_j)$, $j \in \{0, 1\}$, on a, pour les coûts,

$$\begin{aligned}\mathbb{E}[C_j] &= \int_{\mathbb{D}_{C_j}}\int_{\mathbb{D}_{T_{adj_j}}} \chi_j f(\chi_j, y_j)dy_j d\chi_j \\ &\approx \int_{\mathbb{D}_{C_j}}\int_{\mathbb{D}_{T_{adj_j}}} \chi_j c_{\hat{\theta}}(\tilde{F}_{T_{adj}}(y|\hat{\Phi}_{T_{adj}}), \tilde{F}_C(\chi|\hat{\Phi}_C))\tilde{f}_C(y|\hat{\Phi}_C)\tilde{f}_{T_{adj}}(\chi|\hat{\Phi}_{T_{adj}})dy_j d\chi_j\end{aligned}$$

où $\mathbb{D}_{C_j}$ et $\mathbb{D}_{T_{adj_j}}$ sont respectivement les domaines de définition des variables aléatoires $C$ et $T_{adj}$ pour le bras $j$. Pour la variable QALY, on calcule son espérance par

$$\begin{aligned}\mathbb{E}[T_{adj_j}] &= \int_{\mathbb{D}_{T_{adj_j}}}\int_{\mathbb{D}_{C_j}} y_j f(\chi_j, y_j)d\chi_j dy_j \\ &\approx \int_{\mathbb{D}_{T_{adj_j}}}\int_{\mathbb{D}_{C_j}} y_j c_{\hat{\theta}}(\tilde{F}_{T_{adj}}(y|\hat{\Phi}_{T_{adj}}), \tilde{F}_C(\chi|\hat{\Phi}_C))\tilde{f}_C(y|\hat{\Phi}_c)\tilde{f}_{T_{adj}}(\chi|\hat{\Phi}_{T_{adj}})d\chi_j dy_j.\end{aligned}$$

Alors, sachant les coûts espérés et les temps de survie ajustés à la qualité de vie, le ratio incrémental coût-efficacité (ICER) ajusté est estimé par

$$\widehat{ICER} = \frac{\widehat{\mathbb{E}(C_{j=1})} - \widehat{\mathbb{E}(C_{j=0})}}{\widehat{\mathbb{E}(T_{adj_{j=1}})} - \widehat{\mathbb{E}(E_{adj_{j=0}})}}$$

et, grâce au théorème de Fieller ( Fieller,[31] Willan et O'Brien,[14] Chaudhary et Stearns[32]), on obtient l'intervalle de confiance $100(1-\alpha)\%$ tel que

$$\widehat{ICER}(1-z_{1-\alpha/2}^2\hat{\sigma}_{\Delta_C\Delta_{T_{adj}}} \pm \frac{z_{1-\alpha/2}\sqrt{\hat{\sigma}^2_{\Delta_{T_{adj}}} + \hat{\sigma}^2_{\Delta_C} - 2\hat{\sigma}_{\Delta_C\Delta_{T_{adj}}} - z^2_{1-\alpha/2}(\hat{\sigma}^2_{\Delta_{T_{adj}}}\hat{\sigma}^2_{\Delta_C} - \hat{\sigma}^2_{\Delta_C\Delta_{T_{adj}}})}}{1 - z^2_{1-\alpha/2}\hat{\sigma}^2_{\Delta_{T_{adj}}}}).$$

Dans cette formule, $z_{1-\alpha/2}$ représente le $100(1-\alpha/2)$ percentile de la distribution gaussienne standardisée. Par ailleurs, $\hat{\sigma}^2_{\Delta_{T_{adj}}}$ représente la variance de la distribution de QALY où $\Delta_{T_{adj}}$ est la différence entre $T_{adj_{j=1}}$ et $T_{adj_{j=0}}$. Le même schème se présente pour $\hat{\sigma}^2_{\Delta_C}$. Pour ce qui en est de $\hat{\sigma}_{\Delta_C\Delta_{T_{adj}}}$, il ne s'agit de rien de plus que de la covariance estimée des différences, entre $j = 0$ et $j = 1$, pour les coûts et les temps de survie ajustés à la qualité de vie.

On remarque que les raisons pour lesquelles on a construit un intervalle de confiance au sens de Fieller plutôt qu'en appliquant une méthode du type bootstrap proviennent des conclusions

des travaux de Siani et Moatti[33] : en plus d'inconvénients au niveau du support des intervalles de confiance (ce qui peut avoir un impact sur la règle de décision), le bootstrap a l'inconvénient, contrairement à la méthode de Fieller, d'être instable lorsque $\mathbb{E}(\widehat{T_{adj_{j=1}}}) - \mathbb{E}(\widehat{E_{adj_{j=0}}})$ converge vers 0, tout en étant statistiquement différent de 0. Effectivement, dans une telle situation, la densité de l'ICER va être bimodale, ce qui enlève la précision des intervalles de confiance.

### 2.3.6 Bénéfice incrémental net

Le bénéfice incrémental net (INB) ajusté ($\lambda$) est estimé par $\widehat{INB} = \lambda(\mathbb{E}(\widehat{T_{adj_{j=1}}}) - \mathbb{E}(\widehat{T_{adj_{j=0}}})) - (\mathbb{E}(\widehat{C_{j=1}}) - \mathbb{E}(\widehat{C_{j=0}}))$ avec, comme expression de la variance, $\hat{\sigma}^2_{\lambda} = \lambda^2 \hat{\sigma}^2_{\Delta_{T_{adj}}} + \hat{\sigma}^2_{\Delta_C} - 2\lambda \hat{\sigma}_{\Delta_C \Delta_{T_{adj}}}$ où $\lambda$ est la "propension de payer" pour une unité d'efficacité.

### 2.3.7 Analyse de sous-groupes

Il est possible d'accomplir une analyse intra-cohorte en utilisant la procédure présentée dans ce travail. L'idée principale ici est de réaliser une analyse coût-efficacité tout en créant une discrimination entre deux sous-groupes ou plus. Le principe est qu'il existe une variable référentielle $Z_k$, $k \in \{1, 2, ..., d\}$, autant pour les coûts cumulés que pour QALY, qui est, en fait, une variable catégorielle (dichotomique or multichotomique) et pour laquelle on peut déterminer l'$INB$ marginal. De tels sous-groupes doivent être basés sur des covariables cliniquement importantes. Étant donné que ces sous-groupes sont à l'intérieur d'un bras thérapeutique donné, il n'est pas possible d'affirmer qu'ils sont équilibrés. Tel qu'illustré en Nixon et Thompson,[34] Tsai et Peace,[35] une utilisation naïve de l'information provenant de ces sous-groupe sans aucun ajustement peut mener à de sérieux biais d'inférence.

Soient $Z^C_{jki}, Z^{T_{adj}}_{jki}$ signifiant l'allocation ou non d'attributs donnés à une population (e.g. sexe, consommation de tabac, etc.) et étant représentés en tant que covariables d'intérêt agissant sur les coûts et QALY pour un invidu $i$ faisant partie du bras thérapeutique $j$. Dans le but d'illustrer le concept ici, on prend pour exemple que l'on test l'effet d'une thérapeutique sur les fumeurs. Alors, il y aura 4 sous-groupes : fumeurs dans le groupe traité, non-fumeurs dans le groupe traité, fumeurs dans le groupe contrôle et non-fumeurs dans le groupe contrôle.

Soient $T_{adj_{j=1,k=1,i}}$ étant l'efficacité pour les individus $i$, fumeurs, sur le bras traité ; et $T_{adj_{j=1,k=0,i}}$ étant l'efficacité pour les individus $i$, non-fumeurs, sur le même bras. On établit la même notation pour le bras contrôle $j = 0$. Par ailleurs, on établit une notation similaire pour les coûts cumulatifs, $C$. Soient $\mathbb{E}(\overline{T}_{adj_j}) = \mathbb{E}(T_{adj_{j,k=1}}) - \mathbb{E}(T_{adj_{j,k=0}})$ et $\mathbb{E}(\overline{C}_j) = \mathbb{E}(C_{j,k=1}) - \mathbb{E}(C_{j,k=0})$. Alors, l'intérêt de cette discrimination est sur le bénéfice incrémental net marginalisé à la cohorte des fumeurs, qui est $\overline{INB}(\lambda) = \lambda(\mathbb{E}(\overline{T}_{adj_{j=1}}) - \mathbb{E}(\overline{T}_{adj_{j=0}})) - (\mathbb{E}(\overline{C}_{j=1}) - \mathbb{E}(\overline{C}_{j=0}))$. Étant donné que ces sous-groupes sont à l'intérieur d'un bras clinique, la problématique principale est d'établir l'expression de la variance. En ajustant la méthode de Fieller à ce contexte de sous-groupes, on

obtient

$$\begin{aligned}\mathbb{V}ar(\overline{INB}(\lambda)) &= \lambda^2\sigma^2_{\Delta_{\overline{T_{adj}}}} + \sigma^2_{\Delta_{\overline{C}}} - 2\lambda\sigma_{\Delta_{\overline{T_{adj}}}\Delta_{\overline{C}}} \\ &= \lambda^2\mathbb{V}ar(\overline{T}_{adj_{j=1}} - \overline{T}_{adj_{j=0}}) + \mathbb{V}ar(\overline{C}_{j=1} - \overline{C}_{j=0}) \\ &- 2\lambda cov(\overline{T}_{adj_{j=1}} - \overline{T}_{adj_{j=0}}, \overline{C}_{j=1} - \overline{C}_{j=0})\end{aligned}$$

où les variances peuvent être estimées de façon régulière. En ce qui a trait au terme de la covariance, $\sigma_{\Delta_{\overline{T_{adj}}}\Delta_{\overline{C}}}$, il y a deux scénarios possibles. Premièrement, lorsque l'hypothèse de la randomisation entre les sous-groupes à l'intérieur de chaque bras est possible, on a

$$cov(\overline{T}_{adj_{j=1}} - \overline{T}_{adj_{j=0}}, \overline{C}_{j=1} - \overline{C}_{j=0}) = cov(\overline{T}_{adj_{j=1}}, \overline{C}_{j=1}) + cov(\overline{T}_{adj_{j=0}}, \overline{C}_{j=0})$$

qui peut être déterminée facilement en utilisant les techniques statistiques classiques. Deuxièmement, lorsque l'hypothèse de la randomisation entre les sous-groupes n'est pas possible, alors la covariance s'exprime telle que

$$\begin{aligned}cov(\overline{T}_{adj_{j=1}} - \overline{T}_{adj_{j=0}}, \overline{C}_{j=1} - \overline{C}_{j=0}) &= cov(\overline{T}_{adj_{j=1}}, \overline{C}_{j=1}) + cov(\overline{T}_{adj_{j=0}}, \overline{C}_{j=0}) \\ &- cov(\overline{T}_{adj_{j=1}}, \overline{C}_{j=0}) - cov(\overline{T}_{adj_{j=0}}, \overline{C}_{j=1}).\end{aligned}$$

Ici, les termes $cov(\overline{T}_{adj_{j=1}}, \overline{C}_{j=1})$ et $cov(\overline{T}_{adj_{j=0}}, \overline{C}_{j=0})$ sont calculés de façon classique similairement au cas randomisé. Cependant, pour les termes de covariance qui nécessitent de croiser les deux bras thérapeutiques, $cov(\overline{T}_{adj_{j=1}}, \overline{C}_{j=0})$ et $cov(\overline{T}_{adj_{j=0}}, \overline{C}_{j=1})$ ; l'approche que l'on suggère ici est d'estimer les fonctions de répartition jointes $F(\overline{T}_{adj_{j=1}}, \overline{C}_{j=0})$ par $C_{\hat{\theta}}(\hat{F}_{T^-_{adj}}(y_{j=1}), \hat{F}_{\bar{C}}(\chi_{j=0}))$ et $F(\overline{T}_{adj_{j=1}}, \overline{C}_{j=0})$ par $C_{\hat{\theta}}(\hat{F}_{T^-_{adj}}(y_{j=1}), \hat{F}_{\bar{C}}(\chi_{j=0}))$ à partir de la méthodologie présentée dans ce travail, et ensuite, en utilisant la définition de la covariance $cov(\overline{C}, \overline{T_{adj}}) = \mathbb{E}[\overline{T_{adj}}, \overline{C}] - \mathbb{E}[\overline{T_{adj}}]\mathbb{E}[\overline{C}]$, de calculer les covariances désirées à partir de l'estimation de la fonction de densité jointe et de l'estimation des fonctions de densité marginales.

Suivant le raisonnement de Willan et al.,[17] on peut tester l'hypothèse de l'égalité de l'INB entre les sous-groupes, qui peut être rejetée au niveau $\alpha$ si

$$\frac{\left|\overline{INB}(\lambda)\right|}{\sqrt{\mathbb{V}ar(\overline{INB}(\lambda))}}$$

est plus grand que le percentile $z_{1-\alpha/2}$ de la distribution gaussienne standardisée.

## 2.4 Résultats et discussion

Le but de cette section est de donner une illustration de la performance de la copule qui approxime le mieux, en utilisant la méthode présentée dans ce travail, la vraie copule et les fonctions de distribution cumulatives des variables coûts et QALY. Soit la copule exacte $C^{(\star)}_{\theta}(F_{T_{adj}}(y|\Phi_{T_{adj}}), F_C(\chi|\Phi_C))$ et son approximation provenant de la méthode présentée à la section précédente $C^{(i)}_{\hat{\theta}}(\tilde{F}_{T_{adj}}(y|\hat{\Phi}_{T_{adj}}), \tilde{F}_C(\chi|\hat{\Phi}_C))$ où $(i)$ est le modèle de copule sélectionné pour faire une approximation de la distribution jointe à travers l'ensemble des modèles de copules

testés et $(\star)$ représente cette distribution jointe exprimée à travers une copule. Par ailleurs, $\tilde{F}$ est la distribution choisie pour $F$. Alors, l'objectif de ces simulations est de montrer que les biais générés par les approximations de $\theta$ par $\hat{\theta}$, de $\Phi = \Phi_C \cup \Phi_{T_{adj}}$ par $\hat{\Phi}$, la sélection de $C^{(i)}$ en tant que modèle de copule et $\tilde{F}_C, \tilde{F}_{T_{adj}}$ en tant que modèles marginaux paramétriques, est relativement faible.

Pour évaluer la performance de la méthodologie présentée dans ce travail dans des cas non-triviaux, on a réalisé des simulations de Monte-Carlo avec 18 schèmes différents. Le processus était de simuler des données bivariées (représentant les coûts et QALY) à partir de trois copules spécifiques. Pour chaque copule, on a simulé les trois configurations possibles pour les distributions marginales (les coûts sont normallement distribués, les coûts sont log-normallement distribués et les coûts suivent une loi gamma ; alors que QALY est normallement distribué), puis on a appliqué, dans chaque cas, deux différents niveaux de censure aléatoire sur les distributions marginales de QALY. La censure suivait une loi exponentielle telle que $\lambda_{s=15} = 0.041$ et $\lambda_{s=30} = 0.090$ où $s$ représente le pourcentage de censure simulé. Pour chaque procédure de génération de données (PGD), le tau de Kendall était le même et représentait un niveau intermédiaire de dépendance entre les distributions marginales afin d'être au plus près de la réalité : $\tau_K = 0.60$. Puis, on a calculé le paramètre de copule, pour chaque copule, en se basant sur le tau de Kendall. Par ailleurs, on a utilisé des paramètres de moyenne et de variance relativement standard dans l'analyse coût-efficacité, suivant les paramètres suivis par Thompson et Nixon[19] tels que $\mu_C = 1500, \sigma_C = 400$ ; $\mu_E = 4, \sigma_E = 0.75$, et on a paramétrisé chaque distribution marginale afin de demeurer près de ces valeurs. Pour le choix des copules génératrices de données, on a sélectionné une copule elliptique à un seul paramètre (la gaussienne) et les deux copules archimédiennes à un seul paramètre les plus connues pour être différentes dans la forme de la dépendance qu'elles tentent de décrire. Les schèmes des PGD sont présentés au tableau 2.2.

Étant donné que la simulation de covariables linéairement dépendantes et non-censurées pour les coûts peut mener à des biais en notre avantage pour le calcul de la moyenne et de la variance en comparaison à une estimation de ce type dans un contexte clinique classique, on a décidé de défier notre méthode en utilisant l'estimateur de la moyenne par Kaplan-Meier de la fonction de survie et sa variance associée (approche à suivre en l'absence de covariables d'intérêt) au lieu de la procédure présentée et basée sur les covariables. Puis, on a appliqué les étapes présentées dans ce travail pour sélectionner une distribution paramétrique pour les marges, sélectionner une copule paramétrique en utilisant un critère d'information et, finalement, trouver le paramètre de la copule. On a répliqué cette procédure $B = 500$ fois avec $n = 1000$ données et on a collecté l'information sur le nombre de procédures qui furent un succès par rapport à l'inférence sur les marges, l'estimation du tau de Kendall et le choix de la copule.

#### 2.4.0.1 Inférence sur le tau de Kendall

Le chemin proposé pour inférer le tau de Kendall ($\tau_K$) en présence de censure donne des résultats proches au réel $\tau_K$ mesuré sur les données juste avant d'appliquer la censure. À la

| Copule génératrice | Distribution des coûts | Niveau de censure | PGD |
|---|---|---|---|
| Copule gaussienne $\theta \approx 0.809$ | $F_C \sim N(\mu_C = 1500, \sigma_C = 400)$ | 15% | PGD 1 |
| | | 30% | PGD 2 |
| | $F_C \sim \Gamma(shape_C = 12, scale_C = 125)$ | 15% | PGD 3 |
| | | 30% | PGD 4 |
| | $F_C \sim logN(\nu_C = 7.30, \tau_C = 0.25)$ | 15% | PGD 5 |
| | | 30% | PGD 6 |
| Copule de Clayton $\theta = 3$ | $F_C \sim N(\mu_C = 1500, \sigma_C = 400)$ | 15% | PGD 7 |
| | | 30% | PGD 8 |
| | $F_C \sim \Gamma(shape_C = 12, scale_C = 125)$ | 15% | PGD 9 |
| | | 30% | PGD 10 |
| | $F_C \sim logN(\nu_C = 7.30, \tau_C = 0.25)$ | 15% | PGD 11 |
| | | 30% | PGD 12 |
| Copule de Gumbel $\theta \approx 0.809$ | $F_C \sim N(\mu_C = 1500, \sigma_C = 400)$ | 15% | PGD 13 |
| | | 30% | PGD 14 |
| | $F_C \sim \Gamma(shape_C = 12, scale_C = 125)$ | 15% | PGD 15 |
| | | 30% | PGD 16 |
| | $F_C \sim logN(\nu_C = 7.30, \tau_C = 0.25)$ | 15% | PGD 17 |
| | | 30% | PGD 18 |

Tableau 2.2 – Schèmes des 18 différentes procédures de simulations

| Niveau de censure | Moyenne ($\hat{\tau}_K$) | Var ($\hat{\tau}_K$) | Min ($\hat{\tau}_K$) | Max ($\hat{\tau}_K$) |
|---|---|---|---|---|
| censure =15% | 0.6011 | 0.00024 | 0.5416 | 0.6539 |
| censure =30% | 0.6030 | 0.00035 | 0.5319 | 0.6648 |

Tableau 2.3 – Informations sur l'estimation du tau de Kendall pour chaque niveau de censure. Les extrants des 9 simulations avec un niveau de censure de 15% sont joints ensemble dans l'information de la première ligne, et identiquement pour le niveau de censure de 30% à la seconde ligne

figure 2.1 et à la table 2.3, on peut voir la dispersion des tau de Kendall calculés pour, dans un cas, toutes les simulations avec un indice de censure de 15% et, dans un autre cas, toutes les simulations avec un indice de censure de 30%. Ainsi, le vecteur contenant toutes les mesures du tau de Kendall pour un niveau de censure donné a une longueur de 4500 observations. On observe que la dispersion des valeurs pour certains $\hat{\theta}$, en appliquant la méthode d'inversion de $\tau_K$, n'est pas "collée" à la valeur réelle comme peut l'être sa valeur correspondante de $\hat{\tau}_K$ à $\tau_K$ étant donné que le support du tau de Kendall est $[-1, 1]$ alors que, pour certaines copules comme celle de Clayton, par exemple, $\theta$ se situe en $[-1, \infty) \setminus \{0\}$.

#### 2.4.0.2 Inférence sur les distributions marginales de coûts

Pour sélectionner la meilleure distribution marginale au sens de l'adéquation aux données pour chacune des 500 simulations, en chaque PGD, on a utilisé le critère basé sur la déviance tel que proposé précédemment. Alors, à la figure 2.2, on peut voir la performance de ce critère. On

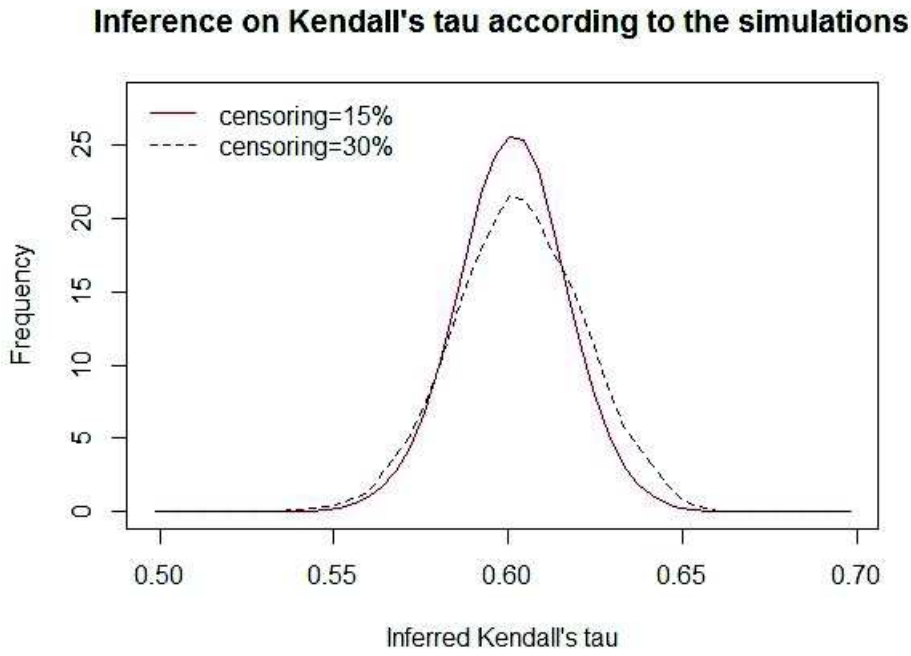


FIGURE 2.1 – Dispersion des tau de Kendall estimés en fonction du niveau de censure.

remarque que, même à un niveau de censure de 30%, la distribution de probabilité choisie est presque toujours correctement estimée.

#### 2.4.0.3 Inférence sur les familles de copules

Il existe une vastitude de familles de copules paramétriques mais, pour simplifier ces simulations, on va limiter notre sélection aux familles les plus présentes dans la littérature : familles gaussienne, de Student, de Clayton, de Gumbel, de Frank et de Joe. Pendant les simulations, pour chaque itération, on a collecté l'information à propos de la famille sélectionnée en utilisant le critère d'information. Sur les tableaux 2.4 et 2.5, on peut voir ces résultats pour chaque PGD. On a indiqué, en gras, la famille de copules ayant été sélectionnée le plus de fois sur les 500 itérations et on affirme qu'il s'agit de la copule à choisir pour la dite PGD. Lorsque la sélection de la famille de copules s'effectue seulement entre celles utilisées pour générer les données (tel qu'illustré à la table 2.4), il est évident que la copule choisie est celle utilisée pour la génération. Autrement, lorsque l'on ajoute des copules intermédiaires (pour lesquelles la forme de la dépendance est proche de celle issue des copules de génération des données) dans la procédure de sélection, les résultats peuvent différer tel que montré au tableau 2.5. En ce qui a trait aux PGD où les coûts étaient simulés suivant une distribution normale (1,2,7,8,13 et 14), le choix de la copule est influencé par la dépendance provoquée par la copule utilisée pour générer les données. En fait, lorsque la distribution marginale des coûts suit la loi $N(\mu_C = 1500, \sigma_C = 400)$ et que la distribution marginale de QALY suit la loi $N(\mu_E = 4, \sigma_E = 0.75)$, en utilisant $\tau_K = 0.60$, si la

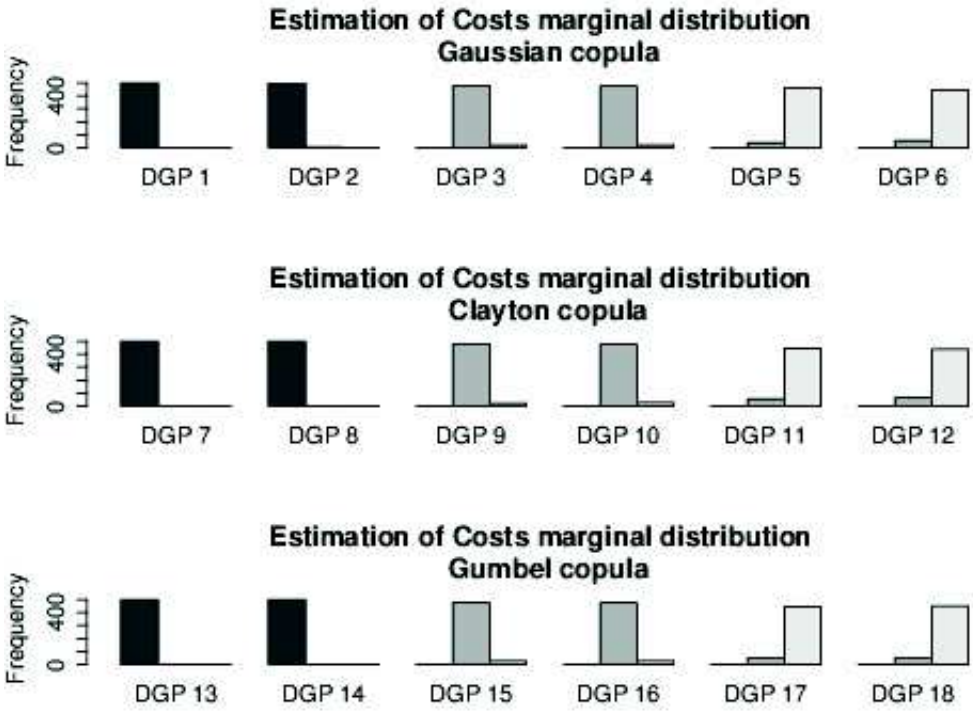


Figure 2.2 – Fréquence de la sélection des distributions marginales (paramétriques) pour les coûts à partir du critère de la déviance. Les bandes noires représentent la sélection de la loi gaussienne, les bandes grises foncées la sélection de la loi Gamma et les bandes grises pâles celles de la loi log-normale.

distribution de la dépendance est pratiquement normale (i.e. provient d'une copule gaussienne ou de Gumbel), il y aura présence d'une queue de distribution modérée, presque inexistante, en chaque extrême sur le graphe de dispersion des données avec un nuage uniforme centré entre ces deux dernières. Cela explique entre autre la meilleure adéquation qu'a la copule de Frank à ce type de données ; copule qui est sélectionnée pour les PGD 1,2, 13 et 14. Toutefois, lorsque, pour une copule de Clayton ayant servi à générer les données, une dépendance stricte est imposée à la queue de distribution inférieure alors qu'une presque totale indépendance est imposée à la queue de distribution supérieure (bien que le terme *queue de distribution* soit plutôt un qualificatif de position ici étant donné son inexistance en cas d'indépendance), seulement une copule de Clayton va être en phase pour modéliser ces données. C'est la raison pour laquelle c'est la copule sélectionnée pour les PGD 7 et 8. Dans la situation où QALY suit la loi $N(\mu_E = 4, \sigma_E = 0.75)$ alors que les coûts suivent une distribution déphasée (loi gamma ou loi log-normale), dans n'importe quel cas, il n'existe pas réellement de queue de dépendance à droite, mais certainement un nuage de points orienté à gauche avec une queue de dépendance lourde à la gauche. C'est la raison pour laquelle, même dans le cas où les données proviennent d'une copule de Clayton, la copule de Student (t) est celle qui est sélectionnée. Ainsi, les simulations ont montré que le critère de sélection bayésien de la famille de copules est en phase avec les propriétés théoriques des copules paramétriques.

Selon la structure des lois marginales, le choix de la copule peut seulement être effectué pour un nombre limité de familles. Ainsi, il est indispensable lorsque l'on cherche la meilleure famille de copules au sens de l'adéquation aux données d'inclure les familles les plus connues et qui couvrent le plus de structures de dépendance possible. Donc, en harmonie avec le tableau 2.5, une copule sélectionnée qui n'est pas celle ayant servi à générer les données est simplement une copule ayant une meilleure adéquation structurelle (sur la forme de la dépendance) entre les marges.

### 2.4.1 Exemple : Données sur l'acupuncture en tant que soin primaire pour les maux de tête chroniques

Pour illustrer ce travail, on utilise les données sur l'acupuncture en tant que soin primaire pour les maux de tête chroniques selon Vickers et al.[36–38] (avec la courtoisie de Pr. Andrew J. Vickers) contenant l'information sur les migraines et les céphalées de tension pour 401 patients âgés entre 18 et 65 ans qui ont rapporté une moyenne d'au moins deux maux de tête par mois. Les sujets ont été recrutés par des médecins généralistes en Angleterre et au Pays de Galles ; et se sont fait allouer jusqu'à 12 séances d'acupuncture sur une période de trois mois. Pour la finalité de l'étude, les interventions d'acupuncture ont été fournies à la communauté par le United Kingdom National Health Service (NHS). L'étude a débuté en 2002 avec un temps d'horizon de 12 mois et a été enregistrée auprès du NHS : $ISRCTN96537534$.

L'information présente dans les données se concentre sur les mesures de l'efficacité en terme de QALY et sur les coûts cumulatifs qui leur sont associés mesurés en livres sterling (£). Les patients

| PGD | copule de Gaussienne | copule de Clayton | copule de Gumbel |
|---|---|---|---|
| DPG 1 | **460** | 0 | 40 |
| PGD 2 | **454** | 0 | 46 |
| PGD 3 | **415** | 0 | 85 |
| PGD 4 | **429** | 0 | 71 |
| PGD 5 | **473** | 0 | 27 |
| PGD 6 | **490** | 0 | 10 |
| PGD 7 | 0 | **500** | 0 |
| PGD 8 | 0 | **500** | 0 |
| PGD 9 | 18 | **482** | 0 |
| PGD 10 | 11 | **489** | 0 |
| PGD 11 | 2 | **498** | 0 |
| PGD 12 | 6 | **493** | 1 |
| PGD 13 | 2 | 0 | **498** |
| PGD 14 | 9 | 0 | **491** |
| PGD 15 | 123 | 0 | **377** |
| PGD 16 | 109 | 0 | **391** |
| PGD 17 | 80 | 0 | **420** |
| PGD 18 | 84 | 0 | **416** |

Tableau 2.4 – Fréquence du choix d'une copule spécifique pour chaque PGD pour 500 itérations avec les trois copules principales. La copule choisie est en caractère foncé.

| PGD | copule Gaussienne | copule de Student | copule de Clayton | copule de Gumbel | copule de Frank | copule de Joe |
|---|---|---|---|---|---|---|
| DPG 1 | 7 | 45 | 6 | 25 | **415** | 2 |
| PGD 2 | 2 | 44 | 4 | 25 | **425** | 1 |
| PGD 3 | 32 | **297** | 0 | 56 | 112 | 3 |
| PGD 4 | 26 | **291** | 0 | 59 | 120 | 4 |
| PGD 5 | 47 | **343** | 1 | 45 | 38 | 26 |
| PGD 6 | 55 | **344** | 0 | 43 | 33 | 25 |
| PGD 7 | 0 | 12 | **364** | 0 | 124 | 0 |
| PGD 8 | 0 | 17 | **372** | 0 | 111 | 0 |
| PGD 9 | 0 | **427** | 56 | 0 | 17 | 0 |
| PGD 10 | 2 | **433** | 55 | 0 | 10 | 0 |
| PGD 11 | 1 | **477** | 19 | 2 | 1 | 0 |
| PGD 12 | 2 | **476** | 18 | 0 | 4 | 0 |
| PGD 13 | 0 | 18 | 0 | 110 | **196** | 176 |
| PGD 14 | 0 | 20 | 0 | 107 | **214** | 159 |
| PGD 15 | 13 | **236** | 0 | 153 | 67 | 31 |
| PGD 16 | 19 | **231** | 0 | 157 | 58 | 35 |
| PGD 17 | 32 | **198** | 0 | 177 | 13 | 80 |
| PGD 18 | 39 | **213** | 0 | 148 | 13 | 87 |

Tableau 2.5 – Fréquence du choix d'une copule spécifique pour chaque PGD pour 500 itérations avec les trois copules principales et trois copules intermédiaires. La copule choisie est en caractère foncé.

ont rapporté eux-mêmes les coûts unitaires associés aux médicaments en vente libre consommés et aux consultations pour des soins de santé en cabinet privé. Le coût de l'intervention étudiée a été estimé à partir du coût horaire pour une consultation en acupuncture avec un professionnel du NHS multiplié par le temps de consultation. Ainsi, les patients dans le bras traité ont eu un temps moyen de 4.2 heures de traitement avec un acupuncteur agrémenté par l'étude.

On précise que l'allocation des patients aux thérapies comparées s'est faite via un algorithme de minimisation des variables âge, sexe, diagnostique (migraine ou céphalées de tension), sévérité des maux de tête au moment de l'inclusion à l'étude (sur une échelle de 0 à 5, 0 étant l'absence totale de maux de tête et 5 étant une présence de maux de tête paralysant), chronicité des maux de tête (nombre d'années durant lesquelles l'individu a remarqué la présence de l'affliction) et taille des effectifs dans les différents centres de soins. Les patients qui ont été alloués au bras thérapeutique de «non-acupuncture» ont reçu les soins standards, selon la sévérité de leurs maux de tête, qu'un médecin généraliste peut leur procurer en situation régulière. Par ailleurs, pour uniformiser le traitement, les acupuncteurs prodiguant les traitements devaient être membre de l'*Acupuncture Association of Chartered Physiotherapist* et avoir complété un minimum de 250 heures d'études suite à l'obtention de leur diplôme universitaire. La mesure de l'efficacité sur les sujets s'est effectuée à partir de deux outils complémentaires, en trois moments : à l'inclusion, trois mois après l'inclusion et un an après l'inclusion. Le premier outil était en fait un journal quotidien sur lequel les patients devaient reporter, quatre fois par jour durant un mois, la sévérité du mal de tête sur une échelle de 0 à 5. Le second outil était un questionnaire sur l'état de santé général de l'individu. Il s'agit du *Questionnaire court d'étude de la santé* SF-36, utilisé internationalement et mesurant l'état de santé général et la qualité de vie d'un individu.

Lors de la collecte et du traitement primaire des données, aucune imputation n'a été effectuée dans l'éventualité où l'un des trois questionnaires n'était pas complètement rempli car, dans ce cas, la variable QALY ne pouvait pas être mesurée. Ainsi, le bras traité a comporté 136 participants alors que le bras contrôle en a comporté 119. À la table 2.6, on voit la procédure de modélisation pour chaque fonction de distribution jointe (pour chaque bras thérapeutique) pour les coûts et QALY. On remarque que dans les deux cas, la dépendance est très faible étant donné que le tau de Kendall se situe entre $-0.10$ et $-0.15$. Pour le choix des lois paramétriques servant à modéliser les coûts, dans chaque bras, la distribution log-normale a été considérée comme étant celle ayant la plus petite déviance. En ce qui a trait à la sélection de la famille de copule en utilisant le DIC, on a comparé, pour chaque bras, les copules gaussienne, de Clayton, de Student, de Frank, de Joe et de Gumbel. Dans le bras traité, la copule de Student a été celle qui a été considérée alors que dans le bras contrôle, il s'est agit de la copule gaussienne. Ainsi, la fonction de distribution jointe du bras traité est estimée par

$$\hat{F}(C_{j=1}, T_{adj_{j=1}}) = C^{(Student)}_{\hat{\theta}=-0.1923} \quad ( \quad F_C \sim logN(\hat{\nu}_C = 5.7111, \hat{\tau}_C = 0.7600),$$
$$F_{T_{adj}} \sim N(\hat{\mu}_{T_{adj}} = 0.7268, \hat{\sigma}_{T_{adj}} = 0.1190))$$

| Procédure de modélisation | Bras contrôle | Bras traité |
|---|---|---|
| tau de Kendall ($\hat{\tau}_K$) | $-0.1065$ | $-0.1232$ |
| statistiques de QALY | $\hat{\mu}_{T_{adj}} = 0.7083$<br>$\hat{\sigma}_{T_{adj}} = 0.1118$ | $\hat{\mu}_{T_{adj}} = 0.7268$<br>$\hat{\sigma}_{T_{adj}} = 0.1190$ |
| distribution de QALY | $T_{adj} \sim N(\hat{\mu}_{T_{adj}} = 0.7083, \hat{\sigma}_{T_{adj}} = 0.1118)$ | $T_{adj} \sim N(\hat{\mu}_{T_{adj}} = 0.7268, \hat{\sigma}_{T_{adj}} = 0.1190)$ |
| statistiques des coûts | $\hat{\mu}_C = 217.20$<br>$\hat{\sigma}_C = 486.00$ | $\hat{\mu}_C = 403.40$<br>$\hat{\sigma}_C = 356.59$ |
| distribution des coûts | $C \sim logN(\hat{\nu}_C = 4.4844, \hat{\tau}_C = 1.3390)$ | $C \sim logN(\hat{\nu}_C = 5.7111, \hat{\tau}_C = 0.7600)$ |
| famille de copules choisie | gaussienne | Student (t) |
| paramètre de copule ($\hat{\theta}$) | $-0.1664$ | $-0.1923$ |

TABLEAU 2.6 – Informations obtenues dans la procédure d'analyse pour les coûts et QALY sur chaque bras

alors que, pour le bras contrôle, l'estimation est donnée par

$$\hat{F}(C_{j=0}, T_{adj_{j=0}}) = C^{(Gaussian)}_{\hat{\theta}=-0.1664} \quad ( \quad F_C \sim logN(\hat{\nu}_C = 4.4844, \hat{\tau}_C = 1.3390), \\ F_{T_{adj}} \sim N(\hat{\mu}_{T_{adj}} = 0.7083, \hat{\sigma}_{T_{adj}} = 0.1118)).$$

En se servant de l'approche présentée dans ce travail se basant sur les densités de copules, le ratio incrémental coût-efficacité est estimé par $\hat{ICER} = 10082.68$£ par unité d'efficacité avec un intervalle de confiance, selon la méthode de Fieller, de $\hat{ICER} + (12.44 \times z^2_{1-\alpha} \pm 1251)$ où $z_{1-\alpha}$ est le $100(1-\alpha/2)$ percentile de la distribution normale standardisée ; ce qui signifie qu'il coûte approximativement 10082.68£ par année pour obtenir une unité d'efficacité additionnelle en utilisant l'acupuncture comme soin primaire pour les maux de tête.

À la figure 2.3, on voit le graphique de l'INB estimé avec ses bandes de confiance à 90 %. L'axe vertical montre la variabilité des coûts et l'intervalle de confiance associé. L'axe horizontal montre l'estimation de l'ICER. Étant donné que ce jeu de données est limité en nombre de covariables, il n'est pas possible de déterminer l'existence de sous-groupes. Si ça avait été le cas, l'approche présentée dans ce travail aurait pu être aisément appliquée.

## 2.5 Alternatives proposées au travail présenté

Le but de cette section est de présenter des résultats qui ont été développés en parallèle du travail de cette thèse. Ainsi, même si ces travaux ont été faits durant le temps de cette thèse, soit ils ne s'inscrivent pas dans sa thématique quant à l'application de fonctions copules pour des données observationnelles ; même s'ils considèrent les données observationnelles dans le cadre

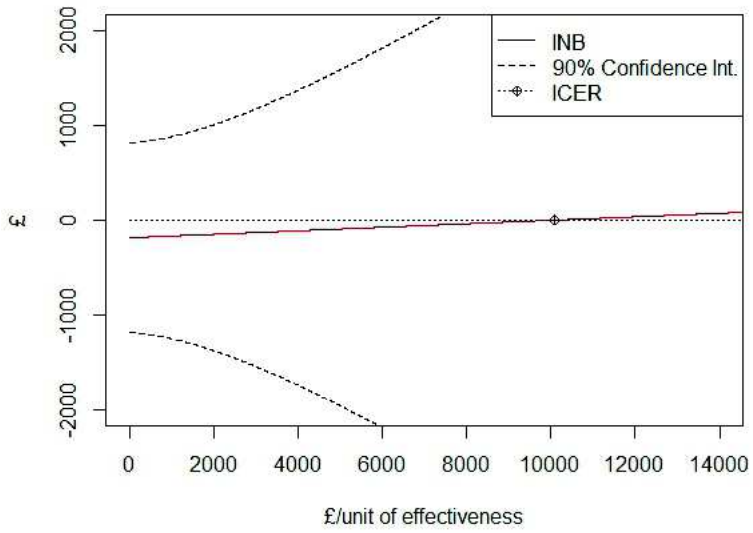


FIGURE 2.3 – Graphique de l'INB versus lambda pour l'acupuncture comme soin primaire pour les maux de tête.

de l'analyse coût-efficacité ; soit ils ont été abordés ultérieurement à la rédaction du présent manuscrit. Effectivement, dans le premier cas, l'approche basée sur l'estimateur de la régression de Buckley et James est développé. Cette approche n'utilise pas la fonction copule, mais bien le développement d'une espérance sur une tribu borélienne. Dans le second cas, l'approche présentée est en fait une généralisation des résultats du chapitre 3 sur le score de propension. En effet, il sera question de réappliquer les résultats proposés pour la probabilité conditionnelle à une réécriture du ratio incrémental coût-efficacité. C'est la raison pour laquelle il ne sera présenté ici que les résultats principaux de ces deux cas.

### 2.5.1 Alternative basée sur l'estimateur de régression de Buckley et James

Ayant la variable aléatoire $T_{adj_j}$ et l'espérance mesurée sur sa sigma-algèbre telle que $Z = \hat{\mathbb{E}}^{\sigma(E_{adj_j},\Delta_{adj_j})}[T_{adj_j}]$ où $\Delta_{adj_j} = \mathbb{1}_{\{T_{adj_j} \leqslant \eta_{adj_j}\}}$, on écrit l'estimateur de l'espérance conditionnelle classique de Buckley et James[22] appliqué à cette variable aléatoire tel que

$$\begin{aligned} \hat{\mathbb{E}}[Z] &= \hat{\mathbb{E}}\left(\hat{\mathbb{E}}^{\sigma(E_{adj_j},\Delta_{adj_j})}[T_{adj_j}]\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} Z_i = \hat{\mathbb{E}}[T_{adj}] \end{aligned}$$

et donc

$$\hat{\mathbb{E}}[T_{adj}] = \frac{1}{n}\left(\sum_{\Delta_{adj_j}=1} T_{adj_{ji}} + \sum_{\Delta_{adj_j}=0} \frac{\int_{\eta_{adj_{ji}}}^{\infty} y d\mathbb{P}_{T_{adj_j}}(y)}{\mathbb{P}(T_{adj_j} > \eta_{adj_j})}\right)$$

où $n$ représente le nombre d'individus présents sur le bras $j$ dans lequel on mesure la variable. Étant donné que la quantité d'intérêt dans l'analyse coût-efficacité est le ratio incrémental coût-efficacité, lui même composé des termes d'espérance pour $T_{adj}$ et $C$, la problématique ici est de déterminer le numérateur de l'ICER ; plus précisément de fixer l'expression de l'espérance au sens de Buckley et James pour les coûts cumulés mesurés sur un bras thérapeutique donné, $C_j$.

Sur un bras fixé, soit le coût cumulé $C$ mesuré sur le temps de survie $T(\omega)$ correspondant à celui ajusté à l'utilité utilisé pour le dénominateur de l'ICER $T_{adj}(\omega)$ tel que

$$C\left[T(\omega)\right] = \int_0^{T(\omega)} c(u)du$$

où $c(u)$ représente les coûts instantanés, ponctuellement mesurés, c'est-à-dire $c(u) = c_k$ si $u \in [\alpha_k, \alpha_{k+1})$ où $0 = \alpha_1 < \alpha_2 < ... < \alpha_{K+1} = \tau, k \leqslant K$ où $\tau$ est le dernier moment de suivi d'un patient. Les coûts cumulés n'étant pas rétractables contrairement aux temps de survie ajustés à la qualité de vie, l'ICER est calculé sur des temps retractés pour être concordants avec le score d'utilité alors que le coût déboursé est entier et ne peut être diminué, on considère les coûts sur la variable $T(\omega)$ au lieu de $T_{adj}(\omega)$.

Par ailleurs, on établit la mesure du coût cumulé sur le temps de censure tel que

$$C[\eta(\omega)] = \int_0^{\eta(\omega)} c(u)du.$$

Ainsi, dans la continuité de la notation utilisée dans le présent chapitre, on introduit $\breve{E} = \min(C[T(\omega)], C[\eta(\omega)])$ et $\breve{\Delta} = \mathbb{1}_{\{C[T(\omega)] \leqslant C[\eta(\omega)]\}}$. Par ailleurs, on peut de façon évidente établir la relation bijective suivante :

$$C[T(\omega)] < C[\eta(\omega)] \Leftrightarrow T < \eta,$$

ce qui est vrai étant donné que $C$ est monotone et croissante.

De cette notation, en utilisant la bijection présentée, on retrouve l'approximation de la valeur de l'espérance pour $C[T(\omega)]$ en utilisant Buckley et James tel que

$$\begin{aligned}
\hat{\mathbb{E}}\left[C[T(\omega)]\right] &= \hat{\mathbb{E}}\left[\mathbb{E}^{\sigma(\breve{E},\breve{\Delta})}[C[T(\omega)]]\right] \\
&= \frac{1}{n}\left(\sum_{\breve{\Delta}_i=1} C[T_i]_i + \sum_{\breve{\Delta}_i=0} \frac{\int_{C[\eta_i]}^{\infty} y d\mathbb{P}(C[T(y)])}{\mathbb{P}(C[T_i] > C[\eta_i])}\right) \\
&= \frac{1}{n}\left(\sum_{\breve{\Delta}_i=1} C(T_i) + \sum_{\breve{\Delta}_i=0} \frac{\int_{C[\eta_i]}^{\infty} y d\mathbb{P}(C[T(y)])}{\mathbb{P}(T_i > \eta_i)}\right) \\
&= \frac{1}{n}\left(\sum_{\breve{\Delta}_i=1} C(T_i) + \sum_{\breve{\Delta}_i=0} \frac{\int_{\eta_i}^{\infty} C(y) d\mathbb{P}_T(y)}{\mathbb{P}(T_i > \eta_i)}\right)
\end{aligned}$$

où l'indice $i$ représente la valeur de la variable au niveau des individus. On note que l'estimation de $\mathbb{P}(T_i > \eta_i)$ s'effectue par une utilisation directe de l'estimateur de Kaplan-Meier. Pour ce qui

en est de $\int_{\eta_i}^{\infty} C(y)d\mathbb{P}_T(y)$, l'utilisation des méthodes d'intégration au sens de Monte Carlo permet d'estimer avec une valeur consistante cette intégration. Ainsi, de cet estimateur de l'espérance des coûts cumulés en présence de censure, on retrouve le numérateur du ratio incrémental coût-efficacité, et par conséquent l'estimation de ce ratio.

### 2.5.2 Alternative basée sur la transformation de probabilité et sur l'utilisation des probabilités conditionnelles

Cette alternative se base sur l'hypothèse que les effectifs entre les deux groupes (traité et contrôle) sont équilibrés et de taille identique. Dans le cas où cette éventualité n'est pas vérifiée, il est recommandé d'effectuer un appariement (*matching*) sur les individus des deux groupes à l'aide d'un score de propension afin de vérifier cette hypothèse et de procéder aux calculs qui suivent, en conservant les jumelages effectués par ce score.

Soit le numérateur de l'ICER $\mathbb{E}\left[C_{j=1}\right] - \mathbb{E}\left[C_{j=0}\right]$ qui, par linéarité de l'espérance, est égal à $\mathbb{E}\left[C_{j=1} - C_{j=0}\right] = \mathbb{E}[C^{\circ}]$ où $C^{\circ}$ est la variable aléatoire issue de la différence entre les variables aléatoires $C_{j=1}$ et $C_{j=0}$. Soit , pour le dénominateur du ratio incrémental coût-efficacité, la même opération de linéarité telle que $\mathbb{E}\left[T^{\circ}_{adj}\right] = \mathbb{E}\left[T_{adj_{j=1}} - T_{adj_{j=0}}\right] = \mathbb{E}\left[T_{adj_{j=1}}\right] - \mathbb{E}\left[T_{adj_{j=0}}\right]$. Alors, on réécrit l'ICER

$$\begin{aligned}
\textbf{ICER} &= \frac{\mathbb{E}\left[C_{j=1} - C_{j=0}\right]}{\mathbb{E}\left[T_{adj_{j=1}} - T_{adj_{j=0}}\right]} \\
&= \frac{\mathbb{E}\left[C^{\circ}\right]}{\mathbb{E}\left[T^{\circ}_{adj}\right]} \\
&\overset{AS}{=} \mathbb{E}\left[\frac{C^{\circ}}{T^{\circ}_{adj}}\right] \\
&= \mathbb{E}\left[Z\right]
\end{aligned}$$

où $Z$ est la variable aléatoire issue du quotient des variables aléatoires $C^{\circ}$ par $T^{\circ}_{adj}$ et où $\overset{AS}{=}$ signifie que l'égalité est vraie dans le cas asymptotique. En admettant la loi jointe $H$ du couple $(C^{\circ}, T^{\circ}_{adj})$ telle que $H_{(C^{\circ},T^{\circ}_{adj})}(u,v) = C_{\theta}(F_{C^{\circ}}(u), G_{T^{\circ}_{adj}}(v))$ où $C_{\theta}$ est la fonction copule et $F_{C^{\circ}}, G_{T^{\circ}_{adj}}$ sont les fonctions marginales des deux variables aléatoires composant $Z$, on trouve la loi de $Z$ par quelques manipulations. En effet, en utilisant l'écriture suivante, on a, pour $h$ étant la densité du couple et $c_{\theta}$ la densité d'une copule paramétrique, $h_{(C^{\circ},T^{\circ}_{adj})}(u,v) = dH_{(C^{\circ},T^{\circ}_{adj})}(u,v)$; alors :

$$\begin{aligned}
\mathbb{P}\left(Z \leqslant z\right) &= \mathbb{P}\left(C^{\circ} < T^{\circ}_{adj}t\right) \\
&= \int_0^{\infty}\int_0^{vz} dH_{(C^{\circ},T^{\circ}_{adj})}(u,v) \\
&= \int_0^{\infty}\int_0^{vz} h_{(C^{\circ},T^{\circ}_{adj})}(u,v)dudv \\
&= \int_0^{\infty}\int_0^{vz} \frac{\partial^2 C_{\theta}(F_{C^{\circ}}(u), G_{T^{\circ}_{adj}}(v))}{\partial F_{C^{\circ}}(u)\partial G_{T^{\circ}_{adj}}(v))}dudv \\
&= \int_0^{\infty} G_{T^{\circ}_{adj}}(v)) \int_0^{vz} c_{\theta}(F_{C^{\circ}}(u), G_{T^{\circ}_{adj}}(v))f_{C^{\circ}}(u)dudv
\end{aligned}$$

qui est valide en acceptant l'hypothèse que les supports de $C^\circ$ et de $T^\circ_{adj}$ sont tels que $C^\circ \in [0,\infty), T^\circ_{adj} \in [0,\infty)$. Cette hypothèse est réaliste dans le cas d'une stratégie où la thérapeutique testée ($j = 1$) est plus coûteuse que la thérapie d'origine ; mais procure une survie ajustée à l'utilitée plus grande. On obtient finalement comme valeur de l'ICER :

$$\begin{aligned}
\textbf{ICER} &= \mathbb{E}[Z] \\
&= \int_0^\infty z d\mathbb{P}(Z \leqslant z) \\
&= -\int_0^\infty z d\mathbb{P}(Z > z) \\
&= -\left[(z\mathbb{P}(Z > z))_0^\infty - \int_0^\infty \mathbb{P}(Z > z)dz\right] \\
&= \int_0^\infty G_{T^\circ_{adj}}(v) \int_0^{vz} c_\theta(F_{C^\circ}(u), G_{T^\circ_{adj}}(v)) f_{C^\circ}(u) du dv
\end{aligned}$$

où l'on réalise aisément l'estimation de cette quantité en considérant les méthodes d'estimation proposées dans le présent chapitre en ce qui a trait au choix des distributions marginales à considérer ; et où l'on peut choisir la famille de copules en se basant sur un critère de distance similairement à ce qui a été considéré au chapitre 3. On note toutefois qu'avec une telle approche, l'estimation du paramètre de la copule à partir de l'inversion du tau de Kendall est computationnellement compliquée : la mesure du tau de Kendall devra tenir compte de l'échangéabilité entre les observations individuelles sur les bras thérapeutiques à l'intérieur des variables $C^\circ$ et $T^\circ_{adj}$. Dans un tel cas, il est recommandé au lecteur d'estimer le paramètre de la copule à l'aide de l'estimateur de vraisemblance pour minimiser l'effort de calcul.

## 2.6 Conclusion

La principale motivation de ce travail a été générée par les limitations des modèles de régression standard appliqués à l'analyse coût-efficacité où la structure de dépendance entre les coûts et l'utilité au fil du temps n'était pas prise en considération. Par ailleurs, l'imposition de la linéarité derrière de tels modèles de régression peut être tout à fait non justifiée dans certains cas. On a présenté ici une procédure, simple, étape-par-étape, afin de trouver le bénéfice incrémental net et le ratio incrémental coût-efficacité, et de déterminer leurs intervalles de confiance ; même dans les cas où une censure sur la variable QALY survient. À la figure 2.4, on peut voir la procédure schématisée qui part des données observationnelles observées sur chaque bras thérapeutique et qui conduit à une analyse coût-efficacité complète. De façon parallèle, on accomplit les étapes 1 et 2, qui consistent en la mesure de la dépendance entre les coûts cumulés et la variable QALY via le tau de Kendall, et la détermination des fonctions de répartition marginales pour chaque variable aléatoire. Puis, à l'étape 3, on génère autant de copules paramétriques que désirées à partir de l'information obtenue aux étapes précédentes et, au sens du critère d'information basé sur la déviance, on sélectionne la copule "la plus proche" de la vraie distribution jointe des deux variables aléatoires. Finalement, on détermine l'INB et l'ICER en utilisant les fonctions

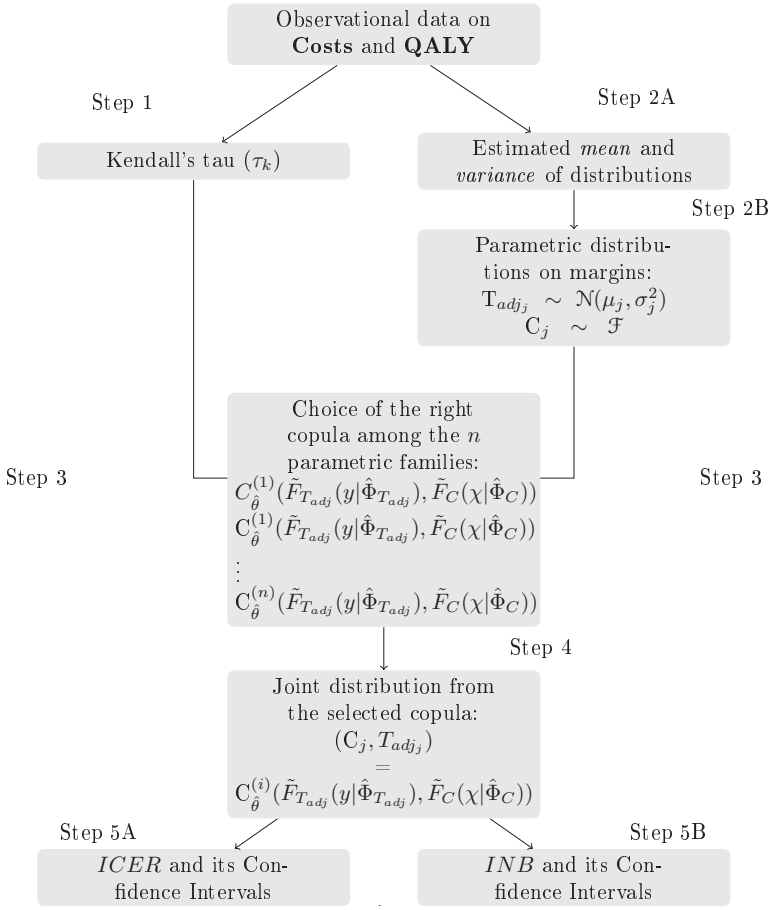


FIGURE 2.4 – Schéma de la procédure à appliquer pour l'analyse coût-efficacité impliquant l'utilisation de copules paramétriques.

jointes de densité et de répartition. En cas d'analyse de sous-groupes, s'il n'y a pas d'hypothèse de randomisation sur ces derniers à l'intérieur d'un bras fixé, on réitère la procédure présentée pour obtenir deux copules supplémentaires et, ultimement, la fonction de répartition jointe des coûts et de l'utilité qui va servir à déterminer le terme de covariance qui croise les deux bras thérapeutiques.

Une perspective future à ce travail serait de pousser l'analyse avec des copules entièrement non-paramétriques, ce qui diminuerait encore davantage le biais d'inférence en adoptant une structure de dépendance épousant parfaitement la forme des données. Étant donné la nature unique de ce type d'analyse sur un jeu de données spécifique permettant d'obtenir une réponse finale à une problématique, dans le sens où l'INB ou l'ICER ne risquent pas d'être utilisés dans d'autres calculs ultérieurs, le risque de surapprentissage sur les données que présente l'utilisation de copules non-paramétriques ne serait pas contraignant.

Pour terminer ce chapitre, il a également été montré ici, dans une perspective moindre, que d'autres approches novatrices sont possibles pour modéliser l'analyse coût-efficacité dans un contexte de censure non-administrative sur les données et de disponibilité de l'information sur l'utilité. L'importance derrière de telles innovations de modélisation n'est pas un simple exercice de réécriture théorique, mais plutôt de fournir les bases à l'implémentation d'un estimateur qui, en plus d'être simple d'utilisation pour un utilisateur néophyte aux statistiques (e.g. practicien hospitalier, etc.), soit un estimateur non-biaisé et à variance minimale ($UMVUE$).

# Chapitre 3

# Copules et données discrètes : cas du score de propension

On débute ce chapitre par une discussion sur les données multivariées discrètes et leurs structures, pour lesquelles on regarde les méthodes actuelles d'analyse et on construit les bases nécessaires pour arriver à en modéliser la dépendance à travers les fonctions copules. La motivation principale de ce chapitre est que dans le cadre de données observationnelles, il est établi que plus que la moitié des données recueillies ne sont pas continues. Alors, le cadre standard des outils statistiques basé sur les distributions continues doit être adapté pour répondre aux exigences d'analyses de ce type de données.

Ce chapitre est organisé comme suit. Dans la section 4.1, on introduit, dans un premier temps, la notion de données discrètes et les types de structures qui en sont issues. On y détaille,entre autre, les cas des données binaires, catégorielles et de comptage. Dans un second temps, on fait état de la littérature quant aux techniques actuellement appliquées dans le cadre de l'analyse de données observationnelles discrètes puis, dans la section 4.2, on établit le cadre proposé afin de permettre l'utilisation de telles données à l'intérieur de copules paramétriques. Dans la section 4.3, on explore les voies proposées dans la littérature statistique utilisant de telles types de données et leurs fonctions de répartition puis, dans la section 4.4 on propose une nouvelle voie basée sur les copules pour travailler avec ce type de données et on l'applique à la probabilité conditionnelle que constitue le score de propension.

## 3.1 Structures de données discrètes et approches actuelles

On s'intéresse aux données discrètes dans le présent ouvrage étant donné que la majeure partie des données observationnelles est de cette nature. On peut prendre, comme exemples, la prise d'un traitement ou non, le niveau de revenu d'un ménage, la nationalité ou bien l'apparition d'effets secondaires dans un délai donné. Par ailleurs, en plus d'être catégorielles ou de comptage, ces variables peuvent être ordonnées ou non ; ce qui en fait un «bon» type de données afin d'y appliquer les statistiques d'ordre.

Sachant que les données discrètes ont pour support l'ensemble $\mathbb{N}$, on note qu'une variable

aléatoire discrète $Z$ peut avoir un schéma de dépendance général ou particulier avec une autre variable d'intérêt. Le schéma de dépendance de telles variables et les conclusions que l'on peut en tirer sont l'objet du présent chapitre.

### 3.1.1 Types de données discrètes couramment rencontrées

On présente ici les types de données données discrètes que l'on rencontre le plus dans le cadre d'études observationnelles. On note que bien que l'on présente des données quantitatives, il est fréquent de rencontrer des données qualitatives sous forme de données de comptage.

*Données binaires :* Les données binaires (ou dichotomiques) sont rencontrées lorsque celles-ci ne peuvent prendre que deux valeurs ; habituellement représentées par les scalaires 0 ou 1 ; où, par convention, 1 représente la survenue d'un événement d'intérêt et 0 la non-survenue. Notons qu'une distribution suivant la loi de Bernoulli est le cas le plus répandu d'un tel type de données. Prenons à type d'exemple la prise d'une thérapie ou non, la présence d'un symptôme donné ou non, ou encore une hygiène de vie exempte de tabac ou non. Ces exemples qui sont tous des mesures fréquentes dans un cadre médico-expérimental illustrent que dans la recherche médicale, et en particulier dans un cadre d'inférence causale et d'analyse du score de propension, les données binaires multivariées sont largement utilisées ; d'autant plus qu'ils minimisent l'espace de stockage lorsqu'ils font partie intégrante de bases de données de grandes dimensions.

*Données catégorielles ordinales :* Une variable ordinale est naturellement ordonnée sur son support, mais les distance entre ses valeurs possibles ne sont pas définies. De ce fait, les données catégorielles ordinales sont fréquentes dans le cadre d'études expérimentales en biostatistiques et particulièrement dans le cadre de l'inférence causale. On prend, par exemple, la classification d'un patient en fonction de son degré d'atteinte par rapport à une maladie (absent, sévère, très sévère,...). Le problème inhérent à de telles données est la difficulté à former une métrique par rapport aux catégories possibles qui forment la variable. En traitant de telles données de façon nominale plutôt qu'ordinale, le traitement est alors possible, mais l'information quant à l'ordonnancement des données est perdue ; ce qui diminue la qualité de l'information tirée de telles variables. On note que pour une variable ordinale, il est habituel de supposer que les catégories ordonnées correspondent à des intervalles exhaustifs et disjoints en l'ensemble $\mathbb{R}$.

*Données de comptage :* Les données de comptage sont, en fait, la mesure du nombre d'occurence d'un événement particulier, n'engageant pas de variable explicative (e.g. temps, traitement) distinguant les dits événements entre-eux ; ce qui fait qu'ils se résument en une simple mesure d'occurences. On note, par exemple, le nombre de lésions cutanées d'une pathologie sur une région particulière, ou encore le nombre de chirurgies subies dans un temps prescrit ; ce qui en font également des données à considérer dans un cadre d'inférence causale.

### 3.1.2 Approches courantes face aux données discrètes

L'objectif de cette sous-section étant de présenter les approches les plus vues dans la littérature afin de traiter les données discrètes, on considère seulement les cas où les variables que l'on appelle discrètes sont uniquement binaires ou catégorielles nominales, dans un but simplifier le texte pour lecteur. On adoptera la même dénomination pour la suite de ce chapitre. Par ailleurs, on considère comme notation $\mathbf{Z} = (Z_1, Z_2, ..., Z_d)^T$ un vecteur de covariables discrètes et $T$ une variable réponse également discrète.

Les modèles présentés ici sont principalement prédictifs et utilisés dans le domaine médical dans des buts diagnostiques ou pronostiques. Un de ces modèles provient du monde des sciences informatiques, mais sa similarité avec les statistiques modernes en fait un modèle à considérer.

**Régression logistique**

On débute cette section en rappelant le modèle standard de la régression linéaire courante avec des données continues. Ayant le vecteur de vecteurs d'observations $(T_i, \boldsymbol{Z}_i)^t$ pour $i = 1, 2, ..., n$ où $n$ est la taille de l'échantillon constituant chaque variable aléatoire, il est possible de déterminer la valeur moyenne de la variable $T$ pour un individu $i$ en suivant le principe de la régression linéaire ; soit en imposant une relation fonctionnelle juxtée à un bruit centré $\epsilon$ entre la variable réponse et les covariables. Par cette relation, il est convenable d'obtenir l'espérance de $T_i$ conditionnelle aux valeurs observées $\mathbf{Z}_i$ des covariables d'intérêt en établissant la droite

$$T_i \quad = \quad \beta_0 + \beta_1 Z_{1(i)} + \beta_2 Z_{2(i)} + ... + \beta_d Z_{d(i)} + \epsilon_i$$

où les coefficients $\beta_0, \beta_1, ..., \beta_d$ sont les paramètres de régression à estimer en utilisant différentes stratégies (voir le chapitre 2 pour plus de détails).

Un tel modèle peut s'avérer intuitif et sembler pratique aux premiers abords. Cependant, en se plaçant dans un contexte de données observationnelles dans le cadre d'une recherche à caractère médical servant à valider ou à infirmer une hypothèse thérapeutique, on est placé dans un contexte où la variable réponse est dichotomique ( prise ou non de la thérapie testée ou encore succès ou échec de la thérapie testée) ; donc discrète, et les covariables d'intérêt sont un mélange de loi continues (e.g. poids, durée temporelle, ...) et discrètes (e.g. sexe, usage du tabac, ...). Ainsi, l'établissement d'une telle relation dans un contexte de régression nécessite un réajustement de la régression linéaire standard aux données discrètes ; ce qui constitue la finalité de la régression logistique. Effectivement, il serait fallacieux d'utiliser la régression linéaire afin de modéliser le comportement de la variable réponse $T$ en terme de la probabilité $\mathbb{P}(\mathbf{Z})$ : dans une telle situation, une fonction de $\mathbf{Z}$ doit avoir un support entre 0 et 1 alors que le fonction linéaire n'est pas bornée.

L'utilisation d'une transformation logarithmique telle que $\log(\mathbb{P}(\mathbf{Z}))$ soit une fonction linéaire de $\mathbf{Z}$ peut sembler avantageuse au sens où un changement dans une covariable multipliera la variable réponse par une quantité fixe. Cependant, elle est également à proscrire pour des raisons relatives au domaine de définition de la variable observée : bien que cette variable puisse prendre valeur sur $\mathbb{R}$, la transformation logarithmique aura pour support un intervalle borné en une direction.

L'intuition de la transformation logarithmique simple mène à chercher un type de transformation, basée sur les logarithmes, qui aura un support non-borné à partir d'une variable à expliquer bornée ; ce qui est assuré par la transformation logit : $\log\left[\mathbb{P}(\mathbf{Z})/(\mathbf{1}-\mathbb{P}(\mathbf{Z}))\right]$. Ainsi, pour un individu $i, i = 1, 2, ..., n$, on définiera le modèle de régression logistique tel que

$$\log \frac{\mathbb{P}(\mathbf{z_i})}{1-\mathbb{P}(\mathbf{z_i})} = \beta_0 + \mathbf{z_i}\beta$$

où $\boldsymbol{\beta}$ représente un vecteur de coefficients de régression et $\beta_0$ un terme d'origine (*intercept*). En isolant $\mathbb{P}(\mathbf{z_i})$, on obtient comme expression de la variable réponse

$$\mathbb{P}(\boldsymbol{z_i}) = \frac{1}{1+\exp\{-(\beta_0+\boldsymbol{z_i\beta})\}} = \frac{\exp\{\beta_0+\boldsymbol{z_i\beta}\}}{1+\exp\{\beta_0+\boldsymbol{z_i\beta}\}}.$$

Il est ainsi évident que pour estimer si la variable réponse dichotomique $T$ prends la valeur de 0 ou de 1, il suffit de déterminer respectivement, pour l'individu $i$, si $\beta_0+\mathbf{z_i}\beta$ est positif ou négatif ; où la frontière de décision sera déterminée par la solution à l'équation $\beta_0 + \mathbf{z_i}\beta = \mathbf{0}$. L'intérêt de l'utilisation de la régression avec la fonction logistique dans un cadre d'études thérapeutiques remonte aux travaux de Berkson[39] et, d'un point de vue théorique, n'a pas subi d'adaptation contextuelle dans son utilisation malgré l'avénement depuis de l'utilisation de statistiques semi et non-paramétriques dans ces domaines d'application.

Ainsi, la régression logistique agit comme un modèle linéaire de classification ; ce qui est en soi une problématique de ce choix de modélisation étant donné que la linéarité entre les prédicteurs ne peut pas être toujours vérifiée. Malgré cette difficulté, la régression logistique demeure l'un des outils statistiques les plus utilisés dans les domaines de recherches appliquées pour plusieurs raisons. Pour commencer, en plus d'être utilisée pour une raison d'habitudes, il faut mentionner que la fonction logit est appelée à jouer un rôle important dans l'analyse des tables de contingences. Effectivement, Barnard[40] a apporté la notion de *log odds*, quantité interpolée par la régression logistique et pour laquelle l'étude des ratios dans un tableau permet de déterminer les probabilités d'intérêts quant à chaque catégorie dans une étude comparative. Ensuite, cette régression demeure liée de près aux distributions de la famille exponentielle ; famille de modèles fréquemment utilisés dans un contexte bayésien en épidémiologie où la probabilité d'un vecteur $s$ est proportionnelle à la quantité $\beta_0 + \sum_{j=1}^{m} f_j(s)\beta_j$. En effet, si une des composantes de $s$ est binaire et que les fonctions $f_j$ sont toutes des indicatrices, on obtient une régression logistique. En réalité, on note que la fonction logistique est une fonction linéaire généralisée appartenant à la famille exponentielle.

Une des problématiques principale auxquelles ce chapitre de thèse tente d'apporter une piste de solution est de déterminer s'il y a un outil statistique stable et relativement parcimonieux arrivant à répondre aux questions auxquelles on soumet les modèles de régression logistique et qui soit uniquement basé sur les données et non sur un modèle imposé par la structure même de ces données. Ainsi, la linéarité supposée entre les variables prédictives que propose le modèle de

régression logistique peut être contournée par un tel outil ; ce qui sera montré dans les sections suivantes.

**Réseaux de neurones artificiels**

Les réseaux de neurones artificiels sont actuellement considérés dans le domaine de la recherche clinique comme étant l'alternative principale à la régression logistique dans une perspective de prédiction d'une réponse binaire sachant un ensemble de covariables d'intérêt. D'ailleurs, Guerriere[41] et Hinton[42] ont présenté cette nouvelle technique comme étant dérivée des sciences informatiques, se configurant davantage sur l'expérience basée sur les similitudes et les divergences dans les données que sur une analyse statistique paramétrée. Ainsi, les réseaux de neurones artificiels sont un groupement d'algorithmes modélisés à l'image du cerveau humain pour lesquels la finalité tout comme les conditions préalables d'application sur les données demeurent les mêmes que dans le cas de la régression logistique.

Un réseau neuronal typique est composé de trois niveaux d'information : le niveau des entrants ; soit le niveau où il y a prise en compte des covariables d'intérêt, le niveau «caché», soit le niveau où le traitement de l'information se fait, et le niveau des extrants : la sortie de la variable réponse. Ce qui est d'intérêt ici est de déterminer la qualité du modèle au niveau des extrants et la fiabilité de prédiction de la variable réponse. Pour y arriver, il faut définir ce qui relie les différents niveaux du modèle, soient les poids de connexion, c'est-à-dire l'équivalent des coefficients $\boldsymbol{\beta}$ de la régression logistique contenant la «connaissance» du modèle après entrainement. Cet entrainement se fait en assignant des poids aléatoires a priori, et en expérimentant le système en cherchant le réseau minimisant l'erreur globale, calculée similairement à la notion statistique d'erreur quadratique moyenne.

Les paramètres que constituent en fait les noeuds de connexion sont toujours en quantité supérieure au nombre de paramètres qui composent une régression logistique. Prenons, par exemple, un système avec deux covariables pour lesquelles l'on veut déterminer une réponse. Avec la régression logistique, il y aura en tout trois paramètres à estimer $(\beta_0, \beta_1, \beta_2)$ alors que pour un réseau neuronal, il y en aura 9. On peut ainsi voir les nets désavantages d'utiliser cet outil informatique : pour commencer, il est clair que la capacité computationnelle exigée pour ce type de système est beaucoup plus imposante que dans un cadre statistique classique. Par ailleurs, en se basant sur des algorithmes de minimisation de l'erreur de prédiction, le risque de surapprentissage à partir des intrants est maximisé ; ce qui cause une difficulté d'adaptation du modèle à des données provenant d'un environnement de saisie différent. Enfin, l'élément causant le plus de craintes quant à l'utilisation des réseaux de neurones artificiels est qu'ils sont en fait des «boîtes noires» avec une capacité limitée à détecter les possibles relations causales. En fait, ce modèle tend à contenir un nombre élevé de variables prédictives non-pertinentes et difficilement détectables pour l'analyste ; ce qui en fait un modèle à proscrire lorsque l'on suppose être en présence de variables de confusion dans un ensemble de données observationnelles.

## 3.2 Données discrètes et fonction copule

L'objectif primaire de cette section est de présenter le cadre limitatif quant à l'utilisation de la fonction copule avec des données discrètes, quel que soit le type de discrétisation. De façon secondaire, cette section veut amener le lecteur à la compréhension de l'impact d'une utilisation stricte de copules classiques sur de telles données et, par conséquent, d'une mauvaise caractérisation de la dépendance de ces données à travers le choix de la modélisation. Afin d'être cohérent avec le reste de cette thèse et dans le but de présenter les pistes de solutions proposées dans le section suivante sans ambivalence, nous nous limiterons à caractériser la dépendance à partir de la mesure du tau de Kendall. Pour toute question d'approfondissement quant à ce sujet, ou pour étendre les résultats présentés au coefficient du rho de Spearman, le lecteur est invité à se référer à Genest et al.[43] qui est considéré comme l'article de référence pour les copules avec les données discrètes.

Débutons par un exemple simple : notons $X$ et $Y$ deux variables aléatoires, de paramètres $\mu$ et $\nu$ respectivement, qui suivent une loi de Poisson pour lesquelles il est légitime de supposer une certaine dépendance. Alors, il est vraisemblable d'émettre la supposition que la copule modélisant la fonction de répartition jointe entre $X$ et $Y$ appartient à la famille de Clayton. Ainsi, on obtient la fonction de répartition jointe

$$\mathbb{P}(X \leqslant x, Y \leqslant t) = \left[F_{\mu}^{-\theta}(s) + F_{\nu}^{-\theta}(t) - 1\right]^{-1/\theta}$$

pour $s, t \in \mathbb{N}$ et $\theta$ étant le paramètre de dépendance de la copule, $\theta \in [-1, \infty)\backslash\{0\}$. Il est donc possible d'estimer les paramètres des fonctions marginales et le paramètre de la copule $\theta$ à partir des observations $(X_j, Y_j)$, $j = 1, 2, ..., n$ par des méthodes statistiques de base telles que celles maximisant la vraisemblance. Sinon, il est possible de déterminer $\theta$ par la méthode d'inversion du tau de Kendall ; ce qui nécessitera justement l'estimation de cette mesure de dépendance. Cependant, plusieurs problèmes sont à considérer pour cette modélisation. Pour commencer, sans ajustement, le tau de Kendall peut prendre une valeur à l'extérieur de son support de définition $[-1, 1]$. Par ailleurs, tel qu'illustré plus loin, un problème d'unicité de la copule se présente pour de telles données.

### 3.2.1 Non-unicité de la copule avec des marges discrètes

Soit $H$ la fonction de répartition jointe entre deux variables aléatoires $X$ et $Y$ de marges $F(x)$ et $G(y)$ telle que , $\forall x, y \in \mathbb{R}$,

$$H(x, y) = \mathbb{P}(X \leqslant x, Y \leqslant y),\ F(x) = \mathbb{P}(X \leqslant x),\ G(y) = \mathbb{P}(Y \leqslant y).$$

L'inverse généralisé pour chaque fonction de distribution marginale, qui est continu à gauche $\forall u \in (0, 1]$, est défini ainsi par

$$F^{-1}(u) = inf\{x \in \mathbb{R} : F(x) \geqslant u\}$$

et de façon similaire pour $G^{-1}(v)$. Afin d'amener le lecteur à bien saisir la problématique analytique des copules discrètes, il sera présenté le concept de *sous-copules* ; concept complémentaire à la présentation générale du cadre des copules de l'introduction.

**Sous-copule :** Une sous-copule de dimension 2 (ou 2-sous-copule) est une fonction notée $C'$ munie des propriétés :

1. Dom $C' = S_1 \times S_2$ où $S_1$ et $S_2$ sont des sous-ensembles du carré unitaire contenants les valeurs de 0 et de 1

2. $C'$ est 2-croissante et à l'instar de la 2-copule, $C'(0,v) = C'(u,0) = 0$ et $C'(1,v) = v$ ; $C'(u,1) = u$

On note que pour tout $(u,v)$ appartenant au domaine de $C'$, $0 \leqslant C'(u,v) \leqslant 1$ et donc que le support de $C'$ est aussi un sous-ensemble du carré unitaire. Cette définition amène à la réécriture du théorème de Sklar en fonction de la sous-copule selon le lemme suivant.

**Lemme 3.2.1.** ***[lemme 2.3.4 de Nelsen, 2006[12]] :*** *Soit $H$ une fonction de distribution cumulative avec des marges $F$ et $G$. Alors, il existe une sous-copule unique $C'$ telle que*

1. *Dom $C'$ = Ran $F$ × Ran $G$ où Ran est l'image (Range)*
2. *Pour tout $x, y \in \bar{\mathbb{R}}$, $H(x,y) = C'(F(x), G(y))$ où $\bar{\mathbb{R}}$ représente la ligne des nombre réels extentionnée (i.e. $[-\infty, \infty]$).*

**Démonstration :**
La distribution jointe $H$ satisfait, lorsque $S_1$=$S_2$=$\bar{\mathbb{R}}$, pour tout point $(x_1, y_1)$ et $(x_2, y_2)$ en $\bar{\mathbb{R}}^2$, l'inéquation

$$|H(x_2,y_2) - H(x_1,y_1)| \leqslant |F(x_2) - F(x_1)| + |G(y_2) - G(y_1)| .$$

Il s'ensuit que si $F(x_1) = F(x_2)$ et $G(y_1) = G(y_2)$, alors $H(x_1,y_1) = H(x_2,y_2)$ et que l'ensemble de couples ordonnés

$$\left\{((F(x), G(y)), H(x,y)) \,|x, y \in \bar{\mathbb{R}}\right\}$$

définit une fonction réelle $C'$ ayant pour domaine *Ran* $F$ × *Ran* $G$. Que cette fonction soit une sous-copule vient directement des propriétés de $H$. Par ailleurs, pour montrer que les conditions de la définition de la sous-copule sont vérifiées, on remarque que pour tout $u \in$ *Ran* $F$, il existe $x \in \bar{\mathbb{R}}$ tel que $F(x) = u$. Alors, $C'(u,1) = C'(F(x), G(\infty)) = H(x,\infty) = F(x) = u$. La vérifications des autres conditions de la définition précédente se fait de façon similaire.

□

## 3.3 Réponses dans la littérature aux limitations de la copule discrète

On commence cette section par le lemme 2.3.5 de Nelsen[12] qui complète le lemme 2.3.4 du même auteur présenté à la section précédente et qui permet de se prononcer sur l'extension par continuité d'une sous-copule.

**Lemme 3.3.1.** ***[lemme 2.3.5 de Nelsen, 2006[12]] :*** *Soit $C'$ une sous-copule. Alors il existe une copule $C$ qui est une extension de la sous-copule d'origine telle que $C(u,v) = C'(u,v) \forall (u,v) \in Dom(C')$. Cette extension est généralement non-unique.*

**Démonstration :**

Soit $Dom(C') = S_1 \times S_2$. En utilisant la propriété Lipschitz de la copule appliquée aux sous-copules, on peut étendre par continuité $C'$ à une fonction $C''$ avec le domaine $Dom(C'') = \bar{S_1} \times \bar{S_2}$ où $\bar{\bullet}$ dénote l'adhérence d'un ensemble. Il est évident que $C''$ est une sous-copule. Ensuite, on peut étendre $C''$ à une copule $C$ ayant pour support $[0,1] \times [0,1]$. Soit $(a,b)$, n'importe quel point de ce support et soient $a_1$ et $a_2$ respectivement le plus grand et le plus petit élément de $\bar{S_1}$ qui peuvent satisfaire $a_1 \leqslant a \leqslant a_2$. Par ailleurs, soient les éléments $b_1$ et $b_2$ également respectivement le plus grand et le plus petit élément de $\bar{S_2}$ qui peuvent satisfaire $b_1 \leqslant b \leqslant b_2$. On remarque que si $a \in \bar{S_1}$, alors $a_1 = a = a_2$ ; et on fait la même remarque pour $b$. Maintenant, on définit

$$\lambda_1 = \begin{cases} \frac{a-a_1}{a_2-a_1} & \text{si } a_1 < a_2, \\ 1 & \text{si } a_1 = a_2; \end{cases}$$

et

$$\mu_1 = \begin{cases} \frac{b-b_1}{b2-b_1} & \text{si } b_1 < b_2, \\ 1 & \text{si } b_1 = b_2. \end{cases}$$

Alors, on obtient l'interpolation bilinéaire $C$ de la sous-copule $C'$

$$\begin{aligned} C(a,b) &= (1-\lambda_1)(1-\mu_1)C''(a_1,b_1) + (1-\lambda_1)\mu_1 C''(a_1,b_2) \\ &\quad +\lambda_1(1-\mu_1)C''(a_2,b_1) + \lambda_1\mu_1 C''(a_2,b_2). \end{aligned}$$

On remarque que cette interpolation est linéaire en chaque point. Enfin, il est aisé de remarquer que $C$ est une copule satisfaisant l'ensemble des conditions du théorème de Sklar.

□

La procédure d'extension d'une sous-copule présentée dans cette démonstration du lemme 2.3.5 de Nelsen est, en fait, la principale réponse dans la littérature aux marges discrètes ; plusieurs publications étant des adaptations de cette dernières.[44,45] Cette interpolation comporte

des difficultés pour l'utilisateur : principalement quant à l'implémentation computationnelle qui nécessite le calcul de 4 sous-copules pour trouver une seule copule dans le cas bivarié.

Une autre voie rencontrée dans la littérature pour ce type de données est l'approche probabiliste de Denuit et Lambert.[44] Soit $X$, une variable aléatoire discrète appartenant à un sous-ensemble non-négatif $X$ de l'ensemble $\mathbb{N}$. Alors, pour $U$, une variable aléatoire continue portée par le domaine $(0,1)$ (la distribution proposée dans ces travaux étant la loi uniforme)$X$, $U$ ayant une fonction de répartition strictement croissante $L_U \in [0,1]$ et indépendante de $X$, il est proposé, par cette approche, de rendre continue la variable aléatoire $X$ en créant la nouvelle variable aléatoire

$$X^\star = X + (U - 1).$$

Ainsi, $X^\star \leqslant X$. En prenant $\langle s \rangle$, la valeur entière de $s \in \mathbb{R}$, $\forall s \in \mathbb{R}$,

$$F^\star(s) = F(\langle s \rangle) + L_U(s - \langle s \rangle) f_{\langle s+1 \rangle}$$

où, $f$ étant une densité, est définie $f_{\langle s+1 \rangle} = \mathbb{P}\left[X = \langle s + 1 \rangle\right]$.

Cette approche probabiliste fait appel à la loi des grands nombres et nécessite des échantillons de taille considérable pour obtenir une convergence exacte de par l'addition d'un bruit provenant de la variable aléatoire $U$, ce qui n'est pas toujours possible dans un contexte de données cliniques.

## 3.4 Alternative proposée et application au score de propension

### 3.4.1 Introduction à l'alternative et généralités

Soit la variable stochastique $T \in \{0,1\}$, une variable binaire représentant l'affectation clinique des individus participant à une étude observationnelle et soumis à un traitement spécifique (i.e. $T = 1$ si un individu prend la thérapeutique visée par l'étude, et $T = 0$ si l'individu a été affecté au bras placebo ou à la thérapie préexistente à l'essai), et soit $\boldsymbol{Z} = (Z_1, Z_2, ..., Z_d)^t$ étant un vecteur de dimension $d$ constitué de $d$ variables d'intérêt. Ces variables d'intérêt peuvent être continues ou discrètes, mais la majeure partie du temps, elles sont discrètes. Étant confronté à une étude observationnelle et, par conséquent, à l'absence de randomisation dans l'assignation des individus à la thérapie objet de l'étude, les cliniciens doivent être en mesure d'équilibrer les deux bras de l'essai thérapeutique afin d'obtenir la plus petite différence possible quant aux informations cliniques pertinentes relatives aux participants entre les deux groupes pour contrôler le biais de sélection. Pour y arriver, Rosenbaum et Rubin[46] proposent l'utilisation d'un score pouvant être calculé pour chaque individu prenant part à l'étude : le score de propension, défini par :

$$e(\boldsymbol{z}) = \mathbb{P}[T = 1 | Z_1 = z_1, Z_2 = z_2, ..., Z_d = z_d].$$

Un tel score est alors utilisé dans une analyse d'équilibration des groupes. Plusieurs types d'analyses d'équilibration peuvent être effectuées tels qu'une utilisation pondérale sur les effectifs, une discrimination non-paramétrique, etc. (voir Guo[47]). Toutefois, on remarque que dans la majeure

partie des cas, c'est un processus d'appariement qui est appliqué. En fait, le principe de l'appariement est de déterminer l'effet traitement moyen (*Average Treatment Effect*, ATE) sachant l'issue clinique (*outcome*) dans chaque bras $\zeta_T$ et le score de discrimination qu'est le score de propension pour conserver uniquement les individus appariés (connaissant un niveau de décision prédéterminé) dans l'estimation de l'ATE tel que :

$$ATE = \mathbb{E}\left[\zeta_{T=1} - \zeta_{T=0} | e(\boldsymbol{z})\right].$$

De cette façon, l'estimation du score de propension affecte directement l'ATE. Rosenbaum et al.,[46] suggèrent l'utilisation du logit de la probabilité prédite au lieu de la dite probabilité elle-même comme score de propension car il est alors possible de le considérer comme une distribution de probabilité suivant une loi logit-normale. Toutefois, on remarque que l'utilisation de la régression logistique peut, en fait, mener à certains problèmes de modélisation. Pour commencer, il est clair que la fonction de régression logistique peut avoir un maximum parce que sa log-vraisemblance est globalement concave. Par contre, il est fréquent de rencontrer des situations où la fonction de vraisemblance n'a aucun maximum (voir Amemiya[48]). Dans un tel cas, les estimateurs du maximum de vraisemblance n'existent pas. Ensuite, l'utilisation de la régression logistique en présence de données observationnelles est inopportune si les variables de confusion dans le modèle sont inconnues. En fait, dans un contexte d'études cliniques, il arrive qu'il y ait des variables explicatives qui apparaissent plus qu'une fois dans le sens où elles peuvent être observées à travers d'autres variables explicatives. En l'absence d'indépendance entre ces covariables, le problème de la colinéarité apparaît ; ce qui, tel que montré dans la littérature, est profondément contraignant.[49–52] Les points principaux ici sont qu'un tel problème va causer une distortion dans l'interprétation du modèle en diminuant son exactitude (biais constaté sur les coefficients de régression). Cette section propose une approche qui tend à éviter complètement l'utilisation de la régression logistique afin de déterminer le score de propension en s'appuyant uniquement sur la dépendance induite entre la variable de l'allocation au traitement et l'ensemble des covariables d'intérêt.

### 3.4.2 Modèle

Le lecteur est invité à remarquer que la notation adoptée dans ce qui suit est basée sur le lemme 3.2.1. Par ailleurs, afin de restreindre le nombre de covariables dans un cadre d'essais thérapeutiques et dans le but de définir un modèle facilement applicable pour le clinicien, on doit émettre les hypothèses suivantes. La première concerne les covariables d'intérêt à utiliser dans le modèle présenté ici et proviennent des conclusions d'Austin.[53]

**Hypothèse 3.4.1.** *Les covariables d'intérêt pour déterminer le score de propension sont seulement celles qui affectent l'issue clinique (outcome) et celles qui affectent à la fois l'assignation au traitement et l'issue clinique.*

Cette hypothèse permet de déterminer le nombre de distributions marginales qui vont être dans le score de propension basé sur la fonction copule. Pour le reste de ce chapitre, les notions sont présentées dans le but d'appliquer la procédure schématisée à la figure 3.3 ; en d'autres mots, pour rendre continue une sous-copule étant unique seulement sur l'image du croisement des fonctions de distributions marginales la constituant, afin de déterminer une copule unique qui joint ces variables discrètes et, finalement, de construire le "réseau" de copules nécessaire pour calculer la probabilité conditionnelle entre les variables $T$ et $\boldsymbol{Z}$ qui constituent le score de propension.

#### 3.4.2.1 Unicité de la sous-copule $C'$ dont les marges ont été rendues continues

Cette sous-section est construite de façon à obtenir une modélisation finale basée sur l'unique copule bivariée $C^{\{\epsilon_1,\epsilon_2\}}$ pour laquelle sa mesure de dépendance est égale à celle attribuable au couple $(T, Z_k), k \in \{1, 2, ..., d\}$ ; e.g. $\tau(T, Z_k) = \tau(C^{\{\epsilon_1,\epsilon_2\}})$. Pour y arriver, on doit introduire un cadre théorique concernant les éléments ayant la plus grande influence sur la structure de la copule : les fonctions de répartition marginales.

Pour commencer, l'hypothèse qui suit concerne la nature des marges pour toutes les variables d'intérêt (variable traitement et covariables affectant l'issue clinique). Cette dernière permet l'implémentation d'une copule unique dans le modèle.

**Hypothèse 3.4.2.** *Chaque variable qui constitue le score de propension peut être approchée par une distribution continue, même s'il s'agit d'une variable discrète.*

Cette hypothèse mène à l'implémentation de nouvelles variables "artificielles" qui sont, en fait, les transformations de variables aléatoires discrètes en variables continues. Soit $Z_k, k \in \{1, 2, ..., d\}$ une covariable d'intérêt pour le score de propension. Soit $\epsilon$ étant n'importe quel élément de $\{\epsilon_1, \epsilon_2\}$, l'ensemble contenant $\epsilon_1$, une fonction de $n$ où $n \to \infty$ ; et $\epsilon_2$, une fonction de $m$ où $m \to \infty$. On remarque que $n$ peut être de la même taille que $m$ sans, toutefois, que $\epsilon_1$ et $\epsilon_2$ ne le soient. Alors, on obtient la nouvelle variable aléatoire continue $Z_k^{\epsilon_2}$ issue de la fonction de distribution cumulative $G_k^{\epsilon_2}(z_k)$ qui est, en fait, une version continue de $G_k(z_k)$ telle que

$$G_k^{\epsilon_2}(z_k) = \begin{cases} G_k(z_k) & \text{si } z_k \in \left(\bigcup_{i=1}^{I}(x_{i-1} + \epsilon_2^2 < z_k < x_i - \epsilon_2)\right); \\ a_i z_k + b_i & \text{si } z_k \in \left(\bigcup_{i=0}^{I}[x_i - \epsilon_2, x_i + \epsilon_2^2]\right), \end{cases}$$

où, tel qu'illustré à la figure 3.1, $x_i$, $i \in \{1, 2, ..., I\}$ et $x_0 = 0$, représente la localisation en abcisse des sauts entre les différentes modalitées de la variable aléatoire, $\epsilon_2$ un terme arbitraire qui tend vers 0 et $a_i z_k + b_i$ est l'équation de la ligne de correction qui lie les modalités de la variable catégorielle dans un voisinage de la locatisation des sauts. Étant donné que $a_i$ et $b_i$ représentent respectivement la pente et l'origine d'une droite, on en déduit aisément leur représentation analytique par quelques manipulations telles que

$$a_i = \frac{\gamma_i}{\epsilon_2^2 + \epsilon_2}$$

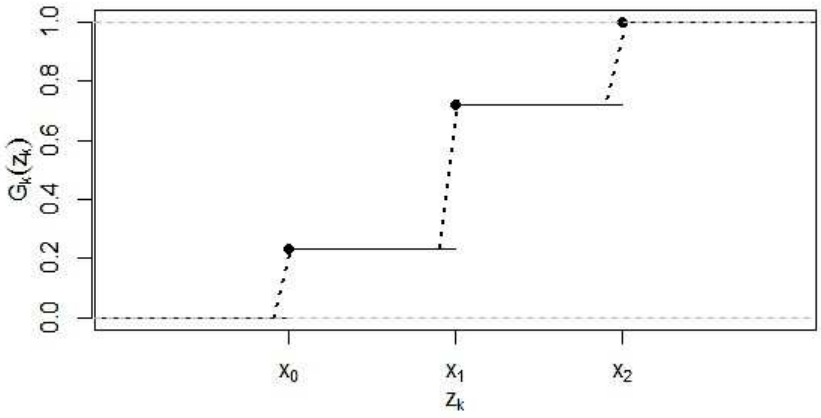


FIGURE 3.1 – Fonction de répartition d'une variable aléatoire simulée $Z$ comportant 3 modalités avec des corrections de continuité pour un $\epsilon$ de $1/250$, créant ainsi la nouvelle variable aléatoire $Z_k^{\epsilon}$.

où $\gamma_i = G_k(x_i^+) - G_k(x_i^-)$ est la hauteur du saut entre $G_k(x_i^-)$, la valeur à gauche de $G_k$ au point $x_i$, et $G_k(x_i^+)$ la valeur à droite de $G_k$ évaluée à ce même point ; et

$$b_i = G_k(x_i^-) - \frac{\gamma_i}{\epsilon_2^2 + \epsilon_2}(x_i - \epsilon_2).$$

On note que la même transformation de continuité est à effectuer avec la variable d'allocation au traitement, $T$, et sa fonction de distribution cumulative $F(t)$. Dans un but de simplicité de la notation et de clarté des explications, pour la suite de cette section, on se limitera au cas où n'il y a qu'une seule covariable d'intérêt (le cas multivarié sera traité plus tard), $Z$, qui est discrète. Alors, la variable aléatoire continue correspondante est $Z^{\epsilon_2}$ et leurs fonctions de distributions cumulatives sont, respectivement, $G(z)$ et $G^{\epsilon_2}(z)$. De façon correspondante, on a, pour la variable discrète d'assignation au traitement $T$, la variable continue $T^{\epsilon_1}$ et sa fonction de répartition $F^{\epsilon_1}(t)$. On précise que l'on utilise une correction de continuité sur l'invervalle $[x_i - \epsilon, x_i + \epsilon^2]$ au lieu de l'intervalle symétrique $[x_i - \epsilon, x_i + \epsilon]$ en raison que la fonction de distribution jointe est càdlàg ; ce qui signifie que lorsque $z_k = x_i$, on s'assure que $G^{\epsilon_2}(z_k)$ converge ponctuellement vers $G(z_k)$. Afin d'arriver à utiliser les nouvelles variables aléatoires $T^{\epsilon_1}$ et $Z^{\epsilon_2}$, on pose la proposition suivante :

**Proposition 3.4.1.** *$G^{\epsilon_2}(z)$ converge en distribution vers $G(z)$ lorsque $\epsilon^2$ converge vers 0.*

**Démonstration :**

On doit commencer par vérifier, pour chaque valeur de $z$, la différence entre la fonction $G$ et la fonction $G^{\epsilon_2}$

$$G^{\epsilon_2}(z) - G(z) = \begin{cases} 0 & \text{si } z \in \left(\bigcup_{i=1}^{I}(x_{i-1} + \epsilon_2^2 < z < x_i - \epsilon_2)\right); \\ \frac{\gamma_i}{\epsilon_2^2+\epsilon_2} z - \frac{\gamma_i}{\epsilon_2^2+\epsilon_2}(x_i - \epsilon_2) - \gamma_i & \text{si } z \in \bigcup_{i=0}^{I}[x_i, x_i + \epsilon_2^2]; \\ \frac{\gamma_i}{\epsilon_2^2+\epsilon_2} z - \frac{\gamma_i}{\epsilon_2^2+\epsilon_2}(x_i - \epsilon_2) & \text{si } z \in \bigcup_{i=0}^{I}[x_i - \epsilon_2, x_i). \end{cases}$$

La convergence de $G^{\epsilon_2}(z)$ vers $G(z)$ est évidente dans le premier cas. Dans le deuxième cas ( lorsque $z \in \bigcup_{i=0}^{I}[x_i, x_i + \epsilon_2^2]$ ), si $z = x_i$, alors considérant la limite lorsque $\epsilon_2$ tend vers 0, on a :

$$\begin{aligned}\lim_{\epsilon_2 \to 0} |G^{\epsilon_2}(z) - G(z)| &= \lim_{\epsilon_2 \to 0} \left| \frac{\gamma_i}{\epsilon_2^2 + \epsilon_2} x_i - \frac{\gamma_i}{\epsilon_2^2 + \epsilon_2}(x_i - \epsilon_2) - \gamma_i \right| \\ &= \left| \gamma_i \lim_{\epsilon_2 \to 0} \frac{\epsilon_2}{\epsilon_2^2 + \epsilon_2} - \gamma_i \right| \\ &= 0.\end{aligned}$$

Si $z = x_i + \epsilon_2^2$, on a alors :

$$\begin{aligned}\lim_{\epsilon_2 \to 0} |G^{\epsilon_2}(z) - G(z)| &= \lim_{\epsilon_2 \to 0} \left| \frac{\gamma_i}{\epsilon_2^2 + \epsilon_2}(x_i + \epsilon_2^2) - \frac{\gamma_i}{\epsilon_2^2 + \epsilon_2}(x_i - \epsilon_2) - \gamma_i \right| \\ &= \left| \gamma_i \left( \frac{\epsilon_2^2 + \epsilon_2}{\epsilon_2^2 + \epsilon_2} \right) - \gamma_i \right| \\ &= 0.\end{aligned}$$

Donc, en utilisant le résultat du lemme de Portmanteau, il est montré que $\forall z \in \bigcup_{i=0}^{I}[x_i, x_i + \epsilon_2^2]$, $\lim_{\epsilon_2 \to 0} |G^{\epsilon_2}(z) - G(z)| = 0$.

Dans le dernier cas, lorsque $z \in \bigcup_{i=0}^{I}[x_i - \epsilon_2, x_i\ )$, on doit évaluer :

$$\begin{aligned}\lim_{\epsilon_2 \to 0} |G^{\epsilon_2}(z) - G(z)| &= \lim_{\epsilon_2 \to 0} \left| \frac{\gamma_i}{\epsilon_2^2 + \epsilon_2} z - \frac{\gamma_i}{\epsilon_2^2 + \epsilon_2}(x_i - \epsilon_2) \right| \\ &= \left| \gamma_i + \gamma_i \lim_{\epsilon_2 \to 0} \frac{z - x_i}{\epsilon_2^2 + \epsilon_2} \right|.\end{aligned}$$

On remarque que $-\epsilon_2 \leqslant z - x_i < 0$. Cependant, il existe un $\epsilon_2^*$ qui tend également vers 0 tel que $-\epsilon_2 \leqslant z - x_i < -\epsilon_2^*$. Alors, $\lim_{\epsilon_2 \to 0} \frac{z - x_i}{\epsilon^2 - \epsilon} = -1$ , et

$$\left| \gamma_i + \gamma_i \lim_{\epsilon_2 \to 0} \frac{z - x_i}{\epsilon_2^2 + \epsilon_2} \right| = 0.$$

Donc, $\forall z \in Dom(G)$,

$$\text{si } \epsilon_2 \to 0 \Rightarrow G^{\epsilon_2}(z) \to G(z).$$

□

Pour l'intuition du lecteur, on rappelle que la fonction de distribution $G(z)$ est càdlàg aux points de discontinuité $x_i$. Ainsi, lorsque $\epsilon_2 \to 0$, au point $x_i$, la limite évaluée à la droite est la valeur de la modalité sise à la droite de ce point.

Par ailleurs, on remarque que les probabilités mesurées avec les nouvelles distributions continues sont équivalentes à celles qui sont originellement d'intérêt. En fait, la différence principale est sur la représentation de la densité : au lieu d'être définie en $k$ intervalles rectangulaires (où

$k$ est le nombre de modalitées de la variable aléatoire discrète originelle), la densité va être représentée sur $I$ intervalles supplémentaires non-rectangulaires de taille $\epsilon$ qui correspondent aux points de discontinuité dans la représentation de la fonction de distribution cumulative.

Suivant le raisonnement de Genest et al.,[43] il est évident que lorsque $F$ et $G$ (fonctions de répartition discrètes) ont des sauts, leurs inverses ont des plateaux. Ainsi, le théorème de Sklar garantit qu'il existe au moins une représentation de copule pour $\mathbb{P}(T \leqslant t, Z \leqslant z)$, mais que cette dernière n'est pas unique. Alors, on note $\boldsymbol{C}$ l'ensemble de toutes les possibles familles de copules qui représentent la fonction de distribution jointe de $F$ et $G$.

**Proposition 3.4.2.** *Soit $C^{\{\epsilon_1,\epsilon_2\}}$, la copule qui converge ponctuellement (faiblement) vers $C'$, qui est la sous-copule qui exprime la fonction de distribution jointe du couple de variables aléatoires $(T, Z)$ ; si le couple $(T^{\epsilon_1}, Z^{\epsilon_2})$ converge ponctuellement vers $(t, z)$ ; alors $C^{\{\epsilon_1,\epsilon_2\}}(t^{\epsilon_1}, z^{\epsilon_2})$ converge vers $C'(t, z)$. On en déduit que $C^{\{\epsilon_1,\epsilon_2\}}$ est une des copules qui expriment la fonction de distribution jointe du couple $(T, Z)$ telle que $C^{\{\epsilon_1,\epsilon_2\}} \in \boldsymbol{C}$.*

**Démonstration :**

Soit $(t, z)$, n'importe quelle valeur appartenant au domaine des fonctions de répartition marginales $F_T, G_Z$ ; et $(t^{\epsilon_1}, z^{\epsilon_2})$ leur réciproque dans le domaine des fonctions de répartitions marginales $F^{\epsilon_1}_{T^{\epsilon_1}}, G^{\epsilon_2}_{Z^{\epsilon_2}}$. On définit la variable aléatoire $T^{\epsilon_1}$ (et similairement $Z^{\epsilon_2}$), avec des limites vers $\epsilon$ telles que $n = k/\epsilon_1, m = s/\epsilon_2$, $k \to 1$, $s \to 1$, et où, à des vitesses de convergence différentes, $n \to \infty$, $m \to \infty$, telle que

- $\mathbb{P}(T^{\epsilon_1} \leqslant t) = F^{\epsilon_1}(t)$ ;
- Si $t \in (-\infty, 0)$, $\mathbb{P}(T^{\epsilon_1} \leqslant t) = 0$ ;
- Si $t \in [t_i - \epsilon, t_i + \epsilon_1^2], i = 0, 1, 2, ...$, $\mathbb{P}(T^{\epsilon_1} \leqslant t) = F^{\epsilon_1}(t) - F^{\epsilon_1}(t - \epsilon)$ ;
- Si $t \in (t_i + \epsilon, t_{i+1} - \epsilon_1), \mathbb{P}(T^{\epsilon_1} \leqslant t) = \mathbb{P}(T \leqslant t) = F(t)$.

En prenant la valeur absolue de la différence entre $C^{\{\epsilon_1,\epsilon_2\}}(t^{\epsilon_1}, z^{\epsilon_2})$ et $C'(t, z)$, on obtient

$$\begin{aligned}\left|C^{\{\epsilon_1,\epsilon_2\}}(t^{\epsilon_1}, z^{\epsilon_2}) - C'(t, z)\right| &= \left|C^{\{\epsilon_1,\epsilon_2\}}(t^{\epsilon_1}, z^{\epsilon_2}) - C^{\{\epsilon_1,\epsilon_2\}}(t, z) + C^{\{\epsilon_1,\epsilon_2\}}(t, z) - C'(t, z)\right| \\ &\leqslant \left|C^{\{\epsilon_1,\epsilon_2\}}(t^{\epsilon_1}, z^{\epsilon_2}) - C^{\{\epsilon_1,\epsilon_2\}}(t, z)\right| + \left|C^{\{\epsilon_1,\epsilon_2\}}(t, z) - C'(t, z)\right|\end{aligned}$$

Cependant, la propriété Lipschitz de la fonction copule stipule que

$$\left|C^{\{\epsilon_1,\epsilon_2\}}(t^{\epsilon_1}, z^{\epsilon_2}) - C^{\{\epsilon_1,\epsilon_2\}}(t, z)\right| \leqslant |t^{\epsilon_1} - t| + |z^{\epsilon_2} - z|\,.$$

Alors, on obtient

$$\left|C^{\{\epsilon_1,\epsilon_2\}}(t^{\epsilon_1}, z^{\epsilon_2}) - C'(t, z)\right| \leqslant |t^{\epsilon_1} - t| + |z^{\epsilon_2} - z| + \left|C^{\{\epsilon_1,\epsilon_2\}}(t^{\epsilon_1}, z^{\epsilon_2}) - C'(t, z)\right|.$$

Actuellement, vu que $\epsilon$ est une fonction de $n$ ou de $m$, $\forall \delta > 0$,

- $t^{\epsilon_1} \to t$ alors, pour $\delta/3$, il existe $n_0$ tel que $n \geqslant n_0 \Rightarrow |T^{\epsilon_1} - t| \leqslant \delta/3$ ;
- $z^{\epsilon_2} \to z$ alors, pour $\delta/3$, il existe $n_1$ tel que $m \geqslant n_1 \Rightarrow |Z^{\epsilon_2} - z| \leqslant \delta/3$.

Étant donné que $C^{\{\epsilon_1,\epsilon_2\}} \to C'$ faiblement, il existe, pour $x$ et $y$ des valeurs arbitraires fixées, $n_2$ tel que
$(n \wedge m) \geqslant n_2 \Rightarrow \left|C^{\{\epsilon_1,\epsilon_2\}}(t,z) - C'(t,z)\right| \leqslant \delta/3.$

Donc, $\forall \delta > 0, \forall (n \wedge m) \geqslant \max(n_0, n_1, n_2)$,

$$\left|C^{\{\epsilon_1,\epsilon_2\}}(t^{\epsilon_1}, z^{\epsilon_2}) - C'(t,z)\right| \leqslant \delta.$$

Ainsi, $C^{\{\epsilon_1,\epsilon_2\}}(t^{\epsilon_1}, z^{\epsilon_2})$ converge vers $C'(t,z)$.

□

On remarque aussi que suivant le raisonnement de la proposition 3.4.1, il est évident que la structure liant $F(t)$ à $G(z)$ est une sous-copule unique sur l'image du croisement de ses marges, soit $Ran\ F \times Ran\ G$. En utilisant le lemme 2.3.5 de Nelsen,[12] on voit qu'il est possible d'étendre, par continuité, la sous-copule $C'(F(t), G(z))$ à une copule $C$ qui appartient à l'ensemble de toutes les familles de copules possibles $\boldsymbol{C}$. Suivant les explications qui suivent, on en déduit que $C^{\{\epsilon_1,\epsilon_2\}}$ est l'une de ces copules.

On remarque aisément que $C^{\{\epsilon_1,\epsilon_2\}}$ est une copule : premièrement, $Dom\ C^{\{\epsilon_1,\epsilon_2\}} = [0,1] \times [0,1]$ ; ensuite $C^{\{\epsilon_1,\epsilon_2\}}(0,v) = C^{\{\epsilon_1,\epsilon_2\}}(u,0) = 0$ et $C^{\{\epsilon_1,\epsilon_2\}}(1,v) = v$ ; $C^{\{\epsilon_1,\epsilon_2\}}(u,1) = u$. Enfin, il est trivial de constater que $C^{\{\epsilon_1,\epsilon_2\}}$ est une fonction 2-croissante et est une interpolation bilinéaire de $C'$ à partir de la définition des structures marginales $F^{\epsilon_1}$ et $G^{\epsilon_2}$.

**Proposition 3.4.3.** *Si $C^{\{\epsilon_1,\epsilon_2\}} \to C'$ faiblement, alors :*

$$\lim_{m\to\infty} \lim_{n\to\infty} \tau(T^{\epsilon_1}, Z^{\epsilon_2}) = \tau(T,Z).$$

*où $\epsilon_1$ est une fonction de $n$ ($\epsilon_s = 1/n$), $\epsilon_2$ est une fonction de $m$ ($\epsilon_2 = k/m$),$s \to 1, k \to 1, n \to \infty, m \to \infty$ à des vitesses de convergence différentes, et $\tau(T^{\epsilon_1}, Z^{\epsilon_2})$ est le tau de Kendall lié à la copule $C^{\{\epsilon_1,\epsilon_2\}}(F^{\epsilon_1}(t), G^{\epsilon_2}(z))$.*

**Démonstration :**
À partir des paires de variables aléatoires $(T,Z)$ et $(T^{\epsilon_1}, Z^{\epsilon_2})$, on a les fonctions de répartition jointes $H_{(T,Z)}$ et $H_{(T^{\epsilon_1},Z^{\epsilon_2})}$.
Si les convergences faibles suivantes existent : $F^{\epsilon_1} \to F$, $G^{\epsilon_2} \to G$ et $H_{(T^{\epsilon_1},Z^{\epsilon_2})} \to H_{(T,Z)}$ ; alors, à partir du théorème 1 de Vandenhende et Lambert,[54] et des axiomes de concordance de Scarsini,[55]

$$\lim_{m\to\infty} \lim_{n\to\infty} \tau(T^{\epsilon_1}, Z^{\epsilon_2}) = \tau(T,Z).$$

Toutefois, on a

$$H_{T^{\epsilon_1},Z^{\epsilon_2}}(r,s) = C^{\{\epsilon_1,\epsilon_2\}}\left(F^{\epsilon_1}(r), G^{\epsilon_2}(s)\right).$$

On doit montrer que si $C^{\{\epsilon_1,\epsilon_2\}} \to C'$ ponctuellement, alors

$$C^{\{\epsilon_1,\epsilon_2\}}\left(F^{\epsilon_1}(r), G^{\epsilon_2}(s)\right) = C'\left(\lim_{n\to\infty} F^{\epsilon_1}(r), \lim_{n\to\infty} G^{\epsilon_2}(s)\right).$$

Alors, en utilisant les propriétés de la convergence faible et le lemme 2.3.4 de Nelsen[12] concernant le théorème de Sklar appliqué aux sous-copules, on obtient

$$H_{(T,Z)}(r,s) = C'\left(\lim_{n\to\infty} F^{\epsilon_1}(r), \lim_{m\to\infty} G^{\epsilon_2}(s)\right)$$

qui est une fonction de distribution cumulative jointe. Ainsi, on a montré que

$$\lim_{m\to\infty}\lim_{n\to\infty} \tau(T^{\epsilon_1}, Z^{\epsilon_2}) = \tau(T,Z).$$

□

Cette proposition assure, en fait, que l'extension de continuité proposée pour les fonctions de distribution marginales discrètes affecte seulement ces distributions. En d'autres mots, la correction de continuité liée à $\epsilon$ n'affecte pas les concordances et les discordances sur les variables $(T,Z)$ et ne va pas affecter le paramètre de dépendance $\theta$ sur les modèles de copules paramétriques qui est, relativement à la famille de copule sélectionnée, une fonction directe du tau de Kendall.

On remarque qu'il n'est pas possible de qualifier le degré d'adéquation de $C^{\{\epsilon_1,\epsilon_2\}}$ sur les données parmi toutes les copules qui composent l'ensemble $\boldsymbol{C}$ en raison du problème d'identifiabilité des éléments qui composent $\boldsymbol{C}$. Toutefois, via des simulations, il est possible de déterminer le comportement de $C^{\{\epsilon_1,\epsilon_2\}}$.

### 3.4.3 Réécriture du score de propension en terme de copules

Suivant la définition du score de propension[46] où $\boldsymbol{Z} = (Z_1, Z_2, ..., Z_d)^t$, à partir des variables aléatoires discrètes originelles, on écrit le PS pour un individu $i$, $i = 1, 2, ..., n$ assigné à une thérapie T=0,1 tel que :

$$\begin{aligned} e(z) &= \mathbb{P}(T=1|\boldsymbol{Z}=\boldsymbol{z}) \\ &= \frac{\mathbb{P}(T=1,\boldsymbol{Z}=\boldsymbol{z})}{\mathbb{P}(\boldsymbol{Z}=\boldsymbol{z})}. \end{aligned}$$

L'utilisation de sous-copules pour calculer ce score peut paraître attractif étant donné la non-nécessité de transformer les fonctions marginales des variables aléatoires. Toutefois, tel que mentionné précédemment dans cette section, ces sous-copules présentent un problème majeur quant à leur identifiabilité et leur unicité sur le carré unitaire. Alors, dans le but d'utiliser plutôt les variables aléatoires continues $(T_1^{\epsilon}, Z_2^{\epsilon})$ et leurs fonctions de répartition continues afin de calculer $e(z)$, on doit introduite $\epsilon_1', \epsilon_2', ..., \epsilon_{d+1}'$ qui sont des scalaires qui tendent vers 0 tels que :

$$\begin{aligned} \mathbb{P}(Z_2^{\epsilon}=t) &\equiv \lim_{\epsilon_2'\to 0} \mathbb{P}(Z_2^{\epsilon} \in [t-\epsilon_2', t+\epsilon_2']) \\ &= \lim_{\epsilon_2'\to 0} \left(G^{\epsilon_2}(t+\epsilon_2') - G^{\epsilon_2}(t-\epsilon_2')\right). \end{aligned}$$

On note que $\epsilon_2'$ peut être différent de $\epsilon_2$, mais ce n'est pas nécessairement le cas. Donc, pour une seule covariable d'intérêt, on écrit le score de propensiton tel que :

$$\begin{aligned} e(z) \equiv \lim_{\epsilon_1' \to 0} \lim_{\epsilon_2' \to 0} \Bigg( & \frac{1}{G^{\epsilon_2}(t+\epsilon_2') - G^{\epsilon_2}(t-\epsilon_2')} \Big[ C^{\{\epsilon_1,\epsilon_2\}}(F^{\epsilon_1}(t+\epsilon_1'), G^{\epsilon_2}(z+\epsilon_2')) \\ & - C^{\{\epsilon_1,\epsilon_2\}}(F^{\epsilon_1}(t+\epsilon_1'), G^{\epsilon_2}(z-\epsilon_2')) \\ & - C^{\{\epsilon_1,\epsilon_2\}}(F^{\epsilon_1}(t-\epsilon_1'), G^{\epsilon_2}(z+\epsilon_2')) \\ & + C^{\{\epsilon_1,\epsilon_2\}}(F^{\epsilon_1}(t-\epsilon_1'), G^{\epsilon_2}(z-\epsilon_2')) \Big] \Bigg). \end{aligned}$$

C'est ainsi qu'on remarque la présence de $2^{d+1}$ fonctions copules (*contraintes*) au numérateur, où $d$ est le nombre de covariables. Donc, dans le cas multivarié, il est nécessaires de réécrire en terme de copules ces $d$ covariables équivalentes à la probabilité jointe

$$e(z) \equiv \lim_{\epsilon' \to 0} \frac{\mathbb{P}\left(T^{\epsilon_1} \in [t-\epsilon_1', t+\epsilon_1'], Z_1^{\epsilon_2,} \in [z_1-\epsilon_2', z_1+\epsilon_2'], ..., Z_d^{\epsilon_{d+1},} \in [z_d-\epsilon_{d+1}', z_d+\epsilon_{d+1}']\right)}{\mathbb{P}\left(Z_1^{\epsilon_2,} \in [z_1-\epsilon_2', z_1+\epsilon_2'], ..., Z_d^{\epsilon_{d+1},} \in [z_d-\epsilon_{d+1}', z_d+\epsilon_{d+1}']\right)}.$$

On note, pour guider l'intuition du lecteur, que l'utilisation des densitées de copules en tant qu'approximation des probabilités jointes pour $T = t$ et pour $Z = z$ au lieu de la procédure présentée ici utilisant les fonctions jointes de distributions cumulatives est à proscrire. En effet, les modèles de copules paramétriques sont bien connues pour le problème de frontières : les densités de copules divergent vers l'infini lorsqu'elles tendent vers les points $[0, 0]$ et $[1, 1]$. Dans le contexte présenté ici, étant donné que la masse de probabilité est hautement concentrée autour de ces points (spécialement dans le cas de variables binaires), il est recommendé d'utiliser uniquement les fonctions de distributions cumulatives qui elles, ne divergent pas. Pour davantage d'informations à propos du problème de frontières, le lecteur est référé à Gijbels et al.[56]

### 3.4.4 Cadres nécessaires à l'estimation des paramètres

Afin d'arriver à travailler de façon efficace avec des mesures de score de propension et des données médicales, il est nécessaire d'être outillé de modèles utiles et applicables dans le but d'obtenir un processus de calcul facilement implémentable sur les outils technologiques pour le score de propension tels que les modèles de régression logistique peuvent l'être. Le but ici est de présenter la méthodologie nécessaire pour estimer tant les familles de copules paramétriques que leurs marges constituantes et leurs paramètres. Ultimement, il sera donc possible d'inférer le score de propension pour chaque sujet participant à une étude donnée. La façon de procéder est donc de procéder à l'estimation des trois structures principales du modèle : les fonctions de répartition marginales, le tau de Kendall et les familles de copules.

#### 3.4.4.1 Fonctions de répartition marginales

On définit $\hat{G}^{\epsilon_2}$ en tant qu'estimateur convergent de $G^{\epsilon_2}$. Alors, il est évident que la seule différence entre $\hat{G}^{\epsilon_2}$ et $G^{\epsilon_2}$ est tributaire de la vitesse de convergence de $\hat{G}^{\epsilon_2}$. Par ailleurs, on

remarque que cette différence existe seulement au niveau des plateaux de $\hat{G}$ telle que :

$$\hat{G}^{\epsilon_2}(z) = \begin{cases} \hat{G}_n(z) & \text{si } z \in \left(\bigcup_{i=1}^{I}(x_{i-1}+\epsilon_2^2 < z < x_i - \epsilon_2)\right); \\ a_i z + b_i & \text{si } z \in \left(\bigcup_{i=0}^{I}[x_i - \epsilon_2, x_i + \epsilon_2^2]\right). \end{cases}$$

Tant l'utilisation d'une modélisation paramétrique que non-paramétrique pour les marges peut être effectuée. Dans le premier cas, il est suggéré d'utiliser une loi de Bernoulli (loi binomiale de paramètres (1,p) ) pour modéliser la variable d'allocation au traitement $T$. En ce qui a trait aux covariables d'intérêt, étant donné que le nombre de modalités de la variable $Z$ peut être supérieur à 2, on suggère également l'utilisation d'une loi multinomiale. L'estimation des paramètres de la loi sélectionnée peut se faire en utilisant l'estimateur du maximum de vraisemblance.

Dans un second cas, s'il est décidé d'utiliser une estimation non-paramétrique des fonctions de répartition marginales, on suggère l'utilisation de la fonction de distribution empirique réechelonnée en tant qu'estimateur de la fonction $G$ où :

$$\hat{G}_n(z) \quad = \quad \frac{1}{n+1}\sum_{j=1}^{n} \mathbb{1}_{\{T_j \leqslant t\}}.$$

En fait, ce rééchelonnement conserve les propriété de la fonction de répartition empirique originale et permet de contourner les problèmes liés aux copules évaluées à leurs frontières. Par ailleurs, il est aisé de démontrer, en utilisant l'inégalité de Chebyshev, que $\hat{G}_n(z)$ converge en probabilités vers $G$. Cette suggestion quant aux choix de modélisation mène à l'hypothèse suivante, permettant d'assurer la convergence des estimateurs avec les vraies lois.

**Hypothèse 3.4.3.** *N'importe quel estimateur $\tilde{G}$ peut être utilisé pour estimer $G$ s'il peut fournir une représentation discrète et satisfait*

$$\tilde{G}(t) = \hat{G}_n(t) + o_p(n^{-1/2})$$

**Taille d'epsilon**

Tel que mentionné précédemment, il est suggéré de sélectionner des valeurs d'epsilon proches de 0 telles que $\epsilon_1 \to 0, \epsilon_2 \to 0$. Étant donné que l'on définit ces valeurs comme $\epsilon_1 = k/n, \epsilon_2 = s/m, k \to 1, s \to 1$, les valeurs d'epsilon ont deux contraintes majeures : elles doivent être assez larges pour ne pas être considérées comme des données aberrantes mais faire partie de la distribution, et suffisament petites pour être considérées proches de 0. Ainsi, un choix raisonnable de ces valeurs est dans le voisinage de 1/100 : par simulations, on peut constater que ce choix est en harmonie avec ces contraintes.

### 3.4.4.2 Utilisation d'un tau de Kendall significatif

Dans le but d'être en mesure d'exprimer une copule possédant une erreur quadratique minimale quant à son adéquation aux données, il est primordial d'inférer le paramètre de dépendance $\theta$ de la copule avec une méthode minimisant le biais dans un contexte de données discrètes.

Pour ce faire, les deux chemins les plus rencontrés dans la littérature sont l'estimation basée sur le maximum de la vraisemblance (MLE), et celle basée sur l'inversion du tau de Kendall. Ces deux méthodes ont leurs lacunes respectives : l'estimateur du maximum de vraisemblance n'est pas consistent pour les données discrètes dans le sens où il fournit des valeurs possibles de $\theta$ que seulement quelques fois (il est fréquent qu'on ne puisse pas calculer les dérivées de la (log) vraisemblance dans le contexte cité). En ce qui a trait au tau de Kendall en lui-même, il est souvent biaisé dans le sens où il possède plusieurs valeurs possibles avec une borne inférieure et une borne supérieure et, parfois, peut prendre valeur à l'extérieur de $[-1, 1]$. Toutefois, il est possible de corriger le tau de Kendall pour en obtenir un estimateur consistant du paramètre de la copule.

Par définition, une expression jointe telle que la fonction de répartition jointe se doit d'exprimer tant les fonctions de répartitions marginales qui la constituent que le niveau de dépendance entre ces dernières ; c'est la raison pour laquelle la fonction copule est un outil d'intérêt dans cette modélisation. Sachant que, dans le cas paramétrique bivarié, le paramètre de la copule est obtenu directement à partir d'une mesure de corrélation entre les variables, le point ici est d'obtenir un paramètre de dépendance basé sur une mesure de concordance robuste.

Un élément capital à inférer dans une modélisation jointe avec des données discrètes est la mesure de dépendance entre les distributions marginales dans le but trouver, par inversion du tau de Kendall (méthode également valide pour le rho de Spearmann, le bêta de Blomqvist, etc.), le paramètre de la copule tributaire à la famille choisie. Dans cette section, on va suivre la voie majeure dans la littérature sur les copules unidimensionnelles : l'utilisation uniquement du tau de Kendall. On remarque que la convergence théorique de $\tau$, telle que montrée à la proposition 2.3, est indépendance du concept de tau de Kendall corrigé en faveur des données discrètes, mais nous montre que l'on peut prendre les mesures de concordance et de discordance directement sur les données brutes des variables aléatoires $T$ et $Z$ si l'on effectue la correction nécessaire ensuite.

Tel que montré dans les articles de Denuit et al.[44] et de Genest et al.,[43] le tau de Kendall ($\tau$) mesuré sur des distributions discrètes perd beaucoup de propriétés et, en particulier, peut avoir une valeur à l'extérieur de son support $[-1, 1]$. C'est la raison pour laquelle l'utilisation du tau de Kendall original en cas de distributions discrètes est indésirable ; une correction doit être effectuée. L'approche de Bouezmarni et al.,[57] basée sur les bornes des distributions des variables aléatoires, sera adoptée ici.

Soient $F(t)$ et $G(z)$ les fonctions de distribution cumulatives pour l'assignation au traitement et pour sa covariable d'intérêt. Alors, la distribution extrême

$$H_{\min}(t, z) = \max[0, F(t) + G(z) - 1]$$

est connue pour être la borne inférieure de Fréchet, et

$$H_{\max}(t, z) = \min[F(t), G(z)]$$

est connue pour être la borne supérieure de Fréchet. De ces bornes, on déduit les valeurs des tau

de Kendall qui leur sont associées, $\tau_{min}$ et $\tau_{max}$, pour lesquelles, si $\tau$ est la version originale du tau de Kendall mesuré directement sur les données discrètes,

$$\tau_{\min} \leqslant \tau \leqslant \tau_{\max}.$$

Ainsi, la version corrigée du tau de Kendall, $\tau_c$, est

$$\tau_c = \begin{cases} \frac{\tau}{\tau_{\max}} & \text{if } \tau \geqslant 0; \\ \frac{-\tau}{\tau_{\min}} & \text{if } \tau < 0. \end{cases}$$

Bouezmarni et al. [57] ont démontré les propriétés asymptotiques de cet estimateur.

#### 3.4.4.3 Distributions jointes

**Modèles et familles de copules**

Tel qu'illustré dans la littérature,[43] il y a une limite quant à l'applicabilité des familles de copules paramétriques pour les données de comptage. Afin d'assurer une certaine consistance dans la modélisation des fonctions jointes, les familles de copules présentées ici seront limitées aux plus couramment rencontrées pour les données discrètes. En fait, la variété de copules qui peuvent être utilisées avec ce type de données et qui conservent leurs propriétés analytiques est réellement limité. Pour l'usage de ce travail de recherche, il a été décidé de limiter notre choix aux familles de copules paramétriques (voir le tableau 3.1) et qui sont associées à un paramètre de dépendance unique.

On remarque que, lorsqu'il y a une évidence sur la nature de la dépendance entre les données, on peut contraindre les copules à une propriété de dépendance spécifique : on suppose que l'on obtient la matrice de probabilités $P$ en discrétisant les variables aléatoires $(T^{\epsilon_1}, Z^{\epsilon_2})$ de façon à ce qu'il existe deux suites croissantes $A_0 < A_1 < ... < A_a; B_0 < B_1 < ... < B_b$ telles que $p_{ij} = \mathbb{P}(T \in ]A_{i-1}, A_i[ \cup Z \in ]B_{j-1}, B_j[), p_{i\cdot} > 0, p_{\cdot j} > 0$. Alors, si le couple $(T, Z)$ suit une propriété de dépendance (e.g. QPD := Quadrant-Positive Dependence, RTI :=Right-tail Increasing), $P$ le fait également. Un tel choix de modélisation peut requérir une procédure de rééchantillonnage soumise à la propriété de dépendance donnée.[58] On note aussi que de la proposition 2.3, en utilisant (Yanagimoto and Okamoto,[59] Tchen[60]), les propriétés de dépendance sont préservées avec la transformation de continuité proposée dans cette section.

**Procédure de sélection de la copule**

Il existe de multiples "zones d'ombres" aux procédures standard de sélection d'une famille de copules lorsque cette dernière provient de l'extension par continuité d'une sous-copule. Les raisons majeures justifiant ces problématiques sont, pour commencer, le problème de sensibilité de la copule à la variation d'une seule observation sur les distributions discrètes ; et, ensuite, la forte masse de probabilité présente sur les valeurs extrêmes des distributions. D'autre part, tel

| Copule type | Fonction $C(u,v)$ | Dépendence |
|---|---|---|
| Joe | $1-[(1-u)^\theta+(1-v)^\theta-(1-u)^\theta(1-v)^\theta]^{1/\theta}$ | $\theta\in(0,\infty)$ |
| Gumbel | $exp\{-[(-\log u)^\theta+(-\log v)^\theta]^{1/\theta}\}$ | $\theta\in[1,\infty)$ |
| Frank | $\frac{\log[1+(\theta^u-1)(\theta^v-1)/(\theta-1)]}{\log\theta}$ | $\theta\in(0,\infty)\backslash\{1\}$ |
| Clayton | $(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$ | $\theta\in[-1,\infty)\backslash\{0\}$ |

TABLEAU 3.1 – Exemple de copules à considérer qui ont un paramètre de dépendance unique et qui peuvent représenter une propriété de dépendance spécifique.

que pointé par Durrleman et al. ,[61] une procédure de sélection d'une famille de copules basée sur un critère d'information (e.g. AIC, BIC, TIC, etc.) va favoriser la copule de Frank.

Dans ce travail, il est suggéré d'utiliser l'approche empirique. Soit la fonction de répartition jointe empirique,[62] telle que :

$$\hat{H}_n(t,z)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\cdot\left\{\hat{F}_n(T)\leqslant t,\hat{G}_n(Z)\leqslant z\right\}$$

où $\hat{F}_n(T)$ and $\hat{G}_n(Z)$ sont les fonctions marginales des variables aléatoires $T$ et $Z$ respectivement. On construit les $K$ copules paramétriques à mettre en compétition en tant que meilleure approximation de $C^{\{\epsilon_1,\epsilon_2\}}$ telles que :

$$\hat{C}_{(1)}^{\{\epsilon_1,\epsilon_2\}}(F^{\epsilon_1}(t),G^{\epsilon_2}(z))$$

$$\hat{C}_{(2)}^{\{\epsilon_1,\epsilon_2\}}(F^{\epsilon_1}(t),G^{\epsilon_2}(z))$$

$$\vdots$$

$$\hat{C}_{(K)}^{\{\epsilon_1,\epsilon_2\}}(F^{\epsilon_1}(t),G^{\epsilon_2}(z)).$$

Pour n'importe quelle famille $j$, $j=1,2,...,K$, la sous-copule $C'(F(t),G(z))$ est approchée par

$$\hat{C}_{(j)}^{\{\epsilon_1,\epsilon_2\}}(F^{\epsilon_1}(t),G^{\epsilon_2}(z)).$$

En se basant sur le tau de Kendall $\tau_c$ ayant subi une correction sur les probabilités de concordance et de discordance et calculé sur les observations $(T,Z)$, on obtient, pour chaque copule $j$ une relation bijective entre le tau de Kendall et le paramètre de copule recherché telle que $(\theta_j)\approx r_j(\tau_c)$ et l'on obtient

$$C^j_{r_j(\tau_c)}\left(F^{\epsilon_1}(t),G^{\epsilon_2}(z)\right)=H^j_\epsilon(t,z)$$

où $H_\epsilon(t,z)$ est la distribution jointe des variables $T$ et $Z$, et $C^j_{r_j(\tau_c)}$ est la distribution jointe continue représentée par la copule $j$ avec un paramètre $\theta_j$. Alors, on a, en utilisant la norme

suprémum

$$\sup_{(x,y)} \left| H^j_{\{\epsilon_1,\epsilon_2\}}(t,z) - \hat{H}_n(t,z) \right| = \delta^j_{\{\epsilon_1,\epsilon_2\}}.$$

Ainsi, on sélectionne la copule $j$ ayant la plus petite valeur de $\delta^j_\epsilon$. On note qu'en utilisant $\epsilon_1 = k/n, \epsilon_2 = s/m$, $k \to 1$, $s \to 1$, pour toute famille $j$, la relation

$$\lim_{m\to\infty} \lim_{n\to\infty} \delta^j_{\{\epsilon_1,\epsilon_2\}} = 0$$

est satisfaite.

**Considération d'une mauvaise spécification**

Évidemment, la modélisation avec une mauvaise spécification de la famille de copule n'est pas problématique dans le cas des données discrètes au même point qu'il l'est avec les données continues. Le fait que la sous-copule ne soit pas unique sur $[0,1] \times [0,1]$ et que la présence (ou l'absence) d'une queue de dépendance entre les distributions marginales soit limitée en raison de la nature discrète des données font que la différence dans la modélisation entre les différentes familles de copules est, en fait, relativement faible ; mais existe bien et est détectable sous un critère de distance muni de la norme suprémum. Ainsi, une estimation de la copule basée sur une distance entre l'ensemble des familles de copules comparées et la fonction de répartition empirique jointe va minimiser le risque d'une mauvaise spécification.

### 3.4.4.4 Extension multivariée

Le calcul du tau de Kendall pour les distributions multivariées (i.e. lorsqu'il y a plus qu'une covariable d'intérêt) est possible, et l'utilisation de $\hat{\theta}$ provenant de la méthode d'inversion du tau de Kendall assure un paramètre de copule consistent.[63] Toutefois, il a été montré dans la littérature qu'une telle méthode est particulièrement onéreuse en termes de performances computationnelles et d'implémentation informatique. Par ailleurs, l'estimation de $\hat{\theta}$ basée sur la maximisation de la pseudo-vraisemblance est clairement l'approche la plus efficace.[29] En se basant sur l'approche de Genest et al.,[64] le principe est d'obtenir la valeur de $\theta$ qui maximise la quantité

$$ll(\theta) \approx \sum_{k=1}^{n} \log\left[\mathbb{P}\left(T^{\epsilon_1} \in [t-\epsilon'_1, t+\epsilon'_1], Z_1^{\epsilon_2,} \in [z_1-\epsilon'_2, z_1+\epsilon'_2], ..., Z_d^{\epsilon_{d+1},} \in [z_d-\epsilon'_{d+1}, z_d+\epsilon'_{d+1}]\right)\right] \quad (3.1)$$

où la probabilité $\mathbb{P}$ est évaluée en calculant les $2^{d+1}$ copules nécessaires. La procédure de sélection de la copule va donc, conséquemment, impliquer tant la sélection des familles de copules multivariées qui sont munies d'un seul paramètre de dépendance (e.g. Clayton, Gumbel, etc.) que celles qui ont plusieurs paramètres de dépendance (e.g. de Student).

### 3.4.5 Simulations

Le but de cette section est, premièrement, de vérifier que la méthode d'estimation du score de propension présentée dans ce travail est consistente tant dans le cas d'une seule variable polytomique que dans le cas de plusieurs covariables d'intérêt dichotomiques et, deuxièmement, de comparer l'adéquation du score de propension basé sur les copules à des données générées à l'adéquation du score de propension basé sur la régression logistique aux mêmes données simulées. Pour y arriver, on va considérer les deux procédures de génération de données (PGD) suivantes :

— **PGD 1 :** Modèle bivarié constitué de deux marges discrètes

1. $F(T) \sim$ une loi de Bernoulli avec $p = 0.5$, une distribution dichotomique ;
2. $G(Z) \sim$ une loi binomiale de paramètres $(3, 0.44)$, i.e. données catégorielles ;
3. $H(F(T, G(Z))$ générée à partir d'une copule de Gumbel de paramètre 6.25 (équivalent à un tau de Kendall de 0.84) avec $n = 2000$ individus.

— **PGD 2 :** Modèle multivarié basé sur des simulations de données cliniques

1. Étude de cohorte hypothétique (n=2000) ;
2. $F(T)$ cdf d'une exposition à un traitement $T$ (binaire) ;
3. $\boldsymbol{G(Z)}$ cdf d'un vecteur aléatoire de 3 covariables d'intérêt (binaires).

#### 3.4.5.1 PGD 1

Le premier élément à vérifier ici est la validité du critère de sélection de la copule. Pour la PGD 1, on a répliqué $B = 500$ fois la génération d'une sous-copule archimédienne (Gumbel) de paramètre relié, par inversion, à un tau de Kendall $\tau = 0.84$ (afin d'imposer une forte structure de dépendance caractéristique de cette famille de copules) avec deux fonctions de répartition marginales telles que décrites précédemment. On note que le choix de la famille de copules utilisée pour générer les données est absolument arbitraire ; n'importe quel choix d'une famille de copules archimédienne avec une dépendance forte entre les données doit être détecté ultérieurement lors de la prodécure de sélection de la copule.

Ensuite, on a estimé les fonctions de distribution marginales pour chaque PGD avec la fonction de distribution empirique, et avons rendu continues ces dernières à l'aide de la méthode présentée dans ce travail. On a noté à chaque fois la copule sélectionnée et avons reporté cela à la table 3.2. Les familles de copules en compétition étaient les quatre citées à la table 3.1 ayant un unique paramètre de dépendance, en plus de la copule gaussienne simplement en raison de sa fréquence d'utilisation dans la littérature en tant que représentante des copules elliptiques. Afin d'éviter tout problème de convergence lors des séquences itératives de simulation, la taille d'epsilon a été fixée à $\epsilon_i = 1/100$, $i = 1, 2$. Toutefois, on a noté que pour de petites valeurs d'epsilon telles que $\epsilon \leqslant 3/2000$, plus qu'une fois sur deux, il n'y avait pas de problèmes computationnels dans la procédure d'application de la copule sur les données rendues continues. Donc, il est possible de voir que le critère de sélection d'une famille de copule a été consistent avec les données

| PGDs | Clayton | Frank | Gumbel | Joe | Gaussian |
|---|---|---|---|---|---|
| PGD 1 | 0 | 2 | **403** | 0 | 95 |

TABLEAU 3.2 – Sélection de la copule $C^{\{\epsilon_1,\epsilon_2\}}$ avec des marges continues pour des données générés à partir d'une sous-copule archimédienne $C'$ avec des marges discontinues

simulées dans plus de 80% des cas. Étant donné que l'impact d'une mauvaise spécification de la copule est faible en présence de données discrètes, on peut affirmer que le critère de sélection d'une famille de copules est satisfaisant.

La deuxième étape de ces simulations a été d'obtenir la probabilité conditionnelle que constitue le score de propension pour chacun des 2000 patients simulés ici en utilisant les deux méthodes de calcul : l'approche basée sur la fonction copule et la régression logistique classique. Dans le but d'éviter tout biais provenant du niveau de dépendance entre les données, on a appliqué la procédure à trois niveaux de dépendance dans la génération des données de PGD 1 : $\tau = 0.20$, $\tau = 0.50$ et $\tau = 0.86$. À la figure 3.2, on observe la dispersion des scores de propension pour les 2000 individus évalués aux trois niveaux à l'aide des deux méthodes. On remarque l'évidence d'une différence significative entre la dispersion de ces scores pour les mêmes populations.

La façon dont nous avons décidé de comparer la méthode de calcul basée sur l'utilisation des fonctions copules et celle basée sur la régression logistique était par l'évaluation du Critère d'Information d'Akaike (AIC) (i.e. 2 fois le nombre de paramètres présents dans le modèle moins 2 fois sa log-vraisemblance). Ainsi, le nombre de paramètres du modèle utilisant les fonctions copules, lorsque la famille spécifiée est de Clayton, de Gumbel, de Joe ou gaussienne, est simplement 1. Cela dit, nous avons réalisé 500 réplications du schème de génération des données et, à chaque fois, avons recueilli l'AIC pour chaque méthode quant à son adéquation aux données. Au tableau 3.3, on peut voir le nombre d'occurences où le modèle basé sur la copule de Gumbel a obtenu un plus petit AIC que celui basé sur la régression logistique. La procédure a été répétée pour les trois différents niveaux de dépendance. Donc, la somme de chaque colonne est égale à 500. Au tableau 3.4, on a performé la même procédure en utilisant une copule possédant une structure de dépendance complètement différente, de type Clayton, pour analyser l'effet d'une mauvaise sélection de copule sur l'adéquation aux données du modèle.

On observe que, pour un très petit niveau de dépendance entre la variable traitement et une covariable d'intérêt, lorsque la sélection de la copule est erronée, le modèle de régression logistique peut obtenir parfois une meilleure adéquation aux données. Cependant, nonobstant le niveau de dépendance entre les variables, si la famille de copules est bien spécifiée, le modèle basé sur ces dernières obtient un plus petit AIC et, conséquemment, une meilleure adéquation. Finalement, on observe que, avec un niveau de dépendance intermédiaire, une mauvaise spécification de la copule n'a pas d'impact significatif sur la qualité de la modélisation : $C^{\{\epsilon_1,\epsilon_2\}}$ obtient ici également une plus petite valeur de critère d'information.

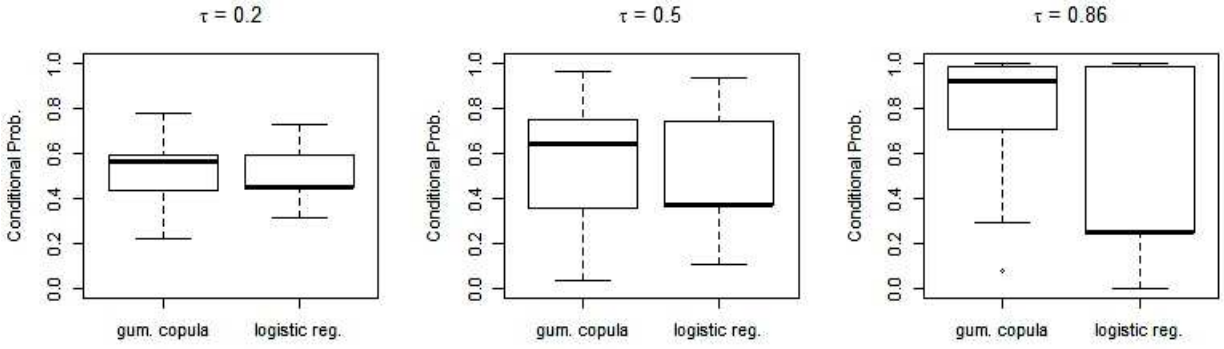


FIGURE 3.2 – Boxplots de la dispersion des scores de propension pour les données générées avec 3 différents niveaux de dépendance.

| Méthode d'estimation de la probabilité | $\tau = 0.2$ | $\tau = 0.5$ | $\tau = 0.86$ |
|---|---|---|---|
| $C^{\{\epsilon_1,\epsilon_2\}}$ de famille de Gumbel | 497 | 500 | 500 |
| Régression logistique | 3 | 0 | 0 |

TABLEAU 3.3 – Comparaison de l'AIC entre la méthode basée sur les copules et la régression logistique pour estimer la probabilité conditionnelle de $T$ sachant $Z$ ; 500 réplications effectuées.

| Méthode d'estimation de la probabilité | $\tau = 0.2$ | $\tau = 0.5$ | $\tau = 0.86$ |
|---|---|---|---|
| $C^{\{\epsilon_1,\epsilon_2\}}$ de famille de Clayton | 330 | 500 | 500 |
| Régression logistique | 170 | 0 | 0 |

TABLEAU 3.4 – Comparaison de l'AIC entre la méthode basée sur les copules et la régression logistique dans le cas d'une mauvaise spécification de la copule ; 500 réplications effectuées.

#### 3.4.5.2 PGD 2

La deuxième PGD consistait, tel que mentionné précédemment, en une variabre binaire d'assignation à un traitement $T$ qui est dépendante de trois covariables $Z_1, Z_2, Z_3$, mais les covariables sont indépendantes entre elles. La construction de ces variables est illustrée dans l'encadré suivant. Par ailleurs, on remarque que le choix des règles de décision pour construire ces variables dichotomiques à partir de variables gaussiennes est inspiré du travail de simulations de Setoguchi.[65]

**Construction des données pour la PGD 2 :**

1. Génération d'une variable aléatoire $R \sim N(0,1)$ ayant $n = 2000$ observations.
2. Pour la variable aléatoire $T$,
   - **si** $R_i \geqslant 1,29$, $i = 1, 2, ..., n$, alors $T_i = 1$ ;
   - **sinon** $T_i = 0$.
3. Pour la variable aléatoire $Z_1$,
   - générer une variable aléatoire $R_{Z_1} \sim N(0,1)$ telle que $cor(R, R_{Z_1}) = 0.6$ ;
   - **si** $R_{Z_{1i}} \geqslant 1,45$, $i = 1, 2, ..., n$, alors $R_{Z_{1i}} = 1$ ;
   - **sinon** $R_{Z_{1i}} = 0$.
4. Pour la variable aléatoire $Z_2$,
   - générer une variable aléatoire $R_{Z_2} \sim N(0,1)$ telle que $cor(R, R_{Z_2}) = 0.35$ ;
   - **si** $R_{Z_{2i}} \geqslant 1$, $i = 1, 2, ..., n$, alors $R_{Z_{2i}} = 1$ ;
   - **sinon** $R_{Z_{2i}} = 0$.
5. Pour la variable aléatoire $Z_3$,
   - générer une variable aléatoire $R_{Z_3} \sim N(0,1)$ telle que $cor(R, R_{Z_3}) = 0.35$ ;
   - **si** $R_{Z_{3i}} \geqslant 1.55$, $i = 1, 2, ..., n$, alors $R_{Z_{3i}} = 1$ ;
   - **sinon** $R_{Z_{3i}} = 0$.

En ce qui concerne la sélection de la famille de copules, en utilisant le critère de sélection décrit dans ce travail, la copule multivariée de Clayton, qui a la forme analytique suivante, a été choisie :

$$C(u_0, u_1, ..., u_d) = \left( d + \sum_{j=0}^{d} u_j^{-\theta} \right)^{-1/\theta}$$

où $d = 3$, le nombre de covariables d'intérêt. Donc, le paramètre de la copule $\hat{\theta}$ a dû être inféré en effectuant la maximisation de la pseudo-vraisemblance décrite à l'équation 3.1.

Dans une voie similaire à celle de la PGD 1, la comparaison entre le modèle de score de propension basé sur la fonction copule et celui basé sur la régression logistique a été effectuée sur la base du critère d'information de l'AIC. Pour le modèle de régression logistique, toutes les covariables ont été incluses et ont été supposées indépendantes entre elles. Le résultat de cette

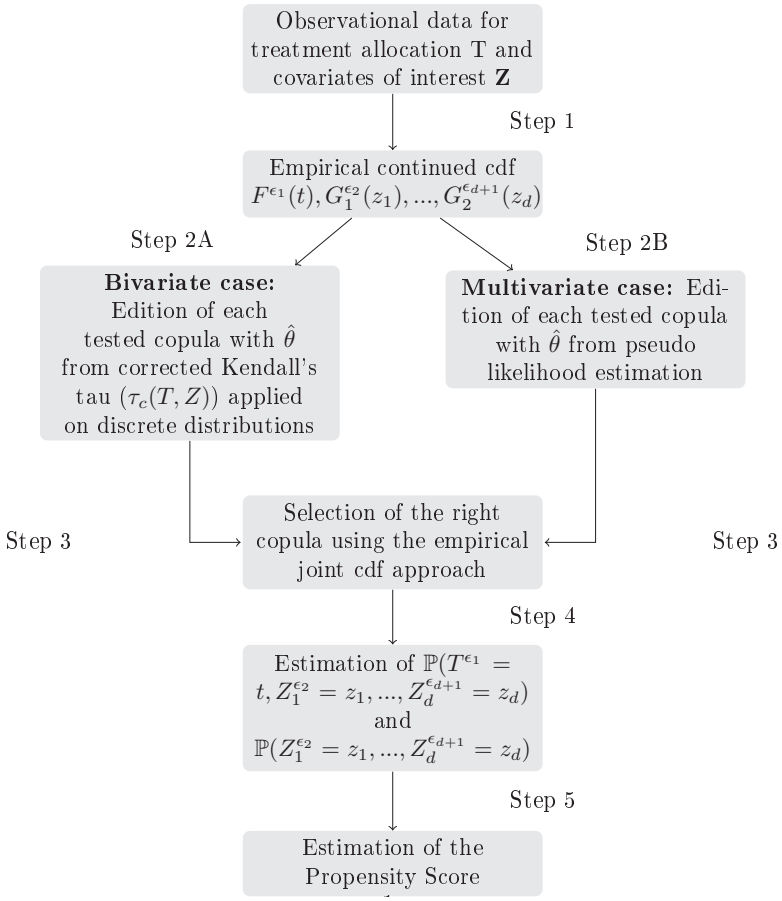


Figure 3.3 – Schéma de la procédure entière pour obtenir le score de propension à partir de la fonction copule.

comparaison a été sans équivoque en faveur du modèle se basant sur la copule de Clayton : des 500 réplications indépendantes de la population, l'AIC a été inférieur pour le score de propension issue des copules 500 fois.

## 3.5 Discussion

Tel qu'illustré dans les simulations, la méthode d'estimation du score de propension basée sur les copules présentée dans ce travail avec une continuation des marges pour les fonctions discrètes est efficiente pour estimer la probabilité conditionnelle que constitue le score de propension étant donné qu'elle n'est pas inhérente à une quelconque hypothèse de linéarité entre les covariables d'intérêt sur le traitement. On a présenté ici deux éléments majeurs : a priori, une approche simple pour prendre en considération les données de comptage avec les fonctions copules basée sur une correction limite des fonctions de répartition marginales avec toute la théorie nécessaire afin d'y arriver. Ensuite, on a proposé l'utilisation de cette approche pour obtenir le score de propension tel qu'elle est résumée sur le schéma 3.3.

Dans une perspective future de travail, l'étape suivante d'une telle procédure d'évaluation du

score de propension est d'appliquer les constructions de copules par paires dans un contexte de copules en vignes (Vine copulas)[66] aux fonctions de répartition marginales rendues continues par correction au lieu de construire, pour chaque probabilité recherchée, un réseau de $2^{d+1}$ copules où $d$ est le nombre de covariables d'intérêt qui affectent l'assignation au traitement.

# Chapitre 4

# Copules et données censurées : cas de la régression

Ce chapitre propose une approche semi-paramétrique d'estimation de la fonction de régression dans le cas où une censure à droite, non-informative, subsiste sur la variable réponse. On propose ici une approche basée sur la fonction copule, dans le cas où cette dernière est paramétrique, menant à une estimation de la régression via l'estimateur de Kaplan-Meier, l'estimateur empirique de la fonction de répartition et l'estimation du paramètre de la copule inhérente au modèle (e.g. : inversion du tau de Kendall, maximum de vraisemblance).

La censure à droite dans un cadre non-informatif sur la variable réponse d'une étude spécifique survient principalement pour des raisons propres aux sujets et conséquemment indépendantes du cadre de l'étude (e.g. déménagement). Ces manques d'information quant à la liaison exacte entre la variables d'intérêt et ses prédicteurs nécessitent l'utilisation de méthodes d'inférence particulières. Dans la littérature, Miller[67] et Koul et al.[68] proposent des modèles d'estimation basés sur le modèle de régression linéaire simple. Malgré la relative simplicité d'utilisation de cette méthodologie, des failles théoriques persistent : l'imposition de la linéarité entre les variables et, surtout, l'imposition d'un cadre restrictif sur la censure. Ensuite, la méthode de Buckley et James,[22] basée sur l'espérance conditionnelle de la variable réponse sachant les covariables et la distribution de la censure, fait appel à l'estimation de coefficients de régression au sens de la méthode des moindres carrés ordinaires. Le problème ici est que la valeur des coefficients de régression $\beta$ est trouvée à partir d'une procédure itérative qui ne converge par toujours. Il arrive que $\beta$ oscille entre deux valeurs numériques. Dans les travaux de Doksum et Yandell,[69] Zheng,[70,71] Leurgans,[72] Zhou[73] et Srinivasan et Zhou,[74] diverses approches basées sur la transformation des données (*data transformation principle*) ont été suggérées. Toutefois, ces approches ont toutes le même inconvénient : la régression linéaire est effectuée avec comme prémisse la non-corrélation entre les données. Autrement dit, l'existence de variables de confusion parmi les covariables tout comme la dépendance entre les données ne peuvent pas être détectées avec une telle méthodologie. Enfin, une méthode non-paramétrique proposée par Fan et Gijbels[75] et basée sur une régression locale linéaire présente une méthodologie flexible étant donné qu'elle ne suppose pas un modèle

paramétrique pour la fonction de régression. Par contre, la problématique majeure est le fléau de la dimensionnalité (*curse of dimensionality*) : l'approche local-linéaire présuppose de travailler sur plusieurs dimensions alors qu'en réalité, les données se situent dans un espace de moins grande dimension.

## 4.1 Censure : informativité et mécanismes

Étant donné qu'en biostatistiques et en épidémiologie, l'objet principal des études menées est l'explication de l'occurence d'un événement d'intérêt (e.g. décès, rejet d'un greffon, présence d'une complication post-opératoire spécifique), toute information disponible doit être décortiquée. Toutefois, en raison du fait que le phénomène de censure est en-soi un cas particulier d'incomplétude des données, les études observationnelles ne présentent que très rarement des données complètes lorsqu'elles se situent dans un cadre d'analyse de la survie. Ainsi, il faut, pour le clinicien, être prompt à utiliser des méthodes statistiques tenant compte des données censurées.

Se plaçant sous l'hypothèse que le temps avant la censure et le temps avant l'événement d'intérêt sont en réalité deux temps dépendants pour diminuer le biais dans l'estimation de la survie, on qualifie trois types de censure : censure à gauche, censure à droite et censure par intervalles. On note que dans la majorité des cas, on se trouve dans le cadre d'une censure à droite. Par ailleurs, la censure peut être informative ou non-informative : en cas de censure informative, il y a dépendance entre le temps de survie et le temps de censure. On prend l'exemple d'un patient perdu de vue : sa rétractation volontaire peut, par exemple, provenir du fait que le patient connait sa fin proche ou décide d'arrêter les traitements pour décéder dans une certaine dignité, sa censure est alors dépendante du temps de décès. Il en va de même pour la censure administrative que l'on considère souvent informative.

## 4.2 Estimateur proposé

Soit $Y$, la variable réponse, une variable aléatoire ayant pour fonction de distribution cumulative $F$ et admettant une fonction de densité $f$ ; et soit $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_d})$ représentant le vecteur de variables prédictives ayant pour fonction de distribution cumulative $\boldsymbol{G(x)} = (G_1(x_1), G_2(x_2), ..., G_d(x_d))$ et admettant pour densité

$$\boldsymbol{g(x)} = (g_1(x_1), g_2(x_2), ..., g_d(x_d)) .$$

À partir de la théorie des copules, pour $X$ et $Y$ étant des variables continues, il est possible de connaître la fonction de répartition jointe du couple $(Y, \mathbf{X})^{\mathbf{T}}$ en tout ensemble de points $(y, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^{\mathbf{d}}$ par $C(F(y), \boldsymbol{G(x)})$ où $C$ est une copule unique. Alors, pour $(u_0, u_1, ..., u_d) \equiv (u_0, u) \in [0, 1] \times [0, 1]^d$, on a

$$C(u_0, u) = \mathbb{P}(U_0 \leqslant u_0, U_1 \leqslant u_1, ..., U_d \leqslant u_d)$$

où $U_0 = F(Y), U_j = G_j(X_j)$ for $j = 1, ..., d$. Ainsi, $C$ est une fonction de distribution avec des marges uniformes sur $[0, 1]$.

En situation de données complètes pour toutes les variables, Noh et al.[76] proposent d'estimer la fonction de régression en conditionnant sur les variables. À partir de la théorie de la densité conditonnelle réécrite en terme de densitées de copules (c'est-à-dire en terme des dérivées partielles par rapport aux marges de la fonction de distribution de la copule), on obtient

$$f_{Y|\boldsymbol{X}=\boldsymbol{x}} = f(y)\frac{c(F(y), \boldsymbol{G(x)})}{\tilde{c}(\boldsymbol{G(x)})}$$

où $\tilde{c}$ est la densité de copule de $\boldsymbol{X}$ ; c'est-à-dire que

$$c(F(y), \boldsymbol{G(x)}) = \frac{\partial^2 C(F(y), \boldsymbol{G(x)})}{\partial F(y) \partial \boldsymbol{G(x)}}.$$

Ainsi, à partir de la définition de l'espérance conditionnelle, selon l'approche de Noh et al., on obtient l'expression de la fonction de régression telle que

$$\begin{aligned} m(x) &= \mathbb{E}(Y|X = x) \\ &= \mathbb{E}(Y w(F(y), G(x))) \\ &= \frac{e(\mathbf{G}(\mathbf{x}))}{\tilde{c}(\mathbf{G}(\mathbf{x}))} \end{aligned} \tag{4.1}$$

où

$$w(u_0, u) = c(u_0, u)/\tilde{c}(u) \quad \text{et} \quad e(u) = \int_0^1 F^{-1}(u_0) c(u_0, u) du_0.$$

Donc, $c$ et $\tilde{c}$ représentent respectivement les densités de $(Y, \mathbf{X})$ et $\mathbf{X}$ et donc, $c(u_0, u) = \dfrac{\partial^{d+1} C(u_0, u_1, u_2, ..., u_d)}{\partial u_0 \partial u_1 \partial u_2 ... \partial u_d}$ et $\tilde{c}(u) = \dfrac{\partial^d C(1, u_1, u_2, ..., u_d)}{\partial u_1 \partial u_2 ... \partial u_d}$. Ainsi, on voit que $w(u_0, u)$ est simplement le ratio de $c(u_0, u)$ par $\tilde{c}(u)$. On remarque que dans le cas où les données sont bivariées (où $d = 1$) tout comme dans le cas où les variables prédictives sont mutuellement indépendantes, on a $\tilde{c} = 1$ et cette fonction de régression se réduit simplement à $\mathbb{E}(Y|X = x) = e(\boldsymbol{G(x)})$.

Dans ce travail, la méthode d'estimation proposée est en fait une adaptation de l'estimateur de Noh et al. aux données dont la variable dépendante est censurée à droite ; cas où la censure est bien entendu non-administrative. On note ainsi le vecteur des observations $Y$ tel que $Y = \min(T, C^0)$ où $Y$ est le temps de survie jusqu'au décès et $C^0$ est la variable aléatoire représentant le phénomène de censure. Par ailleurs, on note la complétude des données par $\delta = \mathbb{1}_{(T \leqslant C^0)}$ où $\mathbb{1}_{.}$ représente la fonction indicatrice. Si, par exemple, on observe uniquement des données complètes sur l'ensemble d'une base de données, alors $\forall i : Y_i = T_i$ et $\delta_i = 1$ où $i = 1, 2, ..., n$ est l'indice des variables d'intérêt au niveau des individus et où $n$ est le nombre d'observations composant chaque variable de la base.

À partir de la formule 4.1, il faut maintenant être en mesure d'obtenir un estimateur robuste des distributions marginales et de la copule paramétrique représentant le mieux le lien de dépendance entre les variables composant les observations dans un contexte de censure omniprésente.

La méthodologie proposée dans ce travail fait appel à l'estimateur de Kaplan-Meier et à une pondération issue directement de cet estimateur. Soit $\Gamma_n$, l'estimateur, au sens de Kaplan-Meier, de la fonction de répartition $F$, inférée à partir des observations $Y_i$. Alors, on définit pour toute valeur $y$, $\Gamma_n$ par

$$\Gamma_n(y) = \begin{cases} 1 - \prod\limits_{i:1\leqslant i\leqslant n, Y_{(i)}<y} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}} & \text{si } y < Y_{(n)}; \\ \\ 0 & \text{autrement.} \end{cases}$$

Par ailleurs, on suppose la complétude sur l'ensemble des variables explicatives. Alors, on propose l'utilisation d'une fonction de distribution empirique rééchelonnée (pour éviter les problèmes de divergence de la densité de copule aux points $(0,0)$ et $(1,1)$) pour les prédicteurs $G_j$, $j = 1, 2, ..., d$ telle que

$$\hat{G}_{j,n}(x_j) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}(X_i \leqslant x_j).$$

Autrement, si les variables sont un mélange de données complètes et censurées, on suggère ici de considérer la distribution empirique rééchelonnée pour les distributions marginales des variables non-censurées, et l'estimateur de Kaplan-Meier pour les distributions marginales des variables censurées.

En ce qui a trait à l'estimation de la famille de copules, ce travail considère uniquement l'utilisation des copules paramétriques pour approcher la loi jointe entre les fonctions marginales étant donné qu'en plus de donner l'opportunité de déterminer l'estimateur des marges, l'inférence de la copule paramétrique se résume, une fois les fonctions marginales déterminées, à inférer uniquement un vecteur de paramètres qui se limite généralement à un maximum de 2 paramètres. Ainsi, on suppose la copule $C(\cdot,\cdot,\theta)$ que l'on estime par $\hat{C}_k(\hat{F}(y), \hat{G}(x), \hat{\theta})$, $k = 1, 2, ..., K$ étant le nombre de familles paramétriques considérées pour l'estimation de la loi jointe. Alors, utilisant la densité de la copule sélectionnée, on propose l'estimateur semi-paramétrique de la régression suivant :

$$\begin{aligned} \hat{m}(x) &= \frac{\int_{\mathbb{D}(\Gamma_n)} y c(\Gamma_n(y), \hat{G}_{1,n}(x_1), ..., \hat{G}_{d,n}(x_d), \hat{\theta}_n) d\Gamma_n(y)}{\tilde{c}(\hat{G}_{1,n}(x_1), ..., \hat{G}_{d,n}(x_d), \hat{\theta}_n)} \\ &= \frac{\sum_{i=1}^{n} Y_i w_i c(\Gamma_n(Y_i), \hat{G}_{1,n}(x_1), ..., \hat{G}_{d,n}(x_d), \hat{\theta}_n)}{\tilde{c}(\hat{G}_{1,n}(x_1), ..., \hat{G}_{d,n}(x_d), \hat{\theta}_n)}. \end{aligned}$$

où $\omega_1 = \Gamma_n(Y_{(1)})$, $\omega_i = \Gamma_n(Y_{(i)}) - \Gamma_n(Y_{(i-1)})$ pour $i = 2, ..., n$ et $\mathbb{D}(\Gamma_n)$ est le domaine de $\Gamma_n$.

**Remarque 4.2.1.** *L'hypothèse qui suit peut être faite dans le cas où l'on utilise n'importe quel autre estimateur que l'estimateur de la fonction de distribution empirique pour les données complètes ou l'estimateur de Kaplan-Meier pour les données censurées pour évaluer $F$ and $G$ respestivement.*

— *Soit $\tilde{G}_j(x)$ dénotant l'estimateur utilisé pour les covariables et $\tilde{\Gamma}_n(y)$ celui pour la variable réponse. Alors, les estimateurs doivent satisfaire la condition qui stipule que, pour*

*toute valeur ponctuelle de $x$, $x \in \mathbb{R}^d$, où l'on désire estimer la fonction de régression, les estimateurs sont tels que :*

— $\tilde{G}_j(x) = \hat{G}_{j,n}(x) + o_p(n^{-1/2})$ *pour* $j = 1, ..., d$;

— $\tilde{\Gamma}_n(y) = \Gamma_n(y) + o_p(n^{-1/2})$.

## 4.3 Cadre théorique

La contribution principale de cette section est de présenter deux résultats asymptotiques de l'estimateur présenté. Le premier établit la convergence uniforme, faible, de cet estimateur et le second établit sa représentation i.i.d (*indépendante et identiquement distribuée*). À partir de ces deux résultats, on déduit la normalité asymptotique de cet estimateur. Par ailleurs, on calcule sa variance asymptotique et on suggère un estimateur pratique de cette variance asymptotique. Pour commencer, on doit établir certaines hypothèses et conditions nécessaires à la consistence de ces résultats. On présente les résultats, dans un premier temps, appliqués au cas bivarié puis, à la section **3.3**, on présente leur généralisation au cas multivarié.

### 4.3.1 Hypothèses sur la copule et ses paramètres

On commence par les suppositions élémentaires suivantes pour s'assurer de la consistance de l'estimateur et introduire certains éléments de notation nouveaux. Ainsi, on doit émettre des hypothèses sur le paramètre de la copule et, à ces fins, on va généraliser certains résultats s'appliquant aux copules bidimensionnelles à des copules multidimensionnelles. Soit $\hat{\theta}_n$, l'estimateur de $\theta_0$, paramètre de la copule réelle égale à la fonction jointe entre les variables observées. Alors, $\hat{\theta}_n$ doit satisfaire

**Hypothèse A :**

$$\hat{\theta}_n - \theta_0 = \frac{1}{n} \sum_{i=1}^{n} \zeta_i + o_p(n^{-1/2})$$

où $\zeta_i$ sont des variables i.i.d. de moyenne nulle et de variance finie.

Dans ce travail, on considère l'estimation de $\theta$ selon la méthologie de Shih et Louis[77] qui est en fait une procédure d'estimation du maximum de vraisemblance en deux étapes paramétriques pour une situation de données censurées à droite. Il s'agit en fait d'un cas particulier de la méthode IFM (*inference for margins*) de Joe et Xu.[78] La validité de cette procédure repose sur la distribution limite de $\hat{\theta}_n$ qui est telle que $\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, B^{-1}(\theta_0))$ où $B$ est la matrice d'information de Godambe.

Afin de pouvoir émettre des hypothèses sur la copule utilisée pour l'inférence, on doit établir la notation suivante[76] :

— $e(\mathbf{u}) = \mathbb{E}(Yc(F(Y), \mathbf{u})) = \displaystyle\int_0^1 F^{-1}(u_0)c(u_0, \mathbf{u})du_0$, $\mathbf{u} = (u_1, ..., u_d)$ ;

— $\partial_j c = \dfrac{\partial c}{\partial u_j}$ for $j = 0, 1, ..., d$ and $\dot{\mathbf{c}} = \left(\dfrac{\partial c}{\partial \theta_1}, ..., \dfrac{\partial c}{\partial \theta_p}\right)^T$.

Avec cette notation, on peut établir l'hypothèse qui suit concernant la continuité des fonctions de masse jointes exprimées à l'aide d'une densité de copule.

**Hypothèse B :**

Soit $q$, une fonction de $\dot{\mathbf{c}}$ ou de $\partial_j c$, $j = 0, 1, ..., d$ et $x \in \mathbb{R}^d$ étant une valeur ponctuelle arbitraire telle que $G(\mathbf{x}) \in (0,1)^d$. Alors,

— $(\mathbf{u}, \theta) \to q_{u_0}(\mathbf{u}, \theta) \equiv q(u_0, \mathbf{u}; \theta)$ est continu en $(G(\mathbf{x}), \theta_0)$ et uniformément pour $u_0 \in [0,1]$

— $u_0 \to q(u_0, G(\mathbf{x}); \theta_0)$ est contenu en [0,1].

Finalement, on doit établir une hypothèse de continuité sur les dérivées partielles afin de s'assurer de la validité des résultats asymptotiques.

**Hypothèse C :**

Soient les éléments de continuité :

— $\mathbb{E}[Y] < \infty$

— $\mathbb{E}[c_j(F(Y), \boldsymbol{G(x)}; \theta_0)]^2, j = 1, 2, ..., d$ est finie, tout comme sa première dérivée par rapport à $\theta_0$

— $\mathbb{E}[Y c_j(F(Y), \boldsymbol{G(x)}; \theta_0)]^2, j = 1, 2, ..., d$ est finie, tout comme sa première dérivée par rapport à $\theta_0$

— $yF(y) \to 0$ lorsque $y \to -\infty$

— $\forall \epsilon \to 0$, $\dot{c}_j$ et $c_j$, $j = 1, 2, 3, ..., d$ sont continus sur les surfaces des parallélépipèdes formés par $[0,1] \times \prod_{j=1}^{d} [G_j(x_j) - \epsilon, G_j(x_j) + \epsilon]$.

### 4.3.2 Résultats principaux

On commence par considérer le cas univarié ; soit en présence de la covariable $X$. On prouve, à la proposition 4.3.1, la convergence uniforme en probabilités de l'estimateur proposé (et par conséquent, la convergence en loi) ; puis on présente sa représentation i.i.d au théorème 4.3.1. Finalement, la proposition 4.3.2 présente la normalité asymptotique vers une moyenne nulle et une variance finie de $\hat{m}$.

**Proposition 4.3.1.** *Sous les hypothèses A et B, on obtient*

$$\sup_x |\hat{m}(x) - m(x)| \xrightarrow{P} 0.$$

Avant de prouver cette proposition, on doit établir certaines notations et résultats asymptotiques de la littérature qui serviront ultérieurement dans cette preuve. Soit

$$q(s) = \int_0^s \left[\mathbb{P}(T_i > v, C_i^0 > v)\right]^{-2} d\mathbb{P}(Y_i \leqslant v, \delta_i = 1)$$

où l'on remarque que $H$ représente la fonction jointe de survie $Y_i$ et $C_i^0$ et $H^1$ représente la sous-distribution des observations complètes parmi l'ensemble des observations. Alors, pour $y, z \in \mathbb{R}^+$, on a

$$\xi(z, \delta, y) = -q(y \wedge z) + \frac{1}{\mathbb{P}(T_i > z, C_i^0 > z)} \mathbb{1}(z \leqslant y, \delta = 1)$$

où le symbole $\wedge$ représente l'élément minimal entre les éléments de part et d'autre de ce symbole. En posant comme notation que $\xi_i(y) = \xi(Y_i, \delta_i, y)$, on note que, $\forall u, v \in Y_i$,

$$\mathbb{E}(\xi_i(y)) = 0 \text{ et } cov(\xi_i(u), \xi_i(v)) = q(u \wedge v).$$

Le lemme qui suit, de Lo et al.,[79] exprime l'estimateur de Kaplan-Meier en tant qu'un processus i.i.d avec un résidu d'ordre négligeable :

**Lemme 4.3.1.** *([79])*
*Pour tout $t \leqslant T$*

$$\Gamma_n(y) - F(y) = n^{-1}\bar{F}(y)\sum_{i=1}^{n}\xi_i(y) + r_n(y),$$

*où le terme résiduel $r_n$ est tel que*

$$\sup_{0\leqslant y\leqslant T}|r_n(y)| = O\left(\frac{log\ n}{n}\right) \quad \textit{presque sûrement,}$$

*et pour tout $\alpha \geqslant 1$,*

$$\sup_{0\leqslant y\leqslant T}\mathbb{E}|r_n(y)|^\alpha = O\left(\left[\frac{log\ n}{n}\right]^\alpha\right).$$

On introduit ici une notation propre à chaque partie de cette décomposition de l'estimateur de Kaplan-Meier :

$$\begin{aligned}\Gamma_n(y) &= F_0(y) + n^{-1}\sum_{i=1}^{n}\xi(Y_i, \delta_i, y) + r_n(y)\\ &= F_0(y) + \kappa(y) + r_n(y).\end{aligned}$$

On remarque ainsi que $\kappa(y) = n^{-1}\sum_{i=1}^{n}\xi(Y_i, \delta_i, y)$. Le lemme suivant fournit une décomposition intéressante de l'estimateur proposé de la fonction de régression évaluée ponctuellement en $x$. Cette décomposition, tout comme la notation présentée ici, sera nécessaire à la démonstration de la proposition 4.3.1 et du théorème 4.3.1.

**Lemme 4.3.2.** *Soit $\hat{m}$, l'estimateur de la fonction de régression tel que présenté précédemment. Sous l'hypothèse B, on a*

$$\hat{m}(x) = m(x) + \sum_{j=1}^{4} I_{n,j} + o_p(n^{-1/2}) \tag{4.2}$$

*où*

$$I_{n,1} = \int_0^\tau y(\Gamma_n(y) - F_0(y))\partial c_0(F(y), G(x), \theta_0)dF(y),$$

$$I_{n,2} = (\hat{G}_n(x) - G(x))\int_0^\tau y\partial c_1(F(y), G(x), \theta_0)dF(y),$$

$$I_{n,3} = \int_0^\tau y(\hat{\theta}_n - \theta_0)^T\dot{\mathbf{c}}(F(y), G(x), \theta_0)dF(y),$$

*et*

$$I_{n,4} = \int_0^\tau yc(F(y), G(x), \theta_0)d\kappa(y).$$

**Démonstration :**

En utilisant la décomposition précédente de Lo et al.,[79] on obtient

$$\begin{aligned}\hat{m}(x) &= \int_0^\tau yc(\Gamma_n(y), \hat{G}_n(x), \hat{\theta}_n)d\Gamma_n(y) \\ &= \int_0^\tau yc(\Gamma_n(y), \hat{G}_n(x), \hat{\theta}_n)dF(y) + \int_0^\tau yc(\Gamma_n(y), \hat{G}_n(x), \hat{\theta}_n)d\kappa(y) \\ &+ \int_{-\infty}^\infty yc(\Gamma_n(y), \hat{F}_n(x), \hat{\theta}_n)dr_n(y) \\ &= I + II + III\end{aligned}$$

• Pour la partie (I) de l'estimateur :

Pour étudier la partie (I) de cet estimateur. on débute en utilisant une série de Taylor d'ordre 1 autour de $(F(y), G(x), \theta_0)$. On a alors

$$I = \int_0^\tau yc(F(y), G(x), \theta_0)dF(y) + V_1 + V_2 + V_3$$

où

$$\begin{aligned}V_1 &= \int_0^\tau y(\Gamma_n(y) - F(y))\partial c_0(\xi_1, \xi_2, \xi_3)dF(y) \\ V_2 &= (\hat{G}_n(x) - G(x)) \int_0^\tau y\partial c_1(\xi_1, \xi_2, \xi_3)dF(y) \\ V_3 &= \int_0^\tau y(\hat{\theta}_n - \theta_0)^T \dot{\mathbf{c}}(\xi_1, \xi_2, \xi_3)dF(y)\end{aligned}$$

dont les termes $\xi_1, \xi_2$ et $\xi_3$ sont

$$\begin{aligned}\xi_1 &= F(y) + t(\Gamma_n(y) - F(y)) \\ \xi_2 &= G(x) + t(\hat{G}_n(x) - G(x) \\ \xi_3 &= \theta_0 + t(\hat{\theta}_n - \theta_0)^T\end{aligned}$$

où $t$ est un scalaire, arbitraire, appartenant à l'intervalle $[0, 1]$ ; et $\partial c_0, \partial c_1, \dot{\mathbf{c}}$ sont respectivement les matrices des dérivées partielles des fonction copules par rapport à $u, v$ et $\theta_0$ lorsque l'on exprime la copule telle que $C_{\theta_0}(u, v)$.

Si l'on étudie $V_2$, on s'apperçoit que l'on peut réécrire ce terme tel que

$$\begin{aligned}V_2 &= (\hat{G}_n(x) - G(x) \int_0^\tau y\partial c_1(\xi_1, \xi_2, \xi_3)dF(y) \\ &= (\hat{G}_n(x) - G(x)) \int_0^\tau y\partial c_1(\xi_1, \xi_2, \xi_3)dF(y) \\ &= (\hat{G}_n(x) - G(x)) \int_0^\tau y\partial c_1(F(y), G(x), \theta_0)dF(y) \\ &+ (\hat{G}_n(x) - G(x)) \int_0^\tau y[\partial c_1(\xi_1, \xi_2, \xi_3) - \partial c_1(F(y), G(x), \theta_0)]dF(y)\end{aligned}$$

où l'on suppose que $[\partial c_1(\xi_1,\xi_2,\xi_3)-\partial c_1(F(y),G(x),\theta_0)]=o_p(1)$. Par ailleurs, grâce au théorème de Donsker,[80] on sait que $sup_x|\hat{G}_n(x)-G(x)|=O_p(n^{-1/2})$. Par conséquent, on obtient

$$\begin{aligned} V_2 &= (\hat{G}_n(x)-G(x))\int_0^\tau y\partial c_1(F(y),G(x),\theta_0)dF(y)+o_p(n^{-1/2}) \\ &\equiv I_{n,2}+o_p(n^{-1/2}). \end{aligned}$$

Ensuite, si l'on étudie $V_3$, on obtient la réécriture

$$\begin{aligned} V_3 &= \int_0^\tau y(\hat{\theta}_n-\theta_0)^T\dot{\mathbf{c}}(\xi_1,\xi_2,\xi_3)dF(y) \\ &= \int_0^\tau y(\hat{\theta}_n-\theta_0)^T\dot{\mathbf{c}}(F(y),G(x),\theta_0)dF(y) \\ &+ \int_0^\tau y(\hat{\theta}_n-\theta_0)^T[\dot{\mathbf{c}}(\xi_1,\xi_2,\xi_3)-\dot{\mathbf{c}}(F(y),G(x),\theta_0)]dF(y) \end{aligned}$$

où l'on suppose que $[\dot{\mathbf{c}}(\xi_1,\xi_2,\xi_3)-\dot{\mathbf{c}}(F(y),G(x),\theta_0)]=o_p(1)$. À partir du travail de Shih et Louis,[77] il est établit que $\sqrt{n}(\hat{\theta}_n-\theta_0)$ converge asymptotiquement vers une variable gaussienne de variance nulle et de variance $B^{-1}(\theta_0)$. Par ailleurs, il y est établit que $\sup|\hat{\theta}-\theta_0|=O_p(n^{-1/2})$. Par conséquent, on peut réécrire $V_3$ comme l'égalité

$$\begin{aligned} V_3 &= (\hat{\theta}_n-\theta_0)^T\int_0^\tau y\dot{\mathbf{c}}(F(y),G(x),\theta_0)dF(y)+o_p(n^{-1/2}) \\ &= I_{n,3}+o_p(n^{-1/2}) \end{aligned}$$

Pour ce qui en est de l'étude de $V_1$, on a la réécriture

$$\begin{aligned} V_1 &= \int_0^\tau y(\Gamma_n(y)-F(y))\partial c_0(\xi_1,\xi_2,\xi_3)dF(y) \\ &= \int_0^\tau y(\Gamma_n(y)-F(y))\partial c_0(F(y),G(x),\theta_0)dF(y) \\ &+ \int_0^\tau y(\Gamma_n(y)-F(y))[\partial c_0(\xi_1,\xi_2,\xi_3)-\partial c_0(F(y),G(x),\theta_0)]dF(y) \end{aligned}$$

Ici, on présume que $[\partial c_0(\xi_1,\xi_2,\xi_3)-\partial c_0(F(y),G(x),\theta_0)]=o_p(1)$. Il a été établi dans le travail de Lo et Singh[81] que $sup_y|\Gamma_n(y)-F(y)|=O_p(n^{-3/4}(log\ n)^{3/4})$. Par ailleurs, à partir de l'hypothèse $C$, pour toute constante $K\geqslant 1$, on déduit la majoration $|\partial c_0(\xi_1,\xi_2,\xi_3)-\partial c_0(F(y),G(x),\theta_0)|\leqslant K|\xi_1-F(y)|\leqslant K|\Gamma_n(y)-F(y)|=O_p(n^{-3/4}(log\ n)^{3/4})$. Par conséquent,

$$\begin{aligned} V_1 &= \int_0^\tau y(\Gamma_n(y)-F(y))\partial c_0(F(y),G(x),\theta_0)dF(y)+O_p(n^{-1}(log\ n)) \\ &= I_{n,1}+o_p(n^{-1/2}). \end{aligned}$$

Donc,

$$(I)=m(x)+I_{n,1}+I_{n,2}+I_{n,3}+o_p(n^{-1/2}).$$

- Pour la partie $(II)$ de l'estimateur :

On étudie

$$II = \int_0^\tau yc(\Gamma_n(y), \hat{F}_{1,n}(x_1), \hat{\theta}_n)d\kappa(y).$$

Par une série de Taylor d'ordre 1 autour de $(F(y), G(x), \theta_0)$, on a

$$II = \int_0^\tau yc(F(y), G(x), \theta_0)d\kappa(y) + W_1 + W_2 + W_3$$

où

$$\begin{aligned} W_1 &= \int_0^\tau y(\Gamma_n(y) - F(y))\partial c_0(\xi_1, \xi_2, \xi_3)d\kappa(y) \\ W_2 &= (\hat{G}_n(x) - G(x)) \int_0^\tau y\partial c_1(\xi_1, \xi_2, \xi_3)d\kappa(y) \\ W_3 &= (\hat{\theta}_n - \theta_0)^T \int_0^\tau y\dot{\mathbf{c}}(\xi_1, \xi_2, \xi_3)d\kappa(y). \end{aligned}$$

Maintenant, on doit montrer $W_1, W_2$ et $W_3$ de la série de Taylor d'ordre 1 évaluée sur (II) convergent vers 0. Pour cela, on montre pour $W_1$ que

$$\begin{aligned} W_1 &= \int_0^\tau y(\Gamma_n(y) - F(y))\partial c_0(\xi_1, \xi_2, \xi_3)d\kappa(y) \\ &= \int_0^\tau y(\Gamma_n(y) - F(y))\partial c_0(F(y), G(x), \theta_0)d\kappa(y) \\ &+ \int_0^\tau y(\Gamma_n(y) - F(y))[\partial c_0(\xi_1, \xi_2, \xi_3) - \partial c_0(F(y), G(x), \theta_0)]d\kappa(y). \end{aligned}$$

Étant donné qu'il a été montré que $\sup_y |\Gamma_n(y) - F(y)| = O_p(n^{-3/4}(\log\ n)^{3/4})$, on admet que $\sup_y |\kappa(y)| = O_p(n^{-3/4}(\log\ n)^{3/4})$. Ainsi,

$$\begin{aligned} \left|\int_0^\tau y(\Gamma_n(y) - F(y))d\kappa(y)\right| &\leqslant \tau \sup_y |\Gamma_n(y) - F(y)| \cdot \left|\int_0^\tau d\kappa(y)\right| \\ &= o_p(n^{-1/2}) \end{aligned}$$

et, en multipliant par $\sqrt{n}$, on obtient

$$\sqrt{n}\left|\int_0^\tau y(\Gamma_n(y) - F(y))d\kappa(y)\right| \leqslant O_p(n^{-3/4}(log\ n))$$

qui converge asymptotiquement vers 0. En utilisant la même technique, on montre que $W_2 = o_p(n^{-1/2})$ et $W_3 = o_p(n^{-1/2})$.

Donc,

$$(II) = I_{n,4} + o_p(n^{-1/2}).$$

• Pour la partie $(III)$ de cet estimateur :

En étudiant la partie (III) , on montre facilement qu'il s'agit d'un terme négligeable. En fait, on a

$$\begin{aligned}\int_{-\infty}^{\infty} yc(\Gamma_n(y), \hat{G}_n(x), \hat{\theta}_n)dr_n(y) &= ||c(\Gamma_n(y), \hat{G}_n(x), \hat{\theta}_n)||_\infty \left|\int_0^\tau dr_n(y)\right| \\ &= O_p(n^{-3/4}(log\ n)^{3/4}) \\ &= o_p(n^{-1/2}).\end{aligned}$$

Donc,

$$(III) = o_p(n^{-1/2}).$$

Ce qui fait la démonstration.

□

À la lumière de ces résultats, on peut finalement prouver la proposition 4.3.1 :

**Démonstration de la Proposition 4.3.1 :**

Pour commencer, on remarque, de la démonstration précédente, que

$$\sup_y |\Gamma_n(y) - F(y)| \text{ et } \sup_x \left|\hat{G}_n(x) - G(x)\right|$$

convergent vers zéro uniformément. Par ailleurs, à partir de l'hypothèse $B$, on remarque que $I_{n,1}$ et $I_{n,2}$ convergent également vers zéro uniformément. Ensuite, à partir des hypothèses $A$ et $B$, on voit que le terme $I_{n,3}$ converge aussi vers zéro uniformément. Finalement, étant donné que $\xi(y)$ et $\kappa = n^{-1}\sum_{i=1}^{n}\xi(Y_i, \delta_i, y)$ sont des variables aléatoires i.i.d de moyenne nulle et de variance finie, le terme $I_{n,4}$ tend uniformément vers zéro. Ce qui complète la démonstration.

□

**Théorème 4.3.1.** *Sous les hypothèses A et B, $\hat{m}$ admet la représentation i.i.d. suivante :*

$$\hat{m}(x) = m(x) + \frac{1}{n}\sum_{i=1}^{n}\eta_i(G(x)) + o_p(n^{-1/2}) \tag{4.3}$$

*où,* $\eta_i = \sum_{j=1}^{4}\eta_{i,j}(G(x))$, *avec*

$$\eta_{i,1}(G(x)) = \int_0^\tau y\xi(y)\partial c_0(F(y), G(x), \theta_0)dF(y),$$

$$\eta_{i,2}(G(x)) = (\mathbb{1}(X_i \leqslant x) - G(x))\int_0^\tau y\partial c_1(F(y), G(x), \theta_0)dF(y),$$

$$\eta_{i,3}(G(x)) = \int_0^\tau y\zeta_i\dot{\boldsymbol{c}}(F(y), G(x), \theta_0)dF(y),$$

*et*

$$\eta_{i,4}(G(x)) = \int_0^\tau \xi(y)d(yc(F(y), G(x), \theta_0))dF(y).$$

**Démonstration :**

Pour commencer, si l'on étudie le paramètre de la copule, à partir de l'hypothèse $A$, on obtient la différence

$$\hat{\theta}_n - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} \zeta_i + o_p(n^{-1/2})$$

où $\zeta_i$ sont des variables aléatoires i.i.d. de moyenne nulle et de variance finie. Ensuite, on remarque pour l'estimation de la variable réponse censurée, à partir de Lo et Singh,[81] que

$$\Gamma_n(y) \quad = \quad F(y) + n^{-1}\sum_{i=1}^{n} \xi(Yi, \delta_i, y) + r_n(y)$$

où $\xi$ représente également une série de variables aléatoires i.i.d. Donc, on utilise directement le lemme 4.3.2, ce qui termine cette preuve.

□

**Proposition 4.3.2.** *Sous les hypothèses $A$ et $B$, $\sqrt{n}(\hat{m}(x) - m(x))$ converge asymptotiquement vers une distribution normale de moyenne nulle et de varialle $\sigma^2 = \mathbb{E}(\eta_1^2)$.*

**Démonstration :**
La proposition 4.3.2 est déduite directement à partir du théorème 4.3.1.

□

### 4.3.3 Prolongement au cas multivarié

On étend ici les résulats montrés dans le cas multivarié, soit le cas où il y a au moins deux covariables ($d \geqslant 2$) affectant la variable réponse.

Tel que présenté précédemment, l'estimateur proposé de la fonction de régression avec données censurées est

$$m(\boldsymbol{x}) = \frac{e(\boldsymbol{G}(\boldsymbol{x}))}{\tilde{c}_{\boldsymbol{X}}(\boldsymbol{G}(\boldsymbol{x}))}. \tag{4.4}$$

où, en supposant une modélisation paramétrique au numérateur, on estime $e(\boldsymbol{G}(\boldsymbol{x}); \boldsymbol{\theta}_0)$ par une expression similaire au cas univarié, soit

$$\hat{e}(\hat{\boldsymbol{G}}(\boldsymbol{x})) = \sum_{i=1}^{n} Y_i w_i c(\Gamma_n(Y_i), \hat{G}_{1,n}(x_1), ..., \hat{G}_{d,n}(x_d), \hat{\theta_n})$$

où $d$ représente le nombre de covariables affectant la variable réponse et où $\hat{\boldsymbol{G}}(\boldsymbol{x}) = (\hat{G}_1(x_1), \ldots, \hat{G}_d(x_d))$.

Suivant le théorème 4.3.1 et sa preuve, il est évident que

$$\hat{e}(\tilde{\boldsymbol{G}}(\boldsymbol{x})) - e(\boldsymbol{G}(\boldsymbol{x})) = n^{-1}\sum_{i=1}^{n} \varphi_i(\boldsymbol{G}(\boldsymbol{x})) + o_p(n^{-1/2}), \tag{4.5}$$

où $\varphi_i = \sum_{j=1}^{4} \varphi_{i,j}$ avec

$$\varphi_{i,1}(\boldsymbol{G}(\boldsymbol{x})) = \int_0^\tau y\xi(Y_i,\delta_i,y)\partial c_0(F(y),\hat{\boldsymbol{G}}(\boldsymbol{x}),\theta_0)dF(y),$$

$$\varphi_{i,2}(\boldsymbol{G}(\boldsymbol{x})) = \sum_{j=1}^{d}(\mathbb{1}(X_{ji}\leqslant x_1) - G(x_1))\int_0^\tau y\partial c_j(F(y),\hat{\boldsymbol{F}}(\boldsymbol{x}),\theta_0)dF(y),$$

$$\varphi_{i,3}(\boldsymbol{G}(\boldsymbol{x})) = \int_0^\tau y\zeta_i^T\dot{\mathbf{c}}(F(y),\hat{\boldsymbol{G}}(\boldsymbol{x}),\theta_0)dF(y),$$

et

$$\varphi_{i,4}(\boldsymbol{G}(\boldsymbol{x})) = \int_0^\tau \xi(Y_i,\delta_i,y)d(yc(F(y),\hat{\boldsymbol{G}}(\boldsymbol{x}),\theta_0)).$$

En se servant de la définition de $\tilde{c}$, soit $\tilde{c}_{\boldsymbol{X}}(\boldsymbol{u}) = \mathbb{E}[c(F(Y),\boldsymbol{u})]$, on propose d'estimer $\tilde{c}_{\boldsymbol{X}}(\boldsymbol{G}(\boldsymbol{x}))$ par

$$\hat{c}_{\boldsymbol{X}}(\tilde{\boldsymbol{G}}(\boldsymbol{x})) = \sum_{i=1}^{n} w_i c(\Gamma_n(Y_i),\hat{\boldsymbol{G}}(\boldsymbol{x});\hat{\boldsymbol{\theta}}).$$

Alors, l'estimateur proposé, $\hat{m}(\boldsymbol{x})$ est donné par

$$\hat{m}(\boldsymbol{x}) = \frac{\hat{e}(\hat{\boldsymbol{G}}(\boldsymbol{x}))}{\hat{c}_{\boldsymbol{X}}(\hat{\boldsymbol{G}}(\boldsymbol{x}))} = \sum_{i=1}^{n} Y_i \frac{w_i c(\hat{F}(Y_i),\hat{\boldsymbol{G}}(\boldsymbol{x});\hat{\boldsymbol{\theta}})}{\sum_{i=1}^{n} w_i c(\hat{F}(Y_i),\hat{\boldsymbol{G}}(\boldsymbol{x});\hat{\boldsymbol{\theta}})}.$$

La représentation asymptotique de $\hat{c}_{\boldsymbol{X}}(\tilde{\boldsymbol{G}}(\boldsymbol{x}))$ est déterminée en utilisant les mêmes arguments que dans le théorème 4.3.1. En fait, suivant ce théorème, il est peu contraignant de vérifier que

$$\hat{c}_{\boldsymbol{X}}(\tilde{\boldsymbol{G}}(\boldsymbol{x})) - c_{\boldsymbol{X}}(\boldsymbol{G}(\boldsymbol{x})) = n^{-1}\sum_{i=1}^{n}\phi_i + o_p(n^{-1/2}), \tag{4.6}$$

où $\phi_i(\boldsymbol{G}(\boldsymbol{x})) = \sum_{j=1}^{4}\phi_{i,j}(\boldsymbol{G}(\boldsymbol{x}))$ avec

$$\phi_{i,1}(\boldsymbol{G}(\boldsymbol{x})) = \int_0^\tau y\xi(Y_i,\delta_i,y)\partial c_0(F(y),\hat{\boldsymbol{G}}(\boldsymbol{x}),\theta_0)dF(y),$$

$$\phi_{i,2}(\boldsymbol{G}(\boldsymbol{x})) = \sum_{j=1}^{d}(\mathbb{1}(X_{ji}\leqslant x_1) - G_1(x_1))\int_0^\tau y\partial c_j(F(y),\hat{\boldsymbol{G}}(\boldsymbol{x}),\theta_0)dF(y),$$

$$\phi_{i,3}(\boldsymbol{G}(\boldsymbol{x})) = \int_0^\tau y\zeta_i^T\dot{\mathbf{c}}(F(y),\hat{\boldsymbol{G}}(\boldsymbol{x}),\theta_0)dF(y),$$

et

$$\phi_{i,4}(\boldsymbol{G}(\boldsymbol{x})) = \int_0^\tau \xi(Y_i,\delta_i,y)d(yc(F(y),\hat{\boldsymbol{G}}(\boldsymbol{x}),\theta_0)).$$

En combinant (4.5) et (4.6), on arrive au résultat principal :

**Théorème 4.3.2.** *Sous l'hypothèse C, si $\tilde{F}$ satisfait l'hypothèse A et $\hat{\boldsymbol{\theta}}$ satisfait l'hypothèse B, on a alors*

$$\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} \frac{1}{c_{\boldsymbol{X}}(\boldsymbol{G}(\boldsymbol{x}))} \left[\varphi_i(\boldsymbol{G}(\boldsymbol{x})) - m(\boldsymbol{x})\phi_i(\boldsymbol{G}(\boldsymbol{x}))\right] + o_p(n^{-1/2}).$$

**Démonstration :**

Application stricte du théorème 4.3.1 et de sa preuve.

□

## 4.4 Sélection du modèle de copule

Lorsque la famille de copules paramétriques est bien spécifiée et que le modèle à l'intérieur de cette famille est adéquat, l'estimateur proposé de la fonction de régression converge vers la vraie régression entre la variable réponse et les covariables l'affectant. Toutefois, la mauvaise spécification d'une densité de copule, dans le cas de données continues, peut mener à de sérieux biais d'inférence (voir Dette et al.[82]). Dans le cas de données complètes, plusieurs tests d'adéquation sont proposés pour sélectionner la copule adéquate (pour une revue exhaustive de ces méthodes, voir Genest et al.[83]). Dans le cas des données censurées ; cas de figure de ce chapitre, Shih[84] et Glidden[85] ont proposé des tests d'adéquation qui ont le problème de se limiter à une classe trop restreinte de copules. En se basant sur l'estimateur présenté dans ce travail, on propose un processus de choix simple et pratique de la fonction copule basé sur un critère de distance.

On considère $c_1, c_2, ..., c_k$ ; $k$ représentant une quantitée finie de modèles de copules paramétriques à comparer. Par exemple, $c_1$ peut être une copule gaussienne, $c_2$ une copule de Gumbel et ainsi de suite. Par ailleurs, on considère l'estimateur de la régression

$$\hat{m}^{c_j}(x) = \sum_{i=1}^{n} y_i c_j(\Gamma_n(y_i), \hat{G}_{i,n}(x_i), \hat{\theta}) w(y)$$

où $j \in \{1, 2, ..., k\}$. Ainsi, on propose de calculer une fonction de différence $\Delta$ pour les $k$ copules possibles telle que

$$\Delta^{c_j} = \sum_{i=1}^{n} (y_i - \hat{m}^{c_j}(x_i))^2 w(y_i), \;\; j = 1, 2, ..., k.$$

Alors, la copule qui procure le $\min(\Delta^{c_1}, \Delta^{c_2}, ..., \Delta^{c_k})$ est celle à sélectionner pour obtenir l'estimateur le plus consistent possible de la fonction de régression. Dans la section suivante, on enquête sur ce critère de sélection via des simulations et on montre la performance de cette méthode.

## 4.5 Simulations

Dans cette section, on montre, dans un premier temps, que le critère de sélection de la copule ayant la meilleure adéquation avec les données, présenté dans ce travail, est valide même dans le

cas d'une censure à droite ; puis on montre la performance de l'estimateur proposé via certaines procédures de générations de données (PGDs) en comparant diverses familles et modèles de copules et ce, à différents degrés de censure.

### 4.5.1 Évaluation du critère de sélection des données

Afin d'évaluer la performance du critère de sélection proposé, on a généré des données à partir de cinq familles différentes de copules : familles de Clayton, de Frank, gaussienne, de Gumbel et de Student. Pour chaque copule utilisée à la génération des données, on a fait usage du même tau de Kendall, $\tau = 0.75$, ainsi que de trois niveaux de censure : 0%, 25% et 32%. La distribution de la censure suivait une loi exponentielle. Dans chaque cas, on a généré $B = 500$ populations bivariées comportant une taille de $n = 200$ observations par variable ; puis sélectionné l'estimateur qui permettait d'obtenir le minimum de $\Delta^{c_j}, j = 1, 2, ..., k$. À la table 4.1, on voit, pour chaque famille de copules, la moyenne et l'écart-type de la valeur $\Delta^c$ obtenue sur l'ensemble des 500 simulations.

Comme il peut être vu à la table 4.1, à un niveau de censure de 0%, le critère de sélection de la copule est robuste et une mauvaise spécification du modèle est donc évitée. À un niveau de censure supérieur ou égal à 25%, le critère est valide pour toutes les copules, sauf la copule de Student qui est moins efficace que la copule gaussienne. Cela s'explique aisément par la proximité entre ces deux copules elliptiques (qui sont les deux seules copules elliptiques du panel de copules à comparer d'ailleurs). Ainsi, similairement aux statistiques inférentielles classiques où, sous certaines conditions, la loi normale peut approximer la loi de Student, la copule normale peut être une approximation de la copule de Student dans certains cas.

### 4.5.2 Performance de l'estimateur proposé

Dans cette section, on s'adonne à deux procédures de génération de données (PGDs) avec un niveau de censure de 25%. On note que la censure est générée à partir d'une loi exponentielle. Les PGDs effectuées dans ce travail sont les suivantes :

— **PGD 1 :** $(F(y), G(x))$ générés à partir d'une copule gaussienne de paramètre $\rho = 0.8$ ; $Y \sim (N)(\mu_Y = 0, \sigma_Y^2 = 1)$

1. La fonction de régression théorique résultante est $m(x) = 0.8\Phi + 1(F_X(x))$ où $\Phi$ est la cdf d'une loi $N(0, 1)$
2. $X$ est généré à partir d'une loi $N(0, 1)$.

— **DGP 2 :** $(F(y), G(x))$ générés à partir d'une copule de Farlie-Gumbel-Morstenstein de paramètre $\theta = 0.8$

1. La fonction de régression théorique résultante est $m(x) = (-\frac{0.8}{\sqrt{\pi}}) + 2(\frac{0.8}{\sqrt{\pi}})\sigma_Y F_X(x)$
2. $Y \sim N(0, 1)$, $X \sim exp(1)$.

| Estimateur \ Générateur | | gaussienne | Clayton | Gumbel | Frank | Student |
|---|---|---|---|---|---|---|
| gaussienne | $c = 0\%$ | **1.671** $\sigma = 0.285$ | 1.630 $\sigma = 0.283$ | 1.703 $\sigma = 0.301$ | 1.459 $\sigma = 0.145$ | 1.776 $\sigma = 0.347$ |
| | $c = 25\%$ | **5.449** $\sigma = 0.864$ | 5.474 $\sigma = 1.028$ | 5.924 $\sigma = 0.797$ | 5.345 $\sigma = 0.971$ | **5.485** $\sigma = 0.891$ |
| | $c = 32\%$ | **5.585** $\sigma = 0.944$ | 5.755 $\sigma = 1.030$ | 6.170 $\sigma = 0.932$ | 6.748 $\sigma = 0.824$ | **5.710** $\sigma = 0.946$ |
| Clayton | $c = 0\%$ | 1.767 $\sigma = 0.290$ | **1.593** $\sigma = 0.278$ | 1.820 $\sigma = 0.320$ | 1.436 $\sigma = 0.165$ | 1.877 $\sigma = 0.359$ |
| | $c = 25\%$ | 5.827 $\sigma = 0.735$ | **5.114** $\sigma = 0.841$ | 5.560 $\sigma = 0.671$ | 5.865 $\sigma = 0.719$ | 5.888 $\sigma = 0.757$ |
| | $c = 32\%$ | 6.065 $\sigma = 0.767$ | **5.318** $\sigma = 0.829$ | 6.910 $\sigma = 0.776$ | 6.820 $\sigma = 0.727$ | 6.177 $\sigma = 0.817$ |
| Gumbel | $c = 0\%$ | 1.677 $\sigma = 0.278$ | 1.969 $\sigma = 0.299$ | **1.702** $\sigma = 0.291$ | 1.494 $\sigma = 0.150$ | 1.776 $\sigma = 0.332$ |
| | $c = 25\%$ | 6.586 $\sigma = 1.255$ | 6.312 $\sigma = 1.401$ | **5.549** $\sigma = 1.218$ | 6.074 $\sigma = 1.690$ | 6.618 $\sigma = 1.266$ |
| | $c = 32\%$ | 6.805 $\sigma = 1.414$ | 6.541 $\sigma = 1.379$ | **5.710** $\sigma = 1.394$ | 7.479 $\sigma = 1.147$ | 6.894 $\sigma = 1.329$ |
| Frank | $c = 0\%$ | 1.685 $\sigma = 0.288$ | 1.594 $\sigma = 0.283$ | 1.718 $\sigma = 0.307$ | **1.445** $\sigma = 0.142$ | 1.795 $\sigma = 0.352$ |
| | $c = 25\%$ | 5.722 $\sigma = 0.908$ | 5.636 $\sigma = 1.049$ | 5.813 $\sigma = 0.840$ | **5.262** $\sigma = 0.940$ | 6.263 $\sigma = 0.939$ |
| | $c = 32\%$ | 5.953 $\sigma = 0.969$ | 5.928 $\sigma = 1.039$ | 6.051 $\sigma = 0.967$ | **6.711** $\sigma = 0.876$ | 6.526 $\sigma = 1.008$ |
| Student | $c = 0\%$ | 1.676 $\sigma = 0.288$ | 1.864 $\sigma = 0.186$ | 1.705 $\sigma = 0.302$ | 1.493 $\sigma = 0.1.541$ | **1.776** $\sigma = 0.348$ |
| | $c = 25\%$ | 6.228 $\sigma = 0.904$ | 5.865 $\sigma = 1.070$ | 6.299 $\sigma = 0.836$ | 5.698 $\sigma = 1.017$ | 5.785 $\sigma = 0.924$ |
| | $c = 32\%$ | 6.429 $\sigma = 0.0.993$ | 6.096 $\sigma = 1.046$ | 6.522 $\sigma = 0.971$ | 7.001 $\sigma = 0.943$ | 6.080 $\sigma = 0.954$ |

TABLEAU 4.1 – Évaluation de la valeur moyenne de $\Delta^c$ pour différentes copules à un tau de Kendall de 0.75. Les valeurs exprimées sont des pourcentages.

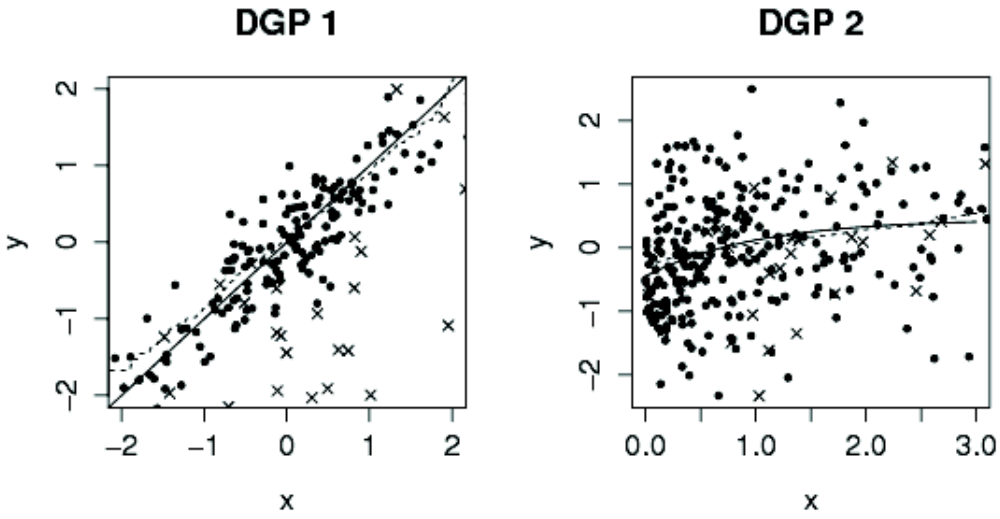


Figure 4.1 – Exemple du diagramme de dispersion des données pour les deux procédures de génération des données avec 300 données censurées exponentiellement à 25%. Un point représente une donnée complète et une croix une donnée censurée. La fonction de régression théorique est représentée par une ligne pleine alors que la régression basée sur les copules est représentée à l'aide d'une ligne pointillée.

| | | n=100 | n=250 | n=500 |
|---|---|---|---|---|
| **PGD 1** | MSE | 0.1148 | 0.1127 | 0.1096 |
| | $Biais^2$ | 0.0151 | 0.0141 | 0.0135 |
| | $\sigma$ | 0.0997 | 0.0986 | 0.0961 |
| **PGD 2** | MSE | 0.1284 | 0.1281 | 0.1267 |
| | $Biais^2$ | 0.0187 | 0.188 | 0.0176 |
| | $\sigma$ | 0.1097 | 0.1093 | 0.1091 |

Tableau 4.2 – Évaluation de l'erreur quadratique moyenne et de l'écart-type sur les 300 réplications des deux procédures de génération de données sous les trois niveaux.

À la figure 4.1, on peut voir un exemple des deux PGD avec leurs données censurées représentées par une croix. Pour chaque PDG, on a effectué $B = 300$ réplications du processus de simulation pour 100, 200 et 500 observations sur chaque variable. Les moyennes sur les 500 réplications du biais au carré, de la variance et, conséquemment, de l'erreur quadratique moyenne de l'estimateur proposé sont consignées au tableau 4.2. On spécifie que ces valeurs ont été calculées en se basant sur la différence entre les valeurs théoriques de la régression et celles trouvées à l'aide de l'estimateur. Ainsi, dans certains cas, il arrive que la valeur théorique de la régression simulée ne représente pas sa valeur réelle.

| Estimateur | Intercept | | Âge | | Âge$^2$ | |
|---|---|---|---|---|---|---|
| | $\hat{\alpha}$ | $SD(\hat{\alpha})$ | $\hat{\beta}_1$ | $SD(\hat{\beta}_1)$ | $\hat{\beta}_2$ | $SD(\hat{\beta}_2)$ |
| Buckley & James | 1.35 | 0.71 | 0.107 | 0.037 | -0.0017 | 0.0005 |

TABLEAU 4.3 – Coefficients pour la régression sur le Stanford heart transplant dataset à partir de la méthode de Buckley et James.

### 4.5.3 Application à des données sur la transplantation cardiaque

À partir des données utilisées par Miller et Halpern,[23] on va comparer l'estimateur de la régression en présence de censure selon ces derniers, à l'estimateur basé sur les copules présenté dans ce travail. Les données que l'on utilise ici proviennent du jeu de données bien connu *Stanford Heart transplant dataset*, provenant de l'étude d'une procédure de transplantation cardiaque ayant débutée en octobre 1967. Les patients étaient présélectionnés pour pouvoir intégrer l'étude (et s'assurer de l'homogénéité entre les cohortes étudiées) et conséquemment recevoir un coeur. Entre le moment de la sélection d'un patient et l'opération de transplantation, certains individus sont décédés, ce qui a mené leur temps de survie à la valeur 0. La date de point de l'étude a été janvier 1980 et, à ce moment, les données concernant la transplantation de 184 patients ont été collectées. Les variables d'intérêt ici étaient le temps de survie (en jours), le statut de la survie *(1 si décès, 0 si survie ; ce qui est équivalent à 1 s'il s'agit s'une donnée complète et 0 si c'est une donnée censurée)*, l'âge du patient au moment de la première transplantation et le score de mauvaise concordance greffon-receveur (*mismatch score*). On en déduit donc qu'il y a présence d'une forte censure administrative (et informative) sur ces données.

Afin de travailler sur exactement les mêmes données que Miller et Halpern, on conserve uniquement les patients qui ont survécu un minimum de 10 jours ; ainsi que ceux pour lesquels le score de mauvaise concordance greffon-receveur n'est pas manquant. Ainsi, on obtient un échantillon de 152 données avec un niveau de censure de 36.18%. Dans leur article, Miller et Halpern constatent que la régression avec la meilleure adéquation à ces données censurées est celle de Buckley et James[22] pour laquelle les paramètres sont présenté au tableau 4.3.

Comme on peut le constater à la figure 4.2, il y a évidence que les deux estimateurs proposent une régression non-linéaire. Toutefois, avec Buckley et James, on doit spécifier que l'on désire une régression de degré 2 dans l'élaboration du modèle ; ce qui n'est pas nécessaire avec la modélisation basée sur les copules. Par ailleurs, l'estimateur basé sur les copules a l'avantage de ne pas être lisse comme celui de Buckley et James ; ce qui peut lui conférer ponctuellement une meilleure estimation de la fonction de régression en considérant toutes les variations locales dans la dispersion des données. On note que pour ce jeu de données, le critère de sélection de la copule proposé dans ce travail propose l'utilisation d'une copule gaussienne.

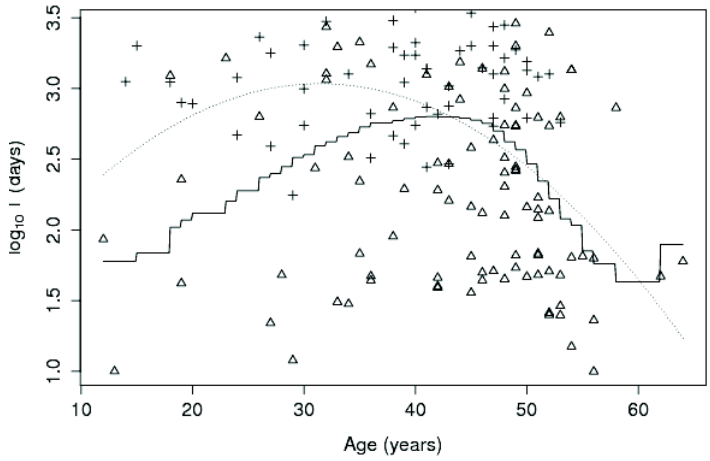


Figure 4.2 – Diagramme de dispersion des temps de survie en fonction de l'âge du receveur pour la transplantation cardiaque de 152 patients. La ligne continue représente l'estimateur de la régression basé sur les copules alors que celle pointillée représente la régression de Buckley et James. On remarque qu'une croix représente une donnée censurée.

## 4.6 Discussion et Conclusion

Dans ce chapitre, une méthodologie a été proposée quant à une modélisation de la fonction de régression en présence d'un phénomène de censure non-informative sur la variable réponse. Cette approche, qui est en fait une réécriture de la régression en tant qu'espérance conditionnelle et, par conséquent, en tant que sommation sur la fonction de densité conditionnelle écrite comme le ratio de Bayes, a l'avantage de ne pas être un modèle imposant une linéarité sur les covariables (e.g. régression selon les moindres carrés ordinaires), mais plutôt un modèle fixant la forme de la dépendance entre la variable réponse et ses covariables via une fonction copule. Comme on peut le voir à la figure 4.3, la loi marginale liée à chaque variable est estimée de façon indépendante, puis il ne reste qu'à déterminer la copule avec la meilleure adéquation aux données et à estimer son paramètre de dépendance afin d'écrire le modèle.

Enfin, les approches paramétriques connues quant à ce type de régression (e.g. Buckley et James (1979)) sont désavantagées quant à l'adéquation aux données censurées par rapport à l'approche basée sur les copules présentée dans ce travail en raison de leur prémisse de linéarité. Toutefois, les approches de régression complètement non-paramétriques ont un net avantage par rapport aux copules paramétriques en ne se limitant pas à un modèle, mais en épousant la relation entre les données dans un voisinage immédiat du point d'évaluation de la fonction de régression. Ainsi, dans une perspective future, il serait intéressant de considérer le cas de l'utilisation d'une copule complètement non-paramétrique pour évaluer la fonction de régression avec des données observationnelles.

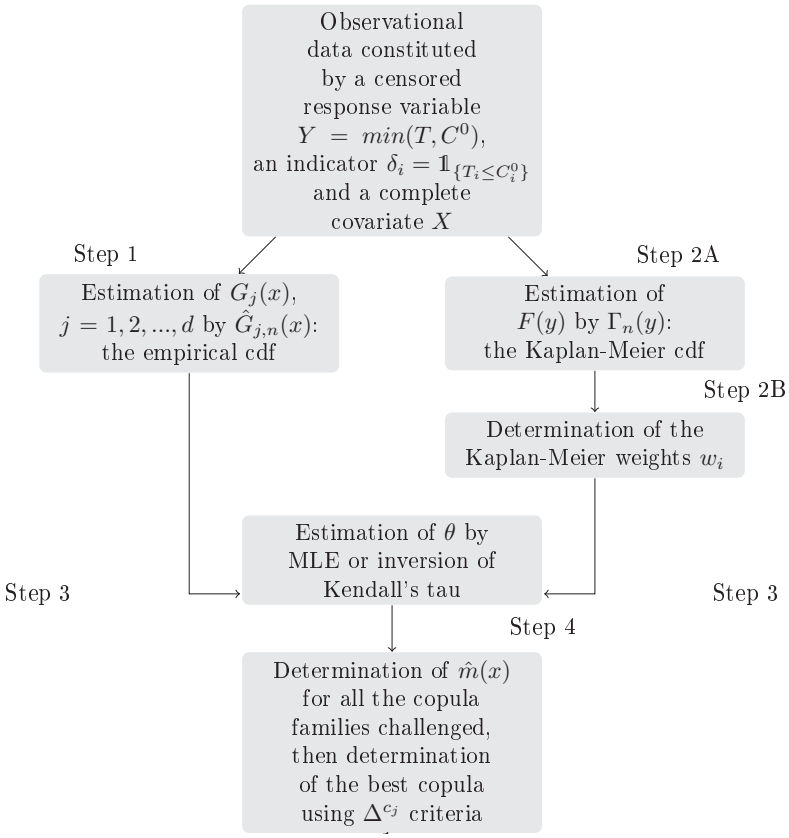


FIGURE 4.3 – Schéma de la modélisation de la fonction de régression en présence de données censurées via l'utilisation de fonctions copules paramétriques.

# Conclusion

L'objectif de ce travail de thèse fut de montrer que malgré la présence de données ayant une structure singulière, recueillies lors d'essais thérapeutiques non-randomisés, il est possible d'écrire les relations entre les variables aléatoires observées à partir des fonctions de répartition jointes et marginales ; puis d'utiliser ces relations pour modéliser directement l'objet de l'étude sans passer par un modèle sous-tendant la linéarité entre les covariables. Ainsi, l'écriture de ces fonctions de répartition jointes à partir de la fonction copule a été la base des modélisations innovantes présentées dans cette thèse.

Pour commencer, on a présenté un rappel des principales propriétés des copules paramétriques et des mesures d'association directement liées à ces copules. Le lien existant entre ces deux concepts se situe au niveau du paramètre de dépendance de la copule : ayant une structure représentant la forme de dépendance entre deux (ou plusieurs) fonctions de répartition marginales, soit la représentation issue de la famille de copule, le paramètre de dépendance va représenter la «force» de cette dépendance. Ainsi, une mesure de dépendance peut être calculée, mis à part via une méthode de maximisation de la vraisemblance, à partir d'une mesure d'association et, pour chaque type de copule, il existe donc une relation fonctionnelle entre le paramètre de la copule et une mesure d'association du type tau de Kendall ou rho de Spearman.

Au chapitre 2, on s'est placé dans un contexte d'analyse coût-efficacité où l'efficacité était mesurée en terme d'utilité. À partir de coûts ponctuellement mesurés, de temps de survie de sujets appartenant à un bras thérapeutique donné et de score de qualité de vie, le principe est de commencer par calculer les coûts cumulatifs puis de déterminer la valeur de QALY en ajustant, par une contraction, les temps de survie aux scores de qualité de vie. Obtenant ainsi les variables $T_{adj}$ et $C$ sur chaque bras clinique, il faut ensuite déterminer les fonctions de répartition marginales et mesurer la dépendance entre ces dernières à partir d'une mesure d'association ; puis déterminer la copule avec la meilleure adéquation aux données. On remarque que bien que l'on ait proposé un critère bayésien pour la sélection du type de copule dans le travail de recherche qui constitue ce chapitre, en raison d'une connaissance a priori sur les paramètres des distributions, un critère de sélection basé sur une distance entre la copule empirique et la copule que l'on compare aurait pu être tout aussi efficace. Cela dit, à partir de la densité de la copule sélectionnée, il est possible d'obtenir l'expression de l'espérance des variables QALY et coûts cumulatifs sur chaque bras et, ainsi, d'obtenir les quantités d'intérêt que sont l'ICER et l'INB, tout comme leurs intervalles de confiance à partir de la méthode de Fieller.

Au chapitre 3, on s'est placé dans un contexte où les fonctions de répartition multivariées sont encore très peu utilisées en raison de leurs limitations théoriques, soit le contexte de données discrètes constituant des données observationnelles ; plus précisément dans le cadre où l'on désire calculer le score de propension sans avoir les contraintes de la linéarité entre les prédicteurs que propose la régression logistique. On a ainsi commencé par proposer une méthode d'extension

d'une sous-copule $C'$ pour obtenir une copule $C$ définie sur le carré unitaire et unique en ce domaine grâce à une réécriture des fonctions de répartition marginales constituée d'interpolations linéaires aux points de discontinuité. On a ensuite montré que le tau de Kendall n'est pas affecté par une telle réécriture, puis on s'est placé dans un cadre d'estimation des paramètres et de la meilleure fonction copule joignant la variable réponse aux covariables d'intérêt. Il est ainsi possible de réécrire le score de propension sous la contrainte d'un nombre limité de covariables d'intérêt.

Au chapitre 4, l'estimation de la fonction de régression avec des données censurées à droite a été abordée. Pour ce faire, on a défini un modèle basé sur une copule semi-paramétrique : le type de copule modélisant la dépendance est paramétrique alors que les fonctions marginales la constituant sont non-paramétriques ( fonction de répartition empirique pour une variable constituée de données complètes et estimateur de Kaplan-Meier pour une variable avec des données censurées). En utilisant la densité de cette copule, il est simple de réécrire la fonction régression en tant qu'espérance conditionnelle entre deux (ou plusieurs) variables et ainsi, d'obtenir un estimateur convergent de la régression. On a, par ailleurs, montré la convergence en probabilité et la décomposition asymptotique de cet estimateur.

Enfin, on remarque que cette thèse laisse place à plusieurs perspectives futures de travail. Pour commencer, l'ensemble des modèles présentés dans ce document reposent sur des copules paramétriques. Par contre, il serait possible de modéliser le tout avec des estimateurs de copules entièrement non-paramétriques (par exemple, l'estimateur empirique de copule basé sur les polynômes de Bernstein). Ainsi, il y aurait une diminution du biais d'inférence en raison que les paramètres des distributions n'auraient pu à être estimés. Ensuite, pour ce qui en est de la modélisation du score de propension qui a été présenté ici, le modèle est restrictif dans le sens qu'il permet difficilement de modéliser ce score dans le cas où il y a plusieurs covariables qui affectent la variable traitement. C'est ainsi qu'il serait intéressant d'intégrer le concept de copules en vignes (Vine copulas) à la modélisation proposée afin de pouvoir tenir compte de plusieurs covariables d'intérêt sans problèmes d'un point de vue computationnel. Pour terminer, on affirme que la copule est un outil puissant avec les données dépendantes ; données qui sont omniprésentes en recherche clinique. C'est pourquoi, sachant qu'il y a encore une panoplie de modèles statistiques d'analyse des données cliniques qui sont basées sur une forme de linéarité, soit entre les variables d'intérêt, soit entre la variable réponse et les variables d'intérêt, il serait intéressant de s'affranchir de cette contrainte en les réécrivant en terme de fonctions de répartition (et de densité) multivariées ; puis en implémentant une procédure permettant leur inférence d'une façon efficace et robuste et qu'elle soit simple d'utilisation pour le clinicien.

# Table des figures

# Liste des tableaux

# Bibliographie


[1] Fisher NI. Copulas. Encyclopedia of statistical sciences. 1997 ;. Cité page 17.

[2] Sklar A. Fonctions de répartition à $n$ dimensions et leurs marges. Publ Inst Statist Univ Paris. 1959 ;8 :229–231. 2 citations pages 17 et 18.

[3] Gaffiot F. Dictionnaire Latin-Française. DIÁLOGOS. 1951 ;25 :00. Cité page 17.

[4] Ali MM, Mikhail N, Haq MS. A class of bivariate distributions including the bivariate logistic. Journal of multivariate analysis. 1978 ;8(3) :405–412. Cité page 22.

[5] Gumbel EJ. Bivariate logistic distributions. Journal of the American Statistical Association. 1961 ;56(294) :335–349. Cité page 22.

[6] Muirhead RJ. Aspects of multivariate statistical analysis. JOHN WILEY & SONS, INC, 605 THIRD AVE, NEW YORK, NY 10158, USA, 1982, 656. 1982 ;. Cité page 23.

[7] Genest C, MacKay J. The joy of copulas : bivariate distributions with uniform marginals. The American Statistician. 1986 ;40(4) :280–283. Cité page 23.

[8] Savu C, Trede M. Hierarchical Archimedean Copulas : International Conference on High Frequency Finance. Konstanz, Germany. 2006 ;. Cité page 27.

[9] Kendall MG. A new measure of rank correlation. Biometrika. 1938 ;30(1/2) :81–93. 3 citations pages 29, 43, et 44.

[10] Durbin J, Stuart A. Inversions and rank correlation coefficients. Journal of the Royal Statistical Society Series B (Methodological). 1951 ;p. 303–309. Cité page 32.

[11] Kruskal WH. Ordinal measures of association. Journal of the American Statistical Association. 1958 ;53(284) :814–861. Cité page 32.

[12] Nelsen RB. An introduction to copulas. Springer Series in Statistics. New York : Springer ; 2006. 5 citations pages 32, 69, 70, 77, et 78.

[13] Willan A, Lin D. Incremental net benefit in randomized clinical trial. Statistics in Medicine. 2001 ;20 :1563–1574. Cité page 37.

[14] Willan A, O'Brien B. Confidence intervals for cost-effectiveness ratios : an application of Fieller's theorem. Health Economics. 1996 ;5 :297–305. 2 citations pages 37 et 45.

[15] Kaplan E, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958 ;53 :457–481. Cité page 38.

[16] Willan A, Chen E, Cook R, Lin D. Incremental net benefit in clinical trials with quality-adjusted survival. Statistics in Medicine. 2003 ;22 :353–362. Cité page 38.

[17] Willan A, Lin D, Manca A. Regression methods for cost-effectiveness analysis with censored data. Statistics in Medicine. 2005 ;24 :131–145. 2 citations pages 38 et 47.

[18] Pullenayegum E, Willan A. Semi-parametric regression models for cost-effectiveness analysis : Improving the efficiency of estimation from censored data. Statistics in Medicine. 2007 ;26 :3274–3299. Cité page 39.

[19] Thompson S, Nixon R. How sensitive are cost-effectiveness analyses to choice of parametric distributions ? Medical Decision Making. 2007 ;4 :416–423. 2 citations pages 41 et 48.

[20] Stamey J, Beavers D, Faries D, Price K, Seaman J. Bayesian modeling of cost-effectiveness studies with unmeasured confounding : a simulation study. Pharmaceutical Statistics. 2014 ;13 :94–100. Cité page 41.

[21] Lin D. Linear regression analysis of censored medical costs. Biostatistics. 2000 ;1 :35–47. 2 citations pages 41 et 42.

[22] Buckley J, James I. Linear regression with censored data. Biometrika. 1979 ;66 :429–436. 4 citations pages 42, 56, 91, et 108.

[23] Miller R, Halpern J. Regression with censored data. Biometrika. 1982 ;69-3 :521–531. 2 citations pages 42 et 108.

[24] Oakes D. A concordance test for independence in the presence of censoring. Biometrics. 1982 ;38 :451–455. Cité page 43.

[25] Oakes D. On consistency of Kendall's tau under censoring. Biometrika. 2008 ;95-4 :997–1001. Cité page 44.

[26] Genest C, Rivest LP. Statistical inference procedures for bivariate Archimedean copulas. Journal of the American Statistical Association. 1993 ;88 :1034–1043. Cité page 44.

[27] Lakhal-Chaieb L. Copula inference under censoring. Biometrika. 2010 ;97-2 :505–512. Cité page 44.

[28] Dos Santos Silva R, Freitas Lopes H. Copula, marginal distributions and model selection : a Bayesian note. Statistics and Computing. 2008 ;18 :313–320. Cité page 44.

[29] Genest C, Neslehova J, Ben Ghorbal N. Estimators based on Kendall's tau in multivariate copula models. Australian and New Zealand Journal of Statistics. 2011 ;53(2) :157–177. 2 citations pages 44 et 84.

[30] El Maache H, Lepage Y. Spearman's rho and Kendall's tau for multivariate data sets. Mathematical Statistics and Applications, Lecture Notes-Monograph Series. 2003 ;42 :113–130. Cité page 44.

[31] Fieller E. Some problems in interval estimation. Journal of the Statistical Royal Society, Series B. 1954 ;16 :175–185. Cité page 45.

[32] Chaudhary N, Stearns S. Confidence intervals for cost-effectiveness ratios : an example from randomized trial. Statistics in Medicine. 1996 ;15 :1447–1458. Cité page 45.

[33] Siani C, Moatti JP. Quelles méthodes de calcul des régions de confiance du ratio coût-efficacité incrémental choisir ? Universites d'Aix-Marseille II et III ; 2002. Cité page 46.

[34] Nixon R, Thompson S. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. Health economics. 2005 ;14 :1217–1229. Cité page 46.

[35] Tsai K, Peace K. Analysis of subgroup data in clinical trials. Sixteenth Annual Biopharmaceutical Applied Statistics Symposium. 2009 ;Savannah, GA. Cité page 46.

[36] Vickers AJ, Rees RW, Zollman CE, McCarney R, Smith CM, Ellis N, et al. Acupuncture for chronic headache in primary care : large, pragmatic, randomised trial. Bmj. 2004 ;328(7442) :744. Cité page 52.

[37] Wonderling D, Vickers A, Grieve R, McCarney R. Cost effectiveness analysis of a randomised trial of acupuncture for chronic headache in primary care. British Medical Journal. 2004 ;328 :747–749. Cité page 52.

[38] Vickers AJ. Whose data set is it anyway ? Sharing raw data from randomized trials. Trials. 2006 ;7(1) :15. Cité page 52.

[39] Berkson J. Application of the logistic function to bio-assay. Journal of the American Statistical Association. 1944 ;39 (227) :357–365. Cité page 66.

[40] Barnard GA. Statistical inference. Journal of the Royal Statistical Society, Series B. 1949 ;11(2) :115–149. Cité page 66.

[41] Guerriere MR, Detsky AS. Neural networks : what are they ? Annals of internal medicine. 1991 ;115(11) :906–907. Cité page 67.

[42] Hinton GE. How neural networks learn from experience. Scientific American. 1992 ;267(3) :145–151. Cité page 67.

[43] Genest C, Neslehova J. A Primer on Copulas for Count Data. ASTIN Bulletin. 2007 11 ;37 :475–515. 4 citations pages 68, 76, 81, et 82.

[44] Denuit M, Lambert P. Constraints on concordance measures in bivariate discrete data. Journal of Multivariate Analysis. 2005 ;93 :40–57. 3 citations pages 70, 71, et 81.

[45] Marshall AW. Copulas, marginals, and joint distributions. Lecture Notes-Monograph Series. 1996 ;p. 213–222. Cité page 70.

[46] Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983 ;70 :41–55. 3 citations pages 71, 72, et 78.

[47] Guo S, Fraser M. Propensity score analysis : statistical methods and applications - 2nd edition. Advanced Quantitative Techniques in the Social Sciences Series. Thousand Oaks, California : SAGE ; 2015. Cité page 71.

[48] Amemiya T. Advanced econometrics. Harvard university press ; 1985. Cité page 72.

[49] Tu YK, Kellett M, Clerehugh V, Gilthorpe MS. Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. British dental journal. 2005 ;199(7) :457–461. Cité page 72.

[50] Miles J, Shevlin M. Applying regression and correlation : A guide for students and researchers. Sage ; 2001. Cité page 72.

[51] Glantz S, Slinker B. Multicollinearity and what to do about it. Primer of Applied Regression & Analysis of Variance 2nd edition McGraw-Hill, Inc New York, NY. 2001 ;p. 185–187. Cité page 72.

[52] Pedhazur EJ. Multiple regression in behavioral research : Explanation and prediction. Coaching Eiticacy and Youth Sport. 1997 ;. Cité page 72.

[53] Austin P, Grootendrost P, Anderson G. A comparaison of the ability of different propensity score models to balance measured variables between treated and untreated subjects : a Monte Carlo study. Statistics in medicine. 2007 ;26 :734–753. Cité page 72.

[54] Vandenhende F, Lambert P. Improved rank-based dependence measures for categorical data. Statistics & probability letters. 2003 ;63(2) :157–163. Cité page 77.

[55] Scarsini M. On measures of concordance. Stochastica : revista de matematica pura y aplicada. 1984 ;8(3) :201–218. Cité page 77.

[56] Gijbels I, Mielniczuk J. Estimating the density of a copula function. Communications in Statistics, Theory and Methods. 1990 ;19(2) :445–464. Cité page 79.

[57] Bouezmarni T, Mesfioui M, Tajar A. On concordance measures for discrete data and dependance properties of Poisson model. Journal of Probability and Statistics. 2009 ;Article ID 895742 :1–15. 2 citations pages 81 et 82.

[58] Gijbels I, Sznajder D. Positive quadrant dependence testing and constrained copula estimation. Canadian Journal of Statistics. 2013 ;41(1) :36–64. Cité page 82.

[59] Yanagimoto T, Okamoto M. Partial orderings of permutations and monotonicity of a rank correlation statistic. Annals of the Institute of Statistical Mathematics. 1969 ;21(1) :489–506. Cité page 82.

[60] Tchen AH. Inequalities for distributions with given marginals. The Annals of Probability. 1980 ;p. 814–827. Cité page 82.

[61] Durrleman V, Nikeghbali A, Roncalli T. Which copula is the right one. Groupe de Recherche Operationnelle du Credit Lyonnais, France. 2000 ;. Cité page 83.

[62] Deheuvels P. La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. Bulletin de la Classe des Sciences de l'Académie Royale de Belgique. 1979 ;01 :65. Cité page 83.

[63] Joe H. Multivariate concordance. Journal of multivariate analysis. 1990 ;35(1) :12–30. Cité page 84.

[64] Genest C, Ghoudi K, Rivest LP. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika. 1995 ;82 :543–552. Cité page 84.

[65] Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation : a simulation study. Pharmacoepidemiology and drug safety. 2008 ;17(6) :546–555. Cité page 88.

[66] Panagiotelis A, Czado C, Joe H. Pair copula constructions for multivariate discrete data. Journal of the American Statistical Association. 2012 ;107(499) :1063–1072. Cité page 90.

[67] Miller R. Least squares regression with censored data. Biometrika. 1976 ;63 :449–464. Cité page 91.

[68] Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right censored data. Annals of Statistics. 1981 ;9 :1276–1288. Cité page 91.

[69] Doksum K, Yandell B. Properties of regression estimates based on censored survival data. In A festschrift for Erich L Lehmann (eds P J Bickel, K Doksum and JL Hodges Jr). 1983 ;p. 140–156. Cité page 91.

[70] Zheng Z. Regression analysis with censored data. PhD Dissertation. 1984 ;Columbia University. Cité page 91.

[71] Zheng Z. A class of estimatord for the parameters in linear regression with censored data. Acta Math Appl Sin. 1987 ;3 :231–241. Cité page 91.

[72] Leurgans S. Linear models, random censoring and synthetic data. Biometrika. 1987 ;74 :301–309. Cité page 91.

[73] Zhou M. Asymptotic normality of the "synthetic data" regression estimator for censored survival data. Annals of statistics. 1992 ;20 :1002–1021. Cité page 91.

[74] Srinivasan C, Zhou M. Linear regression with censoring. Journal of Multivariate Analysis. 1994 ;49 :179–201. Cité page 91.

[75] Fan J, Gijbels I. Censored-regression : local linear approximations and their applications. Journal of the American Statistical Association. 1994 ;89 :560–570. Cité page 91.

[76] Noh H, El Ghouch A, Bouezmarni T. Copula-Based Regression Estimation and Inference. Journal of the American Statistical Association. 2013 ;108 :676–688. 2 citations pages 93 et 95.

[77] Shih JH, Louis TA. Inferences on the association parameter in copula models for bivariate survival data. Biometrics. 1995 ;p. 1384–1399. 2 citations pages 95 et 99.

[78] Joe H, Xu J. The estimation method of inference functions for margins for multivariate models. Technical Reports of Department of Statistics of University of British-Columbia. 1996 ;166 :1–21. Cité page 95.

[79] Lo SH, Mack YP, Wang JL. Density and hazard rate estimation for censored data via strong reptresentation of the Kaplan-Meier estimator. Probability Theory and Relatd Fields. 1989 ;80 :473–473. 2 citations pages 97 et 98.

[80] Donsker MD. Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. The Annals of mathematical statistics. 1952 ;p. 277–281. Cité page 99.

[81] Lo S, Singh K. The Product-Limit Estimator and the Bootstrap : Some Asymptotic Representations. Probability Theory and Related Fields. 1985 ;71 :455–465. 2 citations pages 99 et 102.

[82] Dette H, Van Hecke R, Volgushev S. Some comments on copula-based regression. Journal of the American Statistical Association. 2014 ;109(507) :1319–1324. Cité page 104.

[83] Genest C, Rémillard B, Beaudoin D. Goodness-of-fit tests for copulas : A review and a power study. Insurance : Mathematics and economics. 2009 ;44(2) :199–213. Cité page 104.

[84] Shih JH. A goodness-of-fit test for association in a bivariate survival model. Biometrika. 1998 ;85(1) :189–200. Cité page 104.

[85] Glidden DV. Checking the adequacy of the gamma frailty model for multivariate failure times. Biometrika. 1999 ;86(2) :381–393. Cité page 104.

**Utilisation de copules paramétriques en présence de données observationnelle : cadre théorique et modélisations.**

**Résumé :** Les études observationnelles (non-randomisées) sont principalement constituées de données ayant des particularités qui sont en fait contraignantes dans un cadre statistique classique. En effet, dans ce type d'études, les données sont rarement continues, complètes et indépendantes du bras thérapeutique dans lequel les observations se situent. Cette thèse aborde l'utilisation d'un outil statistique paramétrique fondé sur la dépendance entre les données à travers plusieurs scénarii liés aux études observationnelles. En effet, grâce au théorème de Sklar (1959), les copules paramétriques sont devenues un sujet d'actualité en biostatistique. Pour commencer, nous présentons les concepts de base relatifs aux copules et aux principales mesures d'association basées sur la concordance retrouvées dans la littérature. Ensuite, nous donnons trois exemples d'application des modèles de copules paramétriques pour autant de cas de données particulières retrouvées dans des études observationnelles. Nous proposons d'abord une stratégie de modélisation de l'analyse coût-efficacité basée uniquement sur une réécriture des fonctions de distribution jointes et évitant les modèles de régression linéaire. Nous étudions ensuite les contraintes relatives aux données discrètes, particulièrement dans un contexte de non-unicité de la fonction copule : nous réécrivons le score de propension grâce à une approche novatrice basée sur l'extension d'une sous-copule. Enfin, nous évoquons un type particulier de données manquantes : les données censurées à droite, dans un contexte de régression, grâce à l'utilisation de copules semi-paramétriques.

**Mots-clefs** : Copules paramétriques, Analyse coût-efficacité, Score de propension, Régression semi-paramétrique, Données non-randomisées.

**Use of parametric copulas with observational data : theoretical framework and modelizations.**

**Abstract :** Observational studies (non-randomized) consist primarily of data with features that are in fact constraining within a classical statistical framework. Indeed, in this type of study, data are rarely continuous, complete, and independent of the therapeutic arm the observations are belonging to. This thesis deals with the use of a parametric statistical tool based on the dependence between the data, using several scenarios related to observational studies. Indeed, thanks to the theorem of Sklar (1959), parametric copulas have become a topic of interest in biostatistics. To begin with, we present the basic concepts of copulas, as well as the main measures of association based on the concordance founded on an analysis of the literature. Then, we give three examples of application of models of parametric copulas for as many cases of specific data found in observational studies. We first propose a strategy of modeling cost-effectiveness analysis based essentially on rewriting the joint distribution functions, while discarding the use of linear regression models. We then study the constraints relative to discrete data, particularly in a context of non-unicity of the copula function. We rewrite the propensity score, thanks to an innovative approach based on the extension of a sub-copula. Finally, we introduce a particular type of missing data : right censored data, in a regression context, through the use of semi-parametric copulas.

**Keywords** : Parametric copulas, Cost-effectiveness analysis, Propensity score, Semi-parametric regression, Not-randomized data.