



**Université Assane Seck de Ziguinchor (U.A.S.Z)
U.F.R des Sciences et Technologies (URF-ST)**

Département de Mathématiques

Mémoire de Master

Domaine : Sciences et Technologies
Mention : Mathématiques et Applications
Spécialité : Mathématiques Appliquées
Option : Statistique

TITRE :

**Méthodes d'estimation des paramètres d'une copule bivariée et mise
en œuvre pratique**

Présenté par

Mame Fatou Ndiaye

Sous la direction de : Dr Emmanuel Nicolas CABRAL

Sous la supervision de : Pr Alassane DIEDHIOU

Membres du Jury :

PRESIDENT :

Marie Salomon SAMBOU, Professeur (UFR-ST/U.A.S.Z)

DIRECTEUR :

Emmanuel N. CABRAL, Maître Assistant (UFR-ST/U.A.S.Z)

EXAMINATEURS :

**Alassane DIEDHIOU, Professeur (UFR-ST/U.A.S.Z)
Clément MANGA, Maître Assistant (UFR-ST/U.A.S.Z)**

Soutenu le 21 novembre 2020

Dédicaces

À mes très chers parents,

À Awa NDIAYE,

À Morikéba NDIAYE,

À Assane NDIAYE,

À Adama DIANKO,

À Fatou Sofi BADIANE,

À Awa BADIANE,

À Fatou NDIAYE,

À Lamine NDIAYE

À Alice MBAYE,

À toute la promotion de M.P.I,

À tous mes sœurs et frères en ALLAH,

Remerciements

Je tiens à remercier “Allah” le tout puissant pour toute la volonté et le courage qu’il m’a donné pour l’achèvement de ce travail.

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma reconnaissance.

Je voudrais tout d’abord exprimer ma profonde reconnaissance à Mr Emmanuel Nicolas CABRAL pour avoir accepté de diriger ce travail. Je tiens à le remercier de m’avoir accordé sa confiance en faisant preuve d’une très grande disponibilité à mon égard. Ses conseils, encouragements et remarques minutieuses ont été déterminants et je réalise aujourd’hui à quel point travailler avec lui a été enrichissant.

Je tiens à remercier l’Université Assane SECK, l’Ufr des Sciences et Technologies, le Département de Mathématiques ainsi que tous les professeurs qui nous ont initié et donné les bases de l’étude mathématique pour exceller dans ce domaine.

J’exprime ma gratitude en vers le président du jury le professeur Marie Salomon Sambou qui a accepté de présider ce jury et d’apporter sa contribution pour la présentation de ce mémoire. J’exprime aussi ma gratitude en vers les professeurs Alassane Diédhiou, Clément Manga qui ont bien accepté de m’examiner et d’apporter leurs expertises.

Je renouvelle mes sincères remerciements aux professeurs Edouard DIOUF et Clément MANGA pour leur disponibilité, leur gentillesse et leurs judicieux conseils

Mes plus vifs remerciements à ma très chère mère, mon très cher père, mes sœurs et frères pour leur soutien moral et matériel tout au long de mes études.

Mes remerciements vont également à l’endroit de mes frères doctorants Marcel Sihintoé BADIANE et Pape Modou SARR et de mes promotionnaires pour leur soutien indéfectible.

Résumé

Ces dernières décennies, nous avons assisté à l'émergence du concept de copule en modélisation statistique. Cet essor est justifié par le fait que les copules permettent de faire une analyse séparée des marges et de la structure de dépendance induite par une distribution statistique. Cette séparation facilite l'incorporation de lois non gaussiennes et la prise en compte des dépendances non linéaires entre les variables. La finance et l'hydrologie sont deux exemples de domaines où les copules sont très utilisées, par ce fait les copules représentent un outil innovant pour modéliser la structure de dépendance de plusieurs variables aléatoires.

Son grand intérêt est qu'elles fournissent des expressions relativement simples des structures de dépendance liant les marges d'une loi multidimensionnelle. Plus précisément, pour le cas bidimensionnel, une copule C définie sur $[0, 1]^2$, associée à une distribution H de marges uniformes F et G , permet de représenter la fonction de répartition jointe $H(x, y)$ en fonction de ces marginales $F(x)$ et $G(y)$ par la relation :

$$H(x, y) = C(F(x), G(y)).$$

Table des matières

Introduction	1
1 Rappels	3
1.1 Définitions de quelques notions	3
1.2 Notion d'indépendance	4
1.3 Quelques lois de probabilité	4
1.3.1 La loi normale multidimensionnelle	4
1.3.2 La loi de Student multidimensionnelle	5
1.3.3 Loi de Farlie-Gumbel-Morgenstern	5
1.3.4 Les lois usuelles continues	5
1.4 Nombres aléatoires et pseudo-aléatoires	6
1.5 Simulation des lois de probabilité	7
1.5.1 Cas des lois unidimensionnelles	7
1.5.2 Cas des lois multidimensionnelles	10
1.6 Estimation	12
1.6.1 Estimation ponctuelle	12
1.6.2 Généralisation (Cas de multiples paramètres)	15
1.6.3 Construction d'un estimateur de θ	15
2 Introduction aux copules	17
2.1 Historique des copules	17
2.2 Copules bivariées	18
2.2.1 Définitions et théorèmes	18
2.2.2 Propriétés fondamentales	21
2.2.3 Densité d'une copule	23
2.2.4 Quelques copules populaires et leur représentation graphique	23
2.3 Étude de familles de copules paramétriques	29
2.3.1 Famille des copules elliptiques	29
2.3.2 Famille des copules archimédiennes	30
2.3.3 Famille des copules de valeurs extrêmes	31
2.3.4 Famille des copules archimax	32
2.4 Mesures de dépendance	33
2.4.1 Corrélacion de Pearson	33
2.4.2 Notion de dépendance de queue	34
2.4.3 Tau de Kendall et rho de Spearman	36

3	Estimation des copules	43
3.1	Estimation paramétrique	43
3.1.1	La méthode du maximum de vraisemblance exacte (Full Maximum Likelihood (FML))	43
3.1.2	Inférence sur les marginales (IFM)	44
3.1.3	La méthode des moments	45
3.2	Estimation semi-paramétrique	46
3.2.1	La méthode de pseudo-maximum de vraisemblance (maxi- mum de vraisemblance canonique (CML))	47
3.2.2	Méthode d'inversion d'une statistique d'association	48
3.3	Estimation non-paramétrique	49
3.3.1	Copule empirique	49
3.3.2	Méthode de Genest et Rivest	50
4	Simulation et application à des données réelles	51
4.1	Simulation des copules	51
4.1.1	Méthode des distributions	51
4.1.2	Méthode des distributions conditionnelles	52
4.1.3	Cas particulier des copules archimédiennes	54
4.1.4	Méthode analytique	55
4.2	Le choix de la bonne copule.	57
4.3	Utilisation de données réelles en assurance non-vie	58
4.3.1	L'ajustement des distributions marginales	60
4.3.2	L'ajustement d'une copule aux données bivariées	60
4.3.3	Calcul des primes de réassurance	63
	Conclusion	65

Introduction

Dans la théorie des probabilités, lorsque les variables aléatoires X_1, \dots, X_d sont indépendantes, la fonction de répartition jointe $H(x_1, \dots, x_d)$ s'écrit comme le produit de ses marginales F_1, \dots, F_d , c'est à dire :

$$H(x_1, \dots, x_d) = F_1(x_1) \times \dots \times F_d(x_d).$$

Dans la pratique, les variables aléatoires ne sont pas toujours indépendantes, comme par exemple les risques financiers, donc on ne peut pas négliger cette relation de dépendance. Pour cette raison il faut déterminer la loi jointe de ces variables. Les copules sont des fonctions qui permettent de coupler les lois marginales des variables afin d'obtenir la loi jointe, mesurer et modéliser complètement la dépendance. Ces fonctions de dépendance ont permis de construire une nouvelle théorie, dite "théorie des copules". Cette théorie permet de construire des modèles paramétriques pour la loi jointe de d variables aléatoires réelles dépendantes (X_1, \dots, X_d). En pratique, elle est de plus en plus utilisée en assurance, en finance, en environnement, en fiabilité, en biologie..., pour modéliser de manière plus réaliste la loi jointe de plusieurs variables. C'est pourquoi de nos jours, les études sur l'estimation des paramètres des copules se développent sans cesse et font l'objet de notre mémoire. Ce mémoire est réparti en quatre chapitres.

Le premier chapitre est constitué de rappels de notions de probabilités et de statistiques.

Dans le second chapitre, nous ferons une introduction mathématique aux copules en rappelant les définitions nécessaires concernant les copules bivariées et des propriétés fondamentales associées. Nous énoncerons l'important théorème de Sklar, ainsi que quelques familles de copules les plus utilisées en pratique. À la fin de ce chapitre nous ferons également recours à d'autres indicateurs de dépendance comme le coefficient de corrélation de Pearson, le tau de Kendall et le rho de Spearman.

Le troisième chapitre a pour objectif principal de fournir différentes approches d'estimation des copules à savoir les approches paramétrique, semi-paramétrique et non-paramétrique.

La première approche consiste à estimer le paramètre de la copule par l'une des méthodes suivantes :

- Maximum de vraisemblance exacte (FML),
- Fonctions d'inférence des marginales (IFM),
- Moments.

Dans la seconde approche on utilise :

- Maximum de vraisemblance canonique (CML),
- Inversion d'une statistique d'association.

Quant à la troisième approche on utilise les méthodes suivantes :

- Copule empirique,
- Méthode de Genest et Rivest.

Le quatrième chapitre est dédié à une mise en œuvre pratique des copules. Nous commencerons par la simulation de quelques copules en présentant trois méthodes : la méthode des distributions, la méthode des distributions conditionnelles et la méthode analytique. Nous terminerons ce chapitre par l'utilisation de données réelles obtenues dans une compagnie d'assurance.

Chapitre 1

Rappels

Avant d'aborder les copules à proprement parlé, il nous a semblé nécessaire de devoir effectuer quelques rappels. L'objet de cette partie n'est pas d'entrer dans des problèmes théoriques, c'est pourquoi les définitions présentées seront souvent brèves et donc réductrices.

1.1 Définitions de quelques notions

Définition 1.1.1 (fonction de répartition conjointe) *Considérons un vecteur aléatoire à d -dimension $X = (X_1, X_2, \dots, X_d)$. On appelle alors fonction de répartition conjointe de X , la fonction de d variables réelles définie par :*

$$F(x_1, \dots, x_d) = P(X_1 \leq x_1, \dots, X_d \leq x_d).$$

Définition 1.1.2 (lois marginales) *La fonction de répartition conjointe nous permet d'établir les fonctions de répartition F_1, \dots, F_d des variables aléatoires marginales X_1, \dots, X_d (donc les fonctions de répartition marginales) comme suit :*

$$F_i(x_j) = P(X_j \leq x_j) = \lim_{t \rightarrow +\infty} F(t, \dots, t, x_j, t, \dots, t), 1 \leq j \leq d.$$

Définition 1.1.3 (Lois de probabilité absolument continues) *La fonction de répartition de X est dite absolument continue s'il existe une fonction f telle que*

$$F(x_1, \dots, x_d) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_d} f(s_1, \dots, s_d) ds_1 \dots ds_d.$$

La fonction f est alors appelée densité de probabilité conjointe de la loi de X .

Si la loi de X est absolument continue, alors ses variables aléatoires marginales sont absolument continues et leur densité conjointe est donnée par

$$f(x_1, \dots, x_d) = \frac{\partial^d F(x_1, \dots, x_d)}{\partial x_1 \dots \partial x_d}.$$

De la même façon, on associe ses densités marginales f_1, \dots, f_d aux fonctions de répartition F_1, \dots, F_d par

$$f_j(x) = \frac{d}{dx_j} F_j(x).$$

1.2 Notion d'indépendance

Les composantes du vecteur X sont mutuellement indépendantes si pour tout $(x_1, \dots, x_d) \in \mathbb{R}^d$, on a

$$F(x_1, \dots, x_d) = \prod_{j=1}^d F_j(x_j) \text{ ou si } X \text{ a pour densité } f(x_1, \dots, x_d) = \prod_{j=1}^d f_j(x_j).$$

1.3 Quelques lois de probabilité

1.3.1 La loi normale multidimensionnelle

Le modèle normal multidimensionnel est le modèle le plus fréquent en statistique. Pour le décrire on considère $Z = (Z_1, \dots, Z_d)$, un vecteur dont les composantes sont indépendantes, où $Z_j, 1 \leq j \leq d$, est de loi normale de moyenne nulle et de variance unitaire. Soit Σ une matrice carrée de dimension d , symétrique et définie positive. Alors, la distribution normale multidimensionnelle associée de moyenne $\mu = (\mu_1, \dots, \mu_d)^T$ et de matrice de variance-covariance Σ est définie comme la loi du vecteur

$$X = \Sigma^{1/2}Z + \mu$$

où $\Sigma^{1/2}$ est la racine carrée symétrique de Σ . Puisque les composantes de Z sont indépendantes de loi $N(0, 1)$, leur densité est donnée par :

$$f_Z(z_1, \dots, z_d) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2} \sum_{j=1}^d z_j^2\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2} z^T z\right),$$

où $z = (z_1, \dots, z_d)^T$ et T désigne la transposée. En considérant la transformation X décrite précédemment, on en tire la densité de la loi normale multidimensionnelle, c'est-à-dire :

$$f_X(x) = f_X(x_1, \dots, x_d) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\}.$$

Cas particulier : Pour $d = 2$, soit le couple (X_1, X_2) , la densité de la loi normale de moyennes μ_1, μ_2 , de variances σ_1^2, σ_2^2 et de corrélation ρ s'écrit :

$$f_{X_1, X_2}(x_1, x_2) = \left(\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\right) \times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\right].$$

Les propriétés suivantes de la loi normale bidimensionnelle sont intéressantes :

1. la loi marginale de X_j est $N(\mu_j, \sigma_j^2)$, pour $j = 1, 2$;
2. $Cor(X_1, X_2) = \rho$;
3. la loi de $aX_1 + bX_2$ est normale de moyenne $a\mu_1 + b\mu_2$ et de variance $a^2\mu_1^2 + b^2\mu_2^2 + 2ab\rho\sigma_1\sigma_2$;
4. les variables X_1 et X_2 sont indépendantes si et seulement si $\rho = 0$ (indépendance linéaire).

1.3.2 La loi de Student multidimensionnelle

Un vecteur aléatoire X de \mathbb{R}^d suit une loi de Student de paramètres (k, μ, Σ) s'il admet pour fonction de densité

$$t_d(x; k, \mu, \Sigma) = \frac{\Gamma(\frac{k+d}{2})(\pi k)^{-d/2} |\Sigma|^{-1/2}}{\Gamma(k/2)[1 + ((x - \mu)^T \Sigma^{-1} (x - \mu))/k]^{(k+d)/2}}$$

où $k > 0, \mu, \Sigma$ et Γ désignent respectivement le degré de liberté, le vecteur moyen, une matrice inversible de dimension $d \times d$ et la fonction Gamma d'Euler. Le vecteur X n'a de moyenne (resp. de matrice des covariances) que si $k > 1$ (resp. $k > 2$) et dans ce cas elle coïncide avec μ (resp. avec $\frac{k}{k-2}\Sigma$). Le degré de liberté k contrôle la longueur de queue de la distribution. Plus k est petit plus la queue de la distribution de X est longue et lorsque k tend vers $+\infty$ la loi de X tend en probabilité vers la loi normale de paramètres (μ, Σ) .

1.3.3 Loi de Farlie-Gumbel-Morgenstern

Soient F_1 et F_2 , des fonctions de répartitions univariées. Dans [9] on y décrit une famille de lois bivariées de la forme

$$H(x_1, x_2) = F_1(x_1)F_2(x_2)\{1 + \theta A(F_1(x_1))B(F_2(x_2))\}, \quad \theta \in [-1, 1],$$

où $(X_1, X_2) \in \mathbb{R}^2$ et $A, B : [0, 1] \rightarrow [0, 1]$, sont des fonctions telles que H satisfait les conditions pour être une fonction de répartition bivariée. Comme cas particulier lorsque $A(u) = B(u) = 1 - u$, on a la distribution de Farlie-Gumbel-Morgenstern standard, à savoir

$$H(x_1, x_2) = F_1(x_1)F_2(x_2) + \theta F_1(x_1)F_2(x_2)\{1 - F_1(x_1)\}\{1 - F_2(x_2)\}.$$

Farlie (voir [9]) s'est servi de ces lois pour étudier l'efficacité relative asymptotique de tests d'indépendance basés sur des coefficients de corrélation. En effet, on retrouve l'indépendance lorsque $\theta = 0$.

Nous donnons quelques lois univariées ci-dessous.

1.3.4 Les lois usuelles continues

1. La loi uniforme

On dit que X suit une loi uniforme sur $[a, b]$ et on note $X \rightsquigarrow \mathcal{U}_{[a,b]}$ si elle admet comme densité

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a, b] \\ 0 & \text{sinon.} \end{cases} \quad (1.1)$$

$$E(X) = \frac{a+b}{2} \text{ et } Var(X) = \frac{(b-a)^2}{12}.$$

2. Loi de Gauss ou loi Normale

On dit que X suit une loi de Gauss de moyenne μ et de variance σ^2 et on note $X \rightsquigarrow N(\mu, \sigma^2)$ si elle admet comme densité

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \quad \forall x \in \mathbb{R}. \quad (1.2)$$

On a $E(x) = \mu$; $Var(X) = \sigma^2$.

3. Loi exponentielle

On dit que X suit une loi exponentielle de paramètre $\lambda > 0$ et on note $X \rightsquigarrow \mathcal{E}(\lambda)$ si elle admet pour densité

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{\mathbb{R}_+}. \quad (1.3)$$

$$E(X) = \frac{1}{\lambda}; \quad Var(X) = \frac{1}{\lambda^2}.$$

4. La loi gamma

X est de loi gamma de paramètre $\alpha > 0$ et $\beta > 0$ on note $X \rightsquigarrow \Gamma(\alpha, \beta)$ si elle admet comme densité

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases} \quad (1.4)$$

$$\text{où } \Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt.$$

$$\text{On a } E(X) = \frac{\alpha}{\beta}; \quad Var(X) = \frac{\alpha}{\beta^2}.$$

1.4 Nombres aléatoires et pseudo-aléatoires

Un nombre aléatoire est à priori une des valeurs prises par une certaine variable aléatoire réelle.

Les méthodes algorithmiques de génération de nombres aléatoires se présentent comme des formules mathématiques permettant de disposer d'une suite de nombres qu'on peut considérer comme étant choisis au hasard selon une loi de probabilité donnée.

Du fait de leur caractère mathématique, les méthodes algorithmiques peuvent faire l'objet de programmation informatique, ce qui permet d'obtenir très rapidement autant qu'on veut de nombres aléatoires.

Il convient cependant de noter que de part leur construction, il est impossible d'obtenir avec les méthodes algorithmiques des nombres aléatoires. En effet ces méthodes consistent l'application d'une formule mathématique. La connaissance de cette formule permet ainsi de connaître au préalable le nombre que ces méthodes donnent, ce qui est contraire à la définition de nombres aléatoires. En effet un nombre aléatoire est par définition non prévisible.

Néanmoins, malgré l'impossibilité d'obtenir de véritables nombres aléatoires avec les méthodes algorithmiques, les améliorations continues qu'ont connues ces méthodes ont conduit à l'obtention de nombres qui ressemblent dans plusieurs aspects à des vrais nombres aléatoires.

Ces nombres obtenus sont appelés **nombres pseudo-aléatoires** ou **nombres simulés**.

1.5 Simulation des lois de probabilité

1.5.1 Cas des lois unidimensionnelles

Dans ce mémoire nous utiliserons trois méthodes classiques pour générer des nombres issus d'une loi donnée :

1. la méthode par inversion,
2. la méthode de Box-Muller,
3. la méthode du rejet de Von Neumann.

Toutes ces méthodes nécessitent l'existence d'un générateur de nombres aléatoires suivant la loi uniforme entre 0 et 1. Nous n'étudierons pas les diverses méthodes permettant d'atteindre ce but, nous nous contenterons de l'examen des procédures mises à disposition par les divers langages informatiques. Ici le logiciel R (voir [37]) est censé délivrer des nombres suivant la loi uniforme $\mathcal{U}(0, 1)$.

Soit X une variable aléatoire réelle, on souhaite disposer n valeurs simulées X_1, X_2, \dots, X_n de X .

La méthode par inversion

Définition 1.5.1 Soit F une fonction de répartition définie sur \mathbb{R} . Le quantile (ou l'inverse généralisé ou pseudo-inverse) de F est une fonction, notée F^{-1} , telle que

$$F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t, \text{ pour } t \in [0, 1]\}.$$

Si F est continue et strictement croissante sur \mathbb{R} , l'inverse généralisée coïncide avec l'inverse usuel.

Proposition 1.5.2 Soit X une variable aléatoire de fonction de répartition F .

1. Si U est uniforme sur $[0, 1]$, alors : $F^{-1}(U) \rightsquigarrow F$.
2. Si X est absolument continue alors : $U = F(X) \rightsquigarrow \mathcal{U}(0, 1)$.

Preuve voir [21]. ■

Il est alors possible d'utiliser le changement de variable afin de générer des nombres X suivant une loi caractérisée par F , il suffit d'écrire $X = F^{-1}(U)$.

Utilisation pratique

- Générer n valeurs $u_1, u_2, u_3, \dots, u_n$ de U .
- Déterminer les valeurs simulées x_i de X en utilisant l'expression

$$x_i = F^{-1}(u_i), \quad 1 \leq i \leq n.$$

Exemple 1.5.3 (la loi exponentielle) À partir de la fonction de répartition de la loi exponentielle $F(x) = 1 - e^{-\lambda x} 1_{\mathbb{R}^+}(x)$, on en déduit l'expression de F^{-1}

$$X = F^{-1}(U) = -\frac{1}{\lambda} \ln(1 - U) = -\frac{1}{\lambda} \ln(U).$$

La dernière égalité ne doit pas surprendre car les variables $1 - U$ et U possèdent la même fonction de répartition et sont, par conséquent, égales en loi.

Ainsi, si $u_1, \dots, u_i, \dots, u_n$ est une suite de n nombres pseudo-aléatoires, on en déduit une suite de n valeurs simulées de X comme suit :

$$x_1 = -\frac{\ln(u_1)}{\lambda}, x_2 = -\frac{\ln(u_2)}{\lambda}, \dots, x_i = -\frac{\ln(u_i)}{\lambda}, \dots, x_n = -\frac{\ln(u_n)}{\lambda}.$$

La principale insuffisance de la méthode est qu'elle nécessite la connaissance de l'expression explicite de la fonction de répartition, or plusieurs lois de probabilité dont notamment la loi normale n'ont pas cette propriété.

Cependant elle est en concurrence avec d'autres méthodes assez encombrantes en espace mémoire.

La simulation de la loi normale par la méthode de Box-Muller

On ne connaît pas l'expression de sa fonction de répartition. On ne peut donc théoriquement employer la méthode précédente. On présente ici une méthode qui permet de simuler un couple (X, Y) de variables aléatoires normales, centrées, réduites et indépendantes. On connaît la densité jointe de (X, Y) :

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right).$$

La méthode de Box-Muller est basée sur le théorème suivant.

Théorème 1.5.4 Soient X et Y deux variables aléatoires réelles indépendantes suivant chacune la loi normale centrée réduite $N(0, 1)$. On établit les égalités suivantes :

$$\begin{aligned} X &= \sqrt{-2 \ln U} \cos(2\pi V), \\ Y &= \sqrt{-2 \ln U} \sin(2\pi V) \end{aligned}$$

où U et V sont deux variables aléatoires indépendantes suivant la loi uniforme sur $[0, 1]$.

Preuve voir [21]. ■

Utilisation pratique

Pour simuler $2n$ valeurs d'une variable aléatoire suivant $N(0, 1)$, on procède comme suit :

1. générer u et v suivant chacune la loi uniforme sur $[0, 1]$,
2. calculer x et y par la formule

$$x = \sqrt{-2 \ln u} \cos(2\pi v) \quad \text{et} \quad y = \sqrt{-2 \ln u} \sin(2\pi v),$$

3. répéter ces deux étapes n fois.

La loi normale quelconque

Soit Y suivant $N(\mu, \sigma^2)$ où σ^2 et μ sont données. On sait alors qu'on peut écrire $Y = \sigma X + \mu$ où $X \rightsquigarrow N(0, 1)$. Par conséquent, pour simuler n valeurs y_1, y_2, \dots, y_n de Y on procède comme suit :

- utiliser l'algorithme de Box-Muller pour simuler n valeurs x_1, x_2, \dots, x_n de X ,
- puis déduire y_1, y_2, \dots, y_n par l'expression $y_i = \sigma x_i + \mu$ avec $i = 1$ à n .

La méthode du rejet de Von Neumann

La méthode du rejet consiste à simuler une variable aléatoire X de densité connue f difficile à simuler, à partir d'une autre variable Y de densité g dont la simulation, par exemple par la méthode d'inversion, est facile à mettre en œuvre. La densité g doit pouvoir, après multiplication par une constante M , majorer f pour tout x de son support

$$\forall x, g(x) \neq 0, \frac{f(x)}{g(x)} \leq M.$$

La procédure de simulation est la suivante :

1. on génère Y suivant la loi g , facile à simuler, soit y la valeur trouvée,
2. on génère u d'après la variable U qui suit la loi uniforme entre 0 et 1, les variables Y et U doivent être indépendantes,
3. si $u \leq \frac{1}{M} \frac{f(y)}{g(y)}$, alors $x = y$, sinon on retourne à l'étape 1.

Afin de prouver que la variable X générée de cette façon, suit effectivement la loi de densité f , il faut calculer la probabilité conditionnelle

$$P\left(Y \leq x | U \leq \frac{1}{M} \frac{f(y)}{g(y)}\right)$$

donnant la probabilité de succès du test à l'étape 3 de la procédure et montrer qu'elle est égale à $P(X \leq x)$, voir [21] pour la démonstration.

Remarque 1.5.5 – Les fonctions f et g doivent avoir le même support.

- La méthode de rejet est évidemment plus lente que les autres méthodes de simulation. On l'utilise seulement lorsque les autres méthodes ne sont pas applicables.
- Par hypothèse le nombre M doit être supérieur ou égal à 1 et il n'est pas unique en effet soit $M' > M$ on a aussi $f(x) \leq M'g(x)$. Il convient de choisir M le plus petit possible. En pratique, on le choisit égal au maximum de $h(x) = f(x)/g(x)$.
- On a : $P\left(U \leq \frac{1}{M} \frac{f(y)}{g(y)}\right) = 1/M$, la quantité $1/M$ mesurant la probabilité d'acceptation de U et M présente le nombre moyen de rejet.

Exemple 1.5.6 Soit X une variable aléatoire de densité f telle que :

$$f(x) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) 1_{\{x>0\}}.$$

On propose de simuler X par la méthode de rejet. Connaissant le support de f on peut prendre g tel que : $g(x) = \lambda \exp(-\lambda x) 1_{\{x>0\}}$. Pour trouver M on cherche le maximum de $f(x)/g(x)$ avec $M > 1$ soit

$$h(x) = \frac{f(x)}{g(x)} = \frac{1}{\lambda} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2} + \lambda x\right).$$

Alors en étudiant ces variations, $h(x)$ est maximale si $x = \lambda$ d'où

$$M(\lambda) = \frac{1}{\lambda} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\lambda^2}{2}\right).$$

On peut donc simuler f avec un algorithme de rejet (puisque'il est facile de simuler suivant la densité $g : x \in \mathbb{R} \rightarrow \lambda \exp^{-\lambda x} 1_{]0,+\infty[}(x)$).

1.5.2 Cas des lois multidimensionnelles

Lorsque les composantes d'un vecteur aléatoire sont indépendantes, on est ramené au cas unidimensionnel sur chaque composante. En effet pour simuler un vecteur aléatoire dont les composantes sont indépendantes, en utilisant les méthodes précédentes on simule séparément ses composantes selon leur loi marginale respective. Dans le cas de dépendance, une analyse plus poussée est nécessaire pour simuler le vecteur aléatoire. Dans ce cas nous procéderons suivant le plan ci-dessous :

- lorsque la loi jointe des composantes est connue,
- le cas particulier de vecteur gaussien,
- lorsque seules Les lois marginales sont données.

Simulation d'un vecteur aléatoire de loi jointe connue

Pour simplifier la présentation considérons un vecteur à deux dimensions qu'on note $Z = (X, Y)$.

Bien que ce cas soit très rare, une procédure de simulation du couple (X, Y) doit tenir compte de la liaison probabiliste entre X et Y . Ainsi si X peut être simulé (librement), la simulation de Y différera selon la valeur simulée de X . Soit F_X la fonction de répartition marginale de X et $F_{Y/X}$ la fonction répartition de Y conditionnellement à $X = x$

$$F_X(x) = P(X \leq x) \text{ et } F_{Y/X}(Y \leq y/X = x).$$

Une procédure de simulation d'un couple aléatoire (x, y) de (X, Y) se fait comme suit :

1. on simule X selon sa loi marginale $F_X(x)$ en utilisant une des méthodes précédentes,
2. simuler ensuite Y selon sa loi conditionnelle $F_{Y/X}(y/x)$ où x est la valeur simulée de X à l'étape 1.

Pour simuler Y on utilise également une des méthodes de simulation définie précédemment.

Cette méthode se généralise assez directement pour un vecteur de dimension n .

Exemple 1.5.7 *Si nous voulons simuler un vecteur de dimension 5 prenons $Z = (X, Y, S, T, R)$. On procède comme suit :*

- on simule X selon sa loi marginale et soit x la valeur simulée,
- puis on simule Y selon sa loi conditionnelle à $X = x$, soit y la valeur simulée,
- ensuite on simule S selon sa loi conditionnelle à $X = x$ et $Y = y$, soit s la valeur simulée,
- par la suite, on simule T selon sa loi conditionnelle à $X = x$, $Y = y$ et $S = s$, soit t la valeur simulée,
- et enfin on simule R selon sa loi conditionnelle à $X = x$, $Y = y$, $S = s$ et $T = t$.

Cas particulier des lois normales multidimensionnelles

La simulation d'un vecteur normal peut s'opérer selon la méthode présentée ci-dessus. Cependant l'analyse des propriétés de la loi normale multidimensionnelle conduit à adopter une méthode plus facile à mettre en application.

De la simulation de la loi $N(0, 1)$, on déduit celle de la loi normale d'espérance μ et de variance σ^2 quelconque par une transformation affine. Si X suit la loi $N(0, 1)$ alors, $Y = \mu + \sigma X$ suit la loi $N(\mu, \sigma^2)$. La simulation est analogue en dimension d quelconque.

Tout d'abord, si X_1, \dots, X_d sont indépendantes et de même loi $N(0, 1)$, alors le vecteur (X_1, \dots, X_d) suit la loi normale dans \mathbb{R}^d , d'espérance nulle et de matrice de covariance identité : $N_d(0, I)$. On en déduit la simulation d'une loi normale d -dimensionnelle quelconque par la proposition suivante.

Proposition 1.5.8 *Soit μ un vecteur de \mathbb{R}^d et Σ une matrice carrée de taille d symétrique positive. Soit $X = (X_1, \dots, X_d)$ un vecteur aléatoire de loi $N_d(0, I)$ dans \mathbb{R}^d . Soit A une matrice carrée d'ordre d telle que $AA^T = \Sigma$. Alors le vecteur $Y = AX + \mu$ est un vecteur gaussien de moyenne μ et de matrice de covariance Σ .*

Preuve La preuve est une conséquence directe de la définition d'un vecteur gaussien et de la matrice de covariance d'un vecteur aléatoire. D'un point de vue pratique, il faut donc trouver une matrice A telle que $AA^T = \Sigma$. C'est un problème très classique. Une des réponses est implémentée dans la plupart des bibliothèques d'algèbre linéaire, c'est la décomposition de Cholesky. Dans le cas où Σ est définie positive, cette méthode calcule colonne par colonne une matrice A , triangulaire inférieure telle que $AA^T = \Sigma$. Dans le cas particulier où $k = 2$, la décomposition de Cholesky est explicite. En effet, soit $Y = (Y_1, Y_2)$ un vecteur gaussien de \mathbb{R}^2 . On note $\mu = (\mu_1, \mu_2)$ son espérance et $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ sa matrice de variance-covariance. Dans cette notation, σ_1^2, σ_2^2 représente respectivement la variance de Y_1, Y_2 tandis que ρ est la corrélation entre Y_1 et Y_2 :

$$\rho = \text{Cor}(Y_1, Y_2) = \frac{\text{Cov}(Y_1, Y_2)}{\sigma_1\sigma_2} \text{ avec } |\rho| \leq 1.$$

$$\text{Si on pose } A = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \text{ donc } AA^T = \begin{pmatrix} a^2 & ab \\ ab & b^2 + c^2 \end{pmatrix}$$

d'où par identification avec Σ on obtient : $a = \sigma_1, b = \rho\sigma_2$ et $c = \sigma_2\sqrt{1 - \rho^2}$.

$$\text{On peut alors écrire : } \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1 - \rho^2} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$$

En appliquant la proposition précédente, on constate qu'on peut simuler des réalisations indépendantes de Y en utilisant la représentation en loi suivante dans laquelle X_1 et X_2 sont deux variables aléatoires indépendantes de loi $N(0, 1)$:

$$Y_1 = \mu_1 + \sigma_1 X_1 \text{ et } Y_2 = \mu_2 + \sigma_2(\rho X_1 + \sqrt{1 - \rho^2} X_2).$$

■

Simulation d'un vecteur aléatoire de loi jointe inconnue

Lorsque seules ses lois marginales sont données, la copule étant une fonction qui permet de coupler les lois marginales des variables afin d'obtenir la loi jointe est ici inévitable pour faire la simulation.

Pour mieux approfondir la notion de simulation des lois univariées et multivariées le lecteur pourra se référer à [21].

1.6 Estimation

Dans le langage statistique il existe trois types d'estimations : estimation ponctuelle, estimation par intervalle de confiance et les tests statistiques. Seule la notion estimation ponctuelle sera abordée ici.

1.6.1 Estimation ponctuelle

On considère une variable aléatoire réelle X dont la loi dépend du paramètre θ qui est inconnu sur la base d'information tirée d'un échantillon (X_1, X_2, \dots, X_n) c'est-à-dire n variables aléatoires indépendantes suivant la même loi que X . On cherche parmi les valeurs possibles de θ une quantité numérique pouvant estimer le paramètre θ , on parle alors d'estimation ponctuelle de θ .

Les propriétés de l'échantillon

- On appelle moment empirique d'ordre $k \in \mathbb{N}^*$ de l'échantillon la variable aléatoire noté $\hat{\mu}_k(X)$ définie par

$$\hat{\mu}_k(X) = \frac{1}{n} \sum_{i=1}^n X_i^k.$$

- $\hat{\mu}_1(X)$ noté $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est dite la moyenne empirique.
- On appelle la variance empirique de l'échantillon la variable aléatoire notée

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

- **Fonction de vraisemblance**

On appelle fonction de vraisemblance de l'échantillon la fonction définie par

$$L(\tilde{x}, \theta) = \prod_{i=1}^n f(x_i, \theta)$$

où $\tilde{x} = (x_1, x_2, \dots, x_n)$ et $f(x, \theta)$ est la densité de X .

- **Fonction log-vraisemblance**

On appelle fonction de log-vraisemblance de l'échantillon la fonction définie par

$$l(\tilde{x}, \theta) = \log L(\tilde{x}, \theta) = \sum_{i=1}^n \log f(x_i, \theta)$$

où $\tilde{x} = (x_1, x_2, \dots, x_n)$ et $f(x, \theta)$ est la densité de X .

Estimateur

Définition 1.6.1 Soit X une variable aléatoire de loi dépendant d'un paramètre θ et (X_1, X_2, \dots, X_n) un n échantillon extrait de X . On appelle estimateur de θ toute statistique $\hat{\theta}_n$ fonction de X_1, X_2, \dots, X_n :

$$\hat{\theta}_n = \varphi(X_1, X_2, \dots, X_n).$$

Exemple 1.6.2 $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur de l'espérance de X .

Quelques qualités d'un estimateur

1. Le biais

On appelle biais de l'estimateur $\hat{\theta}_n$ l'expression définie par

$$B(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta.$$

L'estimateur $\hat{\theta}_n$ est dit sans biais si :

$$B(\hat{\theta}_n) = 0 \text{ c'est-à-dire } E(\hat{\theta}_n) = \theta.$$

L'estimateur $\hat{\theta}_n$ est dit asymptotiquement sans biais si :

$$\lim_{n \rightarrow +\infty} B(\hat{\theta}_n) = 0.$$

Exemple 1.6.3 Soit $X \rightsquigarrow N(\mu, \sigma^2)$ et (X_1, X_2, \dots, X_n) extrait de X .

- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur sans biais de μ .

- $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ est un estimateur biaisé de σ^2 de biais $(-\frac{1}{n}\sigma^2)$ et asymptotiquement sans biais.

- $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ est un estimateur sans biais de σ^2 , on l'appelle la variance corrigée.

2. Comparaison

On appelle erreur quadratique moyenne de l'estimateur la quantité

$$R(\hat{\theta}_n) = E[\hat{\theta}_n - \theta]^2.$$

Si $\hat{\theta}_n^{(1)}$ et $\hat{\theta}_n^{(2)}$ sont deux estimateurs de θ , on dit que $\hat{\theta}_n^{(1)}$ est meilleur que $\hat{\theta}_n^{(2)}$ si

$$R(\hat{\theta}_n^{(1)}) \leq R(\hat{\theta}_n^{(2)}).$$

Remarque 1.6.4 $\forall \hat{\theta}_n$ estimateur de θ , $R(\hat{\theta}_n) = \text{Var}(\hat{\theta}_n) + B^2(\hat{\theta}_n)$.

3. Convergence

Définition 1.6.5 $\hat{\theta}_n$ est un estimateur convergent de θ si :

$$\forall \epsilon \geq 0, \lim_{n \rightarrow +\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0.$$

Proposition 1.6.6 Si $\hat{\theta}_n$ est un estimateur sans biais ou asymptotiquement sans biais de θ et $\lim_{n \rightarrow +\infty} \text{Var}(\hat{\theta}_n) = 0$ alors $\hat{\theta}_n$ est un estimateur convergent de θ .

Preuve voir [22]. ■

4. Estimateur exhaustif

Définition 1.6.7 On dit que $\hat{\theta}_n$ est un estimateur exhaustif pour tout θ si la loi conditionnelle de $X = (X_1, \dots, X_n)$ sachant $[\hat{\theta}_n = t]$ ne dépend pas de θ .

5. La quantité d'information de Fisher

Définition 1.6.8 On appelle quantité d'information de Fisher sur le paramètre θ fournie par l'échantillon (X_1, X_2, \dots, X_n) le réel positif noté $I_n(\theta)$ défini par

$$I_n(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \log L(x_1, x_2, \dots, x_n, \theta)\right)^2\right].$$

Théorème 1.6.9 Si le domaine de définition de $f(x, \theta)$ est indépendant de θ alors on a

$$I_n(\theta) = E\left[-\frac{\partial^2}{\partial \theta^2} \log L(x_1, x_2, \dots, x_n, \theta)\right] = nI(\theta)$$

avec $I(\theta)$ est l'information de Fisher de X_1 .

Preuve voir [22]. ■

6. Efficacité

Si $\hat{\theta}_n$ est un estimateur sans biais de θ , $\hat{\theta}_n$ est dit efficace si

$$\text{Var}(\hat{\theta}_n) = \frac{1}{I_n(\theta)}.$$

7. Efficacité asymptotique

Si $\hat{\theta}_n$ est un estimateur asymptotiquement sans biais de θ , $\hat{\theta}_n$ est dit asymptotiquement efficace si

$$\lim_{n \rightarrow +\infty} I_n(\theta) \text{Var}(\hat{\theta}_n) = 1.$$

8. Borne de Cramer-Rao (ou Freshet-Damois-Cramer-Rao)

Soit $\hat{\theta}_n$ un estimateur sans biais pour θ de dimension 1. Sous certaines conditions de régularité on a pour tout $\theta \in \Theta$: $\text{Var}(\hat{\theta}_n) \geq \frac{1}{I_n(\theta)}$. La

borne $\frac{1}{I_n(\theta)}$ est appelée borne de Cramer-Rao.

Alors un estimateur sans biais est dit efficace s'il atteint la borne de Cramer-Rao.

1.6.2 Généralisation (Cas de multiples paramètres)

On suppose que θ symbolise p paramètres inconnus : c'est le vecteur colonne $\theta = (\theta_1, \theta_2, \dots, \theta_p)^T \in \mathbb{R}^p$ $p \geq 1$.

Certaines quantités et inégalités s'adaptent avec l'utilisation de matrices. Les plus importantes d'entre elles sont décrites ci-dessous.

Partant d'un vecteur aléatoire colonne à p composantes, on note E_p son espérance et Σ_p sa matrice de covariance.

- Un estimateur de $\theta = (\theta_1, \dots, \theta_p)^T$ se note $\hat{\theta}_n = ((\hat{\theta}_1)_n, \dots, (\hat{\theta}_p)_n)^T$.
- Le biais de $\hat{\theta}_n$ devient le vecteur colonne à p composantes :

$$B_p(\hat{\theta}_n) = E_p(\hat{\theta}_n) - \theta = \left(E((\hat{\theta}_1)_n) - \theta_1, \dots, E((\hat{\theta}_p)_n) - \theta_p \right)^T.$$

- L'erreur quadratique moyenne de $\hat{\theta}_n$ devient le réel :

$$R_p(\hat{\theta}_n) = E(\|\hat{\theta}_n - \theta\|^2) = \text{Trace}(\Sigma_p(\hat{\theta}_n)) + \|B_p(\hat{\theta}_n)\|^2$$

où $\|\cdot\|$ désigne la norme euclidienne.

- La dérivée partielle 1^{er} de $l(\tilde{x}; \theta)$ en θ est un vecteur à p composantes :

$$\frac{\partial}{\partial \theta} l(\tilde{x}; \theta) = \left(\frac{\partial}{\partial \theta_1} l(\tilde{x}; \theta), \frac{\partial}{\partial \theta_2} l(\tilde{x}; \theta), \dots, \frac{\partial}{\partial \theta_p} l(\tilde{x}; \theta) \right)^T.$$

- La dérivée partielle 2nd de $l(\tilde{x}; \theta)$ en θ est une matrice carrée à p lignes et p colonnes :

$$\frac{\partial^2}{\partial \theta^2} l(\tilde{x}; \theta) = \left(\frac{\partial^2}{\partial \theta_j \partial \theta_k} l(\tilde{x}; \theta) \right)_{j,k=1,\dots,p}.$$

- L'information de Fisher est une matrice carrée à p lignes et p colonnes :

$$I_n(\theta) = \left(-E \left(\frac{\partial^2}{\partial \theta_j \partial \theta_k} l(\tilde{x}; \theta) \right) \right)_{j,k=1,\dots,p}.$$

- La borne de Cramer-Rao devient, pour tout estimateur $\hat{\theta}_n$ de θ sans biais

$$\Sigma_p(\hat{\theta}_n) \geq I_n^{-1}(\theta).$$

où $I_n^{-1}(\theta)$ désigne l'inverse de la matrice $I_n(\theta)$ et le symbole \geq désigne une inégalité matricielle. pour toutes matrices A et B à p lignes et p colonnes, $A \geq B$ signifie que la matrice $A - B$ est positive, c'est-à-dire. pour tout vecteur colonne v à p composantes réels, $v^t(A - B)v \geq 0$.

1.6.3 Construction d'un estimateur de θ

Méthode du maximum de vraisemblance

Définition 1.6.10 Soit $\Theta \in \mathbb{R}^p$ l'ensemble des valeurs possibles du paramètre θ à estimer. On appelle estimateur du maximum de vraisemblance (E.M.V) de θ la statistique $\hat{\theta}_n$ telle que $\forall \tilde{x} = (x_1, \dots, x_n)$,

$$L(\tilde{x}, \hat{\theta}_n) \geq L(\tilde{x}, \theta), \forall \theta \in \Theta.$$

Théorème 1.6.11 Soit $f(x, \theta)$ la densité de X . Si le domaine de définition de f est indépendant de θ alors l'E.M.V de θ est la statistique $\hat{\theta}_n$ solution de l'équation du système :

$$\begin{cases} \frac{\partial}{\partial \theta} \log L(\tilde{x}, \theta) = 0, \text{ pour trouver } \hat{\theta}_n, \\ \frac{\partial^2}{\partial \theta^2} \log L(\tilde{x}, \theta)|_{\theta=\hat{\theta}_n} \leq 0, \text{ pour assurer l'existence du } \max_{\theta \in \Theta} L(\tilde{x}, \theta). \end{cases}$$

Preuve voir [22]. ■

Méthode des moments

La méthode des moments consiste à estimer les paramètres $\theta_1, \theta_2, \dots, \theta_p$, en égalisant les moments empiriques calculés à partir de l'échantillon avec les moments théoriques de même ordre. Soit $\mu_k = E(X^k)$, $k = 1, 2, \dots, p$; moments d'ordre k de la population (théorique) et $\hat{\mu}_k(X) = \frac{1}{n} \sum_{i=1}^n X_i^k$ moment empirique d'ordre k de l'échantillon. La solution du système $\hat{\mu}_k = \mu_k$, $k = 1, \dots, p$; nous donne les estimateurs de $\theta_1, \theta_2, \dots, \theta_p$.

$$\begin{cases} \hat{\mu}_1 = \mu_1, \\ \hat{\mu}_2 = \mu_2, \\ \cdot \\ \hat{\mu}_p = \mu_p. \end{cases} \quad p \text{ équations à } p \text{ inconnus,}$$

Dans la plupart des cas, les estimateurs obtenus par la méthode des moments sont convergents, asymptotiquement normaux mais en général ne sont pas efficaces.

Exemple 1.6.12 Soit (X_1, X_2, \dots, X_n) un n -échantillon issu de $X \rightsquigarrow N(\mu, \sigma^2)$ où μ, σ^2 sont inconnus. Estimons μ et σ^2 par la méthode des moments.

$$\begin{cases} \mu_1 = E(X) \\ \mu_2 = E(X^2) = \text{Var}(X) + (E(X))^2 \\ \hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^n X_i \\ \hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases} \Rightarrow \begin{cases} \mu = \frac{1}{n} \sum_{i=1}^n X_i \\ \mu^2 + \sigma^2 = \frac{1}{n} \sum_{i=1}^n X_i^2. \end{cases}$$

L'estimateur de μ est $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \hat{\mu}^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2. \end{aligned}$$

Chapitre 2

Introduction aux copules

Le but de ce chapitre est de fournir une introduction à la notion de copule. D’abord nous commencerons par un historique des copules puis nous définirons la copule et ses propriétés. Ensuite nous présenterons les copules les plus utilisées accompagnées de leur représentation graphique pour mieux illustrer les propos expliqués. Puis, nous parlerons aussi de quelques familles paramétriques des copules les plus connues et enfin nous ferons recours à d’autres indicateurs de dépendance.

2.1 Historique des copules

Le terme copule (copula) vient du mot latin “*copūlae*”, qui signifie au sens figuré, liaison, lien, alliance ou union. Le concept de copule a été introduit par Sklar en 1959 pour résoudre un problème de probabilité énoncé par Maurice Fréchet. À l’époque, Abe Sklar et Berthold Schweizer travaillent sur les travaux de Karl Menger concernant les espaces métriques aléatoires (Probabilistic Metric Space ou PMS), qui sont une généralisation de l’espace métrique usuel introduit par Fréchet en 1906. Même si les copules occupent une place importante dans l’œuvre de Sklar et Schweizer, elles ne sont pas l’objet central de leurs recherches. L’utilisation des copules par Sklar et Schweizer est assez originale : elles interviennent pour résoudre certains problèmes et ne font pas l’objet véritablement d’études appropriées. Pendant de nombreuses années, les copules sont restées très peu utilisées en statistique, on peut toutefois citer : les travaux de Kimeldorf et Sampson (voir [27]) sur la dépendance, les recherches de Paul Deheuvels (voir [5]) sur la fonction de dépendance empirique pour construire des tests non paramétriques d’indépendance. L’étude systématique des copules débute dans le milieu des années 1980 par quelques statisticiens. Le point de départ est bien sûr *l’article The joy of copulas* de Genest et MacKey (voir [18]) publié dans *The American Statistician*. Suivront de nombreux travaux de Christian Genest avec différents co-auteurs (MacKey, Louis-Paul Rivest,...). Maintenant, les copules sont devenues un outil standard largement utilisé pour étudier la dépendance, les modèles de survie, etc. La théorie des copules permet de construire des modèles paramétriques pour la loi jointe de d variables aléatoires réelles dépendantes (X_1, \dots, X_d) .

Dans ce travail on se restreindra au cas bivarié ($d = 2$) mais les méthodes peuvent aisément se généraliser en dimension supérieure (au moins en théorie).

2.2 Copules bivariées

2.2.1 Définitions et théorèmes

Définition 2.2.1 On appelle copule bivariée toute fonction C définie de $[0, 1]^2 \rightarrow [0, 1]$ qui possède les propriétés suivantes :

i) la copule C est attachée, c'est-à-dire

$$\forall u \in [0, 1]; C(u, 0) = C(0, u) = 0, \quad (2.1)$$

ii) les marges sont uniformes, c'est-à-dire

$$\forall u \in [0, 1]; C(u, 1) = C(1, u) = u, \quad (2.2)$$

iii) C est 2-croissante c'est -à-dire $\forall (u_1, u_2); (v_1, v_2) \in [0, 1] \times [0, 1]$ avec $u_1 \leq u_2$ et $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0. \quad (2.3)$$

Exemple 2.2.2 Soit $(u, v) \in [0, 1]^2$, les fonctions $W(u, v) = \max(u + v - 1, 0)$; $M(u, v) = \min(u, v)$ et $\Pi(u, v) = uv$ sont des copules bivariées. En effet on a :

1. $\forall u \in [0, 1]^2, \min(u, 0) = \min(0, u) = 0 \Rightarrow$ i) est vérifié,
2. $\forall u \in [0, 1]^2, \min(u, 1) = \min(1, u) = u \Rightarrow$ ii) est vérifié,
3. $\forall (u_1, u_2) \in [0, 1]^2, (v_1, v_2) \in [0, 1]^2$ avec $u_1 \leq u_2$ et $v_1 \leq v_2$, on a les cas suivants
 - a. si $u_1 \leq u_2 < v_1 \leq v_2$, on a

$$\min(u_2, v_2) - \min(u_2, v_1) - \min(u_1, v_2) + \min(u_1, v_1) = u_2 - u_2 - u_1 + u_1 \geq 0,$$
 - b. si $v_1 \leq v_2 < u_1 \leq u_2$, on a

$$\min(u_2, v_2) - \min(u_2, v_1) - \min(u_1, v_2) + \min(u_1, v_1) = v_2 - v_1 - v_2 + v_1 \geq 0,$$
 - c. si $u_1 < v_1 < u_2 < v_2$, on a

$$\min(u_2, v_2) - \min(u_2, v_1) - \min(u_1, v_2) + \min(u_1, v_1) = u_2 - v_1 - u_1 + u_1 \geq 0 \Rightarrow$$
 iii) est vérifié.

Par conséquent, M est une copule. De la même façon, on peut montrer que W et Π définissent des copules.

Nous pouvons maintenant énoncer le fameux théorème de Sklar qui permet de représenter toute fonction de répartition conjointe à l'aide d'une copule.

Théorème 2.2.3 (théorème de Sklar) Soit H une fonction de répartition conjointe de marges F et G , alors il existe une copule $C : [0, 1]^2 \rightarrow [0, 1]$ telle que pour tout $(x, y) \in \mathbb{R}^2$

$$H(x, y) = C(F(x), G(y)). \quad (2.4)$$

De plus, cette copule est unique lorsque les marges F et G sont continues.

Ce théorème montre que la loi H d'un couple (X, Y) est composée de trois éléments : les lois marginales F et G , ainsi qu'une fonction C appelée copule qui les lie. Les fonctions F et G décrivent les comportements marginaux de X et Y , leur dépendance est encodée dans la copule C .

Preuve À l'aide de la définition 1.5.1 et de la proposition 1.5.2 du chapitre 1, nous pouvons montrer le théorème 2.2.3. On a :

$$\begin{aligned}
H(x, y) &= P(X \leq x, Y \leq y) \\
&= P(F^{-1}(U) \leq x, G^{-1}(V) \leq y) \\
&= P(U \leq F(x), V \leq F(G(y))) \\
&= C(F(x), G(y)).
\end{aligned}$$

■

À partir du théorème de Sklar on peut extraire une définition de la copule dépendant de la notion de variables aléatoires.

Définition 2.2.4 Soit H une fonction de répartition d'un couple aléatoire (X, Y) . Si les marges F de X et G de Y sont continues, alors la copule C définie sur $[0, 1]^2$ est la loi jointe du couple (U, V) , où $U = F(X)$ et $V = G(Y)$.

À l'inverse du théorème de Sklar, il est également possible d'exprimer une copule à partir d'une loi jointe et des inverses (ou pseudo-inverses) des fonctions de répartition marginale.

Corollaire 2.2.5 (Inverse du théorème de Sklar) Soit H une fonction de répartition conjointe de marges continues F et G . Soient F^{-1} , G^{-1} les pseudo-inverses de F et G , alors pour tout $(u, v) \in [0, 1]^2$, la copule C associée à H est donnée par :

$$C(u, v) = H(F^{-1}(u), G^{-1}(v)). \quad (2.6)$$

Preuve En utilisant le théorème 2.2.3 et en posant $u = F(x)$ et $v = G(y)$, on trouve que

$$\begin{aligned}
C(u, v) &= P(U \leq u, V \leq v) \\
&= P(F(X) \leq u, G(Y) \leq v) \\
&= P(X \leq F^{-1}(u), Y \leq G^{-1}(v)) \\
&= H(F^{-1}(u), G^{-1}(v)).
\end{aligned}$$

■

À partir du corollaire 2.2.5, nous pouvons construire une copule bivariée à partir d'une fonction de répartition conjointe et de lois marginales.

Exemple 2.2.6 1) Soit la distribution logistique bivariée de Gumbel, dont la fonction de répartition conjointe est : $H(x, y) = (1 + e^{-x} + e^{-y})^{-1}$, $\forall (x, y) \in \mathbb{R}^2$. On obtient les fonctions de répartition marginales :

$$F(x) = \lim_{y \rightarrow +\infty} H(x, y) = (1 + e^{-x})^{-1} \text{ et } G(y) = \lim_{x \rightarrow +\infty} H(x, y) = (1 + e^{-y})^{-1},$$

et les fonctions inverses sont respectivement :

$$F^{-1}(u) = -\ln\left(\frac{1}{u} - 1\right) \text{ et } G^{-1}(v) = -\ln\left(\frac{1}{v} - 1\right), \text{ pour tout } (u, v) \in [0, 1]^2.$$

En se basant sur le corollaire 2.2.5, la copule associée à H est :

$$\begin{aligned} C(u, v) &= H(F^{-1}(u), G^{-1}(v)) \\ &= (1 + e^{\ln(\frac{1}{u}-1)} + e^{\ln(\frac{1}{v}-1)})^{-1} \\ &= \frac{uv}{u + v - uv}. \end{aligned}$$

2) Soit (X, Y) un couple de variable aléatoire dont la fonction jointe H est définie par :

$$H(x, y) = \exp \left\{ - \left[(x + y) - (x^{-\theta} + y^{-\theta})^{-\frac{1}{\theta}} \right] \right\}, \forall (x, y) \in [0, +\infty]^2; \theta \geq 1,$$

et admet pour marginales $F(x) = e^{-x}$ et $G(y) = e^{-y}$. Les fonctions inverses sont respectivement :

$$F^{-1}(u) = -\ln(u) \text{ et } G^{-1}(v) = -\ln(v), \text{ pour tout } (u, v) \in [0, 1]^2.$$

D'après le corollaire 2.2.5 la copule associée est exprimée comme suit :

$$C(u, v) = uv \exp \left\{ - \left[(-\ln u^{-\theta}) + (-\ln v^{-\theta}) \right]^{-\frac{1}{\theta}} \right\} \quad \theta \geq 1.$$

C'est la copule de Galambos.

Bornes de Fréchet-Hoeffding

Proposition 2.2.7 Pour toute copule C nous avons :

$$W(u, v) = \max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v) = M(u, v).$$

Preuve On a :

$$\forall (u, v) \in [0, 1]^2 \begin{cases} C(u, v) \leq C(u, 1) = u, \\ C(u, v) \leq C(1, v) = v. \end{cases}$$

Alors $C(u, v) \leq \min(u, v) = M$. (2.7)

$\forall (u, v) \in [0, 1]^2$, en utilisant iii) de la définition 2.2.1 on a :

$$C(u, v) \geq C(u, 1) + C(1, v) - C(1, 1) \geq u + v - 1.$$

Or $\forall (u, v) \in [0, 1]^2, C(u, v) \geq 0$, donc

$$C(u, v) \geq \max(u + v - 1, 0) = W(u, v). \quad (2.8)$$

En confrontant (2.7) et (2.8) il vient : $W(u, v) \leq C(u, v) \leq M(u, v)$. ■

Les copules W et M sont appelées les **bornes de Fréchet-Hoeffding**. W représente la **borne inférieure** ou **copule minimale** et M est la **borne supérieure** ou **copule maximale**.

2.2.2 Propriétés fondamentales

1. Indépendance

Soient X et Y deux variables aléatoires continues de fonctions de répartition respectives F et G , on note C_{XY} la copule associée à (X, Y) . Alors X et Y sont indépendantes si et seulement si

$$C_{XY}(u, v) = F(x)G(y) = \Pi(u, v) = uv.$$

En effet :

▷ si X et Y sont indépendantes alors

$$\forall x, y \in \mathbb{R}, H(x, y) = F(x)G(y)$$

or

$$\begin{aligned} \forall u, v \in [0, 1] \quad C_{XY}(u, v) &= H(F^{-1}(u), G^{-1}(v)) \\ \Rightarrow C_{XY}(u, v) &= F(F^{-1}(u))G(G^{-1}(v)) = uv = \Pi(u, v) \end{aligned}$$

▷ si $C_{XY}(u, v) = \Pi(u, v) = uv$ pour tout $u, v \in [0, 1]$, on a

$$\forall x, y \in \mathbb{R}, H(x, y) = C_{XY}(F(x), G(y)) = \Pi(F(x), G(y)) = F(x)G(y)$$

donc les variables X et Y sont indépendantes.

2. Invariance

Théorème 2.2.8 Soient deux variables aléatoires continues X, Y de marginales F et G respectivement et de copule associée C_{XY} . Si α et β sont deux fonctions strictement croissantes alors

$$C_{\alpha(X)\beta(Y)} = C_{XY}.$$

Preuve Soient F_1, G_1, F_2, G_2 les fonctions de répartition de $X, Y, \alpha(X)$ et $\beta(Y)$ respectivement. Les fonctions α et β sont strictement croissantes alors

$$F_2(x) = P(\alpha(X) \leq x) = P(X \leq \alpha^{-1}(x)) = F_1(\alpha^{-1}(x)), \text{ aussi}$$

$$G_2(y) = P(\beta(Y) \leq y) = P(Y \leq \beta^{-1}(y)) = G_1(\beta^{-1}(y)).$$

$\forall x, y \in \mathbb{R}$ on a :

$$\begin{aligned} C_{\alpha(X)\beta(Y)}(F_2(x), G_2(y)) &= P(\alpha(X) \leq x, \beta(Y) \leq y) \\ &= P(X \leq \alpha^{-1}(x), Y \leq \beta^{-1}(y)) \\ &= C_{XY}(F_1(\alpha^{-1}(x)), G_1(\beta^{-1}(y))) \\ &= C_{XY}(F_2(x), G_2(y)). \end{aligned}$$

■

Par exemple, la copule de la distribution lognormale est la même que celle de la loi normale (en effet la première est une transformation strictement croissante “ $y = \log x$ ” de la seconde). Dans la littérature des copules, ce théorème est connu comme étant le théorème d’invariance sous transformations strictement croissantes. Bien que l’invariance soit assurée seulement pour une croissance stricte des deux fonctions appliquées, il existe un résultat dans le cas où α ou β sont plutôt décroissantes, le comportement de la copule étant ainsi prévisible (voir théorème 2.4.4 de [29]).

3. La symétrie

Si $C(u, v) = C(v, u)$ pour tout $(u, v) \in [0, 1]^2$, on dit que C est symétrique.

4. Ordre

Soient C_1, C_2 deux copules, on dit que C_1 est plus petite que C_2 ou C_2 est plus grande que C_1 et on note $C_1 < C_2$ si

$$C_1(u, v) \leq C_2(u, v) \text{ pour tout } (u, v) \in [0, 1]^2.$$

Exemple 2.2.9 La copule $W = \max(u + v - 1, 0)$ est plus petite que la copule $M = \min(u, v)$.

5. Dérivées partielles

Les dérivées partielles de $C(u, v)$ existent presque sûrement pour tout

$$(u, v) \in [0, 1]^2, \quad 0 \leq \frac{\partial}{\partial u} C(u, v) \leq 1 \text{ et } 0 \leq \frac{\partial}{\partial v} C(u, v) \leq 1.$$

6. Continuité

On note que les copules sont des fonctions continues. Plus précisément, elles vérifient une condition de Lipschitz. $\forall u_1, u_2, v_1, v_2 \in [0, 1]$, avec $u_1 \leq u_2$ et $v_1 \leq v_2$, on a :

$$|C(u_2, v_2) - C(u_1, v_1)| \leq |u_2 - u_1| + |v_2 - v_1|.$$

7. Copule harmonique

Soit C une copule dont les dérivées partielles de second ordre sont continues dans $[0, 1]^2$. On dit C est harmonique dans $[0, 1]^2$ si C satisfait l'équation de Laplace dans $[0, 1]^2$:

$$\nabla^2 C(u, v) = \frac{\partial^2}{\partial u^2} C(u, v) + \frac{\partial^2}{\partial v^2} C(u, v) = 0.$$

Exemple 2.2.10 La copule $\Pi(u, v) = uv$ est une copule harmonique car :

$$\frac{\partial^2}{\partial u^2} (uv) = 0 = \frac{\partial^2}{\partial v^2} (uv).$$

8. Copule homogène

Une copule C est homogène de degré k si $\exists k \in \mathbb{R}^+, \forall u, v, \lambda \in [0, 1]$,

$$C(\lambda u, \lambda v) = \lambda^k C(u, v).$$

Exemple 2.2.11 - La fonction $\Pi(u, v) = uv$ est homogène de degré 2 car

$$(\lambda u)(\lambda v) = \lambda^2 uv.$$

- La fonction $M = \min(u, v)$ est homogène de degré 1 car

$$\min(\lambda u, \lambda v) = \lambda \min(u, v).$$

9. Convexité et concavité

Soient $(a, b), (c, d) \in [0, 1]^2$ et $\lambda \in [0, 1]$, une copule C est concave si on a :

$$C(\lambda a + (1 - \lambda)c, \lambda b + (1 - \lambda)d) \geq \lambda C(a, b) + (1 - \lambda)C(c, d).$$

Elle est dite convexe si on a :

$$C(\lambda a + (1 - \lambda)c, \lambda b + (1 - \lambda)d) \leq \lambda C(a, b) + (1 - \lambda)C(c, d).$$

2.2.3 Densité d'une copule

Supposons que les distributions marginales F , G et la copule C sont différentiables, alors la densité jointe notée h du couple aléatoire (X, Y) prend la forme suivante

$$h(x, y) = f(x)g(y)c(F(x), G(y))$$

où f et g sont les densités marginales respectives de X et Y et c est la densité de la copule C définie par :

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial v \partial u}.$$

Nous constatons ainsi que l'on peut séparer la densité jointe h en deux blocs. Le premier $c(F(x), G(y))$ contient l'information sur la structure de dépendance des variables aléatoires X, Y . Le second est le produit des densités marginales. Cela montre que les copules représentent un moyen d'extraire la structure de dépendance de la distribution jointe et de la séparer des comportements marginaux. Il est possible de caractériser la densité d'une copule entièrement par la densité jointe et les densités marginales. En effet, si la fonction de répartition H du couple (X, Y) admet une densité continue alors la densité de copule de ce couple est donnée par :

$$c(u, v) = \frac{h(F^{-1}(u), G^{-1}(v))}{f(F^{-1}(u))g(G^{-1}(v))}.$$

2.2.4 Quelques copules populaires et leur représentation graphique

De nombreux types de copules ont été proposées dans la littérature statistique. Les plus utilisées sont introduites dans cette section. Elles sont toutes disponibles dans le package `copula` du logiciel R. Une liste plus complète peut être trouvée dans [29].

1. Copule indépendance (ou copule produit)

La copule indépendance est définie par :

$$C(u, v) = \Pi(u, v) = uv.$$

La figure 2.1 représente la copule indépendante "CDF" et sa densité "pdf".

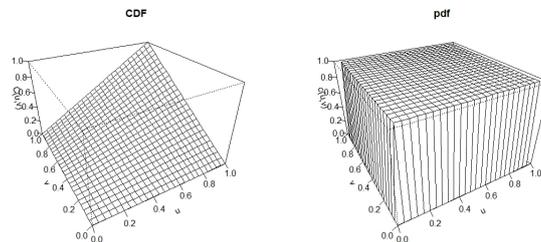


FIGURE 2.1 – Représentation graphique de la copule indépendante et sa densité.

Cette copule est harmonique, homogène de degré 2 (voir les propriétés 6 et 7 de la section 2.2.2).

2. Copule de survie

Cette copule est très intéressante car dans la majorité des applications on s'intéresse à la durée de vie des individus dans une certaine population. Dans le cas univarié la probabilité de survie est définie par :

$$P(X > x) = 1 - F(x) := \bar{F}(x),$$

F représente la fonction de répartition de X .

De même pour le cas bivarié, si H est la fonction de répartition jointe associée au couple aléatoire (X, Y) de distributions marginales F et G respectivement et C sa copule associée, la fonction jointe de survie est définie par :

$$\bar{H}(x, y) = P(X > x, Y > y)$$

et les marginales de survie sont \bar{F} et \bar{G} alors

$$\begin{aligned} \bar{H}(x, y) &= P(X > x, Y > y) \\ &= 1 - P(X \leq x \text{ ou } Y \leq y) \\ &= 1 - [F(x) + G(y) - P(X \leq x, Y \leq y)] \\ &= 1 - F(x) - G(y) + H(x, y). \end{aligned}$$

Et donc $\bar{H}(x, y) = P(X > x, Y > y) = 1 - F(x) - G(y) + H(x, y)$. Ainsi

$$\begin{aligned} \bar{H}(x, y) &= 1 - F(x) - G(y) + C(F(x), G(y)) \\ &= \bar{F}(x) + \bar{G}(y) - 1 + C(F(x), G(y)) \\ &= \bar{F}(x) + \bar{G}(y) - 1 + C(1 - \bar{F}(x), 1 - \bar{G}(y)). \end{aligned}$$

On peut ainsi écrire la distribution de survie \bar{H} en fonction de la copule de survie \tilde{C} de X, Y comme suit

$$\bar{H}(x, y) = \tilde{C}(\bar{F}(x), \bar{G}(y)) = \bar{F}(x) + \bar{G}(y) - 1 + C(1 - \bar{F}(x), 1 - \bar{G}(y)).$$

La copule de survie est donc définie par :

$$\tilde{C}(u, v) = u + v - 1 + C(1 - u, 1 - v).$$

3. La copule gaussienne

La copule gaussienne bidimensionnelle, aussi appelée copule normale, est définie de la façon suivante :

$$C_\rho^{ga}(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))$$

où $\rho \in]-1, 1[$ est le coefficient de corrélation et Φ_ρ représente la distribution normale bidimensionnelle centrée dont la matrice de corrélation est égale à $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ et Φ^{-1} est l'inverse de la distribution normale unidimensionnelle centrée réduite, ainsi on a :

$$C_\rho^{ga}(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{s^2 + t^2 - 2\rho st}{2(1-\rho^2)}\right) ds dt.$$

Il s'agit de la copule la plus couramment utilisée dans la modélisation des dépendances linéaires. Son importance réside dans le fait qu'elle est sous-jacente à la distribution normale multidimensionnelle. En effet, modéliser la structure de dépendance d'un échantillon par une copule gaussienne est cohérent avec la mesure de cette dépendance par le coefficient de corrélation linéaire. Cependant lorsque l'on souhaite modéliser une dépendance non linéaire ou entre événements extrêmes, on va faire appel à d'autres copules que nous expliciterons dans la suite.

En dérivant la formule définissant la copule gaussienne, on peut extraire sa densité bidimensionnelle qui s'écrit :

$$c_{\rho}^{ga}(u, v) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 + y^2 - 2\rho xy}{2(1-\rho^2)} + \frac{x^2 + y^2}{2}\right)$$

avec $x = \Phi^{-1}(u)$ et $y = \Phi^{-1}(v)$.

On retrouve les cas particuliers suivants comme cas limites :

$$\lim_{\rho \rightarrow -1} C_{\rho}^{ga}(u, v) = W(u, v), \lim_{\rho \rightarrow 0} C_{\rho}^{ga} = \Pi(u, v) \text{ et } \lim_{\rho \rightarrow 1} C_{\rho}^{ga} = M(u, v).$$

La figure 2.2 illustre la densité des trois copules gaussiennes bidimensionnelles pour $\rho = 0, \rho = 0.5$ et $\rho = 0.9$.

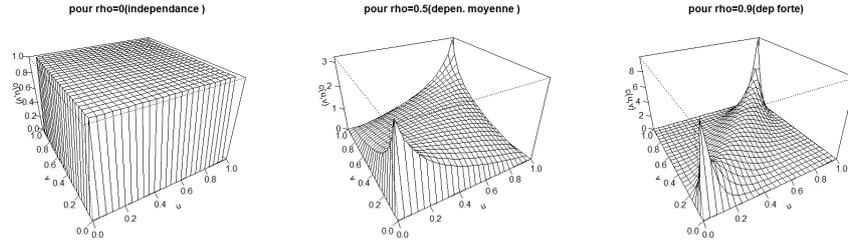


FIGURE 2.2 – Densité des trois copules gaussienne bidimensionnelle pour des valeurs de ρ différentes.

On remarque que pour $\rho = 0$, on a $C_{\rho}^{ga} = \Pi$.

4. La copule de student (ou t-copula)

Dans le cas univarié, la fonction de répartition de Student centrée réduite est définie par

$$T_k(x) = \int_{-\infty}^x \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})} \left(1 + \frac{s^2}{k}\right)^{-\frac{k+1}{2}} ds,$$

avec Γ la fonction d'Euler et $k > 0$ est le nombre de degrés de liberté.

Dans le cas bivarié, soit $\rho \in]-1, 1[$, la distribution de Student bidimensionnelle centrée de matrice de corrélation fonction de ρ et de degré de liberté k , $T_{\rho, k}$ est donnée par :

$$T_{\rho, k}(x, y) = \int_{-\infty}^x \int_{-\infty}^y \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{s^2 + t^2 - 2\rho st}{k(1-\rho^2)}\right)^{-\frac{k+2}{2}} ds dt.$$

La copule de Student bidimensionnelle est définie de la façon suivante :

$$\begin{aligned} C_{\rho,k}^{st}(u,v) &= T_{\rho,k}(T_k^{-1}(u), T_k^{-1}(v)) \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{T_k^{-1}(u)} \int_{-\infty}^{T_k^{-1}(v)} \left(1 + \frac{s^2 + t^2 - 2\rho st}{k(1-\rho^2)}\right)^{-\frac{k+2}{2}} ds dt. \end{aligned}$$

En utilisant la formule définissant la densité de la copule, celle de la copule de Student s'écrit :

$$c_{\rho,k}^{st}(u,v) = \frac{k}{2\sqrt{1-\rho^2}} \frac{\Gamma(k/2)^2}{\Gamma(k+1/2)^2} \frac{\left(1 + \frac{x^2 + y^2 - 2\rho xy}{k(1-\rho^2)}\right)^{-\frac{k+2}{2}}}{\left[\left(1 + \frac{x^2}{k}\right)\left(1 + \frac{y^2}{k}\right)\right]^{-\frac{k+1}{2}}}$$

avec $x = T_k^{-1}(u)$ et $y = T_k^{-1}(v)$, $k > 0$.

La copule de Student (t-copula) est extraite de la même manière que la copule gaussienne mais cette fois-ci à partir de la distribution de Student bidimensionnelle.

La figure 2.3 illustre la densité des trois copules de Student bidimensionnelles pour une corrélation $\rho = 0$, $\rho = 0.5$ et $\rho = 0.9$ et de même degré de liberté $k = 8$.

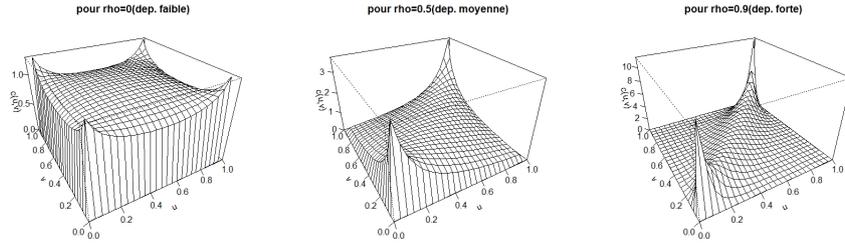


FIGURE 2.3 – Densité des trois copules de student pour des ρ différents et de même degré de liberté.

On remarque que l'indépendance n'est pas caractérisée par une valeur nulle de ρ , car la densité c_{ρ}^{st} , ne vaut pas uniquement 1 sur $[0,1]^2$. Cela sera le cas notamment pour de grandes valeurs du degré de liberté. Contrairement à la copule Gaussienne, la copule de Student possède des dépendances de queue à droite et à gauche (voir [30] et section 2.4, sous section 2.4.2 pour la notion de dépendance de queue) pour $\rho \neq \pm 1$.

Remarque 2.2.12 Si le degré de liberté $k \rightarrow +\infty$ alors la copule de student converge vers la copule Gaussienne (voir [36]).

5. Copule de Farlie-Gumbel-Morgenstern

La copule de Farlie-Gumbel-Morgenstern est la plus connue des copules paramétriques. Cette copule a été étudiée par Morgenstern (voir [28]), Gumbel (voir [20]), et enfin Farlie (voir [9]). L'expression de la copule de Farlie-Gumbel-Morgenstern est donnée par

$$C_w^{FGM}(u,v) = uv - wuv(1-u)(1-v), \quad -1 \leq w \leq 1.$$

En dérivant deux fois, on obtient sa densité associée

$$c_w^{FGM}(u,v) = 1 - w(1-2u)(1-2v).$$

Sa densité est présentée à la figure 2.4, à savoir lorsque $w = 0.25$ et $w = 0.85$.

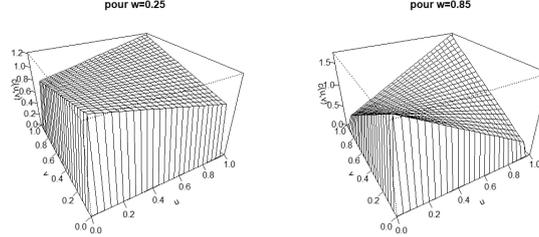


FIGURE 2.4 – Densité de la copule de Farlie-Gumbel-Morgenstern pour des différentes valeurs de w .

6. La copule de Clayton

Connue aussi sous les noms de : copule de Cook Johnson ou de Kimeldorf-Sampson, la copule de Clayton est une copule archimédienne (qu'on verra plus loin) dont le générateur est défini pour $\theta > 0$ et $t \in]0, 1]$ par :

$$\Phi_{\theta}(t) = \theta^{-1}(t^{-\theta} - 1).$$

La copule de Clayton bidimensionnelle s'écrit donc sous la forme :

$$C_{\theta}^{Cl}(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}.$$

Cette copule est différentiable et sa densité est donnée par :

$$c_{\theta}^{Cl}(u, v) = \frac{\partial}{\partial v} \left(\frac{\partial C_{\theta}^{Cl}(u, v)}{\partial u} \right),$$

$$c_{\theta}^{Cl}(u, v) = (\theta + 1)(uv)^{-\theta-1}(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}-2}.$$

La copule de Clayton possède une dépendance asymptotique à gauche (au voisinage de $(0, 0)$) appelée dépendance de queue inférieure mais pas une dépendance asymptotique à droite (au voisinage de $(1, 1)$) appelée dépendance de queue supérieure (voir section 2.4, sous section 2.4.2). Cette copule est plus adéquate aux données qui représentent une forte dépendance à gauche et une faible à droite. Ainsi, il s'agit d'une copule modélisant une relation totalement asymétrique.

La figure 2.5 illustre la densité bidimensionnelle de la copule de Clayton pour $\theta = 5$.

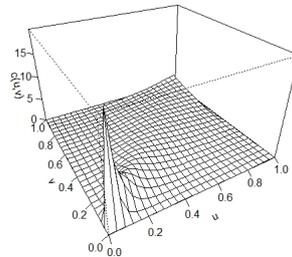


FIGURE 2.5 – Densité de la copule de Clayton pour $\theta = 5$.

7. La copule de Frank

Le générateur de cette copule archimédienne est donné par :

$$\Phi_{\vartheta}(t) = -\ln \frac{\exp(-\vartheta t) - 1}{\exp(-\vartheta) - 1},$$

où $\vartheta \neq 0$ et $t \in]0, 1]$.

La copule de Frank bidimensionnelle s'écrit donc sous la forme :

$$C_{\vartheta}^{Fr}(u, v) = -\frac{1}{\vartheta} \ln \left\{ 1 + \frac{(\exp(-\vartheta u) - 1)(\exp(-\vartheta v) - 1)}{\exp(-\vartheta) - 1} \right\}.$$

La densité de la copule de Frank est égale à :

$$\begin{aligned} c_{\vartheta}^{Fr}(u, v) &= \frac{\partial}{\partial v} \left(\frac{\partial C_{\vartheta}^{Fr}(u, v)}{\partial u} \right) \\ &= \frac{\partial}{\partial v} \left[\frac{-1}{\vartheta} \left(\frac{-\vartheta \exp(-\vartheta u)(\exp(-\vartheta v) - 1)}{(\exp(-\vartheta) - 1) + (\exp(-\vartheta u) - 1)(\exp(-\vartheta v) - 1)} \right) \right] \\ &= \frac{\vartheta(1 - \exp(-\vartheta)) \exp(-\vartheta(u + v))}{[(1 - \exp(-\vartheta)) - (\exp(-\vartheta u) - 1)(\exp(-\vartheta v) - 1)]^2}. \end{aligned}$$

La figure 2.6 correspond à la densité de la copule de Frank bidimensionnelle pour $\vartheta = 2$.

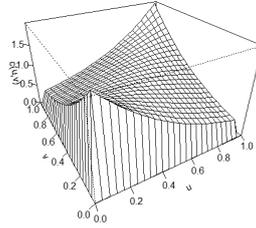


FIGURE 2.6 – Densité de la copule de Frank pour $\vartheta = 2$.

8. Copule de Gumbel-Hougaard (ou copule de Gumbel)

La copule de Gumbel appartient à la famille des copules archimédiennes (voir ci-dessous) et son générateur est :

$$\Phi_r(t) = (-\ln(t))^r.$$

Avec $r > 1$ et $u, v \in]0, 1]$, la copule de Gumbel bidimensionnelle s'écrit donc sous la forme :

$$C_r^{GH}(u, v) = \exp \left[- \{ (-\log u)^r + (-\log v)^r \}^{\frac{1}{r}} \right].$$

La densité de cette copule a une forme compliquée.

Néanmoins, il est possible d'obtenir son expression à l'aide d'un logiciel de calcul symbolique comme *Maple*.

La copule de Gumbel est plus adéquate aux données qui représentent une forte dépendance à droite et une faible à gauche car contrairement à la copule de Clayton, elle possède donc une dépendance de queue supérieure

mais pas une dépendance de queue inférieure (voir section 2.4, sous section 2.4.2).

La figure 2.7 correspond à la densité de la copule de Gumbel bidimensionnelle pour $\theta = 3$.

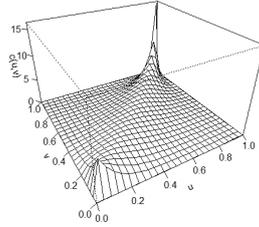


FIGURE 2.7 – Densité de la copule de Gumbel pour $\theta = 3$.

Remarque 2.2.13 Pour les Copules de Clayton, de Gumbel Hougaard et de Frank on retrouve les cas particuliers suivants comme cas limites :

$$\lim_{\theta \rightarrow +\infty} C_{\theta}^{Cl}(u, v) = \Pi(u, v) \quad \text{et} \quad \lim_{\theta \rightarrow 0} C_{\theta}^{Cl}(u, v) = M(u, v),$$

$$\lim_{r \rightarrow 1} C_r^{GH}(u, v) = \Pi(u, v) \quad \text{et} \quad \lim_{r \rightarrow +\infty} C_r^{GH}(u, v) = M(u, v),$$

$$\lim_{\vartheta \rightarrow -\infty} C_{\vartheta}^{Fr}(u, v) = W(u, v), \quad \lim_{\vartheta \rightarrow 0} C_{\vartheta}^{Fr}(u, v) = \Pi(u, v), \quad \lim_{\vartheta \rightarrow +\infty} C_{\vartheta}^{Fr}(u, v) = M(u, v).$$

2.3 Étude de familles de copules paramétriques

Dans cette section, on va présenter les familles de copules paramétriques les plus utilisées en pratique. On commence par les copules elliptiques, ensuite, on introduit les copules archimédiennes, puis les copules de valeurs extrêmes et enfin les copules archimax. Nous donnerons dans chaque cas quelques exemples des copules populaires appartenant à chaque famille.

2.3.1 Famille des copules elliptiques

Les copules elliptiques sont définies à partir des lois elliptiques. Ainsi nous donnons une définition de la distribution elliptique.

Définition 2.3.1 Soit $\mathcal{M}_d(\mathbb{R})$ l'ensemble des matrices carrées réelles de taille $d \times d$. Une loi continue est dite elliptique de paramètre de position $\mu = (\mu_1, \mu_2, \dots, \mu_d)$ et de matrice de forme symétrique définie positive $\Sigma \in \mathcal{M}_d(\mathbb{R})$ si sa densité f peut s'écrire pour tout $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$:

$$f(x) = |\Sigma|^{-\frac{1}{2}} g((x - \mu)\Sigma^{-1}(x - \mu)^T)$$

où g est une fonction à valeurs positives dite génératrice de densité et vérifiant

$$\int_{\mathbb{R}^d} g(xx^T) dx = 1.$$

Cette famille de lois est dite elliptique car les courbes de niveaux de la densité sont des ellipses. La loi d'un vecteur gaussien est un exemple classique de loi elliptique, associé au choix de $g(y) = (2\pi)^{-\frac{d}{2}} \exp(-y/2)$. Pour une présentation détaillée de ces lois, le lecteur pourra se référer à [10]. Par conséquent, une copule est dite elliptique si elle est la copule d'une loi elliptique. Ainsi on appelle copule bivariée elliptique toute copule de la forme $\forall u, v \in [0, 1]$:

$$\begin{aligned} C_\rho(u, v) &= H_\rho(\Phi_{g,1}^{-1}(u), \Phi_{g,2}^{-1}(v)) \\ &= \frac{1}{\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi_{g,1}^{-1}(u)} \int_{-\infty}^{\Phi_{g,2}^{-1}(v)} g\left[-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\right] ds dt. \end{aligned}$$

où H_ρ est la distribution conjointe des variables X et Y , $\Phi_{g,1}^{-1}$ et $\Phi_{g,2}^{-1}$ leur fonction quantile respective et ρ leur coefficient de corrélation. Les deux classes les plus utilisées de copules elliptiques sont les copules gaussiennes et les copules de Student (voir section 2.2.4).

Remarque 2.3.2 *Les copules elliptiques sont des copules symétriques et relativement simples d'utilisation du fait que l'on connaît bien les distributions auxquelles elles sont associées. Elles sont souvent appelées copules implicites car n'ayant pas de forme analytique explicite et s'exprimant par conséquent en fonction de leur distribution bivariée.*

2.3.2 Famille des copules archimédiennes

La classe des copules archimédiennes joue un rôle important. D'une part, elles permettent de construire une grande variété de familles de copules et donc de représenter une grande variété de structures de dépendance. D'autre part, les copules ainsi générées ont des formes analytiques fermées et sont faciles à simuler. En effet, contrairement aux copules gaussienne et de Student, les copules archimédiennes ont le grand avantage de décrire des structures de dépendance très diverses dont notamment les dépendances dites asymétriques, où les coefficients de queue inférieure et de queue supérieure diffèrent. Pour plus d'éléments sur cette famille de copules le lecteur pourra se référer à [29].

Définition 2.3.3 *On appelle copule archimédienne toute copule de la forme :*

$$C_\Phi(u, v) = \Phi^{-1}(\Phi(u) + \Phi(v))$$

où $\Phi : [0, 1] \rightarrow [0, +\infty[$ est une fonction convexe décroissante vérifiant $\Phi(1) = 0$ et Φ^{-1} la pseudo-inverse de Φ . On pose $\Phi(0) = +\infty$ si $\lim_{t \rightarrow 0} \Phi(t) = +\infty$ et $\Phi^{-1}(t) = 0$ si $t \geq \Phi(0)$.

La fonction Φ est dite générateur archimédien de la copule C . Si $\Phi(0) = +\infty$ alors la copule C est dite strictement archimédienne auquel cas Φ^{-1} coïncide avec la fonction réciproque de Φ .

Dans le tableau 2.1, nous donnons les copules archimédiennes bidimensionnelles les plus populaires et leurs générateurs.

nom des copules	copule bivariable $C_\theta(u, v)$	générateur $\Phi_\theta(t)$
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$\theta^{-1}(t^{-\theta} - 1), \theta > 0$
Frank	$-\frac{1}{\theta} \ln\left\{1 + \frac{(\exp(-\theta u) - 1)(\exp(-\theta v) - 1)}{\exp(-\theta) - 1}\right\}, \theta \neq 0$	$-\ln \frac{\exp(-\theta t) - 1}{\exp(-\theta) - 1}$
Joe	$1 - \left[(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta \right]^{\frac{1}{\theta}}$ $\theta \in [1, +\infty]$	$-\log(1 - (1-t)^\theta)$
Gumbel	$C_\theta(u, v) = \exp \left[- \left\{ (-\log u)^\theta + (-\log v)^\theta \right\}^{\frac{1}{\theta}} \right]$ $(\theta \geq 1)$	$(-\ln(t))^\theta$

TABLE 2.1 – Les copules archimédiennes bidimensionnelles les plus populaires.

2.3.3 Famille des copules de valeurs extrêmes

Les copules de valeurs extrêmes permettent de modéliser des phénomènes extrêmes tels que les ouragans ou la crue des eaux. Une copule extrême, noté C_A , est la fonction de dépendance extraite d'une loi bivariable extrême. Elle s'exprime en tout $u, v \in [0, 1]$ sous la forme

$$C_A(u, v) = \exp \left[\ln(uv) A \left(\frac{\ln u}{\ln(uv)} \right) \right]$$

où $A : [0, 1] \rightarrow [0, 1/2]$ est une fonction convexe dite fonction de dépendance de Pickands qui satisfait $A(0) = A(1) = 1$ et $\max(t, 1-t) \leq A(t) \leq 1$ en tout $t \in [0, 1]$. Les copules de valeurs extrêmes ne permettent de modéliser qu'une dépendance positive car pour tout $u, v \in [0, 1]$, elles vérifient la condition

$$uv \leq C_A(u, v) \leq \min(u, v).$$

Lorsque $A(t) = 1$ en tout $t \in [0, 1]$, on retrouve la copule d'indépendance. Ainsi, la copule $\Pi(u, v) = uv$ est un modèle extrême. La copule de Gumbel-Hougaard appartient aussi à cette classe. Cette copule dépend d'un paramètre $\theta \geq 1$ et est engendrée par la fonction de Pickands donnée en tout $t \in [0, 1]$ par la formule

$$A_\theta(t) = [t^\theta + (1-t)^\theta]^{\frac{1}{\theta}}.$$

Il existe une autre définition plus connue des copules de valeurs extrêmes donnée par la définition suivante :

Définition 2.3.4 *Soit k une constante réelle positive, une copule est dite copule de valeurs extrêmes si elle vérifie la relation suivante : $C(u^k, v^k) = C^k(u, v)$.*

Il existe essentiellement deux grandes familles de modèles paramétriques usuels de copules de valeurs extrêmes bivariées : le modèle mixte ou de Tawn (voir [34]) et le modèle logistique ou de Gumbel (voir [19]). Les autres modèles proviennent généralement d'une extension symétrique ou asymétrique de ces modèles.

Nous donnons dans le tableau 2.2 quelques unes d'entre elles en adoptant les notations : $\tilde{u} = -\ln(u)$ et $\tilde{v} = -\ln(v)$.

modèles	copule $C_\theta^A(u, v)$	$A_\theta(t)$
II	uv	$A(t) = 1$
Gumbel A	$\exp\{-\tilde{u}^\theta + \tilde{v}^\theta\}^{\frac{1}{\theta}}, \theta \in [1, +\infty[$	$[t^\theta + (1-t)^\theta]^{\frac{1}{\theta}}$
Gumbel B	$uv \exp\{\theta \frac{\tilde{u}\tilde{v}}{\tilde{u} + \tilde{v}}\}, \theta \in [0, 1]$	$t^2 - \theta t + 1$
Galambos	$uv \exp\{-(\tilde{u}^{-\theta} + \tilde{v}^{-\theta})^{-\frac{1}{\theta}}\}, \theta \in [0, +\infty[$	$1 - [t^{-\theta} + (1-t)^{-\theta}]^{-\frac{1}{\theta}}$
Logistique de Joe	$\exp\{-[(\tilde{u}^{\theta_1} + \tilde{v}^{\theta_1}) - \nu(\tilde{u}, \tilde{v}; \theta)]^{\frac{1}{\theta_1}}\},$ $\theta \in [1, +\infty]^2$	$[t^{\theta_1} + (1-t)^{\theta_1} - \zeta(t; \theta)]^{\frac{1}{\theta_1}}$
Tawn	$uv \exp\{- (1 - \theta_2) + (\theta_1 - \theta_2)\tilde{u} +$ $\chi(\tilde{u}, \tilde{v}; \theta)\}, \theta \in [1, +\infty]^3$	$(1 - \theta_2) + (\theta_2 - \theta_1)t +$ $\xi(t; \theta)$
Marshall-Olikn	$u^{(1-\theta_1)}v^{1-\theta_2} \min(u^{\theta_1}, v^{\theta_2}), \theta \in [0, 1]^2$	$\max(1 - \theta_1 t, \theta_2(1 - t))$

TABLE 2.2 – Quelques copules de valeurs extrêmes.

$$\nu(\tilde{u}, \tilde{v}; \theta) = (\tilde{u}^{-\theta_1\theta_2} + \tilde{v}^{-\theta_1\theta_2})^{-\frac{1}{\theta_2}} \text{ et } \zeta(t; \theta) = [t^{\theta_1\theta_2} + (1-t)^{-\theta_1\theta_2}]^{-\frac{1}{\theta_2}}; \theta = (\theta_1, \theta_2).$$

$$\chi(\tilde{u}, \tilde{v}; \theta) = [(\theta_1\tilde{u})^{\theta_3} + (\theta_2\tilde{v})^{\theta_3}]^{\frac{1}{\theta_3}} \text{ et } \xi(t; \theta) = [(\theta_1 t)^{\theta_3} + (\theta_2(1-t))^{\theta_3}]^{\frac{1}{\theta_3}}; \theta = (\theta_1, \theta_2, \theta_3).$$

2.3.4 Famille des copules archimax

On considère une nouvelle famille de copules introduite par Capéraà et collaborateurs (voir [4]) qui englobe la plupart des familles connues des copules, notamment les copules archimédiennes et toutes les copules de valeurs extrêmes. Cette nouvelle famille offre plus de flexibilité pour la modélisation.

Définition 2.3.5 Une fonction bidimensionnelle est une copule archimax si et seulement si elle est de la forme :

$$C_{\Phi, A}(u, v) = \Phi^{-1}\left[(\Phi(u) + \Phi(v))A\left(\frac{\Phi(u)}{\Phi(u) + \Phi(v)}\right)\right]; \quad u, v \in [0, 1] \text{ avec :}$$

- $A : [0, 1] \rightarrow [\frac{1}{2}, 1]$ tel que $\max(t, t-1) \leq A(t) \leq 1$ pour tout $0 \leq t \leq 1$.
- $\Phi : [0, 1] \rightarrow [0, +\infty[$ est une fonction convexe, décroissante qui vérifie $\Phi(1) = 0$, avec la convention suivante : $\Phi(0) = \lim_{t \rightarrow 0^+} \Phi(t)$ et $\Phi^{-1}(t) = 0$ pour tout $t \geq \Phi(0)$.

Ainsi nous pouvons remarquer que l'ensemble des copules archimax $C_{\Phi, A}$ contient les copules de valeurs extrêmes ainsi que les copules archimédiennes. En effet si l'on pose $\Phi(t) = \ln(t)$, la copule $C_{\Phi, A}$ est alors une copule de valeurs extrêmes, c'est à dire :

$$C_{\Phi, A}(u, v) = C_A(u, v) = \exp\left[\ln(uv)A\left(\frac{\ln u}{\ln(uv)}\right)\right].$$

Si on pose $A(t) = 1$. On retrouve la forme générale des copules archimédiennes :

$$C_{\Phi, A}(u, v) = C_\Phi(u, v) = \Phi^{-1}(\Phi(u) + \Phi(v)).$$

Pour mieux approfondir cette classe de copule le lecteur pourra se référer à [17].

2.4 Mesures de dépendance

En statistique, on calcule souvent une mesure de dépendance entre deux ou plusieurs variables aléatoires. Il s'agit d'une pratique répandue vu son utilité pour décrire et mesurer la nature du lien qui les unit. Il existe une large gamme de mesures de dépendance entre les variables aléatoires. À titre d'exemple, nous citerons dans cette section le coefficient de corrélation de Pearson, qui mesure la dépendance linéaire, la dépendance de queue qui est une mesure locale, le tau de Kendall et le rho de Spearman, qui mesurent une forme de dépendance connue sous le nom de concordance.

2.4.1 Corrélation de Pearson

Définition 2.4.1 Soient X et Y deux variables aléatoires de variances respectives $V(X), V(Y)$ finies. La corrélation de Pearson entre les variables X et Y est donnée par :

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

où $\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(X.Y) - E(X)E(Y)$ est la covariance entre X et Y .

Il s'agit d'un coefficient de corrélation linéaire permettant de mesurer seulement la dépendance linéaire. Il est important de rappeler que la dépendance et la corrélation sont des notions différentes. En effet, si X et Y sont des variables indépendantes elles sont non corrélées mais la réciproque est fautive sauf dans le cas où les variables sont gaussiennes car la dépendance est alors entièrement caractérisée par le coefficient de corrélation. On peut donner le contre-exemple suivant : considérons $X \rightsquigarrow N(0, 1)$ et $Y = X^2$, alors

$$\text{Cov}(X, Y) = E(X.X^2) - E(X)E(X^2) = E(X^3) - E(X)E(X^2) = 0.$$

Donc $\rho(X, Y) = 0$. Bien que X et Y sont des variables aléatoires liées, elles ont une corrélation nulle. Pour d'autres distributions, le coefficient de corrélation linéaire offre une interprétation limitée. La discussion des détails peut être trouvée dans [8] pour les limites du coefficient de corrélation comme mesure de dépendance.

Supposons que nous ayons un échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ du couple (X, Y) , la version empirique du coefficient de corrélation de Pearson est donnée par :

$$\rho_n(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}$$

où \bar{x} et \bar{y} sont les moyennes empiriques de l'échantillon de X et de Y , S_x et S_y sont les écarts-types de l'échantillon de X et de Y tels que

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \text{ et } S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

La corrélation de Pearson peut être exprimée en fonction d'une copule comme l'indique la proposition ci-après.

Proposition 2.4.2 Si C désigne la copule des variables X et Y , la corrélation de Pearson $\rho(X, Y)$ a pour expression :

$$\rho(X, Y) = \frac{1}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \int_0^1 \int_0^1 (C(u, v) - uv) dF^{-1}(u) dG^{-1}(v).$$

Preuve

$$\begin{aligned} \rho(X, Y) &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ &= \frac{1}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [H(x, y) - F(x)G(y)] dx dy \end{aligned}$$

car $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$ où H est la distribution jointe de X et Y . On applique le changement de variable suivant : $u = F(x)$ et $v = G(y)$, on démontre le résultat de la proposition. ■

On remarque que la copule ne permet pas de déduire le coefficient de corrélation. En effet, les distributions marginales sont nécessaires.

2.4.2 Notion de dépendance de queue

La dépendance de queue est une mesure locale, car elle mesure la dépendance au niveau des queues de distribution. Il existe deux coefficients de dépendance de queues, définies comme suit.

Définition 2.4.3 Soient X, Y deux variables aléatoires continues de fonctions de répartition respectives F et G . Le coefficient de dépendance inférieure (lower tail dépendance coefficient) λ_L est définie par

$$\lambda_L(X, Y) = \lim_{u \rightarrow 0^+} P(X \leq F^{-1}(u) | Y \leq G^{-1}(u)).$$

Le coefficient de dépendance supérieure (upper tail dépendance coefficient) λ_U est définie par

$$\lambda_U(X, Y) = \lim_{u \rightarrow 1^-} P(X > F^{-1}(u) | Y > G^{-1}(u)).$$

On peut définir ces mesures en fonction d'une copule C de la façon suivante.

Définition 2.4.4 Soient X, Y deux variables aléatoires continues de copule C , alors nous avons

$$\lambda_L(X, Y) = \lim_{u \rightarrow 0^+} \frac{C(u, u)}{u}.$$

- Quand $\lambda_L \in]0, 1]$, alors C a une dépendance de queue inférieure.
- Quand $\lambda_L = 0$, alors C n'a pas de dépendance de queue inférieure.

Et

$$\lambda_U(X, Y) = \lim_{u \rightarrow 1^-} \frac{1 - 2u + C(u, u)}{1 - u}.$$

- Quand $\lambda_U \in]0, 1]$ alors C a une dépendance de queue supérieure.
- Quand $\lambda_U = 0$ alors C n'a pas de dépendance de queue supérieure.

Nous présentons les coefficients de dépendance de queues de quelques copules, si elles existent dans le tableau suivant.

copule $C_\theta(u, v)$	λ_L	λ_U
Gumbel	0	$2 - 2^{-\frac{1}{\theta}}$
Clayton	$2^{-\frac{1}{\theta}}$	0
Frank	0	0
Gausienne	0	0
F-G-M	0	0

TABLE 2.3 – Dépendance de queue inférieure et supérieure de quelques copules.

Retrouvons λ_L de la copule de Clayton et λ_U de la copule de Gumbel.

– λ_U de la copule de Gumbel

Nous savons que $C_\theta^{GH}(u, u) = u^{2\frac{1}{\theta}}$ donc

$$\lambda_U = \lim_{u \rightarrow 1^-} \frac{1 - 2u + u^{2\frac{1}{\theta}}}{1 - u} = 2 - \lim_{u \rightarrow 1^-} \frac{u^{2\frac{1}{\theta}} - 1}{1 - u}$$

et nous appliquons ensuite la règle de l'Hospital pour obtenir

$$\lambda_U = 2 - \lim_{u \rightarrow 1^-} \frac{\partial(u^{2\frac{1}{\theta}})}{u} = 2 - 2^{\frac{1}{\theta}}.$$

– λ_L de la copule de Clayton

Nous savons que $C_\theta^{Cl}(u, u) = (2u^{-\theta} - 1)^{-\frac{1}{\theta}}$ donc

$$\begin{aligned} \lambda_L &= \lim_{u \rightarrow 0^+} \frac{(2u^{-\theta} - 1)^{-\frac{1}{\theta}}}{u} = \lim_{u \rightarrow 0^+} \frac{1}{u(2u^{-\theta} - 1)^{\frac{1}{\theta}}} \\ &= \lim_{u \rightarrow 0^+} \frac{1}{(2 - u^\theta)^{\frac{1}{\theta}}} = 2^{-\frac{1}{\theta}}. \end{aligned}$$

La copule de Clayton a une dépendance de queue inférieure, contrairement à la copule de Gumbel qui possède une dépendance de queue supérieure. Cette dernière est particulièrement adaptée en assurance et en finance pour étudier l'impact de la survenance d'événement de forte intensité sur la dépendance entre branches d'assurance ou actifs financiers (voir [3]). Pour la copule de Frank il n'existe aucune dépendance de queue ni inférieure ni supérieure, de même que la copule gaussienne (sauf si $\rho = 1$ où $\lambda_L = \lambda_U = 1$).

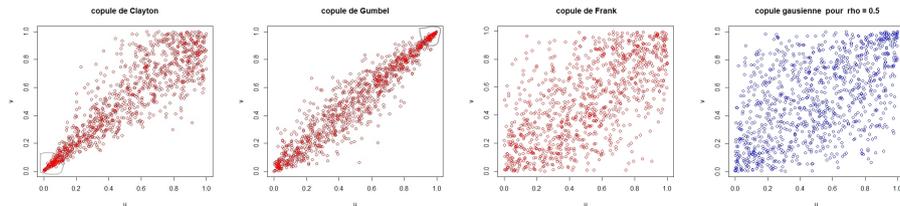


FIGURE 2.8 – Echantillons de taille 1000 de trois copules archimédiennes ($\theta = 5$) et de d'une copule gaussienne.

2.4.3 Tau de Kendall et rho de Spearman

Le tau de Kendall et le rho de Spearman sont deux mesures de concordance bien connues en statistique. Elles donnent une mesure de la corrélation entre les rangs des observations, à la différence du coefficient de corrélation linéaire qui lui mesure la corrélation entre les valeurs des observations. Elles offrent par ailleurs l'avantage de s'exprimer simplement en fonction de la copule associée au couple de variables aléatoires. Pour introduire le tau de Kendall et le rho de Spearman, on donne d'abord une définition de la notion de concordance.

Concept de concordance

Définition 2.4.5 Soit $\{(x_1, y_1), \dots, (x_n, y_n)\}$ un échantillon de n observations d'un couple (X, Y) . Il existe $C_n^2 = n!/2(n-2)!$ paires de distributions distinctes de couples (x_i, y_i) et (x_j, y_j) qui sont dites :

- concordantes si

$$(x_i - x_j)(y_i - y_j) > 0,$$

- discordantes si

$$(x_i - x_j)(y_i - y_j) < 0.$$

Exemple 2.4.6 Soit un échantillon de taille $n = 4$ de données bivariées $\{(x_i, y_i)\}_{i=1}^4$. Les observations x_i et y_i sont présentés dans le tableau suivant

x_i	2	-5	-6	4
y_i	1	2	3	4

À partir de cet échantillon nous donnons les paires concordantes et discordantes :

$(x_i - x_j)(y_i - y_j) > 0$	$(x_1 - x_4)(y_1 - y_4)$	$(x_2 - x_4)(y_2 - y_4)$	$(x_3 - x_4)(y_3 - y_4)$
$(x_i - x_j)(y_i - y_j) < 0$	$(x_1 - x_2)(y_1 - y_2)$	$(x_1 - x_3)(y_1 - y_3)$	$(x_2 - x_3)(y_2 - y_3)$

Définition 2.4.7 Une mesure d'association entre deux variables aléatoires X et Y , notée $k_{X,Y}$ de copule C est une mesure de concordance si elle vérifie les propriétés suivantes :

1. $k_{X,Y}$ est définie pour chaque couple (X, Y) de variables aléatoires continues,
2. $-1 \leq k_{X,Y} \leq 1, k_{X,X} = 1$ et $k_{X,-X} = -1$,
3. $k_{X,Y} = k_{Y,X}$,
4. si X et Y sont indépendantes, alors $k_{X,Y} = 0$,
5. si C_1 et C_2 sont deux copules telles que $C_1 < C_2$, alors $k_{C_1} \leq k_{C_2}$,
6. $k_{-X,Y} = k_{X,-Y} = -k_{X,Y}$,
7. si (X_n, Y_n) est une suite de variables aléatoires continues de copule C_n et C_n converge vers C , alors $\lim_{n \rightarrow +\infty} k_{C_n} = k_C$,
8. si $\alpha(X)$ et $\beta(Y)$ sont des fonctions strictement croissantes, alors

$$k_{\alpha(X),\beta(Y)} = k_{X,Y}.$$

Définition 2.4.8 (Fonction de concordance) La fonction de concordance est la différence entre la probabilité de concordance et celle de discordance entre deux couples (X_1, Y_1) et (X_2, Y_2) . Elle est donnée par

$$Q := P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

Comme les variables aléatoires sont continues, donc

$$\begin{aligned} Q &= P[(X_1 - X_2)(Y_1 - Y_2) > 0] - (1 - P[(X_1 - X_2)(Y_1 - Y_2) > 0]) \\ &= 2P[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1, \text{ avec} \end{aligned}$$

$$P[(X_1 - X_2)(Y_1 - Y_2) > 0] = P[X_1 > X_2, Y_1 > Y_2] + P[X_1 < X_2, Y_1 < Y_2].$$

Le théorème suivant établit une relation entre la fonction de concordance Q et les copules associées aux deux couples.

Théorème 2.4.9 Soient (X_1, Y_1) et (X_2, Y_2) deux couples aléatoires continus, indépendants et dont les fonctions de répartition jointes sont H_1 et H_2 respectivement. Soient F et G les marges associées à X_1, X_2 et à Y_1, Y_2 respectivement. Soient C_1 et C_2 les copules associées à H_1 et H_2 données par

$$H_1(x, y) = C_1(F(x), G(y)),$$

$$H_2(x, y) = C_2(F(x), G(y)).$$

Alors

$$Q = Q(C_1, C_2) = 4 \int_0^1 \int_0^1 C_2(u, v) dC_1(u, v) - 1.$$

Preuve Pour commencer on calcule d'abord $P[X_1 > X_2, Y_1 > Y_2]$ puis $P[X_1 < X_2, Y_1 < Y_2]$ comme suit :

$$\begin{aligned} 1. \quad &P[X_1 > X_2, Y_1 > Y_2] = P[X_2 < X_1, Y_2 < Y_1] \\ &= \int \int_{\mathbb{R}^2} P[X_2 < x, Y_2 < y] dH_1(x, y) \\ &= \int \int_{\mathbb{R}^2} P(X_2 < x, Y_2 < y) dC_1(F(x), G(y)) \\ &= \int \int_{\mathbb{R}^2} C_2(F(x), G(y)) dC_1(F(x), G(y)), \end{aligned}$$

par un changement de variables, $u = F(x)$ et $v = G(y)$ on obtient

$$P(X_1 > X_2, Y_1 > Y_2) = \int_0^1 \int_0^1 C_2(u, v) dC_1(u, v).$$

De façon similaire on a :

$$\begin{aligned} 2. \quad &P[X_1 < X_2, Y_1 < Y_2] = P[X_2 > X_1, Y_2 > Y_1] \\ &= \int \int_{\mathbb{R}^2} P[X_2 > x, Y_2 > y] dH_1(x, y) \\ &= \int \int_{\mathbb{R}^2} \bar{H}_2(x, y) dC_1(F(x), G(y)) \\ &= \int_0^1 \int_0^1 [1 - u - v + C_2(u, v)] dC_1(u, v) \\ &= 1 - \frac{1}{2} - \frac{1}{2} + \int_0^1 \int_0^1 C_2(u, v) dC_1(u, v) \end{aligned}$$

car U, V sont des variables aléatoires uniformes, donc $E(U) = E(V) = 1/2$ et par conséquent

$$P(X_1 < X_2, Y_1 < Y_2) = \int_0^1 \int_0^1 C_2(u; v) dC_1(u, v).$$

En regroupant ces deux résultats on en déduit que :

$$\begin{aligned} Q &= 2P[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1 \\ &= 2(P[X_1 < X_2, Y_1 < Y_2] + P[X_1 > X_2, Y_1 > Y_2]) - 1 \\ &= 2\left(\int_0^1 \int_0^1 C_2(u, v) dC_1(u, v) + \int_0^1 \int_0^1 C_2(u, v) dC_1(u, v)\right) - 1 \\ &= 4 \int_0^1 \int_0^1 C_2(u, v) dC_1(u, v) - 1 \end{aligned}$$

■

Proposition 2.4.10 Soient C_1, C_2 et Q données par le théorème 2.4.9, alors Q est symétrique c'est-à-dire

$$Q(C_1, C_2) = Q(C_2, C_1).$$

Preuve Il suffit de reprendre la démonstration du théorème 2.4.9 en intervertissant les rôles de C_1 et C_2 . ■

Exemple 2.4.11 (Fonctions de concordance des copules usuelles) Les résultats suivants donnent les fonctions de concordances des copules usuelles W, Π et M prises deux à deux

$$Q(W, W) = Q(W, M) = -1; Q(\Pi, \Pi) = 0; Q(\Pi, M) = 1/3; Q(W, \Pi) = -1/3; Q(M, M) = 1.$$

En effet

1. Le support de la copule minimum M est l'ensemble

$$\mathcal{L}_M = \{u, v \in [0, 1] : u = v\}.$$

Si g est une fonction intégrable dans le domaine $[0, 1]^2$, alors

$$\int_0^1 \int_0^1 g(u, v) dM(u, v) = \int_0^1 g(u, u) du.$$

Donc : $M(u, v) = u; W(u, v) = 2u - 1; \Pi(u, v) = u^2$. Alors

$$\begin{aligned} - Q(M, M) &:= 4 \int_0^1 u du - 1 = 1. \\ - Q(M, \Pi) &:= 4 \int_0^1 u^2 du - 1 = 1/3. \\ - Q(M, W) &:= 4 \int_0^1 (2u - 1) du - 1 = -1. \end{aligned}$$

2. De même le support de la copule maximum W est l'ensemble

$$\mathcal{L}_W = \{u, v \in [0, 1] : v = 1 - u\},$$

alors

$$\int_0^1 \int_0^1 g(u, v) dW(u, v) = \int_0^1 g(u, 1 - u) du.$$

Donc : $M(u, v) = \min(u, 1 - u)$; $W(u, v) = 0$; $\Pi(u, v) = u(1 - u)$. Alors

$$- Q(W, W) := 4 \int_0^1 \int_0^1 0 du - 1 = -1.$$

$$- Q(W, \Pi) := 4 \int_0^1 \int_0^1 u(1 - u) du - 1 = -1/3.$$

3. En fin puisque $d\Pi(u, v) = dudv$, alors :

$$Q(\Pi, \Pi) = 4 \int_0^1 \int_0^1 uv dudv - 1 = 0.$$

Le tau de Kendall (τ)

Définition 2.4.12 Soit (X, Y) un couple aléatoire continu ayant comme loi jointe H . Le tau de Kendall de ce couple est la différence entre les probabilités de concordance et de discordance d'une paire (X_1, Y_1) et (X_2, Y_2) de points telle que

$$\tau(X, Y) = P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_2) < 0\}.$$

On peut définir le tau de Kendall en fonction d'une copule C , en utilisant la fonction Q définie dans le théorème 2.4.9. Le théorème suivant représente la relation entre le tau de Kendall et la copule.

Théorème 2.4.13 Soient (X_1, Y_1) et (X_2, Y_2) deux couple aléatoires continus, indépendants et identiquement distribués de fonction de répartition jointe H . Soit C la copule associée à H . Le tau de Kendall du couple (X, Y) a pour expression :

$$\begin{aligned} \tau(X, Y) &= Q(C, C) = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1 \\ &= 4E[C(U, V)] - 1, \end{aligned}$$

car les variables aléatoires $U = F(x)$ et $V = G(y)$ sont des variables aléatoires uniformes.

Preuve Voir la preuve du théorème 2.4.9. ■

Définition 2.4.14 (Tau de Kendall empirique) Soit une série de n observations $\{(x_i, y_i)\}_{1 \leq i \leq n}$ d'un couple (X, Y) . La version empirique du tau de Kendall est définie par :

$$\hat{\tau}_n = \frac{[\text{Nombre de paires concordantes}] - [\text{Nombre de paires discordantes}]}{\text{Nombre total de paires}}$$

comme le nombre total de paires est $C_n^2 = n!/2(n-2)!$ donc le tau de kendall empirique peut s'écrire :

$$\hat{\tau}_n = \frac{4Q'_n}{n(n-1)} - 1$$

où Q'_n est le nombre de paires concordantes.

Exemple 2.4.15 En reprenant l'exemple 2.4.6 on trouve $\hat{\tau}_n = \frac{4 \times 3}{4(4-1)} - 1 = 0$

Remarque 2.4.16 Le tau de Kendall de X et Y d'une copule archimédienne s'obtient simplement à l'aide du générateur de la copule (voir [18]), selon la formule ci-dessous :

$$\tau_C = 1 + 4 \int_0^1 \frac{\Phi(t)}{\Phi'(t)} dt.$$

Exemple 2.4.17 Nous pouvons calculer le tau de Kendall pour la copule de Clayton. On a : $\frac{\Phi(t)}{\Phi'(t)} = \frac{t^{\theta+1} - t}{\theta}$,

$$\begin{aligned} \text{donc } \tau_\theta &= 1 + 4 \int_0^1 \frac{t^{\theta+1} - t}{\theta} dt \\ &= 1 + \frac{4}{\theta} \left(\frac{1}{\theta+2} - \frac{1}{2} \right) = \frac{\theta}{\theta+2}. \end{aligned}$$

Le rho de Spearman (ρ_s)

Définition 2.4.18 Le rho de Spearman de deux variables aléatoires X et Y est égal au coefficient de corrélation entre les variables $F(X)$ et $G(Y)$ distribuées selon la loi uniforme sur $[0, 1]$

$$\rho_s(X, Y) = \rho(F(X), G(Y)).$$

Nous donnons une autre définition plus populaire du rho de Spearman.

Définition 2.4.19 Soient (X_1, Y_1) , (X_2, Y_2) et (X_3, Y_3) trois vecteurs aléatoires indépendants de même loi H . Le rho de Spearman est défini comme étant trois fois la différence de probabilité de concordance et celle de discordance des couples aléatoires (X_1, Y_1) et (X_2, Y_3) . Elle s'écrit comme suit

$$\rho_s(X, Y) = 3 \left(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0] \right).$$

La distribution de (X_2, Y_3) étant Π (car les variables X_2 et Y_3 sont indépendantes) alors d'après ce qui précède on peut énoncé le théorème suivant.

Théorème 2.4.20 (Expression du rho de Spearman en terme de copule)

Si C désigne la copule des variables X et Y , le rho de Spearman ρ_s a pour expression :

$$\begin{aligned} \rho_s(C) &= 3Q(C, \Pi) = 3 \left(4 \int_0^1 \int_0^1 \Pi(u, v) dC(u, v) - 1 \right) = 12 \int_0^1 \int_0^1 uv dC(u, v) - 3 \\ &= 12 \int_0^1 \int_0^1 C(u, v) dudv - 3. \end{aligned}$$

Preuve Soient $U = F(X)$ et $V = F(Y)$, deux variables qui suivent la loi uniforme sur $[0, 1]$. Par conséquent leur espérance et leur variance sont respectivement égales à :

$$E(U) = E(V) = 1/2 \text{ et } Var(U) = Var(V) = 1/12.$$

En utilisant la définition 2.4.1,

$$\begin{aligned} \rho_s(C) = \rho(U, V) &= \frac{Cov(U, V)}{\sqrt{Var(U)Var(V)}} \\ &= 12Cov(U, V) \\ &= 12[E(UV) - E(U)E(V)] \\ &= 12\left(\int_0^1 \int_0^1 uv dC(u, v) - \frac{1}{4}\right) \\ &= 12 \int_0^1 \int_0^1 C(u, v) dudv - 3, \quad \rho_s \in [-1, 1]. \end{aligned}$$

■

Exemple 2.4.21 Soit C une copule de Farlie-Gumbel-Morgenstern, de paramètre θ avec $\theta \in [-1, 1]$ alors :

$$C_\theta^{FGH}(u, v) = uv + \theta uv(1-u)(1-v).$$

$$\begin{aligned} \int_0^1 \int_0^1 C_\theta(u, v) dudv &= \int_0^1 \int_0^1 (uv + \theta uv(1-u)(1-v)) dudv \\ &= \frac{1}{4} + \frac{\theta}{36}. \end{aligned}$$

Le rho de Spearman est : $\rho_s(\theta) = \frac{\theta}{3}$.

Définition 2.4.22 (Version empirique de ρ_s) Soit une série de n observations $\{(x_i, y_i)\}_{1 \leq i \leq n}$ d'un couple (X, Y) . La version empirique du rho de Spearman est définie par

$$\hat{\rho}_n = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (R_i - S_i)^2,$$

où R_i est le rang de l'observation X_i parmi X_1, \dots, X_n et S_i est le rang de l'observation Y_i parmi Y_1, \dots, Y_n .

Exemple 2.4.23 calculons $\hat{\rho}_n$ de l'exemple 2.4.6

R_i	3	1	2	4
S_i	1	2	3	4

$$\text{alors } \hat{\rho}_n = 1 - \frac{6}{4(16 - 1)} \sum_{i=1}^4 (R_i - S_i)^2 = 0.4.$$

Le tableau 2.4 présente le tau de Kendall et le rho de Spearman de quelques copules classiques.

copule	tau de Kendall τ_c	rho de Spearman ρ_s^C
Clayton	$\theta/\theta + 2$	$12I(\theta) - 3$
Gumbel	$(\theta - 1)/\theta$	$12J(\theta) - 3$
Frank	$1 - [4(1 - D_1(\theta))/\theta]$	$1 - [12(D_1(\theta) - D_2(\theta))/\theta]$
Gaussienne	$[2 \arcsin(\rho)]/\pi$	$[6 \arcsin(\rho/2)]/\pi$
F-G-M	$2\theta/9$	$\theta/3$

TABLE 2.4 – Le tau de Kendall et le rho de Spearman de quelques copules classiques.

$$I(\theta) = \int_0^1 \int_0^1 (u^{-\theta} + v^\theta - 1)^{-\frac{1}{\theta}} dudv$$

et

$$J(\theta) = \int_0^1 \int_0^1 \exp[-\{(\log u)^\theta + (-\log v)^\theta\}^{\frac{1}{\theta}}].$$

La fonction de Debye $D_k(\theta)$ est égale à :

$$D_k(\theta) = \frac{k}{\theta^k} \int_0^\theta \frac{t^k}{\exp(t) - 1} dt,$$

et ρ est le coefficient de corrélation.

Chapitre 3

Estimation des copules

Comme les copules sont des fonctions de répartition paramétriques, alors on a besoin d'estimer ces paramètres. Dans ce chapitre, nous allons présenter quelques méthodes d'estimation de copules. Il existe plusieurs méthodes bien connues pour l'estimation d'une copule. Elles sont divisées en approches paramétrique, semi-paramétrique et non-paramétrique.

3.1 Estimation paramétrique

Lorsqu'il s'agit d'une estimation paramétrique d'une copule, on impose un modèle paramétrique pour la copule et pour les distributions marginales. Pour estimer ces paramètres il existe différentes méthodes. Habituellement on utilise la méthode du maximum de vraisemblance exacte pour obtenir des estimateurs des paramètres (voir [31], [24] et [23]). En dehors de cette méthode nous utiliserons d'autres méthodes comme la méthode d'inférence sur les marges et la méthode des moments.

3.1.1 La méthode du maximum de vraisemblance exacte (Full Maximum Likelihood (FML))

Supposons que nous ayons un échantillon $\{(X_i, Y_i)\}_{i=1}^n$ généré à partir d'une distribution H . Les fonctions de distributions marginales sont $F(x; \beta_1)$ et $G(y; \beta_2)$, où β_1 (respectivement β_2) est un paramètre ou un vecteur de paramètres. Les fonctions densités marginales seront notées $f(x; \beta_1)$ et $g(y; \beta_2)$. Nous supposons aussi que la fonction de copule C appartient à une famille paramétrique, par exemple la famille elliptique ou archimédienne vue au chapitre précédent, et sera notée $C(., .; \theta)$ où θ est un paramètre ou un vecteur de paramètres. Le vecteur de paramètres à estimer est $\alpha = (\beta_1, \beta_2, \theta)^T$. D'après le théorème de Sklar, la distribution conjointe s'écrit :

$$H(x, y; \alpha) = C(F(x; \beta_1), G(y; \beta_2); \theta).$$

La fonction de densité conjointe est

$$h(x, y; \alpha) = c(F(x; \beta_1), G(y; \beta_2); \theta) f(x; \beta_1) g(y; \beta_2),$$

où c est la densité associée à la copule C .

La fonction de log-vraisemblance est donnée par

$$l(\alpha) = \sum_{i=1}^n \log c(F(x_i; \beta_1), G(y_i; \beta_2); \theta) + \sum_{i=1}^n (\log f(x_i; \beta_1) + \log g(y_i; \beta_2)).$$

En maximisant la fonction de log-vraisemblance par rapport à α , nous obtenons l'estimateur maximum de vraisemblance,

$$\hat{\alpha}_{FML} = \arg \max_{\alpha \in \Lambda} l(\alpha),$$

où Λ est l'espace des paramètres.

L'estimateur $\hat{\alpha}_{FML} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\theta})^T$ s'obtient à partir de la relation :

$$\left(\frac{\partial l}{\partial \beta_1}, \frac{\partial l}{\partial \beta_2}, \frac{\partial l}{\partial \theta} \right) = 0.$$

Sous certaines conditions de régularité, on peut montrer que l'estimateur $\hat{\alpha}_{FML}$ existe, qu'il est convergent, asymptotiquement efficace et vérifie la propriété de la normalité asymptotique :

$$\sqrt{n}(\hat{\alpha}_{FML} - \alpha_0) \xrightarrow{\text{loi}} N(0, I_n^{-1}(\alpha_0)),$$

avec $I_n^{-1}(\alpha_0)$ est l'inverse de l'information de Fisher et α_0 est l'espérance de $\hat{\alpha}_{FML}$.

La méthode du maximum de vraisemblance exacte peut engendrer des temps de calculs très longs pour une copule multivariée en grande dimension car l'estimation des paramètres des lois marginales et les paramètres de la structure de dépendance se fait d'une façon simultanée. En outre une éventuelle erreur d'estimation des marginales peut rendre erronée l'estimation de la copule car elles interviennent directement dans la fonction de log-vraisemblance.

Cependant Joe et Xu (voir [25]) ont proposé une procédure en deux étapes appelée fonction d'inférence pour les marges (IFM).

3.1.2 Inférence sur les marginales (IFM)

Joe et Xu (voir [23]), ont proposé la méthode IFM (Inference Functions for Margins) et ont estimé les paramètres séparément. Cette méthode repose sur le fait que la représentation en copule permet de séparer les paramètres spécifiques des distributions marginales (β_1, β_2) de ceux de la structure de dépendance θ , cette méthode se compose de deux étapes (voir [31]) :

1. Estimation des paramètres des marginales

$$\hat{\beta}_1 = \arg \max_{\beta_1} \sum_{i=1}^n \log f(x_i; \beta_1) \quad \text{et} \quad \hat{\beta}_2 = \arg \max_{\beta_2} \sum_{i=1}^n \log g(y_i; \beta_2).$$

2. On estime θ en tenant compte des estimateurs précédents par la maximisation de la fonction suivante :

$$l(\theta) = \sum_{i=1}^n \log c(F(x_i; \hat{\beta}_1), G(y_i; \hat{\beta}_2); \theta), \quad \text{ainsi} \quad \hat{\theta} = \arg \max_{\theta} l(\theta).$$

L'estimateur IFM est ainsi donné par

$$\hat{\alpha}_{IFM} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\theta})^T.$$

Là encore, sous certaines conditions de régularité, l'estimateur $\hat{\alpha}_{IFM}$ vérifie la propriété de normalité asymptotique démontré par Joe (voir [24])

$$\sqrt{n}(\hat{\alpha}_{IFM} - \alpha_0) \xrightarrow{\text{loi}} N(0, V^{-1}(\alpha_0)),$$

où $V(\alpha_0)$ est la matrice d'information de Godambe. Si on définit par :

$$j(\alpha) = \left(\frac{\partial l_1}{\partial \beta_1}, \frac{\partial l_2}{\partial \beta_2} \right) \quad \text{où} \quad l_1 = \sum_{i=1}^n \log f(x_i; \beta_1) \quad \text{et} \quad l_2 = \sum_{i=1}^n \log g(y_i; \beta_2),$$

alors la matrice de Godambe est donnée par :

$$V(\alpha_0) = D^{-1}M(D^{-1})^T \quad \text{où} \quad D = E\left[\frac{\partial}{\partial \alpha} j(\alpha)^T\right] \quad \text{et} \quad M = E[j(\alpha)^T j(\alpha)].$$

Cette méthode pourrait présenter l'avantage de reposer sur des calculs plus légers que ceux générés par la méthode du maximum de vraisemblance exacte. Cependant, la détermination de la matrice de Godambe peut s'avérer très complexe en dimension très grande en raison des multiples calculs de dérivées. Là encore, la méthode est sensible à une éventuelle erreur de spécification des marginales pour la même raison que celle évoquée dans la méthode précédente.

3.1.3 La méthode des moments

Cette méthode consiste à estimer séparément les paramètres β_1 et β_2 des lois marginales à celui de la copule c'est-à-dire θ . Pour estimer ce dernier, il revient à se donner une mesure de concordance k_C et à considérer que la valeur du paramètre de la copule C est celle qui égalise la valeur théorique k_C à la valeur estimée \hat{k}_C . Elle n'assure aucune robustesse de l'estimateur. Dans la pratique, le tau de Kendall est le plus souvent utilisé comme mesure de concordance étant donné la simplicité de l'estimation de ce tau. Pour cela on doit suivre les étapes suivantes :

1. Estimer les paramètres des marginales

Soit X une variable aléatoire qui suit une loi de probabilité F de paramètre β_1 (qui peut être un ensemble de plusieurs nombres réels, $\beta_1 = (b_1, b_2, \dots, b_p)$). Il est possible de calculer en fonction de β_1 les p premiers moments de X , notés $\mu_1^n, \mu_2^n, \dots, \mu_p^n$. Ces p équations permettent de calculer l'estimateur de β_1

$$\begin{cases} \mu_1^n := E(X) = f_1(b_1, \dots, b_p) \\ \mu_2^n := E(X^2) = f_2(b_1, \dots, b_p) \\ \vdots \\ \mu_p^n := E(X^p) = f_p(b_1, \dots, b_p) \end{cases} \Rightarrow \begin{cases} b_1 = \Psi_1(\mu_1^n, \mu_2^n, \dots, \mu_p^n) \\ b_2 = \Psi_2(\mu_1^n, \mu_2^n, \dots, \mu_p^n) \\ \vdots \\ b_p = \Psi_p(\mu_1^n, \mu_2^n, \dots, \mu_p^n). \end{cases}$$

L'estimateur de β_1 est obtenu en remplaçant dans ces expressions les moments théoriques par les moments empiriques de X calculés partir de l'échantillon de taille n .

$$\begin{cases} \hat{\mu}_1^n = \frac{1}{n} \sum_{i=1}^n X_i \\ \vdots \\ \hat{\mu}_p^n = \frac{1}{n} \sum_{i=1}^n X_i^p \end{cases} \Rightarrow \begin{cases} \hat{b}_1 = \Psi_1(\hat{\mu}_1^n, \hat{\mu}_2^n, \dots, \hat{\mu}_p^n) \\ \hat{b}_2 = \Psi_2(\hat{\mu}_1^n, \hat{\mu}_2^n, \dots, \hat{\mu}_p^n) \\ \vdots \\ \hat{b}_p = \Psi_p(\hat{\mu}_1^n, \hat{\mu}_2^n, \dots, \hat{\mu}_p^n). \end{cases}$$

Si les fonctions Ψ_i sont continues en $(\mu_1^n, \mu_2^n, \dots, \mu_p^n)$, alors les estimateurs obtenus sont des estimateurs convergents.
On répète cette même étape pour la deuxième marginale Y .

2. Inverser la mesure de concordance pour obtenir le paramètre de la copule

Supposons que la relation entre la mesure de concordance k_C et le paramètre θ de la copule est définie par

$$k_C = J(\theta),$$

où J est une fonction continue et dérivable, alors l'estimateur $\hat{\theta}$ du paramètre θ de la copule est défini par

$$\hat{\theta} = J^{-1}(\hat{k}_C),$$

tel que \hat{k}_C est l'estimateur empirique (non paramétrique) de k_C .

Exemple 3.1.1 (Copule de Clayton) *D'après la remarque 2.4.16, le tau de Kendall de la copule de Clayton est $\tau_\theta = \frac{\theta}{\theta + 2}$ avec $\theta \in]0, +\infty[$. Pour obtenir l'estimateur $\hat{\theta}$ du paramètre θ de la copule on procède comme suit :*

$$\tau_\theta = \frac{\theta}{\theta + 2} \Rightarrow \theta = \frac{2\tau_\theta}{1 - \tau_\theta} \quad \text{ce qui nous donne} \quad \hat{\theta} = \frac{2\hat{\tau}_n}{1 - \hat{\tau}_n}.$$

avec $\hat{\tau}_n$ l'estimateur empirique de τ_θ .

L'estimateur global $\hat{\alpha}_M$ obtenu par la méthode des moments est donné par :

$$\hat{\alpha}_M = (\hat{\beta}_1, \hat{\beta}_2, \hat{\theta})^T,$$

où $\hat{\beta}_1 = (\hat{b}_1, \hat{b}_2, \dots, \hat{b}_p)$ et $\hat{\beta}_2$ à déterminer.

Pour plus de détails sur l'estimation des paramètres de la copule par la méthode des moments le lecteur pourra se référer à [2].

Évidemment, l'intérêt de l'approche paramétrique réside dans le fait que, si le modèle ajusté pour les lois marginales est raisonnable, cette approche permet une réduction de la variabilité et une augmentation de la maniabilité du modèle. En revanche, si le modèle est mal ajusté, les résultats peuvent donner lieu à des interprétations fausses. C'est pour cette raison que l'approche semi-paramétrique est intéressante.

3.2 Estimation semi-paramétrique

L'approche semi-paramétrique suppose un modèle paramétrique pour la copule, $C = C(\cdot, \cdot; \theta)$, et non-paramétrique pour les distributions marginales. Les méthodes d'estimation semi-paramétrique que nous étudierons ici sont la méthode du pseudo-maximum de vraisemblance aussi connue sous le nom de maximum de vraisemblance canonique (Canonical Maximum Likelihood (CML)), et la méthode d'inversion d'une statistique d'association.

3.2.1 La méthode de pseudo-maximum de vraisemblance (maximum de vraisemblance canonique (CML))

Considérons un échantillon $((X_1, Y_1), \dots, (X_n, Y_n))$ généré à partir d'une distribution H . Les fonctions de distributions marginales sont $F(x)$ et $G(y)$, leur fonction de densité marginale seront notées $f(x)$ et $g(y)$. Nous supposons aussi que la fonction copule C appartient à une famille paramétrique, et sera notée $C(\cdot, \cdot; \theta)$. Le paramètre ou le vecteur paramètre à estimer est θ . D'après le théorème de Sklar, la distribution conjointe s'écrit :

$$H(x, y; \theta) = C(F(x), G(y); \theta).$$

La fonction de densité conjointe est

$$h(x, y; \theta) = c(F(x), G(y); \theta)f(x)g(y),$$

où c est la densité associée à C .

Dans le cas où les marges sont connues, la méthode classique du maximum de vraisemblance consiste à maximiser en θ la fonction de log-vraisemblance

$$m(\theta) = \sum_{i=1}^n \log c(F(x_i), G(y_i); \theta) + K$$

où $K = \sum_{i=1}^n \log(f(x_i)g(y_i))$ ne dépend pas de θ . La solution $\hat{\theta}_{CML}$ est l'estimateur du maximum de vraisemblance classique de θ .

Bien que cette méthode soit applicable pour des marges connues, elle a été proposée dans le cas où **les marges sont inconnues**. Pour ce cas, la méthode comporte deux étapes :

1. on remplace les marges F et G par leurs estimateurs naturels (estimateurs empiriques), ils sont définies par :

$$\hat{F}_n(x_i) = \frac{1}{n+1} \sum_{j=1}^n 1_{\{X_j \leq x_i\}} \text{ et } \hat{G}_n(y_i) = \frac{1}{n+1} \sum_{j=1}^n 1_{\{Y_j \leq y_i\}}.$$

Le choix de diviser par $n+1$ plutôt que par n , permet éviter le problème d'absence de limites de la fonction de densité de la copule.

2. Dans le cadre de recherche d'un estimateur issu de la méthode du maximum de vraisemblance canonique, Genest et collaborateurs (voir [15]) ont proposé de maximiser par la pseudo log-vraisemblance

$$l(\theta) = \sum_{i=1}^n \log c(\hat{F}_n(x_i), \hat{G}_n(y_i); \theta).$$

L'estimateur $\hat{\theta}_{CML}$ est donné par :

$$\hat{\theta}_{CML} = \arg \max l(\theta).$$

En remarquant que $(n+1)\hat{F}_n = R_i$ et $(n+1)\hat{G}_n = S_i$, où R_i et S_i sont les rangs de X_i et Y_i dans leur échantillon univarié respectif, donc $l(\theta)$ devient

$$l(\theta) = \sum_{i=1}^n \log c\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}; \theta\right).$$

Cette méthode présente le grand avantage de procéder à une estimation paramétrique de la copule totalement indépendante de la spécification des lois marginales. En outre, elle génère des temps de calculs limités. Ce sont deux atouts majeurs qui la rendent très attractive.

Kim et collaborateurs (voir [26]) ont comparé les méthodes IFM, FML et CML pour l'estimation des copules, leur étude a montré que :

- Les méthodes IFM et FML ne sont pas robustes contre les erreurs de spécification des distributions marginales.
- La méthode CML est conceptuellement presque la même que l'IFM, mais elle surmonte sa non robustesse contre une mauvaise spécification des distributions marginales.
- En termes de calculs statistiques et d'analyse des données, la CML est aussi facile à implémenter que la méthode IFM.
- Un avantage de plus pour la CML sur l'IFM, est que la CML n'impose pas de modèles paramétriques pour les distributions marginales.
- Leurs résultats de simulation montrent que l'estimateur de CML est meilleur que ceux de IFM et FML dans la plupart des situations pratiques.

3.2.2 Méthode d'inversion d'une statistique d'association

Cette méthode consiste à estimer le paramètre de la copule en utilisant certaines mesures d'association telles que le tau de Kendall ou le rho de Spearman, car il existe une relation entre ces mesures et le paramètre de dépendance de la copule.

Soit (X, Y) un couple de variables aléatoires dont la copule est $C(., .; \theta)$ de paramètre θ , tel que $\theta \in \Theta \subset \mathbb{R}$.

1. Estimation basée sur le tau de Kendall

Supposons que la relation entre le tau de Kendall et le paramètre θ est définie par l'égalité suivante

$$\tau(X, Y) = g(\theta),$$

puisque le tau de Kendall s'écrit en fonction de la copule $C(., .; \theta)$, donc il s'écrit également en fonction du paramètre θ , comme suit :

$$\tau(X, Y) = g(\theta) = 4 \int_0^1 \int_0^1 C(u, v; \theta) dC(u, v; \theta) - 1,$$

où g est une fonction continue et dérivable. Un estimateur $\hat{\theta}_{TK}$ de θ est défini par :

$$\hat{\theta}_{TK} = g^{-1}(\hat{\tau}_n),$$

tel que $\hat{\tau}_n$ est l'estimateur empirique de τ . Cet estimateur $\hat{\theta}_{TK}$ est asymptotiquement normale

$$\sqrt{n}(\hat{\theta}_{TK} - \theta) \xrightarrow{\text{loi}} N(0, \hat{\sigma}_\tau^2)$$

où $\hat{\sigma}_\tau^2 = \sigma_\tau^2 / \{g'(\hat{\tau}_n)\}^2$ avec $\sigma_\tau^2 = V(\tau)$ car $\hat{\tau}_n$ est asymptotiquement normale de moyenne nulle et de variance $\sigma_\tau^2 = V(4C(u, v) - 1)$. Une étude de l'efficacité de cette estimation basée sur le tau de Kendall est présentée dans [14].

2. Estimation basée sur le rho de Spearman

De la même manière que le tau de Kendall, le rho de Spearman s'écrit en fonction de la copule $C(\cdot, \cdot; \theta)$, donc il s'écrit en fonction du paramètre de la copule. Leur relation est définie par :

$$\rho_s(X, Y) = f(\theta) = 12 \int_0^1 \int_0^1 C(u, v; \theta) dudv - 3,$$

où f est une fonction continue et dérivable. Un estimateur $\hat{\theta}_{RS}$ de θ est défini par :

$$\hat{\theta}_{RS} = f^{-1}(\hat{\rho}_n),$$

tel que $\hat{\rho}_n$ est l'estimateur empirique de ρ_s . Cet estimateur $\hat{\theta}_{RS}$ est asymptotiquement normale de la même manière que le tau de Kendall.

Pour des détails le lecteur pourra consulter [14].

3.3 Estimation non-paramétrique

Si le modèle paramétrique de la copule est mal spécifié, les approches paramétrique et semi-paramétrique conduisent généralement à un mauvais ajustement aux données. Dans ce cas les méthodes d'estimation non paramétrique peuvent être une bonne alternative.

3.3.1 Copule empirique

La notion de copule empirique a été introduite par Deheuvels (voir [5]) sous l'appellation de fonction empirique de dépendance pour construire des tests non paramétriques d'indépendance.

Supposons que nous ayons un échantillon $((X_1, Y_1), \dots, (X_n, Y_n))$ généré à partir d'une distribution H . Nous construisons la fonction de répartition empirique comme suit :

$$\hat{H}_n(x, y) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq x, Y_i \leq y\}}, \quad -\infty < x, y < +\infty.$$

Soient \hat{F}_n et \hat{G}_n les distributions marginales qui lui sont associées :

$$\hat{F}_n(x) := \hat{H}_n(x, \infty) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq x\}} \text{ et } \hat{G}_n(y) := \hat{H}_n(\infty, y) = \frac{1}{n} \sum_{i=1}^n 1_{\{Y_i \leq y\}}.$$

En utilisant le corollaire 2.2.5, on peut construire la copule empirique \hat{C}_n en fonction des distributions empiriques \hat{F}_n, \hat{G}_n et \hat{H}_n comme ceci :

$$\hat{C}_n(u, v) := \hat{H}_n(\hat{F}_n^{-1}(u), \hat{G}_n^{-1}(v)), \quad 0 \leq u, v \leq 1.$$

Définition 3.3.1 Soit $((x_1, y_1), \dots, (x_n, y_n))$ un échantillon de taille n d'un couple de variables aléatoires (X, Y) . La copule empirique est la fonction \hat{C}_n définie par :

$$\hat{C}_n\left(\frac{j}{n}, \frac{k}{n}\right) = \frac{\text{nb de paires } (x, y) \text{ dans l'échantillon tels que } x \leq x_{(j)} \text{ et } y \leq y_{(k)}}{n}.$$

Par conséquent, la copule empirique \hat{C}_n peut s'écrire comme suit :

$$\hat{C}_n\left(\frac{j}{n}, \frac{k}{n}\right) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq x_{(j)}, Y_i \leq y_{(k)}\}}$$

telles que $x_{(j)}$ et $y_{(k)}$ représentent les statistiques de rang de l'échantillon (X_1, X_2, \dots, X_n) et (Y_1, Y_2, \dots, Y_n) respectivement.

La fonction de densité empirique de la copule \hat{C}_n parfois dite copule de fréquence empirique notée \hat{c}_n est donnée par

$$\hat{c}_n\left(\frac{j}{n}, \frac{k}{n}\right) = \begin{cases} \frac{1}{n} & \text{si } (x_{(j)}, y_{(k)}) \in \{(x_i, y_i)\}_{1 \leq i \leq n} \\ 0 & \text{sinon.} \end{cases}$$

Il existe une relation entre \hat{C}_n et \hat{c}_n donnée par : $\hat{C}_n\left(\frac{j}{n}, \frac{k}{n}\right) = \sum_{p=1}^j \sum_{q=1}^k \hat{c}_n\left(\frac{p}{n}, \frac{q}{n}\right)$

et

$$\hat{c}_n\left(\frac{j}{n}, \frac{k}{n}\right) = \hat{C}_n\left(\frac{j}{n}, \frac{k}{n}\right) - \hat{C}_n\left(\frac{j-1}{n}, \frac{k}{n}\right) - \hat{C}_n\left(\frac{j}{n}, \frac{k-1}{n}\right) + \hat{C}_n\left(\frac{j-1}{n}, \frac{k-1}{n}\right).$$

La plupart des méthodes non-paramétriques est basée sur la copule empirique. Celle-ci se base sur le rang des observations pour extraire ensuite la structure de dépendance.

3.3.2 Méthode de Genest et Rivest

Dans article de Genest et Rivest (voir [16]), une méthode originale d'identification d'une copule archimédienne a été proposée. Ils considèrent la fonction K définie par :

$$K(t) = P(C(U, V) \leq t).$$

Ils montrent que dans le cas d'une copule archimédienne de générateur Φ , $K(t)$ est :

$$K(t) = t - \Phi(t)/\Phi'(t).$$

La connaissance de K permet de spécifier complètement la copule archimédienne, puisque K contient toute "l'information" sur Φ .

Soit $((X_1, Y_1), \dots, (X_n, Y_n))$ un échantillon, C la copule associée au générateur Φ .

Un estimateur non-paramétrique de $K(t)$ est alors donné par :

$$\hat{K}_n(t) = \frac{1}{n} \sum_{i=1}^n 1_{\{T_i \leq t\}} \quad \text{où} \quad T_i = \frac{1}{n-1} \sum_{j=1}^n 1_{\{X_j \leq x_i, Y_j \leq y_i\}}.$$

La fonction K est liée au τ de Kendall par : $\tau = 4 \int_0^1 (1 - K(t))dt - 1$.

Pour plus de détails sur cette méthode, voir [16].

Remarque 3.3.2 *Il existe d'autres méthodes d'estimation non-paramétrique de la copule qui ne sont pas présentées dans ce mémoire, comme entre autres :*

- la méthode d'estimation à noyau par Fermanian et Scaillet (voir [11]),
- la méthode d'estimation de Bench-mark,
- la méthode d'estimation bayésienne,
- la méthode d'estimation à distance minimale (voir et [35]).

Chapitre 4

Simulation et application à des données réelles

4.1 Simulation des copules

Simuler une copule bivariée C signifie simuler les arguments u et v de cette fonction, tirés d'un couple aléatoire (U, V) de loi uniforme sur $[0, 1]$. Ceci permet alors de déterminer la simulation d'un couple aléatoire (X, Y) dont la structure de dépendance est définie par C et de distribution jointe H dont les marges sont F et G . Pour cela on utilise la transformation (2.6) du corollaire 2.2.5 du chapitre 2.

Toute la difficulté est de simuler des nombres aléatoires issus de C . Par exemple, la simulation de la copule indépendance est relativement simple : on simule des réalisations de loi uniforme sur $[0, 1]$ de manière indépendante. Mais pour la majorité des copules, cela s'avère plus complexe. Nous présentons dans cette section trois méthodes de simulation des copules qui sont décrites dans [13] ou [7].

4.1.1 Méthode des distributions

On suppose que l'on se trouve dans une situation où la loi jointe du couple (X, Y) est plus facile à simuler directement que la copule C , c'est par exemple le cas de la copule gaussienne (ou copule normale). Un vecteur gaussien de dimension 2 est aisé à simuler (via la décomposition de Cholesky de la matrice de variance-covariance, voir le chapitre 1, sous section 1.5.2), alors que la copule gaussienne n'est pas simple à simuler directement.

Simuler des réalisations (u, v) du couple aléatoire (U, V) de distribution C revient à :

1. simuler des réalisations (x, y) du couple aléatoire (X, Y) de distribution H ,
2. appliquer la transformation $(u, v) = (F(x), G(y))$.

C'est aussi le cas d'une copule de Student.

Exemple 4.1.1 (Simulation de la copule normale) *Il est possible de simuler une paire de variables aléatoires (U, V) ayant pour distribution jointe la copule normale. La première étape consiste à simuler une paire de variables*

aléatoires (X, Y) ayant pour distribution la loi normale bivariée $N_2(0, R)$; R étant la matrice de corrélation ayant ρ comme élément non diagonal, et bien sûr 1 sur la diagonale. En prenant ensuite $U = \Phi(X)$ et $V = \Phi(Y)$, alors (U, V) aura pour distribution la copule normale.

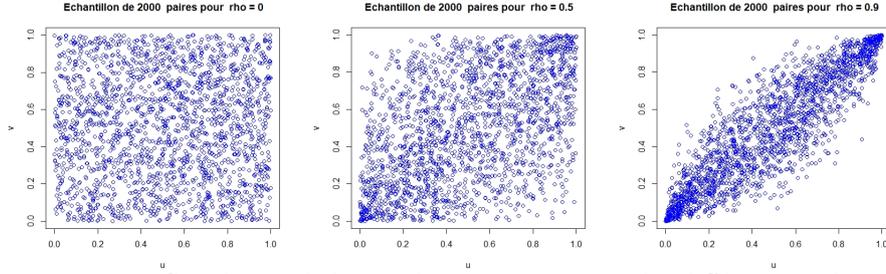


FIGURE 4.1 – Simulation de la copule gaussienne pour des différentes valeurs de ρ .

4.1.2 Méthode des distributions conditionnelles

Cette méthode consiste à simuler directement les marges (U_1, U_2) de la copule. L'algorithme de cette méthode est le suivant :

1. simuler deux variables aléatoires indépendantes U_1 et V_2 , uniformes sur $[0, 1]$, soient u_1 et v_2 des nombres simulés,
2. déterminer la distribution conditionnelle de la copule C de U_2 sachant $U_1 = u_1$, cette dernière se définit de la manière suivante :

$$\begin{aligned}
 C_{U_2|U_1}(u_1, u_2) &= P(U_2 \leq u_2 | U_1 = u_1) \\
 &= \lim_{\Delta u \rightarrow 0^+} \frac{P(U_2 \leq u_2, u_1 \leq U_1 \leq u_1 + \Delta u)}{P(u_1 \leq U_1 \leq u_1 + \Delta u)} \\
 &= \lim_{\Delta u \rightarrow 0^+} \frac{P(U_1 \leq u_1 + \Delta u, U_2 \leq u_2) - P(U_2 \leq u_2, U_1 \leq u_1)}{P(U_1 \leq u_1 + \Delta u) - P(U_1 \leq u_1)} \\
 &= \lim_{\Delta u \rightarrow 0^+} \frac{C(u_1 + \Delta u, u_2) - C(u_1, u_2)}{C(u_1 + \Delta u, 1) - C(u_1, 1)} \\
 &= \lim_{\Delta u \rightarrow 0^+} \frac{C(u_1 + \Delta u, u_2) - C(u_1, u_2)}{\Delta u - 0} \\
 &= \frac{\partial C(u_1, u_2)}{\partial u_1},
 \end{aligned}$$

3. inverser la distribution conditionnelle

$$C_{U_2|U_1}^{-1}(u, u_1) = \{u_2 : C_{U_2|U_1}(u_1, u_2) = u\},$$

4. remplacer u par v_2 pour obtenir l'expression de u_2

$$u_2 = C_{U_2|U_1}^{-1}(v_2, u_1).$$

On obtient alors (U_1, U_2) les marges de la copule.

Nous prenons comme exemples la copule de Frank et la copule de Gumbel.

Exemple 4.1.2 (Copule de Frank) Dans le cas de la copule de Frank, la méthode des distributions conditionnelles peut être mise en œuvre simplement. L'expression de cette copule est donnée par :

$$C(u_1, u_2) = -\frac{1}{\vartheta} \ln \left\{ 1 + \frac{(\exp(-\vartheta u_1) - 1)(\exp(-\vartheta u_2) - 1)}{\exp(-\vartheta) - 1} \right\}; \vartheta \in \mathbb{R}^*.$$

On en déduit que

$$C_{U_2|U_1}(u_1, u_2) = \frac{\partial C(u_1, u_2)}{\partial u_1} = \frac{\exp(-\vartheta u_1)(\exp(-\vartheta u_2) - 1)}{\exp(-\vartheta) - 1 + (\exp(-\vartheta u_1) - 1)(\exp(-\vartheta u_2) - 1)}.$$

On inverse alors $C_{U_2|U_1}(u_1, u_2)$ en résolvant l'équation $C_{U_2|U_1}(u_1, u_2) = u$ en u_2 , pour obtenir

$$C_{U_2|U_1}^{-1}(u, u_1) = -\frac{1}{\vartheta} \ln \left\{ 1 + \frac{u(\exp(-\vartheta) - 1)}{u + (1 - u)\exp(-\vartheta u_1)} \right\}.$$

Donc

$$u_2 = C_{U_2|U_1}^{-1}(v_2, u_1) = -\frac{1}{\vartheta} \ln \left\{ 1 + \frac{v_2(\exp(-\vartheta) - 1)}{v_2 + (1 - v_2)\exp(-\vartheta u_1)} \right\}.$$

La simulation peut donc être mise en œuvre simplement.

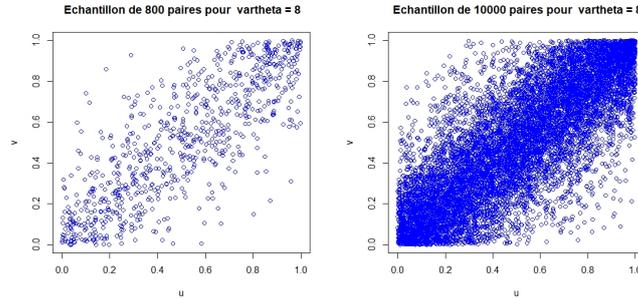


FIGURE 4.2 – Simulation de la copule de Frank pour des échantillons différents.

Exemple 4.1.3 (Copule de Gumbel) L'expression de cette copule est

$$C(u_1, u_2) = \exp \left[- \left\{ (-\log u_1)^\theta + (-\log u_2)^\theta \right\}^{\frac{1}{\theta}} \right], \theta \geq 1.$$

On en déduit que

$$\begin{aligned} C_{U_2|U_1}(u_1, u_2) &= \frac{\partial \exp \left[- \left\{ (-\log u_1)^\theta + (-\log u_2)^\theta \right\}^{\frac{1}{\theta}} \right]}{\partial u_1} \\ &= \frac{1}{u_1} \left[1 + \left(\frac{\ln u_2}{\ln u_1} \right)^\theta \right]^{-1 - \frac{1}{\theta}} \exp \left[- \left\{ (-\log u_1)^\theta + (-\log u_2)^\theta \right\}^{\frac{1}{\theta}} \right]. \end{aligned}$$

L'inverse de cette expression n'est pas aisé à calculer, et on fera ici recours à des méthodes numériques.

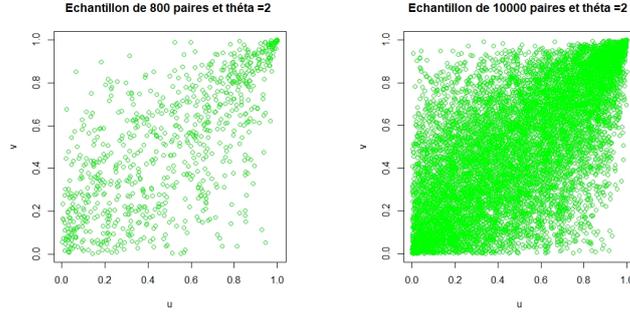


FIGURE 4.3 – Simulation de la copule de Gumbel pour des échantillons différents.

4.1.3 Cas particulier des copules archimédiennes

Dans le cas particulier d'une copule archimédienne, la distribution conditionnelle s'exprime à l'aide du générateur de la copule. Cette méthode de simulation de copules archimédiennes est introduite par Genest et Mackay (voir [18]). Si C est une copule archimédienne de générateur Φ , on a vu que C s'écrit sous la forme

$$C_{\Phi}(u_1, u_2) = \Phi^{-1}\{\Phi(u_1) + \Phi(u_2)\}, \quad u_1, u_2 \in [0, 1].$$

Par définition même de $C_{U_2|U_1}(u_1, u_2)$, on a

$$C_{U_2|U_1}(u_1, u_2) = \frac{\partial}{\partial u_1} C_{\Phi}(u_1, u_2) = \frac{\Phi'(u_1)}{\Phi'[\Phi^{-1}\{\Phi(u_1) + \Phi(u_2)\}]}.$$

Ainsi, l'inverse de $C_{U_2|U_1}(u, u_1)$ est donné en tout $u \in [0, 1]$ par

$$C_{U_2|U_1}^{-1}(u, u_1) = \Phi^{-1}\left[\Phi\left[(\Phi')^{-1}\left\{\frac{\Phi'(u_1)}{u}\right\}\right] - \Phi(u_1)\right].$$

Pour simuler les marges (U_1, U_2) d'une copule archimédienne C_{Φ} , on procède comme suit :

1. on génère deux variables aléatoires de façon indépendantes U_1, V_2 qui sont uniformes sur l'intervalle $[0, 1]$, soient u_1 et v_2 des nombres simulés,
2. on calcule $C_{U_2|U_1}^{-1}(u, u_1) = \Phi^{-1}\left[\Phi\left[(\Phi')^{-1}\left\{\frac{\Phi'(u_1)}{u}\right\}\right] - \Phi(u_1)\right]$,
3. ainsi pour $u = v_2$, on calcule u_2 à partir de l'expression

$$u_2 = C_{U_2|U_1}^{-1}(v_2, v_1) = \Phi^{-1}\left[\Phi\left[(\Phi')^{-1}\left\{\frac{\Phi'(v_1)}{v_2}\right\}\right] - \Phi(v_1)\right].$$

Le couple (U_1, U_2) est alors de loi C_{Φ} . Cette procédure fonctionne bien pour les copules de Clayton et de Frank, mais pour la copule de Gumbel, il n'y a pas une formule analytique pour $(\Phi')^{-1}$.

Exemple 4.1.4 (copule de Clayton) L'expression de la copule de Clayton bidimensionnelle s'écrit donc sous la forme :

$$C_{\theta}^{Cl}(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}.$$

Le générateur de la copule de Clayton est défini, pour $\theta > 0$ et $t \in]0, 1]$, par :

$$\Phi_{\theta}(t) = \theta^{-1}(t^{-\theta} - 1).$$

Les expressions $\Phi_\theta^{-1}(t)$, $\Phi'_\theta(t)$ et $(\Phi'_\theta)^{-1}(t)$ sont données par :

$$\Phi_\theta^{-1}(t) = \left(\frac{1}{\theta t + 1}\right)^{\frac{1}{\theta}}, \Phi'_\theta(t) = -t^{-\theta-1} \text{ et } (\Phi'_\theta)^{-1}(t) = \left(-\frac{1}{t}\right)^{\frac{1}{\theta+1}}.$$

L'algorithme de simulation pour une copule de Clayton de paramètre $\theta > 0$ est le suivant :

1. on génère deux variables aléatoires indépendantes U_1, V_2 qui sont uniformes sur l'intervalle $[0, 1]$, soient u_1 et v_2 des nombres simulés,
2. l'inverse $C_{U_2|U_1}^{-1}(u, u_1)$ est donné par

$$C_{U_2|U_1}^{-1}(u, u_1) = \left[u_1^{-\theta} u^{-\frac{\theta}{\theta+1}} - u_1^{-\theta} + 1 \right]^{-\frac{1}{\theta}},$$

3. pour $u = v_2$, on obtient $u_2 = \left[u_1^{-\theta} v_2^{-\frac{\theta}{\theta+1}} - u_1^{-\theta} + 1 \right]^{-\frac{1}{\theta}}$.

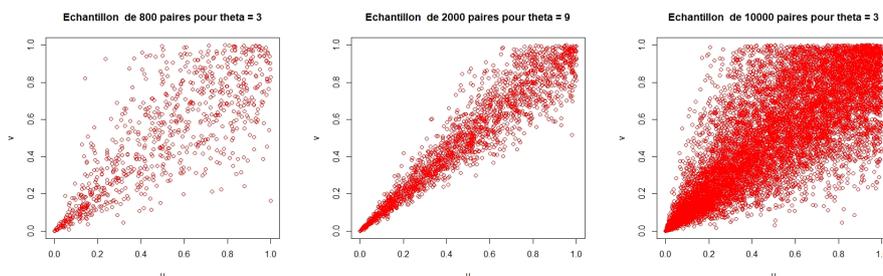


FIGURE 4.4 – Simulation de la copule de Clayton pour des échantillons de tailles différentes et des valeurs de θ différentes.

4.1.4 Méthode analytique

Une méthode analytique est une méthode de simulation spécifique à chaque copule.

Exemple 4.1.5 (Copule de Clayton) On choisit de présenter ici la méthode analytique de simulation d'une copule de Clayton proposée par Devroye (voir [6]). L'expression de la copule de Clayton est

$$C_\theta^{Cl}(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}.$$

Cette méthode revient à :

1. simuler deux variables aléatoires S et T indépendantes et uniformes sur $[0, 1]$, soient s et t des nombres simulés à partir desquelles on détermine $x = -\ln(s)$ et $y = -\ln(t)$,
2. simuler une variable aléatoire Z de loi Gamma de paramètres 1 et θ : $Z \rightsquigarrow \Gamma(1, \theta)$, soit z le nombre simulé,
3. déterminer les réalisations de la copule de Clayton à partir des expressions suivantes :

$$u = \left(1 + \frac{x}{z}\right)^{-\theta} \text{ et } v = \left(1 + \frac{y}{z}\right)^{-\theta}.$$

Exemple 4.1.6 (copule de Marshall-Olikn) L'expression de la copule de Marshall-Olikn est

$$C_{\theta}^{MO}(u, v) = u^{(1-\theta_1)}v^{(1-\theta_2)} \min(u^{\theta_1}, v^{\theta_2}); \quad \theta = (\theta_1, \theta_2)$$

de générateur

$$A_{\theta}(t) = \max(1 - \theta_1 t, \theta_2(1 - t)).$$

D'après Devroye (voir [6]), il existe deux méthodes.

Algorithme 1 :

1. Générer trois variables aléatoires indépendantes uniformes P, W et R sur $[0, 1]$, soient p, w et r des nombres simulés.
2. Soient

$$x = \min\left(\frac{-\ln p}{\lambda_1}, \frac{\ln r}{\lambda_{12}}\right) \quad \text{et} \quad y = \min\left(\frac{-\ln w}{\lambda_2}, \frac{\ln r}{\lambda_{12}}\right); \quad \text{avec } \lambda_1, \lambda_2 \text{ et } \lambda_{12} \geq 0.$$

3. La paire souhaitée simulée (u, v) est obtenue à partir de

$$u = \exp\{-(\lambda_1 + \lambda_{12})x\} \quad \text{et} \quad v = \exp\{-(\lambda_2 + \lambda_{12})y\}.$$

Algorithme 2 :

1. Générer des variables indépendantes exponentielles

$$Z_1 \rightsquigarrow \mathcal{E}(\lambda_1); Z_2 \rightsquigarrow \mathcal{E}(\lambda_2) \quad \text{et} \quad Z_{12} \rightsquigarrow \mathcal{E}(\lambda_{12}); \quad \text{avec } \lambda_1, \lambda_2 \text{ et } \lambda_{12} \geq 0.$$

2. Soient

$$U = 1 - \exp\{-(\lambda_1 + \lambda_{12}) \max(Z_1, Z_{12})\} \quad \text{et} \quad V = 1 - \exp\{-(\lambda_2 + \lambda_{12}) \max(Z_2, Z_{12})\}.$$

3. Soient

$$a = \frac{\lambda_{12}}{\lambda_1 + \lambda_{12}} \quad \text{et} \quad b = \frac{\lambda_{12}}{\lambda_2 + \lambda_{12}}.$$

Dans la figure 4.5, on présente les nuages de points de la copule de Marshall-Olikn pour des échantillons de tailles $n = 1000$ et $n = 10000$.

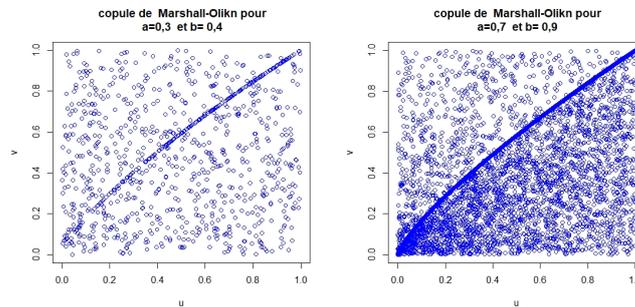


FIGURE 4.5 – Simulation de la copule de Marshall-Olikn.

4.2 Le choix de la bonne copule.

Au vu de toute la théorie exposée précédemment, nous sommes en droit de nous demander quelle copule pourrait le mieux correspondre à une certaine série de données. Nous allons donc, dans cette section, donner une méthode permettant de faire le tri parmi les copules afin de ne choisir que celles qui pourraient présenter des caractéristiques semblables à celles de notre série de données. En effet, il existe plusieurs fonctions, dont certaines ont été exposées par Mr Gary Venter dans son article “Tails of Copulas” et qui auront des caractéristiques tout à fait différentes selon la copule choisie. Ces fonctions pouvant également être établies de façon empirique, uniquement à partir du rang de chaque observation, une simple comparaison graphique nous permettra de ne retenir qu’une ou deux copules pour la poursuite de notre étude. Dans cette partie nous exposerons seulement la fonction $K(z)$ et pour les autres fonctions le lecteur pourra se référer à [1].

Fonction $K(z)$

Cette fonction, qui a déjà été évoquée dans notre présentation de l’estimation par la méthode de Genest et Rivest, n’est rien d’autre que la fonction de répartition de la variable aléatoire $C(U, V)$. Il a été démontré que pour une copule de type archimédien, cette fonction se définissait comme suit :

$$K(z) = z - \frac{\Phi(z)}{\Phi'(z)}.$$

Dans le cadre des copules archimédiennes présentées auparavant, cette fonction $K(z)$ est donc la suivante :

Gumbel : $K_\theta(z) = z(1 - \frac{1}{\theta} \ln(z))$, **Frank** : $K_\theta(z) = z + \frac{1}{\theta} \ln \left[\frac{1 - \exp(-\theta z)}{1 - \exp(-\theta)} \right]$

et **Clayton** : $K_\theta(z) = z + \theta z(1 - z^{\frac{1}{\theta}})$.

Les graphiques de ces différentes fonctions et pour différentes valeurs du tau de Kendall sont présentées ci-après.

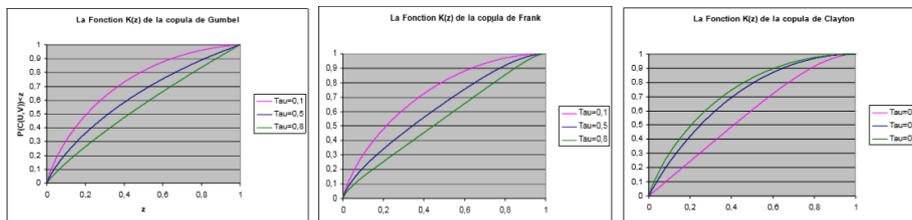


FIGURE 4.6 – Les fonctions K_θ des copules archimédiennes et pour différentes valeurs du Tau de Kendall.

Supposons maintenant que nous disposions d’un échantillon d’observations $((x_1, y_1), \dots, (x_n, y_n))$ issu d’un vecteur aléatoire (X, Y) . Pour établir un estimateur non-paramétrique de la fonction K à partir de cet échantillon, la procédure à suivre consiste à :

1. définir la pseudo-observation z_i pour chaque $i = 1, \dots, n$;

$$z_i = \frac{\text{nombre de paires } (x_j, y_j) \text{ telles que } x_j < x_i \text{ et } y_j < y_i}{n - 1}$$

$$= \frac{1}{n - 1} \sum_{j=1}^n 1_{\{x_j \leq x_i, y_j \leq y_i\}},$$

où x_j et y_j représentent les statistiques d'ordre associées à l'échantillon,

2. définir l'estimateur non-paramétrique de K comme suit :

$$\hat{K}_n(z) = \frac{\text{nombre de } z_i < z}{n} = \frac{1}{n} \sum_{i=1}^n 1_{\{z_i \leq z\}}.$$

Cette estimation non-paramétrique de K pourra ensuite être comparée graphiquement aux versions paramétriques de K pour les différentes copules archimédiennes.

Le paramètre θ de la copule pourra être établi, par exemple, à partir du coefficient de Kendall empirique de l'échantillon. En effet, nous avons vu qu'il existe une relation directe entre le coefficient de Kendall et le paramètre de la copule. Il suffit donc simplement de résoudre une équation pour déterminer le paramètre de la copule, même si cette opération peut se révéler un peu plus délicate dans le cas de la copule de Frank, par exemple. Une autre méthode pour déterminer ce paramètre pourra être celle du maximum de vraisemblance que nous avons exposée précédemment. Nous pourrions ensuite être en mesure d'effectuer une comparaison graphique entre K_θ et l'estimateur non-paramétrique \hat{K}_n de K calculé à partir de l'échantillon. En règle générale, cette simple comparaison graphique mettra assez nettement en évidence la copule archimédienne à utiliser.

4.3 Utilisation de données réelles en assurance non-vie

Cette section illustre les méthodes d'ajustement des copules archimédiennes aux demandes d'indemnisation des compagnies d'assurance. Les données Loss-ALAE contiennent 1500 réclamations générales choisies aléatoirement dans un bureau des services d'assurances. Chaque réclamation comprend un paiement d'indemnité (Loss, X) et une charge d'ajustement de perte attribuée (ALAE, Y), des exemples d'ALAE sont les honoraires versés aux avocats, aux experts et aux enquêteurs, utilisés pour défendre les réclamations (pour plus de détails voir [12]). Les données Loss-ALAE sont disponibles avec le package `copula`, `evd`, ou `VGAM` du logiciel R.

Notre objectif est de décrire la répartition conjointe de Loss et ALAE. La distribution conjointe permet de calculer via la simulation les primes d'une stratégie de réassurance en présence de limites de police et de rétentions de l'assureur. L'estimation de la distribution conjointe de Loss et ALAE est compliquée du fait de la présence de la censure, une caractéristique commune des données sur Loss. Spécifiquement, en plus des informations sur Loss et ALAE, pour chaque réclamation, nous avons un enregistrement de la limite de police, le montant maximal de la réclamation. Avec la présence de la limite de police, la variable

Loss (X) est censurée quand le montant de la réclamation dépasse la limite de police indiquée. Pour certaines réclamations, la limite de police était inconnue et pour celles-ci, nous avons supposé qu'il n'y avait pas de limite de police. Le tableau 4.1 résume les données.

(\$ US)	ALAE	Loss	limite de police	Loss non censuré	Loss censuré
Nombre	1500	1500	1352	1466	34
Moyenne	12588	41208	559098	37110	217491
Médiane	5471	12000	500000	11048	100000
Écart-type	28146	102748	418649	92513	258205
Minimum	15	10	5000	10	5.000
Maximum	501863	2173595	7500000	2173595	1000000
1 ^{ere} Quartile	2.333	4000	300000	3750	50000
3 ^{eme} Quartile	12577	35000	1000000	32000	300000

TABLE 4.1 – Statistiques sommaires de Loss et ALAE

Ici, seulement 34 des 1500 réclamations ont des montants qui égalent la limite de police donc les indemnités correspondantes sont considérées comme censurées. Cependant, les indemnités censurées ne peuvent être ignorées, par exemple, l'indemnité moyenne de revendications censurées est beaucoup plus élevée que la moyenne correspondante pour les revendications non censurées (217491 \$ US contre 37110 \$ US).

La figure 4.7 est un diagramme de dispersion de Loss en fonction d'ALAE sur les échelles logarithmiques.

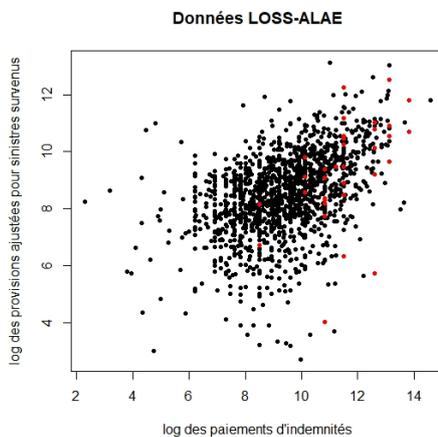


FIGURE 4.7 – Diagramme de dispersion de Loss en fonction de l'ALAE.

Les Loss importantes ont tendance à être associées aux ALAE importantes, comme prévu. De plus, les points de données censurées (représentés par des points rouges sur la figure 4.7) se regroupent clairement à droite.

Le tau de kendall et le rho de spearman empirique de l'échantillon sont respectivement 0.3154175 et 0.451872.

4.3.1 L'ajustement des distributions marginales

Pour appliquer une fonction copule, nous devons identifier les marginaux appropriés pour X et Y . Normalement, nous devons ajuster différentes distributions à nos données empiriques et choisir les distributions qui offrent les meilleurs ajustements. Cependant, les données Loss-ALAE sont fortement asymétriques et ont une queue lourde. Deux distributions typiques utilisées en actuariat pour ce type de données sont les distributions de Burr et de Pareto. Ici, nous utilisons la distribution de Pareto pour les deux composantes. Soit H la distribution conjointe du couple (X, Y) , F et G leur distribution marginale respective.

$$F(x) = 1 - \left(\frac{\lambda_1}{\lambda_1 + x} \right)^{\gamma_1} \text{ et } G(y) = 1 - \left(\frac{\lambda_2}{\lambda_2 + y} \right)^{\gamma_2}$$

où $\lambda_1, \lambda_2, \gamma_1$ et γ_2 sont des paramètres. On estime ces paramètres en utilisant le maximum de vraisemblance classique.

On obtient numériquement

$$\gamma_1 = 1.135; \lambda_1 = 14453 \text{ et } \gamma_2 = 2.223; \lambda_2 = 15133.$$

La qualité de l'ajustement des distributions marginales peut être examinée avec une comparaison graphique de la fonction de distribution ajustée par rapport aux versions empiriques de Loss et ALAE, comme indiqué dans les figures suivantes. En raison de la censure, nous avons utilisé la fonction de distribution empirique de Kaplan-Meier pour la variable Loss (voir [15]).

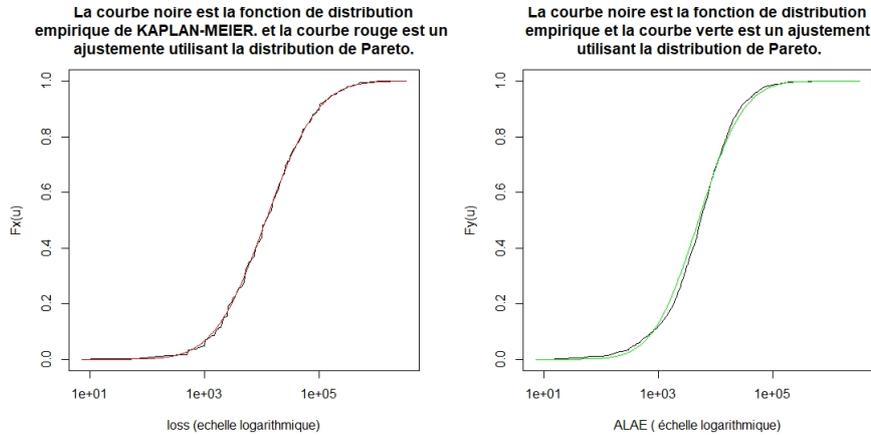


FIGURE 4.8 – Les fonctions de distribution ajustées et empirique de Loss et ALAE.

Pour ALAE, l'ajustement n'est pas aussi bon que celui obtenu sur Loss, mais est acceptable.

4.3.2 L'ajustement d'une copule aux données bivariées

Sélection d'une copule

Nous utilisons la procédure développée précédemment pour sélectionner une copule appropriée. Selon la procédure, nous examinons graphiquement le degré

de proximité des versions non-paramétrique ($\hat{K}_n(z)$) et paramétrique ($K_\theta(z)$) de la fonction de distribution $K(z)$, pour choisir la bonne copule. Ici, les copules à comparer sont les copules de Gumbel-Hougaard et de Cook-Johnson (ou Clayton). Les graphes de ces fonctions sont représentés ci-dessous.

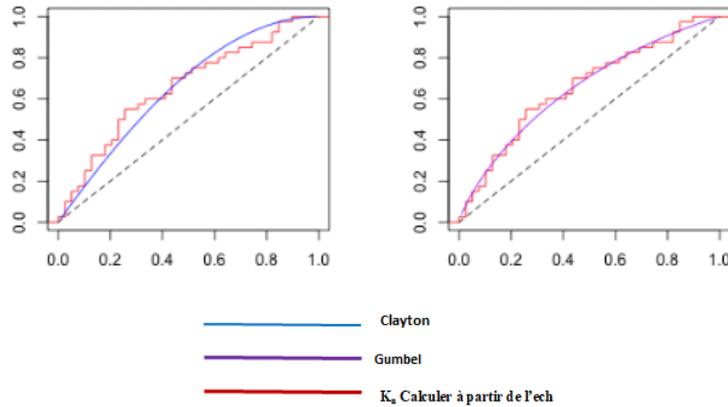


FIGURE 4.9 – Comparaison graphique entre K_θ des copules et \hat{K}_n calculé à partir de l'échantillon.

La procédure suggère alors l'utilisation de la copule de Gumbel, car il ressort des figures que la copule de Gumbel offre le meilleur ajustement.

Estimation d'une copule en utilisant la méthode du maximum de vraisemblance exacte

Maintenant, pensons à un modèle bivarié avec la copule de Gumbel. Nous avons déjà vu que cette copule est la plus adéquate à notre série de données. Mais cette fois, considérons le paramètre global à estimer $\alpha^T = (\theta, \beta_1, \beta_2)$ avec θ le paramètre de la copule C , $\beta_1 = (\lambda_1, \gamma_1)$ et $\beta_2 = (\lambda_2, \gamma_2)$. Notons :

$$\begin{cases} H(x, y; \alpha) = C(F(x), G(y); \theta) \text{ avec } F(x) = F(x; \beta_1), G(y) = G(y; \beta_2), \\ c = \frac{\partial C(F(x), G(y); \theta)}{\partial x \partial y}, c_1 = \frac{\partial C(F(x), G(y); \theta)}{\partial x}, c_2 = \frac{\partial C(F(x), G(y); \theta)}{\partial y}. \end{cases}$$

Pour développer la méthode, nous ferons la distinction entre les cas censurés et non censurés.

Nous prenons δ comme l'indicateur de censure.

Si la variable Loss n'est pas censurée, $\delta = 0$ et la fonction densité conjointe est

$$h(x, y; \alpha) = f(x)g(y)c[F(x), G(y); \theta]. \quad (a)$$

Si la variable Loss est censurée, alors $\delta = 1$ et la probabilité commune est donnée par :

$$P(X \geq x, Y \leq y) = G(y) - H(x, y; \alpha).$$

Ainsi, la fonction densité conjointe lorsque la variable Loss (X) est censurée est

$$g(y) - \frac{\partial H(x, y; \alpha)}{\partial y} = g(y) - g(y)c_2[F(x), G(y); \theta] = g(y) \left[1 - c_2[F(x), G(y); \theta] \right]. \quad (b)$$

En combinant les équations (a) et (b), pour une seule observation, la fonction du maximum de vraisemblance est :

$$L(x, y; \delta, \alpha) = \prod_{i=1}^n \left\{ h(x_i, y_i; \alpha)^{(1-\delta)} \times \left\{ g(y_i) \left[1 - c_2[F(x_i), G(y_i); \theta] \right] \right\}^\delta \right\}.$$

On obtient ainsi la fonction de log-vraisemblance l ,

$$l(\alpha) = \sum_{i=1}^n \left\{ (1-\delta) \log h(x_i, y_i; \alpha) + \delta \left\{ \log g(y_i) + \log \left[1 - c_2[F(x_i), G(y_i); \theta] \right] \right\} \right\}.$$

L'estimateur du vecteur de paramètres α est ensuite déterminé en maximisant la fonction de log-vraisemblance :

$$\hat{\alpha}_{FML} = \arg \max_{\alpha \in \Lambda} l(\alpha),$$

où Λ est l'espace des paramètres.

Les résultats de l'estimation du maximum de vraisemblance exacte correspondant à la copule de Gumbel-Hougaard, dont les dérivées partielles sont calculables, sont résumés dans le tableau 4.2.

	paramètres	Distribution bivariable		Distribution univariée	
		estimation	erreur standard	estimation	erreur standard
Loss (X)	λ_1	14036	1298	14453	1397
	γ_1	1.122	0.062	1.135	0.066
ALEA(Y)	λ_2	14219	1426	15133	1633
	γ_2	2.118	0.153	2.223	0.175
Dépendance	θ	1.453	0.034		

TABLE 4.2 – Les résultats des estimations des paramètres.

Ici, nous voyons que les estimations des paramètres des distributions marginales sont légèrement différentes lorsque nous comparons l'estimation univariée à l'estimation bivariable. Les erreurs standards sont plus petites dans l'estimation bivariable, ce qui indique une plus grande précision des estimations des paramètres.

L'estimateur du paramètre de dépendance θ est significativement différent de 1. Cela fournit des preuves statistiques solides que Loss et ALAE ne sont pas indépendantes.

Après estimation des paramètres nous simulons la copule de Gumbel en utilisant la méthode des distributions conditionnelles vue dans la section 4.1.

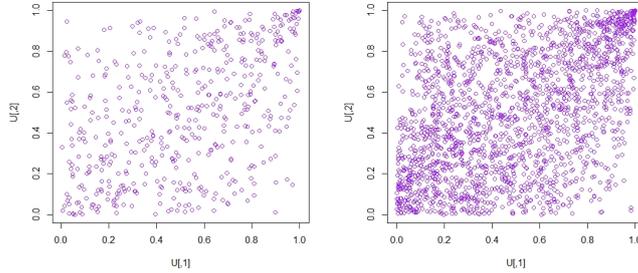


FIGURE 4.10 – Simulation d'échantillons de tailles 500 et 2000 de la copule de Gumbel après estimation des paramètres.

4.3.3 Calcul des primes de réassurance

Après avoir identifié la meilleure copule qui exprime la distribution conjointe de (U, V) telles que $U = F(X)$ et $V = F(Y)$, nous pouvons examiner la distribution de toute fonction connue de X et Y , par exemple $g(X, Y)$. Ainsi, nous calculons le paiement attendu sur une police avec limite L et rétention de l'assureur R (voir [12]). En supposant un partage des dépenses au prorata, nous avons l'expression suivante du paiement :

$$g(X, Y) = \begin{cases} 0 & \text{si } X < R \\ X - R + \frac{X - R}{X}Y & \text{si } R \leq X < L \\ L - R + \frac{L - R}{L}Y & \text{si } X \geq L \end{cases}$$

Le paiement prévu $E[g(X, Y)]$ pourrait être calculé à l'aide de l'intégrale numérique lorsque la densité conjointe des Loss et ALAE est disponible. Cependant, la simulation est un outil d'évaluation numérique plus simple. L'idée avec la simulation est de générer une séquence de données bivariées $\{(u_{1i}, u_{2i})\}_{i=1}^n$ à partir de la copule de Gumbel (voir section 4.1). À partir des nombres (u_{1i}, u_{2i}) , il suffit ensuite d'utiliser des transformations marginales pour générer une séquence $\{(x_i, y_i)\}_{i=1}^n$. Pour cela, on inverse les fonctions de distribution marginale. Pour la distribution de Pareto, il n'est pas difficile de vérifier que

$$x_i = F^{-1}(u_{1i}) = \lambda_1[(1 - u_{1i})^{-\frac{1}{\gamma_1}} - 1] \quad \text{et} \quad y_i = G^{-1}(u_{2i}) = \lambda_2[(1 - u_{2i})^{-\frac{1}{\gamma_2}} - 1].$$

Alors, la valeur estimée pour le paiement prévu du réassureur est

$$\hat{g}^*(L, R) = \frac{1}{n} \sum_{i=1}^n g(x_i, y_i) \quad \text{avec une erreur standard}$$

$$es(\hat{g}^*(L, R)) = \sqrt{\frac{\frac{1}{n} \sum_{i=1}^n g(x_i, y_i)^2 - \hat{g}^*(L, R)^2}{n}}.$$

Une étude de simulation avec un échantillon de taille $n = 100000$ a été effectuée. Les résultats sont résumés dans le tableau suivant.

limite de police	Ratio de rétention de l'assureur à la limite de police (R/L)				
	0.00	0.25	0.50	0.75	0.95
10000	15636 (640)	11232 (480)	7220 (320)	3498 (160)	684 (32)
100000	34264 (655)	17965 (493)	10003 (328)	4425 (164)	819 (33)
500000	49367 (733)	17457 (544)	9234 (359)	4007 (179)	739 (36)
1000000	55683 (818)	16762 (597)	8740 (390)	3716 (193)	672 (38)

TABLE 4.3 – Primes de réassurance basées sur la simulation (les erreurs standard de simulation sont entre parenthèses).

Les résultats du tableau 4.3 fournissent les primes ajustées que le réassureur aurait évaluées pour couvrir les coûts des indemnités et des dépenses selon diverses limites de police et de ratio R/L . Sur une base brute, ces résultats semblent logiques. Par exemple, d'après les statistiques sommaires du tableau 4.1, la moyenne limite de police est de 559098, avec une moyenne des indemnités plus moyenne des dépenses de 53796. Sans rétention, nos résultats indiquent une prime ajustée de 49367, avec une erreur standard de 733, pour une limite de police de 500000.

De plus, les résultats sont intuitivement attrayants parce que, comme prévu, nous observons :

- une prime plus élevée pour des limites de police plus importantes ,
- une prime plus faible lorsque le ratio R/L est plus élevé, c'est-à-dire que l'assureur conserve une plus grande quantité de l'indemnité.

Conclusion

Dans ce mémoire nous avons abordé la notion de copules. Nous avons présenté trois approches d'estimation de la copule. La première approche est paramétrique, elle impose un modèle paramétrique pour la copule et pour les distributions marginales. La seconde approche est semi-paramétrique, elle suppose un modèle paramétrique pour la copule et non-paramétrique pour les distributions marginales. La troisième approche d'estimation de la copule est non-paramétrique, celle-ci ne fait pas de restriction sur les modèles.

Nous avons également présenté trois méthodes de simulation de la copule. La première est la méthode des distributions, on l'utilise si la distribution jointe au quelle la copule est associée est connue et facile à simuler. La seconde est la méthode des distributions conditionnelles qui consiste à simuler la copule directement. La troisième est la méthode analytique qui est propre à chaque copule. Enfin nous avons illustré toute cette théorie par une application en assurance non-vie, avec utilisation des données Loss-ALAE qui ont été enregistrées par une compagnie d'assurance. Ces données sont disponibles avec le package `copula` ou `evd` du logiciel R.

Malgré sa relative simplicité mathématique, l'outil copule n'est pas toujours utilisé dans toutes ses potentialités. En pratique, la théorie des copules doit pouvoir être étendue à plus de deux dimensions. Mais, la plupart des articles parus à ce jour se contentent d'utiliser les copules pour des travaux en deux dimensions. Lorsque le nombre de dimensions est supérieur à deux, les auteurs se ramènent toujours à une copule elliptique (Normale ou de Student) pour la simple et bonne raison que les densités de ces copules sont aisées à calculer et que ces copules sont simulables de manière simple, même lorsque le nombre de dimensions est très élevé. Or, ce type de copules n'est pas forcément le plus adapté dans certains domaines, par exemple dans l'étude de la sinistralité en assurance. Même s'il existe des extensions, et encore sous certaines conditions, des copules archimédiennes à plus de deux dimensions, établir les dérivées nécessiterait des calculs tout à fait fastidieux.

En espérant que les prochains travaux sur les copules seront multidimensionnels et sauront mettre en évidence leur apport dans l'assurance. Par exemple, des contrats d'assurance comportant plusieurs garanties (incendie et pertes d'exploitation) ou impliquant le paiement de plusieurs types de sinistres (en santé les remboursements pour différents actes comme les consultations et visites ainsi que la pharmacie ...) pourraient faire également l'objet d'une modélisation par les copules.

Bibliographie

- [1] Belguise O. (2001). *Tempêtes : Etude des dépendances entre les branches Auto et Incendie avec la théorie des copulas*. Université Louis pasteur STRASBOURG
- [2] Brahimi B., Necir A. (2012). *A semiparametric estimation of copula models based on the method of moments*. *Statistical Methodology*, 9 (4), 467-477.
- [3] Cadoux D., Loizeau J.M (2004). *Copules et dépendances : application pratique à la détermination du besoin en fonds propres d'un assureur non vie*. Institut des Actuaire. Paris.
- [4] Capéraà P., Fougères A.L., Genest C. (2000). *Bivariate Distributions with Given Extreme Value Attractor*. *Journal of Multivariate Analysis*.72, 30-49.
- [5] Deheuvels P. (1979). *La fonction de dépendance empirique et ses propriétés. Un test non-paramétrique d'indépendance*. *Académie Royale de Belgique Bulletin de la Classe des Sciences*, 65 (5), 274-292.
- [6] Devroye L.,(1986). *Non-Uniform Random Variate Generation*. Springer, New York.
- [7] Drouet, M.D., Kotz, S. (2001). *Correlation and dependence*. Imperial College Press, London.
- [8] Embrechts p., McNeil A. , Lindskog F. (2003) . *Modelling dependence with copulas and applications to risk management*. Rachev (Ed.), *Handbook of Heavy Tailed Distributions in Finance*. Elsevier, New York, 329-384.
- [9] Farlie D. J. G., 1960. *the performance of some correlation coefficients for a general bivariate distribution*. *Biometrika* 47, 307-323.
- [10] Fang K., Kotz S. (1990) *Symmetric multivariate and related distributions*. Chapman and Hall.
- [11] Fermanian J. D., Scaillet O. (2002). *Non parametric estimation of copulas for time series*. *Journal of risks* 5, 25-54.
- [12] Frees E.W.,Valdez E.A. (1998). *Understanding relationships using copulas*. *North American Actuarial Journal* 2, 1-25.
- [13] Frederic P. (2017). *MODÈLES FINANCIERS EN ASSURANCE ET ANALYSES DYNAMIQUES*. grande école d'acturiat et de gestion des risque institut de science financière et d'assurance.
- [14] Genest C., Quessy J., and Remillard B. (2006). *Goodness-of-t procedures for copula models based on the integral probability transformation*. *Scandinavian Journal of Statistics*. 33, 337-366.
- [15] Genest C., Ghoudi K., Rivest L. P. (1995). *A semiparametric estimation procedure of dependence parameters in multivariate families of distributions*. *Biometrika* 82, 543-552.

- [16] Genest C., Rivest L. (1993). *Statistical inference procedures for bivariate Archimedean copulas*. Journal of American Statistical Association. 88, 423.
- [17] Genest C., Rivest L.P. (1989). *A characterization of Gumbel's family of extreme value distributions*. Statistical and probability Letters. 8, 207-211.
- [18] Genest C., Mackay R. J. (1986). *Copules archimédiennes et famille de loi bidimensionnelle dont les marges sont données*. Canadian Journal of statistics. 14, 145-159.
- [19] Gumbel E.J. (1960). *Distributions des valeurs extrême en plusieurs dimensions*, Institut de Statistique de l'Université de Paris. 9, 171-173.
- [20] Gumbel E. J. (1960). *Bivariate exponential distributions*. Journal of American Statistical Association 55, 698-707.
- [21] Hassen MALOUTHI. *support de cours de simulation ESSAIT*. année universitaire 2014-2015, Université de Cartage
- [22] Jean-Yves DAUXOIS, *support de cours de Statistique Inférentielle*. Master Enseignement des Mathématiques, Année 2011-2012, Université de Franche-Comté
- [23] Joe H. (2005). *Asymptotic efficiency of the two-stage estimation method for copulabased models*. Multivariate Analysis Journal 94, 401-419.
- [24] Joe H. (1997). *Multivariate models and dependence concepts*. Chapman and Hall, London
- [25] Joe H., Xu J.J. (1996). *The Estimation Method of Inference Functions for Margins for Multivariate Models*. Dept. of Statistics University of British Columbia, Tech. Rept. 166.
- [26] Kim G., Silvapulle J., Silvapulle P., (2007). *Comparison of semi-parametric and parametric methods for estimating copulas* Computational Statistics and Data Analysis 51, 2836-2850.
- [27] Kimeldorf G., Sampson A., (1975). *One-parameter families of bivariate distributions with fixed marginals*. Comm. Statist. 4, 293-301.
- [28] Morgenstern D. (1956). *Simple examples of two-dimensional distributions*. Statist. Math. Doc. 8, 234-235.
- [29] Nelsen R.B.(2006). *An Introduction to copula*. Springer Verlag, New York.
- [30] Pierre Clauss, *Théorie des copules*. Notes de cours, École nationale de la statistique et de l'analyse de l'information (Ensa), Filière Gestion des risques et ingénierie financière, 2007-2010.
- [31] Shih J.H., Louis T.A. (1995) *Inference on the association parameter in copula models for bivariate survival data*. Biometrics. 51,1384-1399.
- [32] Sklar A. (1999) *Fonction de répartition à n dimension et leurs marges*. Publication de l'institut de statistique de l'université de paris. 8,229-231.
- [33] Source('SimulationVecteurGaussien2D.r')
- [34] Tawn J.A. (1988). *Bivariate extreme value theory : Models and estimation*. Biometrika. 75, 397-415.
- [35] WeiB G. (2011). *Copula parameter estimate by maximum- Likelihood and minimum-Distance estimators -A Simulation Study*. Comput. Stt. 26,3154.
- [36] Yannick M., Sornette D. (2006). *Extreme financial risks : from dependence to risk management*. Birkhuser.
- [37] [http ://www.r-project.org/](http://www.r-project.org/)