Prise en compte d'information pour l'estimation de quantiles agrégés.

Véronique Maume-Deschamps [1]& Andrés Cubéros [2] & Esterina Masiello [1].

[1] *Université de Lyon, Université Lyon 1, Institut Camille Jordan ICJ UMR 5208 CNRS, esterina.masiello@univ-lyon1.fr, veronique.maume@univ-lyon1.fr.*
[2] *Université de Lyon, Université Lyon 1, Laboratoire SAF EA 2429, SCOR SE, acuberos@scor.com.*

Résumé. L'estimation de quantiles (de niveau porche de 1) de variables agrégées (essentiellement des sommes ou des sommes pondérées) est essentielle pour la gestion des risques (financiers, assurantiels, environnementaux ...). Cette problématique est très présente dans la littérature mais de nouvelles méthodes efficaces sont toujours utiles, notamment en grande dimension. Nous proposons une méthode d'estimation basée sur la copule *checkerboard* qui permet d'obtenir de bonnes estimations à partir d'un petit échantillon de la loi multivariée et de la connaissance des lois marginales. Cette situation est réaliste dans de nombreuses applications, notamment en assurance. Par ailleurs, de l'information auxiliaire (connaissance de la loi d'un sous-vecteur, connaissance de probabilités extrêmes) peut être injectée dans la copule checkerboard et améliorer les estimations.

Mots-clés. Quantiles agrégés, copule checkerboard, information auxiliaire.

Abstract. Estimating high level quantiles of aggregated variables (mainly sums or weighted sums) is crucial in risk management for many application fields such as finance, insurance, environment... This question has been widely treated but new efficient methods are always welcome; especially if they apply in high dimension. We propose an estimation procedure based on the *checkerboard copula*. It allows to get good estimations from a (quite) small sample of the multivariate law and a full knowledge of the marginal laws. This situation is realistic for any applications, mainly in insurance. Moreover, we may also improve the estimations by including in the checkerboard copula some additional information (on the law of a sub-vector or on extreme probabilities).

Keywords. Aggregated quantiles, checkerboard copula, auxiliary information.

1 Définitions et problématique.

Considérons un vecteur aléatoire $X = (X_1, \dots, X_d)$ dont les lois marginales sont connues. On souhaite, à partir d'un échantillon indépendant de X de taille réduite n, estimer les quantiles de $S = \sum_{i=1}^{d} X_i$ pour des niveaux α proches de 1.

On rappelle que si les lois des X_i sont absolument continues, la structure de dépendance de X est uniquement déterminée par la copule définie par le théorème de Sklar :

$$F_X(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)),\ \forall (x_1, \ldots, x_d) \in \mathbb{R}^d,$$

où F_X est la fonction de répartition de X et F_i la fonction de répartition de X_i. C est ainsi la fonction de répartition du vecteur $(F_1(X_1), \ldots, F_d(X_d))$ dont toutes les marges sont uniformes.
Pour $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, on notera $[0, x] \subset \mathbb{R}^d$ le produit des intervalles $[0, x_i]$.
On remarque que si μ est une mesure de probabilité sur $[0, 1]^d$, elle définit une copule en posant $C(x) = \mu([0, x])$ si et seulement si $\mu([0, x]) = x_k$ si $x_j = 1$ pour tout $j \neq k$.

1.1 La copule checkerboard

La copule checkerboard a été introduite dans Mikusinski (2010), c'est une approximation de la copule C.
Soit $m \in \mathbb{N}^*$, considérons $(I_{i,m})_{i \in \{1,\ldots m\}^d}$ la partition (modulo un ensemble de mesure nulle) de $[0, 1]^d$ formées par les m^d carrés :

$$I_{i,m} = \prod_{j=1}^{d} \left[\frac{i_j - 1}{m}, \frac{i_j}{m}\right],\ i = (i_1, \ldots, i_d).$$

La copule cherckerboard C_m d'ordre m est associée à la mesure de probabilité μ_m sur $[0, 1]^d$:

$$\mu_m([0, x]) = \sum_i m^d \mu(I_{i,m}) \lambda([0, x] \cap I_{i,m}).$$

On remarque que $\mu_m([0, x]) = x_k$ si $x_j = 1$ pour $j \neq k$, et donc C_m est bien une copule. Elle approxime C :

$$\sup_{x \in [0,1]^d} |C_m(x) - C(x)| \leq \frac{d}{m}.$$

1.2 Information sur un sous-vecteur

Si on dispose d'informations supplémentaires comme la loi d'un sous-vecteur, celles-ci peuvent être introduites dans la copule checkerboard.
Supposons que la copule d'un sous-vecteur $\mathbf{X}^J$, $J \subset \{1, \ldots, d\}$, C^J est connue, $|J| = k < d$. On considère μ^J la mesure de probabilité sur $[0, 1]^k$ associée à C^J.
Pour $i = (i_1, \ldots, i_d)$, soit $x = (x_1, \ldots, x_d) \in [0, 1]^d$, $x^J = (x_j)_{j \in J}$, $x^{-J} = (x_j)_{j \notin J}$ et

$$I^J_{i,m} = \left\{ x \in [0, 1]^d \ / \ x_j \in \left[\frac{i_j - 1}{m}, \frac{i_j}{m}\right],\ j \in J \right\},$$

$$I_{i,m}^{-J} = \left\{ x \in [0,1]^d \ / \ x_j \in \left[\frac{i_j - 1}{m}, \frac{i_j}{m} \right], \ j \notin J \right\}.$$

La copule checkerboard avec information est associée à la mesure de probabilité :

$$\mu_m^J([0,x]) = \sum_{i \subset \{1,\ldots,d\}} \frac{m^{d-k}}{\mu^J(I_{i,m}^J)} \mu(I_{i,m}) \lambda([0,x^{-J}] \cap I_{i,m}^{-J}) \mu^J([0,x^J] \cap I_{i,m}^J).$$

$C_m^J(x) = \mu_m^J([0,x])$. On montre que

$$\mu_m^J([0,x]) = x_\ell \text{ si } x_j = 1 \ j \neq \ell.$$

Ainsi, C_m^J est une copule. On montre aussi qu'elle approxime C.

2 La procédure d'estimation.

Lorsque l'on dispose d'un *petit* échantillon indépendant de X (de taille $n = 30$ pour $d = 2$ ou $n = 150$ pour $d = 10$ par exemple), l'estimation de C n'est pas très bonne, surtout dans la queue de distribution. Néanmoins, on peut estimer la copule checkerboard et l'utiliser pour estimer les quantiles de S. Rappelons que la copule empirique a été définie par Deheuvels (1979) à partir des rangs.

Définition 1 *Soient $X^{(1)}, \ldots X^{(n)}$, n copies indépendantes de X et $R_i^{(1)}, \ldots, R_i^{(n)}$, $i = 1, \ldots, d$ leurs rangs marginaux, i.e.,*

$$R_i^{(j)} = \sum_{k=1}^{n} 1\{X_i^{(j)} \geq X_i^{(k)}\}, \ i = 1, \ldots, d, \ j = 1, \ldots, n.$$

La copule empirique $\widehat{C}$ est définie par :

$$\widehat{C}(u) = \frac{1}{n} \sum_{k=1}^{n} 1 \left\{ \frac{1}{n} R_1^{(k)} \leq u_1, \ldots, \frac{1}{n} R_d^{(k)} \leq u_d \right\}.$$

$\widehat{\mu}$ est la mesure de probabilité sur $[0,1]^d$ associée à $\widehat{C}$.

2.1 La copule checkerboard empirique

On peut alors estimer μ par $\widehat{\mu}$ puis construire la copule checkerboard empirique :

$$\widehat{C}_m(x) = \sum_i m^d \widehat{\mu}(I_{i,m}) \lambda([0,x] \cap I_{i,m}).$$

De même dans le cas où on dispose de la loi d'un sous-vecteur :

$$\widehat{C}_m^J(x) = \sum_{i \subset \{1,\ldots,d\}} \frac{m^{d-k}}{\mu^J(I_{i,m}^J)} \widehat{\mu}(I_{i,m}) \lambda([0,x^{-J}] \cap I_{i,m}^{-J}) \mu^J([0,x^J] \cap I_{i,m}^J).$$

2.2 Estimation

La procédure d'estimation consiste à générer générer des échantillons bootstrap suivant la copule checkerboard empirique puis, en composant par les fonctions quantile des marges, à obtenir un échantillon indépendant de S et à faire l'estimation des quantiles sur cet échantillon. Plus précisément : pour $N >> n$,

- Simuler un échantillon indépendant de taille N suivant la copule $\widehat{C}_m$, (ou $\widehat{C}_m^J$ dans le cas ou on dispose de la loi d'un sous vecteur) :

$$(u_1^{(1)}, \dots, u_d^{(1)}), \dots, (u_1^{(N)}, \dots, u_d^{(N)})$$

- En utilisant les lois marginales, construire un échantillon de S :

$$\sum_{i=1}^{d} F_i^{\leftarrow}(u_i^{(1)}), \dots, \sum_{i=1}^{d} F_i^{\leftarrow}(u_i^{(N)}).$$

- Estimer les quantiles de S à l'aide de l'échantillon ci-dessus.

Les résultats de convergence de la copule empirique (voir par exemple Fermanian et al. (2004) (2004)) garantissent, si S est absolument continue et C possède des dérivées partielles continues, si m est choisi tel que $A\sqrt{n} \leq m \leq n$, pour $A > 0$,

$$\|F_S - F_S^*\|_\infty = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right),$$

où F_S est la fonction de répartition de S et F_S^* la fonction de répartition de la somme des coordonnées de X^* dont les marginales sont le même que X et la dépendance est donnée par la copule checkerboard empirique.
Les simulations montrent que même avec n relativement faible, l'information contenue dans la copule checkerboard empirique est suffisante pour obtenir de bonnes estimations des quantiles de S.

3 Quelques simulations.

On présente ici des simulations en dimension 2 et 10, pour le modèle *Pareto - Clayton* pour lequel les valeurs exactes des quantiles de la somme sont calculables (voir Cuberos et al. (2014)).

3.1 Le modèle Pareto - Clatyon

On considère $X = (X_1, \ldots, X_d)$ avec

$$\mathbb{P}(X_1 > x_1, \ldots, X_d > x_d \,|\, \Lambda = \lambda) = \prod_{i=1}^{d} e^{-\lambda x_i}.$$

Autrement dit, conditionnellement à la variable aléatoire Λ, les marges de X sont indépendantes et distribuées suivant une loi exponentielle.
Si Λ suit une loi Gamma, alors les X_i suivent une loi de Pareto et sont liés par une copule de survie de Clayton.
Si Λ suit une loi de Levy, alors les X_i suivent une loi de Weibull est sont liés par une copule de survie de Gumbel.
Ces modèles ont été étudiés par de nombreux auteurs et notamment introduits dans Oakes (1989) et Yeh (2007). Ils ont été utilisés, par exemples, pour obtenir des formules explicites de probabilités de ruine et d'indicateurs de risque multi-variés, dans Albrecher et al. (2011) , Maume-Deschamps et al. (2014), Cénac et al. (2014) et dans Dacorona et al. (2014). Nous montrons ci-dessous que dans le cas Pareto-Clayton, les quantiles de la somme peuvent être calculés explicitement.
On se place dans le cas où $\Lambda \rightsquigarrow \Gamma(\alpha, \beta)$. Alors les X_i suivent des lois de Pareto (α, β) et sont liés par une copule de survie de Clayton de paramètre $1/\alpha$.
Dubey (1970) montre qu'alors S suit une loi Beta prime, i.e. sa fonction de répartition est donnée par :

$$F_S(x) = F_\beta\left(\frac{x}{1+x}\right).$$

où F_β est la fonction de répartition d'une loi $(d\beta, \alpha)$.
La fonction quantile $F_S^{\leftarrow}$ s'exprime alors en terme de de la fonction quantile de la loi Beta :

$$F_S^{\leftarrow}(p) = \frac{F_\beta^{\leftarrow}(p)}{1 - F_\beta^{\leftarrow}(p)}.$$

Les quantiles sont ainsi calculables explicitement.

3.2 Résultats en dimension 2

On considère le modèle Pareto-Clayton avec $\beta = 1$, $\alpha = 1$, l'échantillon de la loi multivariée est $n = 30$, on prend $N = 1000$. Le tableau ci-dessous compare la méthode *checkerboard* en l'estimation directe à partir de la fonction de répartition empirique de S, on donne l'estimation moyenne et l'écart-type relatif. On remarque que la méthode checkerboard est plus performante, notamment en terme de stabilité et pour les quantiles les plus élevés.

	VaR 80%	VaR 90%	VaR 95%	VaR 99%	VaR 99.5%	VaR 99.9%
Exact value	8.5	18.5	38.5	198.5	398.5	1998.5
Empirical	9 (40%)	19.5 (55%)	45 (105%)	293.5 (547%)	341.4 (566%)	379.7 (717%)
Checkerboard	8.8 (19%)	19.4 (23%)	41.8 (26%)	224.5 (10%)	432.6 (11%)	2049.7 (11%)

3.3 Résultats en dimension 10

On considère le modèle Pareto-Clayton avec $\beta = 1$, $\alpha = \frac{1}{2}$, l'échantillon de la loi multi-variée est $n = 75$ puis $n = 175$, on prend $N = 1000$. Le tableau ci-dessous compare la méthode *checkerboard* en l'estimation directe à partir de la fonction de répartition empirique de S, on donne l'estimation moyenne et l'écart-type relatif. On remarque que la méthode checkerboard est plus performante, notamment en terme de stabilité et pour les quantiles les plus élevés.

	VaR 80%	VaR 90%	VaR 95%	VaR 99%	VaR 99.5%	VaR 99.9%
Exact value	12.2	19.2	29	70.1	100.8	230.5
Empirical, $n = 75$	12.6 (12%)	20 (15%)	29.9 (19%)	62.2 (39%)	75.8 (58%)	86.7 (71%)
Checkerboard, $n = 75$	12.5 (10%)	20.1 (13%)	31.2 (14%)	74.8 (20%)	92.4 (20%)	152.6 (16%)
Empirical, $n = 150$	12.4 (8%)	19.6 (11%)	30.3 (14%)	67.3 (27%)	89.9 (38%)	121 (59%)
Checkerboard, $n = 150$	12.4 (7%)	19.6 (9%)	29.8 (12%)	75.4 (16%)	107.6 (21%)	173.9 (19%)

Remarquons que les échantillons de la loi multi-variée sont trop petits pour espérer faire l'estimation à l'aide de le théorie des valeurs extrêmes. C'est tout l'intérêt de la méthode : fournir des estimations acceptables, même avec de petits échantillons.
Les simulations en utilisant l'information auxiliaire sont en cours.

References

[1] Hansjörg Albrecher, Corina Constantinescu, and Stéphane Loisel. Explicit ruin formulas for models with dependence among risks. *Insurance: Mathematics and Economics*, 48:265–270, 2011.

[2] Peggy Cénac, Stephane Loisel, Véronique Maume-Deschamps, and Clémentine Prieur. Risk indicators with several lines of business: comparison, asymptotic behavior and applications to optimal reserve allocation. *Annales de l'ISUP*, 58(3), 2014.

[3] Andrés Cuberos, Esterina Masiello, and Véronique Maume-Deschamps. High level quantile approximations of sums of risks. 2014.

[4] Michel Dacorogna, Leila El Bahtouri, and Marie Kratz. Explicit diversification benefit for dependent risks. preprint, 2014.

[5] Paul Deheuvels. La fonction de dépendance empirique et ses propriétés. *Acad. Roy. Belg. Bull. Cl. Sci.*, 65(5):274 – 292, 1979.

[6] Jean-David Fermanian, Dragan Radulovic, and Marten Wegkamp. Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847–960, 2004.

[7] Véronique Maume-Deschamps, Didier Rullière, and Khalil Saïd. On the optimal capital allocation by minimizing multivariate risk indicators. 2014.

[8] Piotr Mikusinski and Michael D Taylor. Some approximations of n-copulas. *Metrika*, 72(3):385–414, 2010.

[9] David Oakes. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493, 1989.

[10] Hsiaw-Chan Yeh. The frailty and the Archimedean structure of the general multivariate Pareto distributions. *Bulletin Institute of Mathematics Acedemia Sinica*, 2(3):713–729, 2007.