Estimation des paramètres d'une copule pour un modèle de risque collectif

Sous la supervision de
Étienne Marceau

Rapport des travaux réalisés

Préparé par

Alexandre Lepage,
Diamilatou N'diaye,
Amedeo Zito

le 06 juin 2019

Faculté des sciences et de génie
École d'actuariat
Université Laval
Automone 2018

Table des matières

Liste des illustrations

Liste des tableaux

1 Introduction

Le présent projet consiste à estimer les paramètres d'un modèle collectif du risque incorporant une structure de dépendance entre la variable aléatoire de dénombrement (fréquence) et les variables aléatoires représentant les montants individuels de sinistre.

Tout d'abord, la section sur les notions préliminaires expose le modèle de risque avec dépendance, puis introduit la notion d'estimation de paramètres par la méthode du maximum de vraisemblance. Par la suite, on présente les copules archimédiennes hiérarchiques. Finalement, la dernière section expose les résultats d'estimations sur des données simulées.

2 Notions préliminaires

2.1 Modèle collectif du risque

Dans [Cossette et al., 2019b], on propose un modèle collectif du risque dont les composantes sont dépendantes entre elles. Ce modèle s'exprime comme suit.

Soient N, la v.a. du nombre de sinistres, tel que $N \in \mathbb{N}$, et la suite de v.a. $\{X_i, i \in \mathbb{N}_+\}$ représentant les montants de sinistres tel que $X_i \sim X \in \mathbb{R}_+$. On a

$$S = \sum_{i=1}^{\infty} X_i \times \mathbb{1}_{\{N \geq i\}}, \tag{1}$$

où $\{N, X_1, X_2, \ldots, X_k\}$ est une suite de v.a. dépendantes. Pour chaque $k \in \mathbb{N}_+$, la loi multivariée de $(N, X_1, X_2, \ldots, X_k)$ est définie par une copule archimédienne hiérarchique.

Maintenant, on cherche à estimer les paramètres afférents à un tel modèle. Pour y arriver, [Cossette et al., 2019a] propose une approche de calcul de la vraisemblance par décomposition hiérarchique et une autre méthode, dite plus classique, qui prend l'ensemble du modèle. Dans le cas présent, compte tenu que la variable dénombrante est discrète et que les variables de sinistres sont continues, il est plus simple d'utiliser la méthode de vraisemblance complète. Le présent rapport présente donc les résultats d'une telle approche.

2.2 Fonction du maximum de vraisemblance

Soit une copule de Clayton multivariée dénotée C et représentée par

$$C(u_0, \ldots, u_k; \alpha) = (u_0^{-\alpha} + \cdots + u_k^{-\alpha} - k)^{-\frac{1}{\alpha}}, \tag{2}$$

$k \in \mathbb{N}_+$. Avec le modèle posé dans la section 2.1, on peut trouver la fonction de densité de la copule C de la manière décrite comme suit.

Soient λ et β les paramètres des lois de fréquence et de sévérité respectivement, ainsi que α, le paramètre de la copule de Clayton. La densité conjointe des v.a. N et $(X_1, \ldots X_N)$ est donnée par

$$\begin{aligned} f_{N,X_1,\ldots,X_n}(n, x_1, \ldots, x_n; \lambda, \beta, \alpha) = {} & \frac{\partial^n}{\partial x_1 \ldots \partial x_n} C(F_N(n; \lambda), F_{X_1}(x_1; \beta), \ldots, F_{X_n}(x_n; \beta); \alpha) \\ & - \frac{\partial^n}{\partial x_1 \ldots \partial x_n} C(F_N(n-1; \lambda), F_{X_1}(x_1; \beta), \ldots, F_{X_n}(x_n; \beta); \alpha), \end{aligned} \tag{3}$$

$n \in \mathbb{N}_+$. On observe que $f_{N,X_1,\ldots,X_n}$ n'est pas une fonction de densité au sens strict du terme, car N est une v.a. discrète et $X_1, \ldots, X_n$ sont des v.a. continues.

Puisque la dérivation en chaîne nécessaire pour trouver (3) est excessivement longue à calculer, on peut utiliser `R` à l'aide de la fonction `Deriv` [1] du package du même nom pour y arriver. Cependant, il est

1. https ://cran.r-project.org/web/packages/Deriv/Deriv.pdf

important de noter que, même avec cet outil, si la loi de fréquence admet des valeurs supérieures à 5, le temps de calcul peut grimper très rapidement.

Soient n_j, le nombre d'observations où $N = j$, n_{total}, le nombre d'observations total tel que $\sum_{j=1}^{k} n_j = n_{total}$, et $x_{i,j}$, le j-ième sinistre de la i-ième observation, la fonction de vraisemblance du modèle énoncé dans la section 2.1 est donné par

$$\begin{aligned}\mathcal{L}(\lambda, \beta, \alpha) = (\Pr(N = 0; \lambda))^{n_0} & \\ \times \prod_{i=1}^{n_1} f_{N,X_1}(1, x_{i,1}; \lambda, \beta, \alpha) & \\ \times \prod_{i=1}^{n_2} f_{N,X_1,X_2}(2, x_{i,1}, x_{i,2}; \lambda, \beta, \alpha) & \\ \dots & \\ \times \prod_{i=1}^{n_k} f_{N,X_1,\dots,X_k}(k, x_{i,1}, \dots, x_{i,k}; \lambda, \beta, \alpha). &\end{aligned}$$

Par la suite, afin d'estimer les paramètres, il faut avoir recours à une méthode d'optimisation numérique afin de maximiser le log de la fonction de vraisemblance.

On obtient

$$\begin{aligned}\ell(\lambda, \beta, \alpha) = n_0 \times \ln(\Pr(N = 0; \lambda)) & \\ + \sum_{i=1}^{n_1} \ln\left(f_{N,X_1}(1, x_{i,1}; \lambda, \beta, \alpha)\right) & \\ + \sum_{i=1}^{n_2} \ln\left(f_{N,X_1,X_2}(2, x_{i,1}, x_{i,2}; \lambda, \beta, \alpha)\right) & \\ \dots & \\ + \sum_{i=1}^{n_k} \ln\left(f_{N,X_1,\dots,X_k}(k, x_{i,1}, \dots, x_{i,k}; \lambda, \beta, \alpha)\right). &\end{aligned} \tag{4}$$

Finalement, on calcule numériquement les paramètres en utilisant la fonction `R` `constrOptim`[2]. Cette dernière cherche à minimiser une fonction passée en argument. La fonction qui sera donc insérée en argument sera le négatif de (4). Afin de s'assurer que le processus d'optimisation numérique soit le plus précis possible, il faut entrer en argument de cette fonction des valeurs initiales adéquates pour chacun des paramètres à estimer. Pour ce faire, la méthode du maximum de vraisemblance de la distribution marginale de chacune des variables est un choix judicieux. Dans le cas où cette méthode s'avère trop laborieuse, on peut compenser avec la méthode des moments ou des quantiles (voir [Klugman et al., 2012]).

2.3 Copule archimédienne hiérarchique

Comme les copules archimédiennes hiérarchiques offrent une bonne flexibilité dans la modélisation de la dépendance, la présente section se penchera sur ce sujet d'intérêt.

Les copules archimédiennes hiérarchiques avec des distributions multivariées composées sont décrites dans [Cossette et al., 2017]. La représentation graphique du modèle collectif du risque tel que décrit dans la section 2.1 et expliquée dans [Cossette et al., 2019b] est présentée dans l'illustration 1.

2. https ://stat.ethz.ch/R-manual/R-devel/library/stats/html/constrOptim.html

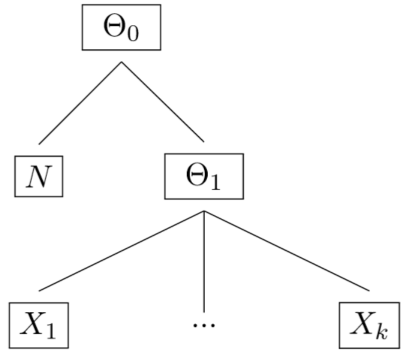


Illustration 1 – Arbre hiérarchique à un niveau.

Ici, Θ_0 représente une variable aléatoire définie sur $\mathbb{N}_+$ qui sert à créer un lien de dépendance entre la variable N et les $\{X_i, i = 1, \ldots, k\}$. Ensuite, on a $\theta_1 = \sum_{i=1}^{\theta_0} B_i$ qui sert à créer un lien de dépendance entre les X_i. Les B_i sont i.i.d. et indépendants de θ_0. B_i peut appartenir à R_+ comme à N_+.

Sous cette représentation, on obtient une copule archimédienne hiérarchique s'exprimant comme

$$C(u_0, u_1, \ldots, u_n) = \mathscr{L}_{\theta_0}\left(\mathscr{L}_{\theta_0}^{-1}(u_0) - \ln\left(\mathscr{L}_B\left(\sum_{i=1}^{n} \mathscr{L}_B^{-1}\left(\exp\left(-\mathscr{L}_{\theta_0}^{-1}(u_i)\right)\right)\right)\right)\right), \tag{5}$$

où $\mathscr{L}$ correspond à la transformée de Laplace-Stieltjes d'une variable aléatoire.

À ce point-ci, afin de trouver les paramètres d'une telle copule, il suffit d'appliquer (3) et (4), puis d'optimiser numériquement.

3 Résultats

La présente section explique les scénarios testés ainsi que les résultats obtenus avec la méthodologie expliquée dans la section 2.2.

Pour les fins du présent travail, les estimations sont faites sur des données simulées. Pour faire ces simulations, le module `R` nommé `copula` [3] offre une fonction `rCopula` avec laquelle, il est possible de simuler différentes copules connues, dont la copule de Clayton.

Pour les copules archimédiennes hiérarchique, le module `nCopula` [4] offre une fonction nommée `rCompCop` qui permet de simuler de la même façon que `rCopula`, mais en entrant comme argument la structure hiérarchique.

3.1 Copule de Clayton

Scénario 1 ***Binomial :***

Pour débuter gentiment, prenons un modèle binomial-exponentiel avec une copule de Clayton dont la représentation est exprimée en (2). Soient $N \sim Binom(n, q)$ et $X_i \sim X \sim Exp(\beta)$ et le paramètre de dépendance de la copule est dénotée α.

Pour le premier scénario, on utilise des données simulées avec les paramètre initiaux présentés dans le tableau 1.

3. https ://cran.r-project.org/web/packages/copula/copula.pdf
4. https ://cran.r-project.org/web/packages/nCopula/nCopula.pdf

n	q	α	β
5.00	0.40	6.00	0.01

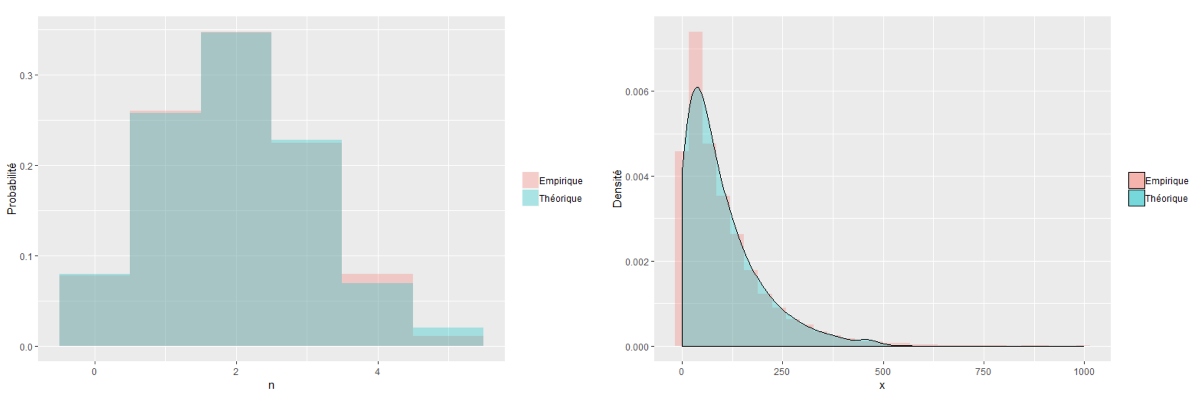
Tableau 1 – Paramètres initiaux pour la copule de Clayton avec $N \sim Bin(n, q),\ X_i \sim X \sim Exp(\beta)$.

Le sommaire des données simulées aux fins de l'estimation sont présentées dans le tableau 2. On y voit que la moyenne du nombre de sinistres est de 2 et que la moyenne des x_i est très proche de 100 ; ce qui correspond aux attentes. De plus, les quantiles des X_i sont très proches l'un de l'autre ; ce qui signifie que les variables X_i sont identiquement distribués. Visuellement parlant, l'illustration 2 présente l'adéquation des données empiriques avec les lois théoriques. On voit donc que les variables simulées ont un comportement très similaire à la distribution marginale théorique des variables.

N	X_1	X_2
Min. :0.000	Min. : 0.0046	Min. : 0.0035
1st Qu. :1.000	1st Qu. : 28.5104	1st Qu. : 28.2379
Median :2.000	Median : 68.3689	Median : 68.8120
Mean :2.001	Mean :100.8227	Mean : 100.1348
3rd Qu. :3.000	3rd Qu. :139.9337	3rd Qu. : 136.6506
Max. :5.000	Max. :997.1353	Max. :1186.4911

X_3	X_4	X_5
Min. : 0.0032	Min. : 0.0035	Min. : 0.0028
1st Qu. : 28.5687	1st Qu. : 28.2418	1st Qu. : 28.7874
Median : 68.7141	Median : 68.1072	Median : 68.0692
Mean : 99.0552	Mean : 100.5849	Mean :100.6532
3rd Qu. :136.9080	3rd Qu. : 138.4138	3rd Qu. :139.0188
Max. :857.5771	Max. :1026.7730	Max. :952.8392

Tableau 2 – Sommaire de 10 000 simulations de $\{N, X_1, \dots, X_5\}$ en supposant que $F_{N,X_1,\dots,X_5}$ est définie avec une copule de Clayton($\alpha = 6$), $N \sim Bin(5, 2/5)$ et $X_i \sim X \sim Exp(1/100)$, pour $i = 1, \dots 5$.



(a) Fréquence

(b) Sévérité

Illustration 2 – Comparaisons de la distribution marginale des données simulées avec les distributions théoriques pour le scénario 1.

Pour ce qui est du comportement entre les variables, l'illustration 3 présente les nuages de points représentant les corrélations. On y voit que le ρ de Spearman est le même pour tous les X_i. De plus, la forme des nuages de points pour les variables continues est caractéristique de la copule de Clayton. Cela indique que la simulation sort des résultats adéquats.

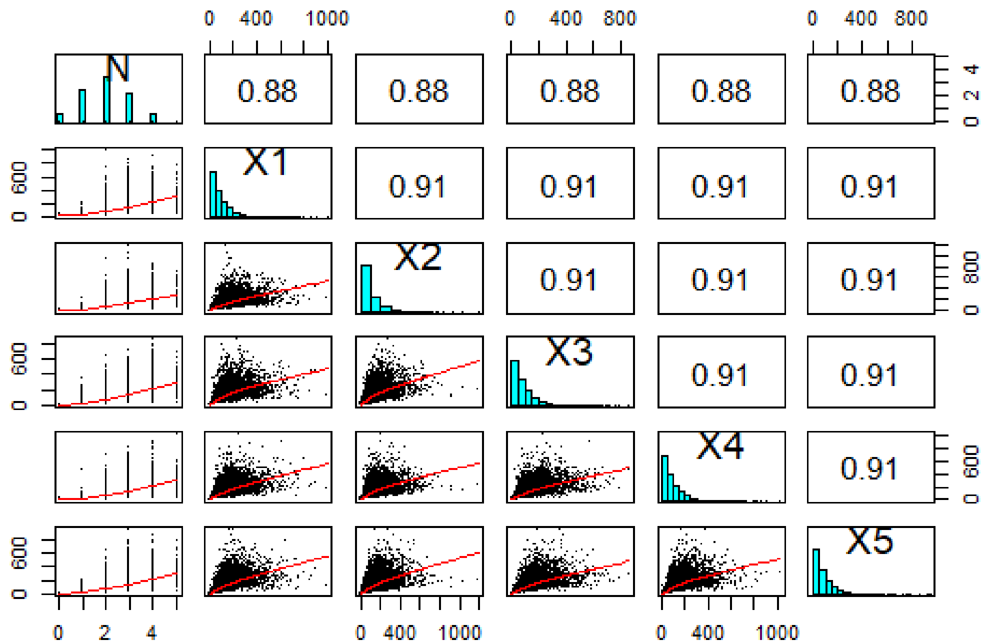


Illustration 3 – Nuages de points avec copule de Clayton ($\alpha = 6$), $N \sim Binom(5, 2/5)$ et $X_i \sim X \sim Exp(1/100)$: En bas de la diagonale se trouve les nuages de points illustrant la corrélation des différentes variables. En haut, on indique les coefficients de corrélation de Spearman. À noter que ce graphique est produit avec la fonction R `pairs.panels` du module `psych`. Dans cette dernière, le calcul des coefficients de corrélation sont calculés selon l'hypothèse que les variables sont continues. Cela implique que la première ligne n'est pas valide.

Avec les données simulées, on désire maintenant estimer les paramètres. On pose donc les valeurs de départ pour l'optimisation à l'estimateur du maximum de vraisemblance des distributions marginales des variables aléatoires. On a donc que $q \approx E[N]/n$, $\beta \approx 1/E[X]$. Pour ce qui est du α de la Clayton, il faut regarder les nuages de points de l'illustration 3. Plus le nuage est dense, plus le paramètre est grand. En revanche, si le nuage est large, plus α est près de 0.

Pour ce scénario, on a comme valeur de départ $q \approx 0.400\,280, \beta \approx 0.009\,975$. Pour α, posons la valeur initiale à 1. On obtient donc les estimateurs présentés dans le tableau 3. On y voit que les résultats de l'estimation sont relativement très précises. De plus, pour une fréquence maximale de 5, les temps de dérivation et d'estimation sont très rapides.

	q	beta	alpha		
Estimateurs	0.4024	0.0099	5.9539	temps de dérivation	4.22
Vrais paramètres	0.4000	0.0100	6.0000	temps d'estimation	180.23

Tableau 3 – Résultats de l'estimation des paramètres avec une copule de Clayton, $N \sim Binom(n, q)$ et $X \sim Exp(\beta)$ suite à 10 000 simulations.

Scénario 2 ***Poisson* :**

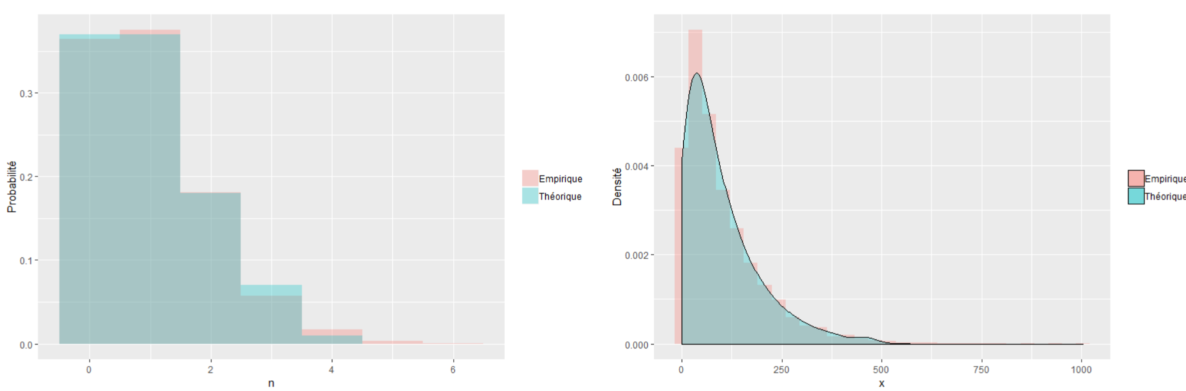
À la suite des résultats concluants du premier exemple, un modèle avec une loi de fréquence ayant un support infini peut ajouter un défi en terme de temps de calcul. Pour ce scénario, posons que $N \sim Pois(\lambda = 1)$ et $X_i \sim X \sim Exp(\beta = 1/100)$ et que le paramètre de dépendance de la Copule de Clayton est $\alpha = 6$.

Comme pour le scénario 1, un sommaire des données simulées est présenté dans le tableau 4, l'illustrations 5 compare la distribution marginale des données empiriques avec celle des lois théorique et l'illustration 5 présente les nuages de points afin de voir les corrélations.

N	X_1	X_2	X_3	X_4
Min. :0.0000	Min. : 0.0093	Min. : 0.012	Min. : 0.0129	Min. : 0.0172
1st Qu. :0.0000	1st Qu. : 29.7317	1st Qu. : 29.593	1st Qu. : 29.4194	1st Qu. : 29.5410
Median :1.0000	Median : 70.2592	Median : 70.060	Median : 70.3962	Median : 70.1612
Mean :0.9989	Mean : 101.2211	Mean :101.325	Mean :100.4977	Mean : 101.1802
3rd Qu. :2.0000	3rd Qu. : 140.8462	3rd Qu. :141.488	3rd Qu. :136.6166	3rd Qu. : 139.6666
Max. :6.0000	Max. :1003.4804	Max. :884.530	Max. :875.3387	Max. :1071.0681

X_5	X_6	X_7	X_8
Min. : 0.0121	Min. : 0.0118	Min. : 0.01	Min. : 0.0105
1st Qu. : 29.3078	1st Qu. : 29.7289	1st Qu. : 29.23	1st Qu. : 29.6958
Median : 70.1336	Median : 70.0210	Median : 70.47	Median : 70.2837
Mean : 100.2858	Mean :100.6869	Mean :100.69	Mean : 100.3833
3rd Qu. : 137.5781	3rd Qu. :138.3483	3rd Qu. :136.65	3rd Qu. : 138.1448
Max. :1089.6821	Max. :952.6193	Max. :810.01	Max. :1029.2935

Tableau 4 – Sommaire de 10 000 simulations de $\{N, X_1, \ldots, X_5\}$ en supposant que $F_{N,X_1,\ldots,X_5}$ est définie avec une copule de Clayton$(\alpha = 6)$, $N \sim Pois(1)$ et $X_i \sim X \sim Exp(1/100)$, pour $i = 1, \ldots 5$.



(a) Fréquence

(b) Sévérité

Illustration 4 – Comparaisons de la distribution marginale des données simulées avec les distributions théoriques pour le scénario 2.

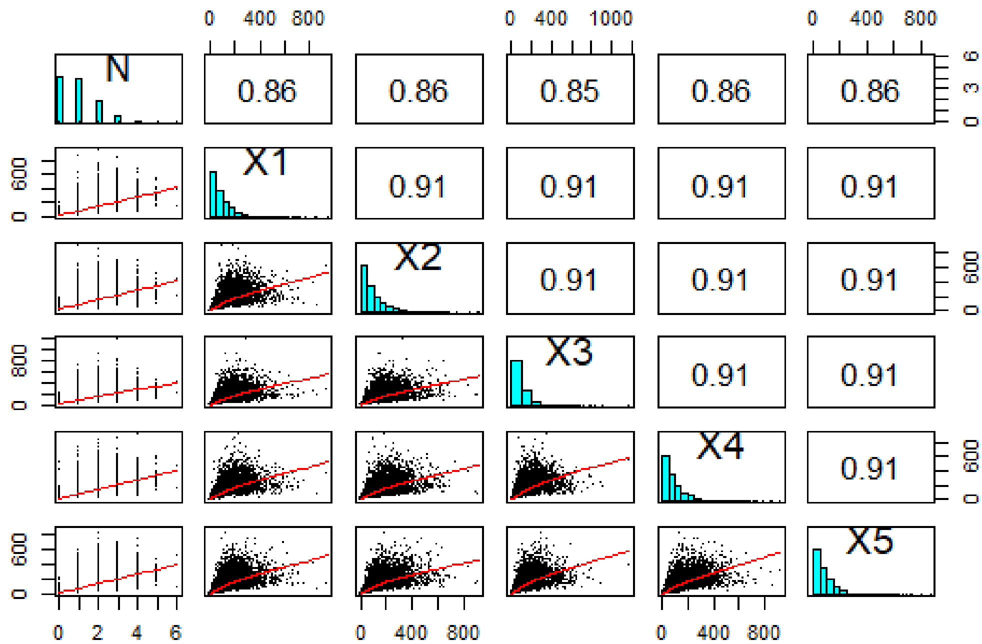


Illustration 5 – Nuages de points avec copule de Clayton ($\alpha = 6$), $N \sim Pois(1)$ et $X_i \sim X \sim Exp(1/100)$: En bas de la diagonale se trouve les nuages de points illustrant la corrélation des différentes variables. En haut, on indique les coefficients de corrélation de Spearman. À noter que ce graphique est produit avec la fonction `R` `pairs.panels` du module `psych`. Dans cette dernière, le calcul des coefficients de corrélation sont calculés selon l'hypothèse que les variables sont continues. Cela implique que la première ligne n'est pas valide.

Comme pour le scénario 1, le tableau 4 et les illustrations 5 démontrent que les résultats de la simulation sont adéquats.

Pour ce qui est de l'estimation des paramètres, commençons par définir les paramètres initiaux. On pose donc $\lambda = E[N] = 0.998\,900, \beta = 0.009\,922$ et comme pour l'exemple précédent, prenons 1 comme valeur de départ pour α. Les estimations obtenus sont présentés dans le tableau 5.

	lambda	beta	alpha		
Estimateurs	1.0006	0.0099	5.9160	temps de dérivation	13.29
Vrais paramètres	1.0000	0.0100	6.0000	temps d'estimation	86.24

Tableau 5 – Résultats de l'estimation des paramètres avec une copule de Clayton, $N \sim Pois(\lambda)$ et $X \sim Exp(\beta)$ suite à 10 000 simulations.

Dans cet exemple, l'estimation n'est pas significativement plus longue puisque le paramètre de la poisson n'est pas très grand. Cependant, si on l'augmente un tant soit peu, les quantiles de la variable de fréquence, on obtient des quantiles élevés qui font en sorte que le temps de dérivation augmente de façon exponentielle.

Par exemple, si $N \sim Pois(2)$, le temps de calcul dépasse 12 heures puisque le 99,9999 percentile de cette loi est de 12. Ainsi la méthode du maximum de vraisemblance nécessite de générer une liste de douze fonctions qui sont dérivées jusqu'à douze fois. Même avec la fonction `Deriv` de `R`, ce processus est extrêmement long. À cet effet, l'illustration 6 présente le temps de dérivation selon le nombre de dérivées partielles à effectuer sur une copule de Clayton.

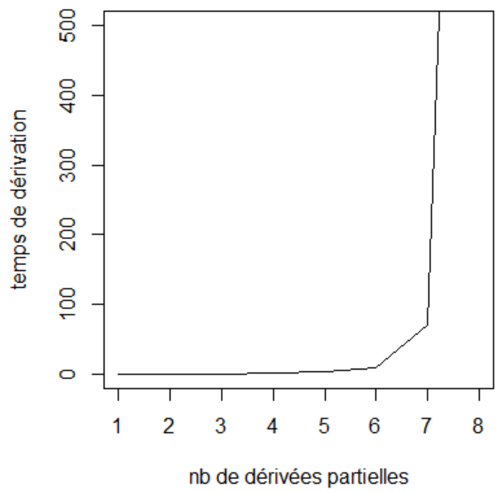


Illustration 6 – Temps de dérivation d'une copule de Clayton en fonction du nombre de dérivées partielles à effectuer.

Ce que l'on peut observer avec ces deux résultats, c'est, d'une part, qu'avec 10 000 simulations, on obtient des résultats très adéquats. D'autre part, on peut observer que le temps de calcul augmente significativement si N peut prendre des valeurs supérieures à 7 et si le nombre de paramètres à estimer est grand. Le temps de calcul est donc un enjeu important.

3.2 Copule archimédienne hiérarchique

Désormais, prenons la copule qui nous intéresse vraiment : la copule archimédienne hiérarchique. Avec $\theta_0 \sim \textit{logarithmique}\,(\gamma = 1 - \exp(-\alpha_0))$, $B \sim \textit{Gamma}(1/\alpha_1, 1)$, $N \sim \textit{Binom}(5, q)$ et $X_i \sim X \sim \textit{Exp}(\beta)$ on obtient les résultats présentés dans le tableau 6.

		α_0	α_1	β	q
Estimateurs		0.50	5.17	0.01	0.40
Vrais Paramètres		0.50	5.00	0.01	0.40
Temps de calcul	821.4 sec.				

Tableau 6 – Résultats de l'estimation des paramètres avec une copule archimédienne hiérarchique, $N \sim \textit{Binomiale}(5, q)$ et $X \sim \textit{Exp}(\beta)$ suite à 10 000 simulations.

Dans le tableau 6, on note que le temps de calcul est significativement supérieur à celui présenté dans le tableau 3. Ce phénomène est explicable du fait qu'il y a plus de paramètres à estimer et que la copule contient plusieurs fonctions imbriquées qui nécessitent plus d'étapes dans le processus de dérivation en chaîne.

4 Conclusion

Pour conclure, le modèle collectif du risque présenté dans [Cossette et al., 2019b] présente un défi dans l'estimation des paramètres par la méthode du maximum de vraisemblance tel que présentée dans la section 2.2 puisque la dérivation en chaîne sur un grand nombre de variables pose un problème de temps de calcul.

À cet effet, une solution envisageable pourrait être d'utiliser la vraisemblance par décomposition hiérarchique proposé dans [Cossette et al., 2019a] afin de trouver le paramètre de dépendance entre N et

X_1, puis de se limiter à un nombre restreint de X_i (disons 5) afin d'estimer le paramètre de dépendance entre les X_i. Cependant, avec cette méthode le fait de travailler avec des variables continues et discrètes peut causer un problème. Pour ce qui est de trouver les paramètres des lois de N et X, la méthode du maximum de vraisemblance classique (de façon univariée) pourrait être envisagée.

Références

[Cossette et al., 2017] Cossette, H., Gadoury, S.-P., Marceau, E., and Mtalai, I. (2017). Hiearchical archimedean copulas trough multivariate compound distributions. *Insurance : Mathematics and Economics*, 76.

[Cossette et al., 2019a] Cossette, H., Gadoury, S.-P., Marceau, E., and Robert, C. Y. (2019a). Composite likelyhood estimation method for hierarchical archimedean copulas defined with multivariate compound distributions. *Journal of Multivariate Analysis*, 172.

[Cossette et al., 2019b] Cossette, H., Marceau, E., and Mtalai, I. (2019b). Collective risk models with dependance.

[Klugman et al., 2012] Klugman, S., Panjer, H., and Willmot, G. (2012). *Loss Models : From Data to Decisions*. Wiley, fourth edition.