Analyse de scénarios à risque élevé au moyen de copules balayant tout le spectre de dépendance d'extrémité

Une commandite de la Section conjointe CAS/ICA/SOA sur la gestion du risque

par
Lei Hua
Michelle Xia

Décembre 2013

Analyse de scénarios à risque élevé au moyen de copules balayant tout le spectre de dépendance d'extrémité*

Lei Hua† Michelle Xia‡

Le 12 décembre 2013

Résumé. Les copules ayant la propriété de balayer tout le spectre de dépendance d'extrémité peuvent mesurer l'ensemble des dépendances positives dans l'extrémité et ainsi permettre la construction d'un modèle de régression qui tienne compte des schémas de dépendance dynamique entre les variables. Nous proposons un modèle qui intègre à la fois la régression sur chacune des distributions marginales des variables de réponse bidimensionnelles et la régression sur le paramètre de dépendance des variables de réponse. La copule ACIG, qui possède la propriété de modéliser tout le spectre de dépendance d'extrémité, est intégrée à l'analyse de régression. Nous établissons des comparaisons entre les analyses de régression basées sur la copule ACIG et la copule de Gumbel, qui indiquent que la copule ACIG donne généralement de meilleurs résultats que la copule de Gumbel lorsqu'il y a dépendance intermédiaire dans l'extrémité supérieure (*intermediate upper tail dependence*). Nous réalisons une étude de simulation pour montrer qu'il est possible de prendre en compte les structures de dépendance dynamique dans l'extrémité entre les sinistres et les frais de règlement imputés, en utilisant la copule ACIG à un paramètre. Enfin, nous appliquons les modèles de régression ACIG et de Gumbel à un ensemble de données tiré du Medical Expenditure Panel Survey (MEPS) des États-Unis. L'analyse empirique révèle que le modèle de régression avec la copule ACIG améliore l'analyse des scénarios à risque élevé, surtout dans le cas de risques dépendants agrégés.

Mots clés : copule ACIG, ordre de l'extrémité (*tail order*), dépendance dynamique, frais de règlement imputés, ensemble de données MEPS, analyse de régression.

---

* Ce projet de recherche a été commandité par la Society of Actuaries (SOA), la Casualty Actuarial Society (CAS) et l'Institut canadien des actuaires (ICA). Nous leur sommes très reconnaissants pour leur aide financière ainsi qu'aux bénévoles pour le temps précieux qu'ils ont consacré à l'étude des propositions et au bon déroulement du projet. Le premier auteur tient à remercier M. Peng Shi (Ph.D.) pour ses commentaires utiles au sujet du projet. Nous assumons la responsabilité de toute erreur pouvant subsister.

† Auteur correspondant, hua@math.niu.edu, département de statistiques, Université de Northern Illinois, DeKalb, IL, 60115, États-Unis.

‡ cxia@niu.edu, département de statistiques, Université de Northern Illinois, DeKalb, IL, 60115, États-Unis.

1 Introduction

L'analyse de régression avec divers modèles linéaires généralisés (MLG, qui dans la présente analyse incluent les modèles additifs généralisés) s'est avérée utile pour tirer des conclusions statistiques au sujet de distributions à une variable. En particulier, ils sont couramment employés par les sociétés d'assurances IARD pour établir les tarifs. Toutefois, ces modèles servent surtout à la tarification et ne conviennent pas toujours à la gestion quantitative du risque, la principale raison étant que le choix des modèles de tarification et l'étalonnage des paramètres sont basés sur des variables explicatives qui prennent des valeurs relativement et hautement probables. Or, les valeurs extrêmes, bien qu'étant relativement moins probables, ont une incidence directe sur la stabilité financière. Dans le domaine de la gestion quantitative du risque, l'étude des distributions des sinistres dans l'extrémité fait l'objet d'une plus grande attention que jamais depuis la dernière crise financière qui a débuté en 2007. Entre autres débats qui ont suivi, il a été question des limites des copules gaussiennes. Pour une analyse des inconvénients liés à l'utilisation d'une copule gaussienne pour modéliser les risques de crédit, se reporter à Donnelly et Embrechts (2010).

Tous les modèles statistiques présentent certaines limites. En matière de gestion quantitative du risque, l'une des limites majeures des copules gaussiennes est qu'elles ne prennent pas en compte le comportement simultané des risques importants dans l'extrémité de distribution lorsqu'il existe effectivement des groupes de sinistres importants. Plus précisément, le paramètre de dépendance d'extrémité $\lambda = \lim_{x \to \infty} \mathbb{P}[X_1 > x] = 0$, où X_1 et X_2 sont des variables aléatoires identiquement distribuées ayant pour structure de dépendance la copule gaussienne. L'absence de dépendance d'extrémité peut être contournée par l'utilisation d'une copule possédant une dépendance dans l'extrémité supérieure, telle que la copule de Gumbel et la copule *t* de Student, dont le paramètre de dépendance d'extrémité supérieure est $\lambda > 0$. Toutefois, on pourrait réagir de façon excessive face à ce problème en utilisant des copules possédant une dépendance d'extrémité, alors qu'en fait la dépendance dans l'extrémité supérieure n'est pas assez forte. C'est pourquoi il nous faut une copule qui prenne en compte les deux possibilités, soit lorsque le paramètre de dépendance dans l'extrémité supérieure est $\lambda > 0$ ou $\lambda = 0$.

Lorsqu'une famille de copules est en mesure de prendre en compte les deux possibilités $\lambda > 0$ et $\lambda = 0$ pour représenter la dépendance positive dans l'extrémité, nous la désignons sous le nom de copule balayant tout le spectre de dépendance d'extrémité; on notera qu'ici, la « dépendance d'extrémité » est une description générique de la dépendance dans l'extrémité, et non pas le concept exposé à la section 2.1.10 de Joe (1997). Pour ce dernier, nous emploierons l'expression « dépendance d'extrémité usuelle » pour le cas où $\lambda > 0$.

Lorsque $\lambda > 0$, il s'agit de la dépendance d'extrémité usuelle, et la valeur de λ elle-même sert à quantifier le degré de dépendance. Mais lorsque $\lambda = 0$, il nous faut plus d'information pour quantifier le degré de dépendance. Dans Hua et Joe (2011), la notion d'« ordre de l'extrémité » sert à quantifier le degré de dépendance dans le cas où $\lambda = 0$. Nous proposons ensuite le concept de « dépendance d'extrémité intermédiaire » pour prendre en compte la dépendance d'extrémité positive avec $\lambda = 0$.

Les résultats obtenus par Hua et Joe (2014) témoignent également de l'importance d'utiliser une copule balayant tout le spectre de dépendance d'extrémité pour analyser les scénarios à risque élevé. Dans un contexte de régression, si l'on veut tirer des conclusions au sujet des mesures de risque ayant les formes suivantes $\mathbb{E}(Y_1|Y_2 > t, \mathbf{X} = \mathbf{x})$ ou $\mathbb{E}(Y_1|Y_2 = t, \mathbf{X} = \mathbf{x})$, où t est la valeur à risque (VaR) de Y_2, la force de la dépendance d'extrémité entre Y_1 et Y_2 , sachant que $\mathbf{X} = \mathbf{x}$, devient très importante, car les mesures de risque y sont très sensibles. Dans ce cas, l'utilisation d'une copule balayant tout le spectre de dépendance d'extrémité permet d'améliorer la modélisation de façon significative.

Le présent article a pour but d'étudier comment une copule balayant tout le spectre de dépendance d'extrémité pourrait être utile à l'évaluation des risques extrêmes dans un contexte de régression. Nous présenterons tout d'abord les concepts d'ordre de l'extrémité et de copules balayant tout le spectre de dépendance d'extrémité, ainsi qu'une copule à un paramètre qui balaie tout le spectre de dépendance d'extrémité et que l'on désigne sous le nom de copule ACIG. Plus particulièrement, nous comparerons cette dernière avec la copule de Gumbel d'utilisation courante, et nous insisterons sur la comparaison des différents comportements de ces deux copules dans l'extrémité supérieure. Nous montrerons que la copule ACIG balaie une plus grande partie de la dépendance d'extrémité que ne le fait la copule de Gumbel, et qu'elle comporte moins d'erreurs de prévision estimées par validation croisée que la copule de Gumbel dans le contexte d'une étude de simulation. Puis nous développerons un modèle de régression basé sur des copules pour lequel la variable de réponse est bidimensionnelle et les degrés de dépendance de celle-ci peuvent changer en fonction des valeurs des variables explicatives. Nous utiliserons un ensemble de données simulées sur les sinistres et les frais de règlement pour montrer les structures de dépendance dynamique entre les variables de réponse et pour montrer comment une copule balayant tout le spectre de dépendance d'extrémité permet de prendre en compte l'évolution des schémas de dépendance dans l'extrémité de distribution. Enfin, nous procéderons à une analyse empirique de l'ensemble de données du MEPS, pour lequel la copule ACIG balayant tout le spectre de dépendance d'extrémité permet d'améliorer le modèle d'analyse des scénarios à risque élevé.

Le principal apport de notre étude est le suivant : nous proposons le concept de spectre complet de dépendance d'extrémité pour modéliser les schémas de dépendance dynamique d'extrémité, ce qui permet de surmonter le problème des modèles actuels qui permettent de balayer soit uniquement l'indépendance d'extrémité, soit uniquement la dépendance d'extrémité, et nous avons étudié une copule à un paramètre qui balaie tout le spectre de dépendance d'extrémité et l'avons intégrée à des modèles de régression afin d'améliorer le modèle d'analyse des scénarios à risque élevé. Le modèle de régression avec copule balayant tout le spectre de dépendance d'extrémité permet d'améliorer les évaluations des risques non seulement dans le cas des distributions marginales mais aussi dans celui des sinistres agrégés.

Le document se divise comme suit : à la section 2 sont présentés les concepts de base que sont l'ordre de l'extrémité et la copule balayant tout le spectre de dépendance d'extrémité. À la section 2.3, nous utilisons une erreur de prévision estimée par validation croisée pour comparer la performance de modélisation de la copule ACIG et de la copule de Gumbel. Des modèles de régression utilisant une copule qui balaie tout le spectre de dépendance d'extrémité sont construits à la section 3. À la section 4, nous réalisons une étude de simulation qui utilise la copule ACIG pour modéliser les structures de dépendance dynamique entre les sinistres et les frais de règlement en assurance automobile. À la section 5, nous utilisons un ensemble de données du MEPS pour présenter le modèle de régression construit avec la copule ACIG, que nous comparons au modèle de Gumbel. Enfin, à la section 6, nous présentons nos observations finales, tandis qu'à la section 7, nous présentons des considérations d'ordre technique relatives à l'application des modèles.

2 Ordre de l'extrémité et structure complète de dépendance d'extrémité

2.1 Structure complète de dépendance d'extrémité

Pour un vecteur aléatoire $(X_1, \ldots, X_d)$ ayant une fonction de distribution cumulative (fdc) conjointe F et des fdc unidimensionnelles F_i, $i = 1, \ldots, d$, il existe une fonction copule C : $[0, 1]^d \rightarrow [0, 1]$ telle que pour tout $\mathbf{x} = (x_1, \ldots, x_d)$ faisant partie du support de F, $F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$. Lorsque les F_i sont continues, la copule C est déterminée de façon unique. Soit $\hat{C}$ la copule de survie correspondante de C; c'est-à dire que $\hat{C}(u_1, \ldots, u_d) = \bar{C}(1 - u_1, \ldots, 1 - u_d)$, où $\bar{C}$ représente la fonction de survie de C et se définit comme étant $\bar{C}(u_1, \ldots, u_d) := 1 + \sum_{\emptyset \neq I \subseteq I_d} (-1)^{|I|} C_I(u_i, i \in I)$ avec C_I la copule pour la I-marginale, $I_d := \{1, \ldots, d\}$, et $|I|$ représente le nombre d'éléments dans l'ensemble I; lorsque $|I| = 1$, la notation $C_I(u_i) = u_i, i \in I$.

Le comportement de C dans l'extrémité supérieure est simplement le comportement de $\hat{C}$ dans l'extrémité inférieure, et vice versa. Vu que c'est habituellement l'extrémité supérieure qui est utile à l'évaluation des risques si les variables aléatoires représentent des montants de sinistres, nous nous intéressons uniquement dans le présent article à l'extrémité supérieure d'une copule. Par conséquent, ci-après, nous considérons l'extrémité inférieure de $\hat{C}$ lorsque nous définissons des concepts pertinents au sujet de l'extrémité, et nous pouvons aussi définir des concepts similaires relativement à l'extrémité supérieure de $\hat{C}$.

Si $\hat{C}(u, \ldots, u) \sim \lambda u$ avec $0 < \lambda \leq 1$, lorsque $u \to 0^+$, C a la dépendance usuelle dans l'extrémité supérieure et le paramètre de dépendance est λ, où « $\sim$ » signifie asymptotiquement équivalent; c'est-à-dire que, $g(x) \sim h(x), x \to x_0 \iff \lim_{x \to x_0} g(x)/h(x) = 1$. Si $\hat{C}(u, \ldots, u) \sim u^{\kappa} \ell(u)$, lorsque $u \to 0^+$, avec ℓ une fonction à variation lente[1] et $1 \leq \kappa$, κ est désigné sous le nom d'ordre de l'extrémité supérieure de la copule C. À l'évidence, si $\hat{C}$ a la dépendance usuelle dans l'extrémité supérieure, son ordre de l'extrémité $\kappa = 1$ et $\lambda = \lim_{u \to 0^+} \ell(u)$. Plus la valeur de κ est élevée, plus la dépendance dans l'extrémité est faible. Lorsqu'il y a dépendance d'extrémité usuelle, c'est-à-dire lorsque $\lambda > 0$, nous utilisons la valeur de λ pour quantifier le degré de dépendance dans l'extrémité; lorsque $\kappa > 1$, et qu'en conséquence $\lambda = 0$, nous utilisons la valeur de κ pour quantifier le degré de dépendance dans l'extrémité. Lorsque $1 < \kappa < d$, on dit que l'extrémité correspondante de C a une dépendance intermédiaire d'extrémité avec certaines conditions de régularité. Le lecteur intéressé à obtenir de plus amples détails sur la notion d'ordre de l'extrémité est invité à consulter Hua et Joe (2011, 2013).

Lorsqu'une famille C de copules symétriques par permutation est en mesure de prendre en compte $1 \leq \kappa \leq d$, nous la désignons sous le nom de copule balayant tout le spectre de dépendance d'extrémité. Par exemple, dans le cas d'une copule gaussienne bidimensionnelle avec coefficient de corrélation $|\varrho| \neq 1$, l'ordre de l'extrémité $1 < \kappa = 2/(1 + \varrho) < \infty$ (exemple 1, Hua et Joe, 2011), et il ne s'agit pas d'une copule balayant tout le spectre de dépendance d'extrémité puisque $\kappa \neq 1$. Dans les familles de copules paramétriques existantes, il n'existe que quelques familles de copules non triviales qui ont la propriété de balayer tout le spectre de dépendance d'extrémité. Un exemple en est la copule archimédienne construite à l'aide de la transformation de Laplace de la loi inverse-gamma (désignée sous le nom de copule ACIG, (exemple 4, Hua et Joe, 2011)). Un autre exemple

[1] On dit qu'une fonction mesurable g est à variation lente au point x_0 si, pour tout $t > 0$, $\lim_{x \to x_0} g(tx)/g(x) = 1$. Par exemple, $\log(x)$ est à variation lente à l'infini.

est la copule archimédienne construite par mélange d'une gamma généralisée et d'un simplexe (appelée en anglais copule GGS, (exemple 2, Hua, 2013)).

Dans les deux sous-sections du point 2, nous analyserons la copule ACIG et nous la comparerons avec la copule de Gumbel, car la copule ACIG nous sera plus utile que la copule GGS pour modéliser les structures de dépendance existant dans les ensembles de données à analyser dans ce document. Pour obtenir de plus amples détails sur la copule GGS et ses applications pour modéliser la dépendance entre la fréquence des sinistres et la sévérité des sinistres, se reporter à Hua (2013).

2.2 Copule ACIG

On peut construire comme suit une copule archimédienne au moyen d'un générateur archimédien ψ.

$$C(u_1, \ldots, u_d) = \psi(\psi^{-1}(u_d) + \cdots + \psi^{-1}(u_d)), \qquad u_i \in [0, 1], \qquad i = 1, \ldots, d.$$

Souvent, on peut choisir le générateur archimédien ψ comme étant la transformation de Laplace d'une variable aléatoire positive; c'est-à-dire que $\psi(s) = \int_0^\infty e^{-sx} F(dx)$, où F est la fdc d'une variable aléatoire positive X. On a $\psi(0) = 1$ $\psi(\infty) = 0$, et ψ est complètement monotone. Pour obtenir de plus amples informations sur la construction de copules archimédiennes à l'aide de la transformation de Laplace d'une variable aléatoire positive, se reporter à Joe (1997).

À l'image de l'exemple 4 de Hua et Joe (2011), nous posons $Y = X^{-1}$, et soit X une variable suivant une distribution gamma $(\alpha, 1)$, où α est le paramètre de forme. La transformation de Laplace de la variable aléatoire Y qui suit une loi inverse-gamma est donnée par :

$$\psi(s; \alpha) = \frac{2}{\Gamma(\alpha)} s^{\alpha/2} K_\alpha(2\sqrt{s}), \quad s \geq 0, \ \alpha > 0, \tag{1}$$

où $K\alpha$ est la fonction modifiée de Bessel de seconde espèce. En ce qui concerne la copule bidimensionnelle $C(u, v) := \psi(\psi^{-1}(u) + \psi^{-1}(v))$, la densité est donnée par :

$$c(u, v) = \psi''(\psi^{-1}(u) + \psi^{-1}(v)) \cdot [\psi'(\psi^{-1}(u))]^{-1} \cdot [\psi'(\psi^{-1}(v))]^{-1}. \tag{2}$$

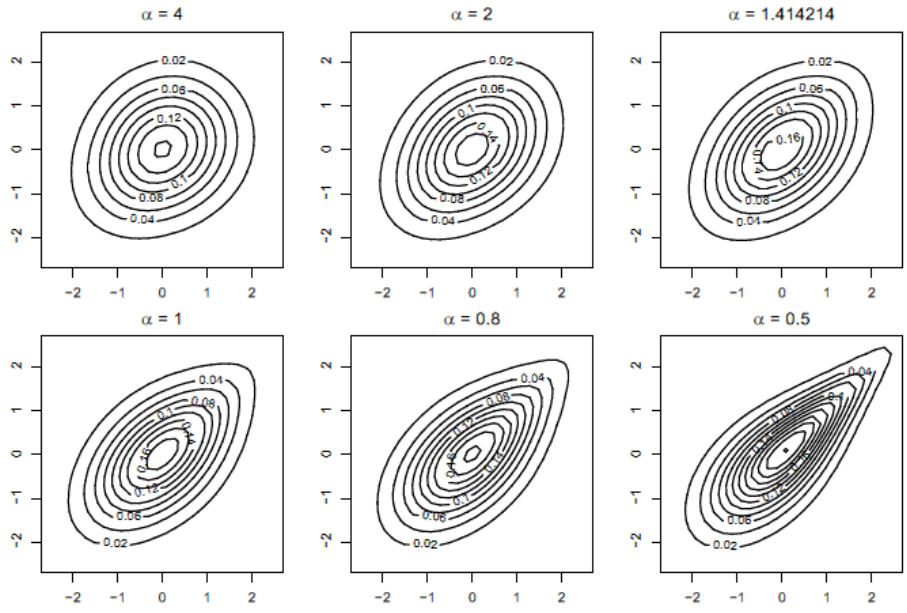
Pour calculer la densité, il nous faut trouver $\psi\rho$, $\psi\rho\rho$ et ψ^{-1}. Les deux premiers termes peuvent être obtenus de façon analytique comme suit, et ψ^{-1} s'obtient numériquement.

$$\psi'(s) = -2s^{(\alpha-1)/2}K_{\alpha-1}(2\sqrt{s})/\Gamma(\alpha)$$
$$\psi''(s) = 2s^{(\alpha-2)/2}K_{\alpha-2}(2\sqrt{s})/\Gamma(\alpha)$$

En ce qui concerne l'analyse de données par des copules archimédiennes construites au moyen d'un ψ strictement décroissant, nous faisons remarquer ici que tant qu'il est possible de calculer les dérivées du générateur ψ de façon analytique, la copule est souvent utile à des applications réelles, car une méthode numérique est habituellement très rapide et efficiente pour obtenir ψ^{-1} pour une telle fonction strictement décroissante ψ.

La figure 1 illustre les courbes de niveau normalisées de la copule ACIG. Ces courbes s'obtiennent par la transformation distincte de chaque copule marginale en une loi normale centrée réduite; c'est-à-dire en transformant (u, v) par $(\Phi^{-1}(u), \Phi^{-1}(v))$, puis en traçant les paires de données de cette dernière, où Φ représente la fdc de la loi normale centrée réduite. La figure 1 indique clairement que la copule ACIG est en mesure de prendre en compte un bien plus large spectre de dépendance dans l'extrémité supérieure avec un seul paramètre de dépendance α. Cette propriété utile de la copule nous offre un moyen convenable de modéliser la dépendance dynamique entre les sinistres et les frais de règlement, sachant que diverses valeurs des covariables se sont réalisées.

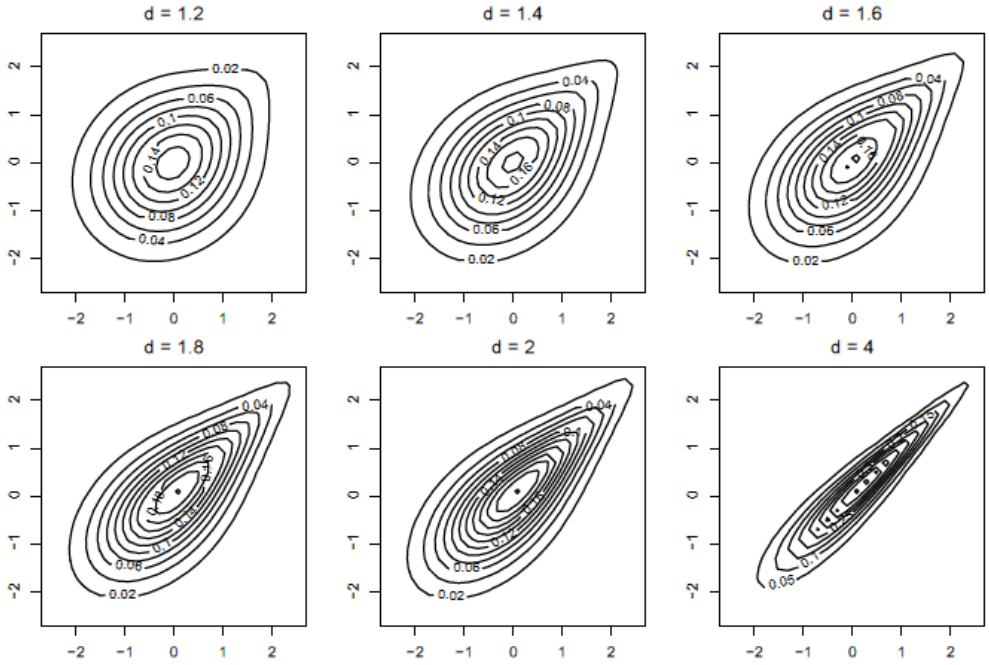
Figure 1 : Courbes de niveau normalisées de la copule ACIG



2.3 Comparaisons

En ce qui concerne l'ensemble bien connu de données sur les sinistres et les frais de règlement imputés qui a été étudié dans nombre d'articles, tels que ceux de Frees et Valdez (1998) et Klugman et Parsa (1999), on peut ajuster la copule de Gumbel aux structures de dépendance. La figure 2 montre les courbes de niveau normalisées de la copule de Gumbel. Au vu des courbes de niveau des copules ACIG et de Gumbel, lorsqu'il y a dépendance d'extrémité usuelle, les deux copules sont très semblables; toutefois, lorsque la dépendance est plus faible, la copule de Gumbel présente toujours une bosse dans l'extrémité supérieure, tandis que l'extrémité supérieure de la copule ACIG est en mesure de prendre en compte une structure de dépendance relativement moindre.

Figure 2 : Courbes de niveau normalisées de la copule de Gumbel



Afin de comparer ces deux copules en ce qui concerne leur capacité d'ajustement à une variété de dépendances dans l'extrémité supérieure, nous appliquons comme critère les erreurs de prévision estimées par validation croisée[2] (désignées par CPVE, en anglais), tel qu'il a été proposé dans Acar et coll. (2011). Pour un échantillon aléatoire (u_i, v_i), $i = 1, \ldots, n$, où $0 \leq u_i, v_i \leq 1$, nous ajustons la copule C à l'échantillon. La CVPE relative à la copule C, après suppression d'une observation (*leave-one-out*), se définit comme suit :

$$\mathrm{CVPE}(C) = \sum_{i=1}^{n} \left[\left\{ u_i - \hat{\mathbb{E}}^{(-i)}(U_i|v_i) \right\}^2 + \left\{ v_i - \hat{\mathbb{E}}^{(-i)}(V_i|u_i) \right\}^2 \right],$$

où

$$\hat{\mathbb{E}}^{(-i)}(U_i|v_i) = \int_0^1 uc(u, v_i|\hat{\theta}^{(-i)})du, \quad i = 1, \ldots, n,$$

et $c(\cdot, \cdot|\widehat{\emptyset}^{(-i)})$ représente l'estimation de la densité de la copule lorsque la i^e observation (u_i, v_i) est supprimée.

Nous comparons maintenant la performance de ces deux copules en ce qui concerne leur capacité d'ajustement aux ensembles de données simulées générées respectivement par la copule ACIG et la copule de Gumbel. Trois différents niveaux de dépendance sont étudiés en termes du β de Blomqvist; plus la valeur du β est élevée, plus grande est la dépendance positive. Le β de Blomqvist d'une copule bidimensionnelle se définit par $\beta = 4 \times C(1/2, 1/2) - 1$. Nous posons ensuite $\beta 1 = 0{,}1$, $\beta 2 = 0{,}3$ et $\beta 3 = 0{,}5$, et les valeurs correspondantes des paramètres de la copule ACIG et de la copule de Gumbel sont respectivement égales à 4,66639, 1,20143, 0,50171 et 1,11453, 1,43406, 1,99664. Nous générons des ensembles de données au moyen de copules de Gumbel (désignées par M_0 = G), puis nous ajustons les copules de Gumbel et ACIG (désignées par G et A) aux ensembles de données et calculons pour chaque copule les CVPE *leave-one-out*. Nous générons aussi des ensembles de données au moyen de copules ACIG, puis nous ajustons les copules de Gumbel et ACIG aux ensembles de données et calculons une fois de plus pour chaque copule les CVPE. Nous répétons cette procédure 50 fois. Les moyennes et les écarts-types de ces CVPE sont indiqués au tableau 1 dans le cas d'un petit échantillon de taille n = 100 et d'un gros échantillon de taille n = 1 000, et ce pour chaque copule.

Tableau 1 : Erreurs de prévision estimées par validation croisée (CVPE) relatives aux copules ACIG et de Gumbel

β de Blomqvist		n = 100				n = 1 000			
		M_0 = G		M_0 = A		M_0 = G		M_0 = A	
		G	A	G	A	G	A	G	A
β_1 = 0,1	CVPE	12,02	11,65	13,89	12,66	120,16	118,09	134,43	122,77
	é.-t.	2,45	3,19	2,15	2,34	8,35	12,82	8,94	8,64
β_2 = 0,3	CVPE	5,55	4,66	6,28	5,48	55,56	45,70	65,06	56,03
	é.-t.	1,29	1,07	1,51	1,33	4,63	3,72	4,83	4,30
β_3 = 0,5	CVPE	1,90	1,88	1,82	1,83	19,19	19,14	19,55	19,36
	é.-t.	0,49	0,36	0,51	0,37	1,89	1,35	1,92	1,47

Au vu du tableau 1, la copule ACIG est généralement meilleure ou comparable à la copule de Gumbel. En ce qui concerne la copule ACIG, β_1 = 0,1, 0,3, 0,5 correspondent respectivement à l'indépendance d'extrémité quadratique (*tail quadrant independance*), l'indépendance d'extrémité intermédiaire, et l'indépendance d'extrémité usuelle. Lorsque la dépendance dans l'extrémité supérieure n'est pas aussi forte que dans le cas de l'indépendance d'extrémité usuelle, la copule ACIG est généralement meilleure que la copule de Gumbel en termes de CVPE. En particulier, dans le cas de l'indépendance d'extrémité intermédiaire, la copule ACIG est nettement meilleure. Comme il fallait s'y attendre, lorsque la dépendance dans l'extrémité supérieure est relativement plus forte, il est difficile de distinguer la copule ACIG de la copule de Gumbel.

3 Régression aux fins de dépendance

3.1 Dépendance dynamique

La modélisation de la dépendance entre les sinistres et les dépenses a fait l'objet de nombreuses recherches en actuariat depuis la parution des articles de Frees et Valdez (1998) et Klugman et Parsa (1999). Des copules ont été appliquées pour modéliser les schémas de dépendance non gaussienne existant entre les sinistres et certaines dépenses s'y rattachant, sans toutefois faire intervenir de covariables en raison probablement de l'absence d'ensembles de données pertinentes. En ce qui concerne la régression sur des variables de réponse bidimensionnelles, le paramètre de dépendance de la variable de réponse a été étudié dans des articles comme ceux de Leon et Wu (2011) et Czado et coll. (2011). Dans ce dernier article, les auteurs ont modélisé la dépendance entre la fréquence des sinistres et la sévérité des sinistres, où l'on a supposé que les distributions marginales dépendent des valeurs des covariables mais que le paramètre de dépendance est homogène. Acar et coll. (2011) ont appliqué une méthode non paramétrique pour étalonner les paramètres de dépendance selon les covariables, où le paramètre de dépendance peut varier en fonction des covariables.

Par une analyse préliminaire des données, nous observons que, si les schémas de dépendance basés sur les différentes valeurs connues des covariables semblent être très différents, le modèle d'ajustement pourrait être amélioré par l'ajout d'une analyse de régression qui tienne compte des structures de dépendance dynamique. En ce qui concerne un certain ensemble de données d'assurance automobile sur les dommages corporels (LOSS) et les frais de règlement imputés (ALAE), nous avons remarqué que les structures de dépendance entre les sinistres et les frais de règlement imputés variaient de façon significative selon la durée de l'étude des causes à l'origine des sinistres. Plus la durée de

l'étude est longue, plus forte est la dépendance dans l'extrémité supérieure, ce qui est logique puisque une étude de longue durée peut s'expliquer par le fait que l'étude et les causes des sinistres sont plus complexes et occasionnent donc de plus grandes dépenses telles que les honoraires d'avocat, et, par conséquent, une dépendance plus forte entre les sinistres et les frais de règlement imputés. Pour des raisons de confidentialité, nous utilisons dans la figure 3 uniquement des données simulées pour illustrer l'idée, et les schémas paraissent similaires à ceux représentant des ensembles de données réelles. À la section 5, nous effectuons une analyse empirique qui repose sur un ensemble de données réelles tiré du MEPS des États-Unis.

Les différents schémas de dépendance indiquent que la structure de dépendance locale dans l'extrémité supérieure peut couvrir un très large spectre, depuis l'indépendance jusqu'à une forte dépendance positive. Toutefois, les familles de copules couramment utilisées comme la copule de Gumbel ne peuvent prendre en compte qu'un très faible éventail de schémas de dépendance d'extrémité, soit la dépendance d'extrémité usuelle. Pour surmonter cet obstacle, il nous faut considérer une famille de copules ayant une extrémité supérieure plus flexible et comportant moins de paramètres de dépendance qui peuvent être liés aux covariables.

3.2 Modèles de régression

En règle générale, le modèle proposé peut servir à modéliser des variables de réponse de n'importe quelle dimension. Nous utilisons ici le cas à deux variables pour illustrer l'idée. Les distributions marginales conditionnelles des variables de réponse Y_1 et Y_2 sont chacune modélisée par des modèles linéaires généralisés ou d'autres modèles de régression à une variable. Elles sont désignées respectivement par $F_1(\cdot|\mathbf{X}_1 = \mathbf{x}_1)$ et $F_2(\cdot|\mathbf{X}_2 = \mathbf{x}_2)$. Il est à noter que l'ensemble de covariables $\mathbf{X}_1$ et $\mathbf{X}_2$ n'est pas nécessairement le même. Mais les covariables pour le paramètre de dépendance devraient aussi être incluses dans les covariables pour les variables de réponse. Ci-après nous employons le même ensemble de covariables $\mathbf{X}$ pour simplifier la notation.

Soit $Y_i|\mathbf{X} = \mathbf{x}$, $i = 1, 2$ des variables aléatoires continues sachant que $\mathbf{X} = \mathbf{x}$ et qui ont pour fcd $F_{i|\mathbf{X}}(\cdot|\mathbf{x})$, $i = 1, 2$. La copule de ces $Y_i|\mathbf{X} = \mathbf{x}$, $i = 1, 2$ est appelée copule conditionnelle de Y_i sachant que les covariables $\mathbf{X} = \mathbf{x}$. Nous utiliserons une copule bidimensionnelle à un paramètre, telle que la copule ACIG qui balaie tout le spectre de dépendance dans l'extrémité supérieure, pour modéliser la dépendance conditionnelle entre les variables de réponse. La fdc conjointe conditionnelle est donc définie par :

$$F(y_1, y_2|\,\alpha(\mathbf{x}), \boldsymbol{\theta}_1(\mathbf{x}), \boldsymbol{\theta}_2(\mathbf{x})) = C(F_1(y_1|\boldsymbol{\theta}_1(\mathbf{x})), F_2(y_2|\boldsymbol{\theta}_2(\mathbf{x}))|\,\alpha(\mathbf{x})) \qquad (3)$$

où C (·, ·| α (**x**)) est la copule conditionnelle, $\boldsymbol{\theta}_i(\mathbf{x})$, i = 1, 2 sont les paramètres de chacune des deux distributions marginales, et α(**x**) est le paramètre de dépendance qui est aussi lié aux covariables **x**.

Vu que le paramètre de dépendance α de la copule ACIG est positif, nous utilisons log(·) comme fonction de lien. Par ailleurs, nous appliquerons aux covariables continues certaines fonctions non linéaires comme les splines cubiques naturelles, car nous ne savons pas quels sont les liens fonctionnels existant entre les covariables et le paramètre α. C'est-à-dire que :

$$\log(\alpha(\mathbf{x})) = \sum_{i=1}^{k} \mathrm{ns}_i(x_i) + \gamma_0 + \sum_{i=k+1}^{p} \gamma_i x_i,$$

où $x_1, \ldots, x_k$ sont des covariables continues, et $x_{k+1}, \ldots, x_p$ des covariables catégorielles. Les formes de $\boldsymbol{\theta}_i(\mathbf{x})$, i = 1, 2 dépendent de la façon dont les modèles de régression marginale sont choisis.

Une fois spécifiées les distributions paramétriques marginales F_1, F_2 et la copule paramétrique C, nous pouvons obtenir la fonction de densité de (3) :

$$f(y_1, y_2|\alpha(\mathbf{x}), \boldsymbol{\theta}_1(\mathbf{x}), \boldsymbol{\theta}_2(\mathbf{x})) = c(F_1(y_1|\boldsymbol{\theta}_1(\mathbf{x})), F_2(y_2|\boldsymbol{\theta}_2(\mathbf{x}))|\alpha(\mathbf{x})) \cdot f_1(y_1|\boldsymbol{\theta}_1(\mathbf{x})) \cdot f_2(y_2|\boldsymbol{\theta}_2(\mathbf{x})), \tag{4}$$

où c, f_1 et f_2 sont les densités conditionnelles correspondantes de la copule et des distributions marginales. Nous estimons ensuite les coefficients β associés aux splines naturelles et les coefficients γ_i associés aux variables discrètes en maximisant la fonction suivante de vraisemblance totale :

$$L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \alpha|y_1, y_2, \mathbf{x}) = \prod_{i=1}^{n} f(y_{1i}, y_{2i}|\alpha(\mathbf{x}_i), \boldsymbol{\theta}_1(\mathbf{x}_i), \boldsymbol{\theta}_2(\mathbf{x}_i)),$$

où (y_{1i}, y_{2i}, $\mathbf{x}_i$) est la i^e observation de l'échantillon aléatoire de taille n.

4 Étude de simulation

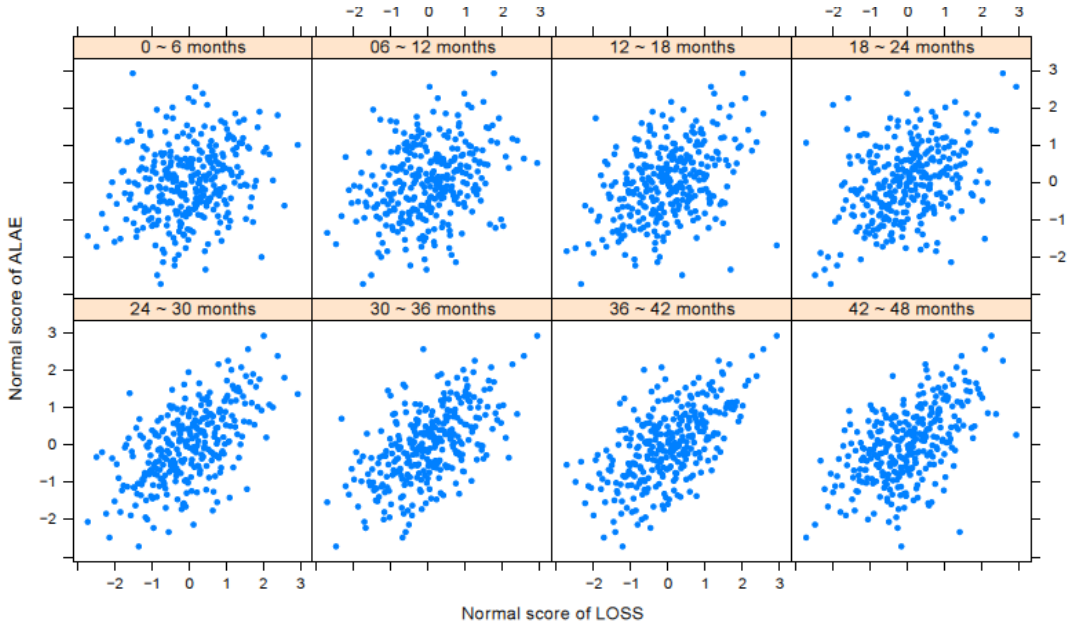
Nous appliquons ici le modèle de régression évoqué à la section 3.2 à un ensemble de données portant sur les sinistres et les frais de règlement imputés, afin d'illustrer l'idée d'utiliser une copule balayant tout le spectre de dépendance d'extrémité pour prendre en compte la dépendance dynamique entre les sinistres et les frais de règlement imputés en

fonction des covariables.

Nous générons un échantillon aléatoire de taille $n = 2\,400$ au moyen de la copule ACIG, et la taille de l'échantillon à chaque mois est de 50. On suppose que le paramètre de dépendance est fonction de la durée des sinistres exprimée en mois ($x \in \{1, \ldots, 48\}$) selon la formule suivante, et la figure 3 représente les nuages de points normalisés pour chaque période de six mois :

$$\alpha(x) = 4{,}62 - 3{,}60 \times 10^{-1}x + 1{,}33 \times 10^{-2}x^2 - 2{,}30 \times 10^{-4}x^3 + 1{,}54 \times 10^{-6}x^4. \qquad (5)$$

Figure 3 : La dépendance dans l'extrémité supérieure augmente en fonction de la durée



Nous lions maintenant le paramètre de dépendance α de la copule ACIG à la covariable « durée » X. Vu que α doit être positif, nous utilisons log() comme fonction de lien. La fonction spline cubique naturelle est utilisée pour la durée, car nous ne savons pas quels sont les liens fonctionnels entre α et X. Posons

$$\log(\alpha(x)) = \mathrm{ns}(x) = \beta_1 b_1(x) + \cdots + \beta_p b_p(x),$$

où $b_1(x), \ldots, b_p(x)$ sont les bases de la spline et les β_i les coefficients à estimer. Soit u_i, v_i, $i = 1, \ldots, n$ l'échantillon aléatoire généré par la copule ACIG. La fonction de vraisemblance L et la fonction de log-vraisemblance l, exprimées en termes du générateur archimédien ψ, sont données ci-après.

$$L(\beta_1, \ldots, \beta_p | u_i, v_i, i = 1, \ldots, n) = \prod_{i=1}^{n} c\left(u_i, v_i | \alpha_i = \exp(\beta_1 b_1(x_i) + \cdots + \beta_p b_p(x_i)\right).$$

$$\begin{aligned} & l(\beta_1, \ldots, \beta_p | u_i, v_i, i = 1, \ldots, n) \\ = & \sum_{i=1}^{n} \left\{ \log \psi''(\psi^{-1}(u_i) + \psi^{-1}(v_i)) - \log[\psi'(\psi^{-1}(u_i))] - \log[\psi'(\psi^{-1}(v_i))] \right\}. \end{aligned}$$

Nous faisons correspondre un nœud à la médiane de la « durée » de la fonction spline cubique naturelle. Nous pouvons ensuite ajuster une courbe entre le paramètre de dépendance α et la « durée », comme à la figure 4. La centaine de lignes grises sert à indiquer dans quelle mesure la droite α estimée peut varier. Ces lignes grises ont été générées à l'aide des coefficients simulés de la fonction spline cubique naturelle, et leurs valeurs ont été générées de façon aléatoire par une distribution normale à plusieurs variables ayant pour moyennes les estimations du maximum de vraisemblance des β, tandis que la matrice des covariances est l'inverse de la matrice hessienne.

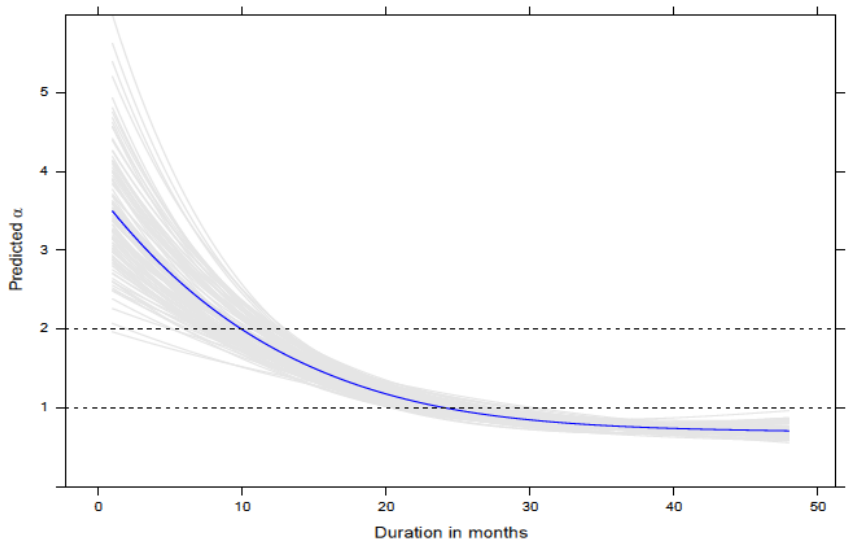
Au vu de la figure 4, nous constatons que lorsque la durée est courte (≤ 10 mois), la dépendance dans l'extrémité supérieure est proche de l'indépendance ($\alpha \geq 2$); lorsque la durée est longue (≥ 24 mois), la dépendance dans l'extrémité supérieure semble être usuelle; et entre 10 et 24 mois, il y a dépendance d'extrémité intermédiaire. Ce schéma ne peut être caractérisé par une copule d'utilisation courante telle que la copule de Gumbel.

5 Étude empirique

À la section 4, nous avons effectué une étude de simulation afin de montrer que la copule qui balaie tout le spectre de dépendance d'extrémité peut être utilisée pour prendre en compte les schémas de dépendance d'extrémité dynamique. Nous allons ici appliquer la méthode à un ensemble de données tiré du MEPS des États-Unis, où les distributions marginales et la dépendance sont liées aux covariables.

Les données du MEPS à analyser constituent un sous-ensemble des données consolidées pour l'ensemble de l'année 2010.

Figure 4 : La dépendance dans l'extrémité supérieure augmente en fonction de la durée



Les variables de réponse bidimensionnelles représentent les visites à l'urgence, ce qui comprend les dépenses associées à la facturation séparée des médecins (ERDEXP10: Y_1) et celles liées aux équipements (ERFEXP10: Y_2). Nous prenons en considération des covariables telles que l'âge (AGELAST) et la garantie d'assurance (INSCOV10). Nous excluons tous les enregistrements pour lesquels au moins une des variables de réponse est nulle. Entre autres méthodes utilisées pour traiter des variables de réponse comportant une masse en zéro (*zero-inflated*), citons une régression logistique sur l'occurrence de valeurs nulles, ainsi que les modèles tels qu'une régression de Poisson ou une régression binomiale négative comportant une masse en zéro. La taille de notre échantillon est de 2 381. On trouvera au tableau 2 des statistiques descriptives des variables à l'étude. La figure 5 montre le nuage de points normalisés correspondant aux données, où chaque marge a été transformée en une loi normale centrée réduite; c'est-à-dire que nous avons transformé chaque paire d'observations (y_{1i}, y_{2i}), $i = 1, \ldots, 2381$ en $(\Phi^{-1}(\text{rank}(y_{1i})/2381{,}5), \Phi^{-1}(\text{rank}(y_{2i})/2381{,}5)$, où Φ^{-1} désigne l'inverse de la fdc de la loi normale centrée réduite. Au vu de la figure 5, nous constatons que la dépendance dans l'extrémité supérieure est plus forte que dans l'extrémité inférieure.

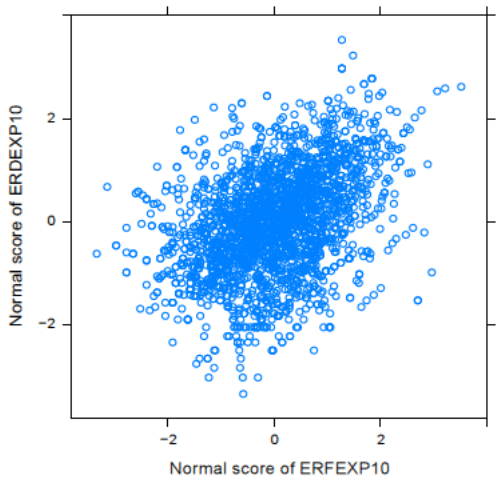
Tableau 2 : Résumé des variables

	Min	1er quantile	Médiane	Moyenne	3e quantile	Max
ERDEXP10	1	65	143	294,5	306	7 579
ERFEXP10	1	222	520	1 255	1 245	50 900
AGELAST	0	17	36	37,87	58	85

INSCOV10 (nombre d'observations)	Tout type de régime privé (1) : 1 126 Régime public seulement (2) : 1 064 Non assuré (3) : 191

Nous ajustons tout d'abord séparément les deux variables de réponse unidimensionnelles à l'aide d'une régression à une variable.

Figure 5 : Nuage de points normalisés représentant ERDEXP10 et ERFEXP10

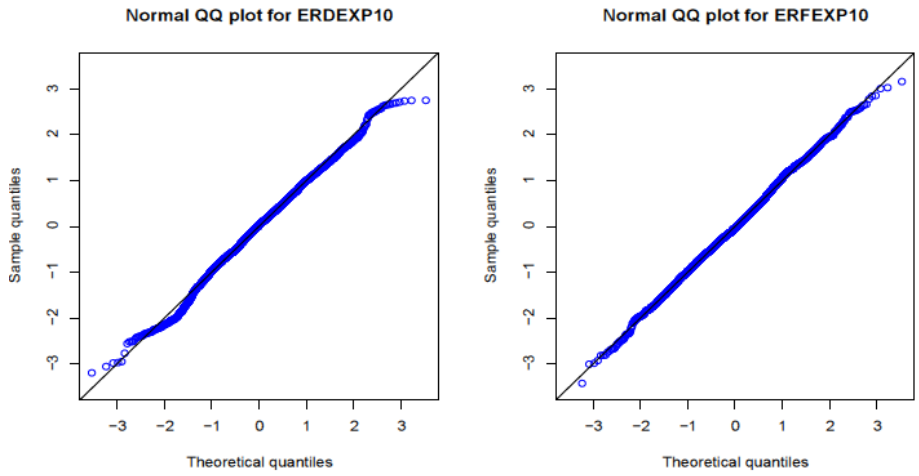


Parmi les nombreuses et différentes distributions à une variable, la distribution t de Student généralisée, caractérisée par des paramètres de position et d'échelle, ajuste très bien les données log-transformées. Si la variable aléatoire X suit une distribution t de Student standard à ν degrés de liberté, $Y := \mu + \sigma X$ suit une distribution t de Student généralisée de paramètre de position μ et de paramètre d'échelle σ. On notera que la variance de Y est donnée par $\sigma^2(\nu/(\nu - 2))$ lorsque $\nu > 2$. Nous choisissons des covariables significatives pour ajuster les données, et leur estimation du maximum de vraisemblance est indiquée au tableau 3; les écarts-types s'obtiennent à partir de la matrice d'information observée, soit l'inverse de la matrice hessienne. Aux fins de l'analyse de régression à une variable, nous pouvons utiliser un progiciel Gamlss, écrit en langage R (Rigby et Stasinopoulos, 2005), qui exécute de nombreuses distributions à une variable, dont la distribution t de Student généralisée. Nous construisons les diagrammes Quantile-Quantile normalisés résiduels afin de poser un diagnostic relativement à chaque régression marginale (voir la figure 6). Pour connaître dans le détail les procédures d'exécution des

diagrammes Q-Q normalisés résiduels, se reporter à Dunn et Smyth (1996). Selon Dunn et Smyth (1996), lorsque les paramètres sont estimés de façon cohérente, les quantiles résiduels convergent vers la loi normale centrée réduite. Au vu des diagrammes Q-Q présentés à la figure 6, la distribution *t* de Student généralisée ajuste très bien les données. Il est à noter que toutes les estimations et les comparaisons de modèles indiquées ci-après se rapportent à des variables de réponse transformées par le logarithme naturel.

Après l'ajustement des modèles à une variable, nous pouvons utiliser les estimations des coefficients des modèles de régression à une variable comme valeurs initiales des coefficients utilisés dans les modèles de régression de copules. Ensuite, nous pouvons utiliser la méthode des fonctions d'inférence sur les marginales (méthode IFM, en anglais) pour estimer les coefficients associés aux covariables liées au paramètre de dépendance. C'est-à-dire que nous allons transformer chacune des variables de réponse en leurs probabilités correspondantes obtenues à partir des fonctions de répartition marginales ajustées, puis nous ajusterons une copule à ces données transformées.

Figure 6 : Diagrammes Q-Q normalisés résiduels de la régression de la distribution *t* de Student généralisée sur les marginales



Cette méthode est rapide et donne de bonnes valeurs initiales du paramètre de dépendance au moment de maximiser la vraisemblance totale des distributions marginales et des structures de dépendance. Pour de plus amples informations sur la méthode IFM, se reporter à Joe (1997) et aux ouvrages de référence qui y sont cités. Nous employons la méthode IFM pour trouver les valeurs initiales aux fins de la maximisation, puis nous utilisons la méthode de la vraisemblance totale pour estimer simultanément les paramètres des distributions marginales et des copules. Les paramètres d'échelle obtenus par régression marginale sont conservés pour les modèles de régression de copules, vu que les

paramètres d'échelle ne devraient pas altérer les structures de dépendance des copules et que cela permet d'accélérer les calculs et de les rendre plus stables. En pratique, il est possible également d'utiliser des paramètres de position différents de ceux obtenus par régression marginale.

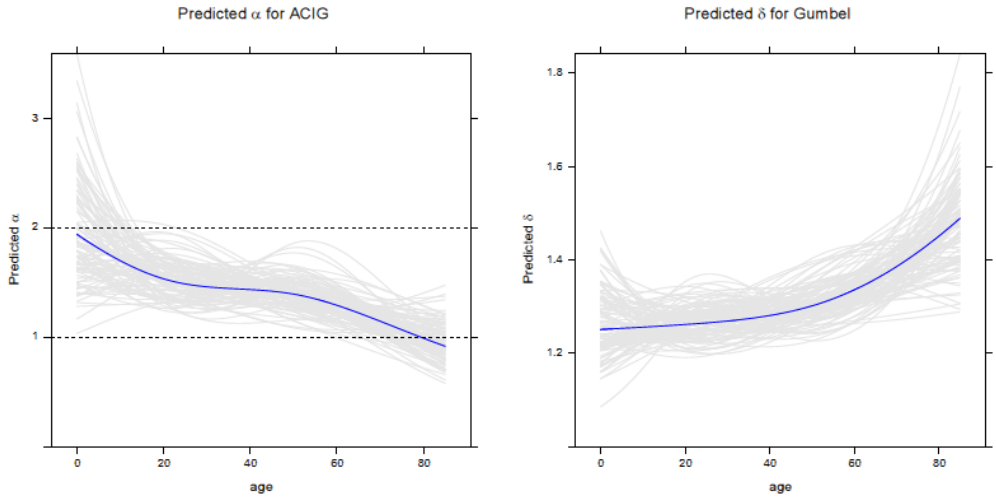
Pour montrer comment les covariables peuvent être liées au paramètre de dépendance, nous utilisons les splines cubiques naturelles de la variable « âge » afin de permettre l'existence de liens flexibles entre l'« âge » et le paramètre de dépendance. Nous choisissons les percentiles 33,3 % et 66,7 % de la variable âge comme étant les deux nœuds de la spline cubique naturelle, et il nous faut maintenant estimer quatre coefficients β_i, i = 1, 2, 3, 4 afin d'approcher la relation existant entre la covariable « âge » et le paramètre de dépendance de la copule. La vraisemblance maximale des coefficients est indiquée au tableau 3, et les écarts-types s'obtiennent à partir de l'inverse de la matrice hessienne.

La figure 7 représente une estimation de la relation existant entre l'âge et les paramètres de dépendance respectivement des modèles ACIG et de Gumbel. Au vu des graphiques, la relation est non linéaire, et les copules ACIG et de Gumbel permettent de voir que la dépendance existant entre les dépenses associées à la facturation séparée des médecins et les dépenses liées aux équipements augmente en fonction de l'âge.

Tableau 3 : Estimations relatives aux modèles ACIG, de Gumbel et de régression marginale

		ACIG	é.-t.	Gumbel	é.-t.	Marginale	é.-t.
ERD	(Point d'interception)	5,004	0,050	5,000	0,049	5,000	0,050
	âge	0,007	0,001	0,007	0,001	0,007	0,001
	ins2	-0,595	0,047	-0,585	0,047	-0,589	0,048
	ins3	-0,248	0,093	-0,241	0,090	-0,260	0,091
	ln(v)	1,847	0,093	1,878	0,092	1,919	0,146
	ln(σ)	-0,008	–	-0,008	–	-0,008	0,026
ERF	(Point d'interception)	6,152	0,054	6,155	0,054	6,144	0,054
	âge	0,012	0,001	0,012	0,001	0,012	0,001
	ins2	-0,703	0,052	-0,699	0,051	-0,679	0,052
	ins3	-0,123	0,098	-0,136	0,097	0,099	0,097
	ln(v)	2,572	0,167	2,516	0,155	2,716	0,272
	ln(σ)	0,129	–	0,129	–	0,129	0,023
Dépendance	β_1	0,226	0,241	-1,122	0,263	–	–
	β_1	0,160	0,221	-0,657	0,237	–	–
	β_3	0,813	0,263	-2,569	0,258	–	–
	β_4	-0,557	0,247	0,369	0,247	–	–

Figure 7 : La dépendance dans l'extrémité supérieure est fluctuante : les modèles ACIG et de Gumbel prennent tous deux en compte la structure dynamique, et le modèle ACIG peut même prendre en compte les cas de dépendance d'extrémité intermédiaire.



Afin de comparer les modèles de régression ACIG et de Gumbel, nous pouvons appliquer le test de Vuong (Vuong, 1989) à ces deux modèles non imbriqués. La statistique du test se définit comme suit :

$$Z := \frac{\sqrt{n} \times \overline{m}}{\sqrt{\sum_{i=1}^{n}(m_i - \overline{m})^2}},$$

où n est la taille de l'échantillon, $m_i = l_i^A - l_i^G$, $i = 1, \ldots, n$, représente la différence entre la log-vraisemblance ponctuelle des modèles ACIG et de Gumbel, et $\overline{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$. La statistique du test, Z, est asymptotiquement distribuée suivant une loi normale centrée réduite lorsque $n \to \infty$. Pour cet ensemble de données, $Z = -0{,}086$, ce qui indique l'absence de différence significative entre les modèles ACIG et de Gumbel sur le plan de l'ajustement global; ces deux modèles sont assez similaires dans le sens du test de Vuong. Toutefois, lorsqu'il s'agit d'analyser des scénarios à risque élevé pour lesquels les risques se produisent dans les extrémités supérieures des distributions, l'utilisation d'un critère pour l'ajustement global n'est peut-être pas convenable pour déterminer le meilleur modèle. Un critère que nous pouvons utiliser est l'erreur quadratique moyenne (EQM) au-delà d'un certain percentile. Si $\hat{y}_i$, $i = 1, \ldots, n$, sont les valeurs prévues et y_i, $i = 1, \ldots, n$, sont les valeurs observées, une EQM au-delà d'un percentile p ($p \in [0, 1]$) peut se définir par

$$\mathrm{MSE}_p = \frac{1}{|I_p|}\sum_{i \in I_p}(\hat{y}_i - y_i)^2,$$

où $I_p := \{i \in \{1, \ldots, n\} : y_i > \text{VaR}_p(y_i, i = 1, \ldots, n)$, le percentile p des $y_i\}$, et $|I_p|$ est le nombre d'éléments de l'ensemble indiciel I_p. La valeur de l'EQM$_p$ indique dans quelle mesure un modèle ajuste bien l'extrémité au-delà du percentile p.

Pour cet ensemble de données, bien que le test de Vuong ne puisse indiquer quel est le meilleur modèle, nous pouvons utiliser l'EQM$_p$ pour comparer les modèles et conclure que le modèle de régression ACIG modélise mieux l'extrémité supérieure et produit une EQM$_p$ plus petite lorsque p est suffisamment grand. Les résultats, présentés au tableau 4, nous indiquent que : 1) le modèle ACIG est généralement meilleur que le modèle de Gumbel pour analyser les scénarios à risque élevé, et les avantages sont plus nets lorsque p est relativement grand; 2) les modèles de régression avec ACIG peuvent améliorer l'ajustement des extrémités supérieures pour chaque distribution marginale, ce qui justifie l'avantage de combiner deux modèles de régression à une variable au moyen d'une copule bien choisie; 3) l'avantage du modèle ACIG devient relativement plus important lorsqu'il s'agit d'analyser des risques dépendants agrégés.

On notera que, selon l'EQM$_p$, l'amélioration permise par le modèle de Gumbel par rapport au modèle indépendant est faible, ce qui laisse penser que l'évaluation de la somme des deux variables de réponse de l'ensemble de données n'est pas très sensible aux structures de dépendance existant entre les variables de réponse. Cela peut se produire du fait que l'évaluation dépend non seulement des structures de dépendance, mais également des distributions marginales. Cela dit, l'amélioration permise par le modèle ACIG par rapport au modèle de Gumbel est relativement plus importante, quoique la différence des EQM soit due à la propriété de la copule ACIG de balayer tout le spectre de dépendance de l'extrémité, de sorte que l'extrémité supérieure puisse être mieux prise en compte. Vu que les deux variables de réponse dépendent l'une de l'autre et que leurs liens de dépendance ne peuvent être complètement expliqués par les covariables disponibles, une modélisation adéquate de leur structure de dépendance pourrait mieux expliquer le comportement de chaque variable de réponse.

Tableau 4 : EQM_p relatives aux modèles ACIG, de Gumbel et de régression marginale. L'EQM_p relative au modèle ACIG est généralement plus petite, particulièrement pour l'évaluation de la somme des dépenses.

	p	0,1	0,3	0,5	0,7	0,9	0,95	0,99	0,995	0,999
	Nombre d'observations	2 142	1 666	1 190	714	238	118	24	12	3
ERD	ACIG	0,92	0,93	1,23	1,93	4,12	6,00	11,98	13,55	14,27
	Gumbel	0,93	0,94	1,24	1,96	4,16	6,04	12,05	13,63	14,35
	Indép.	0,93	0,95	1,25	1,97	4,17	6,06	12,10	13,68	14,35
ERF	ACIG	1,13	1,15	1,48	2,30	4,57	6,36	10,86	12,66	17,60
	Gumbel	1,13	1,15	1,49	2,31	4,59	6,39	10,91	12,71	17,70
	Indép.	1,13	1,16	1,49	2,33	4,63	6,43	10,93	12,74	17,65
ERD	ACIG	3,08	3,05	3,88	5,99	12,59	17,80	33,56	40,76	57,56
+	Gumbel	3,08	3,07	3,92	6,05	12,72	17,95	33,78	40,98	57,92
ERF	Indép.	3,08	3,07	3,94	6,09	12,79	18,04	33,86	41,07	58,00

Nous pouvons utiliser le diagramme Q-Q résiduel (Dunn et Smyth (1996)) pour évaluer l'ajustement de la quantité d'intérêt. Nous considérons ici le diagramme Q-Q de la somme de ERDEXP10 et ERFEXP10. Soit $S \stackrel{d}{=} Y_1 + Y_2$, où Y_1 et Y_2 sont des variables aléatoires continues ayant pour support $(-\infty, \infty)$. Soit F la fonction de répartition de S, F_1 et F_2 les fonctions de répartition respectives de Y_1 et Y_2, et $C(u_1, u_2)$ la copule pour Y_1 et Y_2. Posons

$$F(s) = \int_0^1 C_{1|2}\left(F_1(s - F_2^{-1}(u)), u\right) du \tag{6}$$

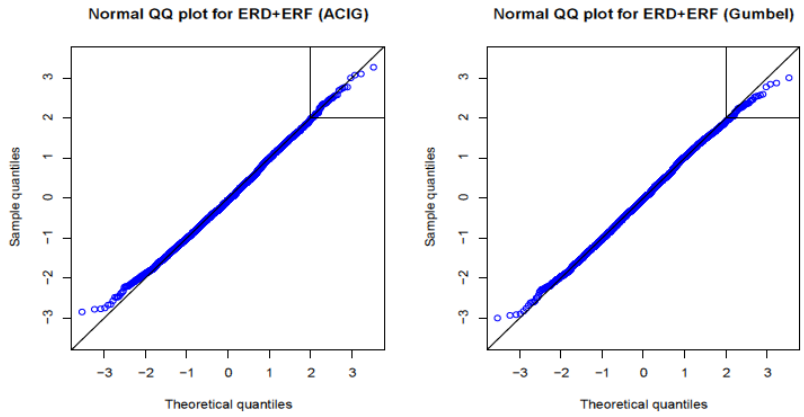
où $C_{1|2}(u_1, u_2) = \partial C(u_1, u_2)/\partial u_2 = \psi'(\psi^{-1}(u_1) + \psi^{-1}(u_2))/\psi'(\psi^{-1}(u_2))$.

À l'aide de (6) et des estimations des paramètres de la copule et des distributions marginales, nous pouvons évaluer les $\hat{F}(s)$ estimées. Ensuite, pour chaque paire de variables de réponse observées, nous pouvons calculer la somme des deux valeurs, désignées par s_i, $i = 1, \ldots, n$, où n est la taille de l'échantillon. Puis nous insérons les s_i dans $z_i = \Phi^{-1}\hat{F}(s_i))$ pour obtenir les z_i. La représentation des z_i en fonction de la loi normale centrée réduite nous conduit vers le diagramme Q-Q normalisé résiduel pour la somme de Y_1 et Y_2. Dans un contexte de régression, la fonction de répartition F, la copule C et les distributions marginales dans (6) peuvent être simplement remplacées par leurs versions conditionnelles qui dépendent des valeurs des covariables, et le diagramme Q-Q demeure toujours valable. Pour plus de détails, se reporter à Dunn et Smyth (1996). L'intégrale dans (6) peut être calculée numériquement.

Les diagrammes Q-Q normalisés résiduels pour la somme de ERDEXP10 et ERFEXP10 sont présentés à la figure 8 et nous indiquent que : 1) les performances globales des

modèles ACIG et de Gumbel sont similaires; 2) le modèle de régression basé sur la copule ACIG donne de meilleurs résultats dans l'extrémité supérieure, là où les risques se produisent, tandis que le modèle de Gumbel donne lieu à une extrémité plus lourde de la somme, ce qui surestime les sinistres; 3) en ce qui concerne le modèle ACIG, le meilleur ajustement dans l'extrémité supérieure a un prix, à savoir le relativement moins bon ajustement dans l'extrémité inférieure; toutefois, celle-ci ne revêt pas d'importance pour l'évaluation des risques.

Figure 8 : Diagrammes Q-Q normalisés résiduels pour ERD+ERF



6 Mot de la fin

Le projet de recherche avait pour but d'observer les structures de dépendance dynamique entre les sinistres et les frais de règlement imputés à partir d'un ensemble de données d'assurance automobile sur les dommages corporels. Lorsque des covariables sont disponibles, on peut les utiliser pour expliquer non seulement les distributions marginales, mais aussi la structure de dépendance elle-même. Étant donné que la dépendance dans l'extrémité supérieure peut être relativement faible ou forte sachant les différentes valeurs des covariables, il vaut mieux utiliser une copule balayant le plus large spectre de dépendance dans l'extrémité supérieure lorsqu'il s'agit de modéliser la structure de dépendance entre ces deux variables de réponse, et laisser les données révéler le degré de dépendance. La copule de Gumbel d'utilisation courante balaie la dépendance d'extrémité usuelle. En d'autres termes, quel que soit le degré de dépendance dans l'extrémité supérieure des données, une copule de Gumbel ajustée révèle toujours une dépendance locale relativement plus forte dans l'extrémité supérieure. À cette fin, le modèle ACIG est un bon candidat car l'extrémité supérieure est très flexible, et le seul paramètre de dépendance de la copule prend en compte l'information la plus importante qui nous intéresse, à savoir

la dépendance dans l'extrémité supérieure, à l'origine de la différence des structures globales de dépendance.

Un critère statistique basé sur l'ajustement global ne convient pas toujours lorsqu'il s'agit de comparer des modèles servant à analyser des scénarios à risque élevé. Dans notre étude empirique, le modèle de régression basé sur la copule ACIG, bien qu'il ne donne pas de meilleurs résultats en termes d'ajustement global, produit un meilleur ajustement pour l'extrémité supérieure, surtout lorsqu'il est utilisé pour évaluer le risque extrême que représentent les sinistres dépendants agrégés.

Les covariables expliquent bien la dépendance existant entre les variables de réponse, par leurs effets sur les distributions marginales. Lorsqu'elles ne peuvent le faire complètement, il pourrait être utile de procéder à une régression sur les paramètres de dépendance. Par ailleurs, une forte structure de dépendance d'extrémité qui peut être observée marginalement en l'absence de condition sur les covariables peut quand même nécessiter une copule pour la dépendance d'extrémité relativement plus faible pour pouvoir effectuer la modélisation lorsque des covariables sont disponibles. Par exemple, l'ensemble bien connu de données susmentionné sur les sinistres et les frais de règlement imputés, qui a fait l'objet d'études en actuariat, semble afficher une dépendance d'extrémité supérieure relativement plus forte. Or, si des covariables sont disponibles pour cet ensemble de données, la dépendance d'extrémité supérieure relativement plus forte pourrait s'expliquer en partie par des effets communs des covariables sur les distributions marginales, et il pourrait être nécessaire d'avoir une copule pour la dépendance d'extrémité plus faible. À cette fin, il serait utile d'avoir une copule balayant tout le spectre de dépendance d'extrémité, telle que la copule ACIG. Selon le degré de dépendance d'extrémité pouvant être partiellement expliqué par les covariables, la copule utilisée pour prendre en compte la dépendance dans les extrémités devrait être assez flexible pour caractériser les structures de dépendance dynamique, qui pourraient être très faibles ou très fortes.

Le modèle de régression peut aussi être étendu au cas où les variables de réponse ont plus de deux dimensions. La tâche cruciale consiste alors à bien modéliser la dépendance entre les variables de réponse, ce qui est souvent difficile lorsque le nombre de dimensions devient important. La modélisation de la dépendance dans le cas de variables aléatoires à plusieurs dimensions connaît aujourd'hui un grand essor. Le lecteur intéressé à prendre connaissance de quelques-unes des nouvelles méthodes est invité à se renseigner sur la copule *Vine* (Bedford et Cooke, 2002; Kurowicka et Joe, 2011), la copule à facteurs (Krupskii et Joe, 2013; Oh et Patton, 2012), et les copules utilisées dans les modèles graphiques probabilistes (Elidan, 2013).

7 Considérations d'ordre numérique

La vitesse de calcul revêt souvent de l'importance au moment d'instaurer un nouveau modèle statistique. Les fonctions R (avec routines C) ont été utilisées pour les modèles de régression avec la copule ACIG et la copule de Gumbel. Une fois la copule ACIG mise en application à l'aide des codes du langage C, les calculs s'y rapportant gagnent nettement en vitesse, et la performance globale est satisfaisante pour traiter des applications réelles.

Nous présentons ci-après des informations numériques prises en considération au moment de la mise en application des modèles. Afin d'améliorer la vitesse pour les fonctions utilisées avec la copule ACIG, nous nous sommes servis des codes du langage C pour les très longs calculs.

7.1 Copule ACIG

1. Pour obtenir l'inverse de la transformation de Laplace de la loi inverse-gamma, nous prenons le logarithme naturel des deux membres de $\psi(s) = t$ pour $0 < t \leq 1$, puis nous appliquons la méthode de Newton pour trouver la racine, la raison étant que le logarithme d'une fonction gamma peut traiter de grands arguments, alors que la fonction gamma elle-même comporte souvent des erreurs numériques pour les grands arguments.

$$\psi(s;\alpha) = \frac{2}{\Gamma(\alpha)} s^{\alpha/2} K_\alpha(2\sqrt{s}), \quad s \geq 0,\ \alpha > 0.$$

En conséquence, il reste à résoudre

$$g(s) := \frac{\alpha}{2}\ln(s) + \ln(K_\alpha(2\sqrt{s})) - \ln(t) - \ln(\Gamma(\alpha)) + \ln(2) = 0.$$

Par ailleurs,

$$\begin{aligned} g'(s) &:= \frac{\alpha}{2s} - \frac{K_{\alpha-1}(2\sqrt{s}) + K_{\alpha+1}(2\sqrt{s})}{2\sqrt{s}K_\alpha(2\sqrt{s})} \\ &= \frac{\alpha}{2s} - \exp\{\ln(K_{\alpha-1}(2\sqrt{s})) - \ln(2\sqrt{s}) - \ln(K_\alpha(2\sqrt{s}))\} \\ &\quad - \exp\{\ln(K_{\alpha+1}(2\sqrt{s})) - \ln(2\sqrt{s}) - \ln(K_\alpha(2\sqrt{s}))\}. \end{aligned}$$

Ici les fonctions $\ln K_\alpha()$ et $\ln \Gamma()$ peuvent traiter des arguments relativement plus grands et sont plus stables numériquement.

2. En ce qui concerne la fonction de densité conjointe de F_1 et F_2 où la dépendance est modélisée par la copule ACIG C, vu que

$$\psi'(s) = -2s^{(\alpha-1)/2}K_{\alpha-1}(2\sqrt{s})/\Gamma(\alpha);$$
$$\psi''(s) = 2s^{(\alpha-2)/2}K_{\alpha-2}(2\sqrt{s})/\Gamma(\alpha),$$

nous posons $s := \psi^{-1}(F_1(x_1)) + \psi^{-1}(F_2(x_2))$, $s_1 := \psi^{-1}(F_1(x_1))$ et $s_2 := \psi^{-1}(F_2(x_2))$, et nous obtenons

$$\begin{aligned}\ln f(x_1,x_2) &= \ln\psi''(s) + \ln f_1(x_1) + \ln f_2(x_2) - \ln(-\psi'(s_1)) - \ln(-\psi'(s_2))\\ &= \ln K_{\alpha-2}(2\sqrt{s}) - \ln K_{\alpha-1}(2\sqrt{s_1}) - \ln K_{\alpha-1}(2\sqrt{s_2})\\ &\quad + \frac{\alpha-2}{2}\ln(s) - \frac{\alpha-1}{2}[\ln(s_1)+\ln(s_2)]\\ &\quad + \ln f_1(x_1) + \ln f_2(x_2) - \ln(2) + \ln\Gamma(\alpha),\end{aligned}$$

où ln $K\alpha()$ et ln $\Gamma()$ peuvent traiter des arguments relativement plus grands et sont plus stables numériquement.

3. En ce qui concerne $C_{1|2}(u_1, u_2)$, nous supposons que $s_1 = \psi^{-1}(u_1)$, $s_2 = \psi^{-1}(u_2)$, et s = s_1 + s_2, puis nous appliquons le logarithme

$$\begin{aligned}\log C_{1|2}(u_1,u_2) &= \log(-\psi'(s)) - \log(-\psi'(s_2))\\ &= \frac{\alpha-1}{2}(\log(s) - \log(s_2)) + \log K_{\alpha-1}(2\sqrt{s}) - \log K_{\alpha-1}(2\sqrt{s_2})\end{aligned}$$

7.2 Copule de Gumbel

1. Dans le cas de la copule de Gumbel, ψ (transformée de Laplace) possède les propriétés suivantes :

$$\begin{aligned}\psi(s) &= \exp\{-s^{1/\theta}\}, \quad \theta \geq 1;\\ \psi'(s) &= -\exp\{-s^{1/\theta}\}s^{1/\theta-1}/\theta, \quad \theta \geq 1;\\ \psi''(s) &= \exp\{-s^{1/\theta}\}\left(s^{1/\theta-1}/\theta\right)^2 - \exp\{-s^{1/\theta}\}(1/\theta-1)s^{1/\theta-2}/\theta, \quad \theta \geq 1;\\ \psi^{-1}(s) &= (-\ln(s))^{\theta}, \quad \theta \geq 1.\end{aligned}$$

Maintenant, nous posons $s := \psi^{-1}(F_1(x_1))+\psi^{-1}(F_2(x_2))$, $s_1 := \psi^{-1}(F_1(x_1))$ et $s_2 := \psi^{-1}(F_2(x_2))$, et nous obtenons

$$\begin{aligned}\ln f(x_1, x_2) &= \ln \psi''(s) + \ln f_1(x_1) + \ln f_2(x_2) - \ln(-\psi'(s_1)) - \ln(-\psi'(s_2)) \\ &= -s^{1/\theta} + \ln\left[s^{2/\theta-2}/\theta^2 - (1/\theta^2 - 1/\theta)s^{1/\theta-2}\right] + \ln f_1(x_1) + \ln f_2(x_2) \\ &\quad - \ln(F_1(x_1)) + 2\ln(\theta) - (1-\theta)[\ln(-\ln(F_1(x_1)))] \\ &\quad - \ln(F_2(x_2)) - (1-\theta)[\ln(-\ln(F_2(x_2)))].\end{aligned}$$

2. En ce qui concerne $C_{1|2}(u_1, u_2)$, nous supposons que $s_1 = \psi^{-1}(u_1)$, $s_2 = \psi^{-1}(u_2)$, et $s = s_1 + s_2$, et nous obtenons comme suit le logarithme naturel :

$$\log C_{1|2}(u_1, u_2) = \log(-\psi'(s)) - \log(-\psi'(s_2)) = (1 - 1/\theta)[\log(s_2) - \log(s)] + s_2^{1/\theta} - s^{1/\theta}.$$

Ouvrages de référence

Acar, E., R. Craiu, et F. Yao. (2011). « Dependence calibration in conditional copulas: A nonparametric approach », *Biometrics*, vol. 67, 2011, p. 445-453.

Bedford, T., et R. M. Cooke. « Vines—a new graphical model for dependent random variables », *The Annals of Statistics*, vol. 30, nº 4, 2002, p. 1031-1068.

Czado, C., R. Kastenmeier, E. Brechmann, et A. Min. « A Mixed Copula Model for Insurance Claims and Claim Sizes », *Scandinavian Actuarial Journal*, 2011.

de Leon, A., et B. Wu. « Copula-based regression models for a bivariate mixed discrete and continuous outcome », *Statistics in Medicine*, vol. 30, 2011, p. 175-185.

Donnelly, C., et P. Embrechts. « The devil is in the tails: actuarial mathematics and the subprime mortgage crisis », *Astin Bulletin*, vol. 40, nº 1, 2010, p. 1-33.

Dunn, P. K., et G. K. Smyth. « Randomized quantile residuals », *Journal of Computational and Graphical Statistics*, vol. 5, nº 3, 1996, p. 236-244.

Elidan, G. « Copulas in machine learning », *Copulae in Mathematical and Quantitative Finance*, Springer, 2013, p. 39-60.

Frees, E. W., et E. A. Valdez. « Understanding relationships using copulas », *North American Actuarial Journal*, vol. 2, nº 1, 1998, p. 1-25.

Hua, L. « Tail negative dependence and its applications for aggregate loss modeling », *Technical Report*, 2013.

Hua, L., et H. Joe, H. « Tail order and intermediate tail dependence of multivariate copulas », *Journal of Multivariate Analysis*, vol. 102, 2011, p. 1454-1471.

Hua, L., et H. Joe. « Intermediate tail dependence: a review and some new results », dans H. Li et X. Li, éditeurs, *Stochastic Orders in Reliability and Risk: In Honor of Professor Moshe Shaked*, Springer, 2013, chapitre 15, p. 291-311.

Hua, L., et H. Joe. « Strength of tail dependence based on conditional tail expectation », *Journal of Multivariate Analysis,* 2014, p. 143-159.

Joe, H. « Multivariate Models and Dependence Concepts », *Monographs on Statistics and Applied Probability*, vol. 73, Chapman & Hall, London, 1997.

Klugman, S. A., et R. Parsa. « Fitting bivariate loss distributions with copulas », *Insurance: Mathematics and Economics*, vol. 24, nº 1 et 2, 1999, p. 139-148. Travaux du 1er congrès IME, Amsterdam, 1997.

Krupskii, P., et H. Joe. « Factor copula models for multivariate data », *Journal of Multivariate Analysis*, vol. 120, 2013, p. 85-101.

Kurowicka, D., et H. Joe. *Dependence Modeling: Vine Copula Handbook*, World Scientific Publishing Company, Singapour, 2011.

Oh, D., et A. Patton. « Modelling dependence in high dimensions with factor copulas », *Technical report*, université Duke, 2012.

Rigby, R. A., et D. M. Stasinopoulos. « Generalized additive models for location, scale and shape » (suivi d'une analyse), *Applied Statistics*, vol. 54, 2005, p. 507-554.

Vuong, Q. H. « Likelihood ratio tests for model selection and non-nested hypotheses », *Econometrica: Journal of the Econometric Society*, 1989, p. 307-333.