Estimation non paramétrique de la fonction de dépendance extrémale

Baptiste Klehammer

Université de Strasbourg

UFR de Mathématique Informatique

MASTER

Mention Mathématiques et Applications

Spécialité Statistique

Estimation non paramétrique de la fonction de dépendance extrémale

---

Présenté par :

Baptiste KLEHAMMER

Sous la direction de :

Laurent GARDES,

Université de Strasbourg & CNRS, IRMA
7 rue René Descartes
67084 Strasbourg Cedex

E-mail : gardes@unistra.fr

Remerciements

Je tiens à exprimer ma profonde gratitude à Laurent Gardes pour l'honneur singulier qu'il m'a accordé en ayant accepté d'être mon tuteur de mémoire. Je lui suis particulièrement reconnaissant pour ses conseils et pistes fournies, ainsi que sa disponibilité quasi-permanente qui va de pair avec la promptitude de ses réponses.

Résumé

La finalité de ce mémoire est la simulation de l'estimation de la fonction de dépendance de queue dans le cas non paramétrique. Pour ce faire, deux notions mais essentielles à la théorie, à savoir l'estimation par la méthode des noyaux et la théorie des noyaux, seront introduites de manière générale, ce qui permettra d'avoir une meilleure vue d'ensemble de la troisième partie qui introduira les notions clés de la dépendance de queue. La première partie introduira un paramètre essentiel à l'estimation de la fonction de dépendance de queue, à savoir la fenêtre de lissage, alors que la seconde partie présentera quelques résultats généraux et les distributions des valeurs extrêmes. Enfin, la dernière partie sera consacrée à la simulation et son implémentation en langage R.

Table des matières

Estimation par la méthode du noyau

Introduction

L'estimation par noyau (ou méthode de Parzen-Rozenblatt) est une méthode non paramétrique d'estimation de la densité d'une variable aléatoire. Cette méthode permet d'obtenir une densité continue et constitue en ce sens une généralisation de la méthode de l'histogramme. En effet, la fonction indicatrice utilisée pour l'histogramme est ici remplacée par une fonction continue (le noyau) et une somme de fonctions continues reste continue.

Définition

Si x_1, …, x_N est un échantillon indépendant et identiquement distribué d'une variable aléatoire de densité f continue, alors l'estimateur non paramétrique de la densité par la méthode du noyau est le suivant :

$$\hat{f}_h(x) = \frac{1}{(Nh)} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

Où h > 0 est appelé fenêtre ou constante de lissage et K est un noyau (fonction positive telle que $\int_{\mathbb{R}} K(u)du = 1$ et souvent symétrique).

Par ailleurs, si y_1, …, y_N est un échantillon indépendant et identiquement distribué d'une variable aléatoire, alors on a l'estimation suivante :

$$\hat{f}_h(y|x) = \frac{1}{(Nh)} \sum_{i=1}^{N} \mathbb{I}\{Y_j \geq y\} K\left(\frac{x - x_i}{h}\right)$$

Enfin, on peut aussi estimer la fonction de répartition par cette méthode à l'aide de la formule suivante :

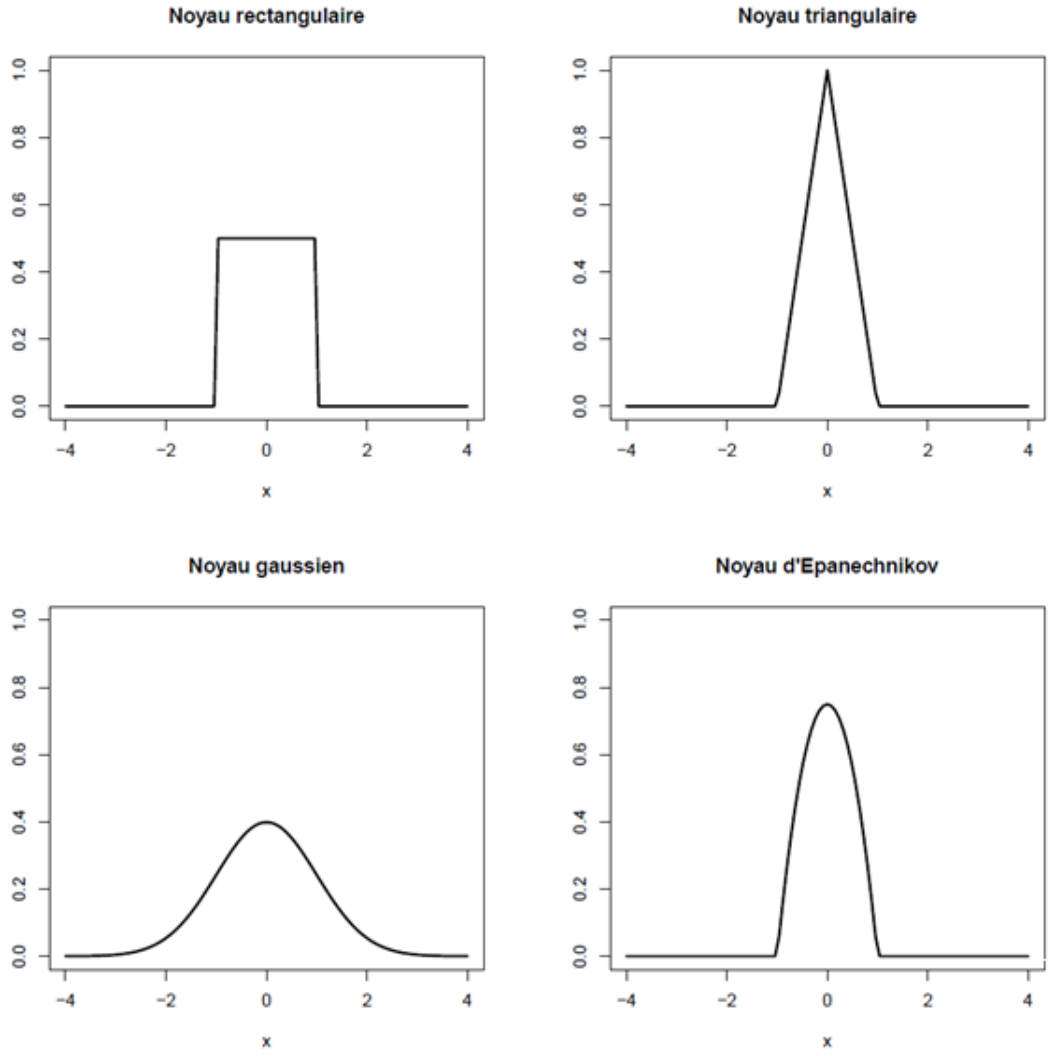
$$F_n(y|x) = \frac{\sum_{i=1}^{N} \mathbb{I}\{Y_j \leq y\} K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)}$$

Généralités

Sous les hypothèses précédentes, l'estimateur par noyau est l'estimateur non paramétrique qui converge le plus vite, ie à la vitesse $n^{-4/5}$ et pour un choix de choix de h proportionnel à $n^{-1/5}$.

Les noyaux les plus couramment utilisés sont les suivants :

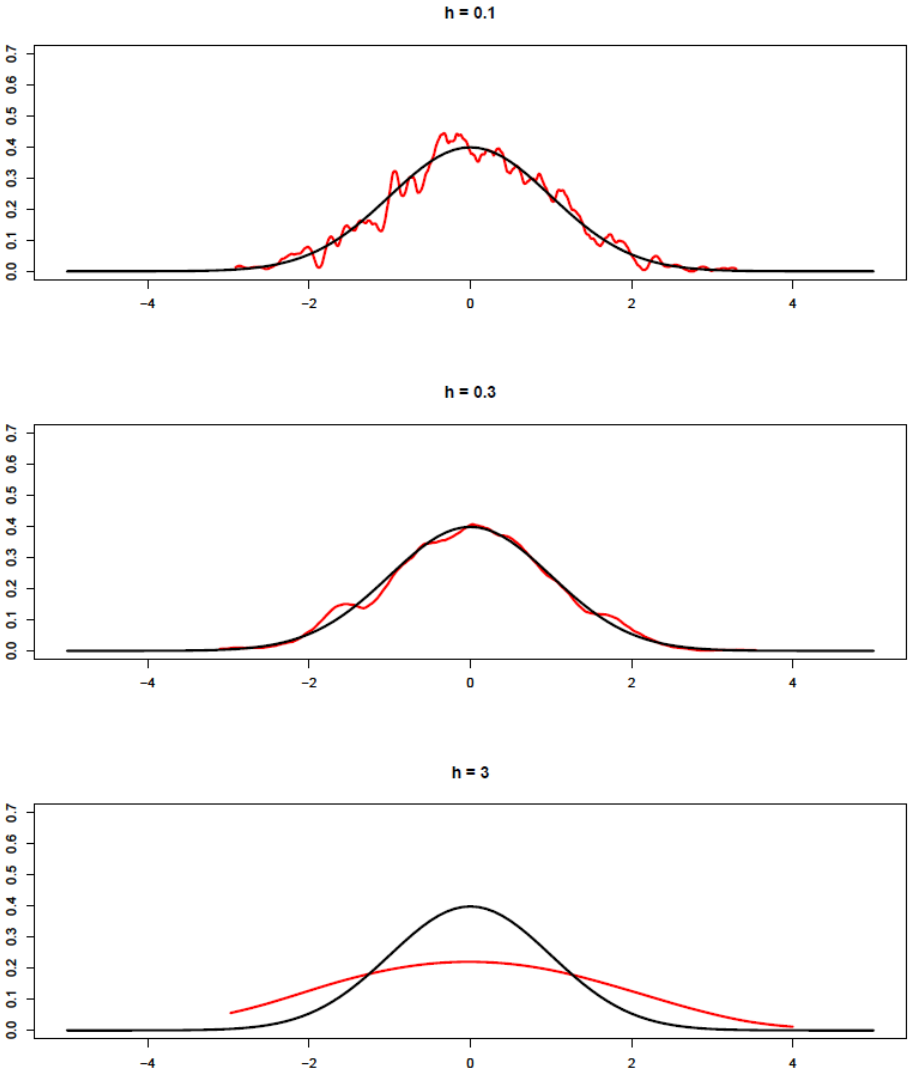
- Noyau rectangulaire : $K(u) = \frac{1}{2}\mathbb{I}_{[-1;1]}(u)$
- Noyau triangulaire : $K(u) = (1 - |u|)\mathbb{I}_{[-1;1]}(u)$
- Noyau gaussien : $K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$
- Noyau d'Epanechnikov : $K(u) = \frac{3}{4}(1 - u^2)\mathbb{I}_{[-1;1]}(u)$



Le noyau d'Epanechnikov est apprécié puisqu'il est celui qui a la meilleure efficacité (mais les autres noyaux présentés sont à peine moins efficaces).

Choix de h

Le choix de h est une étape importante lors de l'estimation par noyau dans le sens où $h_{optimal}$ nécessite de connaître la densité que l'on cherche à estimer, ce qui n'est en pratique pas le cas. Par ailleurs, si h est trop petit, le biais de l'estimateur devient petit devant sa variance et l'estimateur trop fluctant. On obtient un phénomène de sous-lissage. Dans le cas contraire, lorsque h est trop grand, le biais prend l'ascendant sur la variance et l'estimateur varie peu : on obtient un phénomène de sur-lissage.



L'illustration ci-dessus présente respectivement un cas de sous-lissage, un cas « normal » et un cas de sur-lissage. (Noyau d'Epanechnikov, N = 1000, loi normale centrée réduite.)

En pratique, on utilise souvent la méthode de la validation croisée pour choisir automatiquement h. Cette méthode demande dans un premier temps de trouver un estimateur sans biais de $J(h) = MISE\left(\hat{f}_h^K\right) - \|f\|_2^2$ et de minimiser ensuite cet estimateur.

On trouve que $\hat{J}(h) = \|f\|_2^2 - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right)$ convient. Il suffit alors de minimiser cette fonction et l'on obtient $\hat{h}_{opt} = argmin_{h>0}\{\hat{J}(h)\}$.

Théorie des valeurs extrêmes

Introduction

La théorie des valeurs extrêmes permet de prendre le relais là où la théorie probabiliste classique n'est plus capable de prédire des évènements rares (crise financière, tsunami, choc pétrolier…) et est en ce sens de plus en plus utilisée en pratique. En particulier, elle permet de prédire des phénomènes extrêmes.

Points clés de la théorie

On dit que deux fonctions de répartions G et H sont de même type s'il existe a>0 et b tels que pour tout réel x, G(x) = H(ax+b).

On appelle domaine d'attraction d'une fonction de répartition non dégénérée la quantité

$$D(G) = \{F\ tq\ \exists a_n > 0\ et b_n\ tq\ F^n(a_n x + b_n) \to G(x), \forall x > 0\}$$

Théorème de Fisher-Tippett

Soient X_1, …, X_n des variables aléatoires indépendantes et de même fonction de répartition F, en notant $X_{1,n} \leq \ \ldots \leq X_{n,n}$, telles qu'il existe des suites (a_n) avec a_i>0 et (b_n) vérifiant

$$\lim_{n\to\infty} P(\frac{X_{n,n} - b_n}{a_n} \leq x) = G(x), x \in \mathbb{R}$$

Alors G est une fonction de répartition non dégénérée de l'un des 3 types suivants :

- $G_0(x) = \exp(-e^{-x}), x \in \mathbb{R}$
- $G_{1,\alpha}(x) = \exp(-x^{-\alpha}), x \geq 0, \alpha > 0$
- $G_{2,\alpha}(x) = \exp(-(-x)^{-\alpha}), x \leq 0, \alpha < 0$

Les distributions des valeurs extrêmes

Le premier type de loi G_0, est appelé loi de Gumbel ou encore loi à « queue légère » dans le sens où $\bar{F}(x) \underset{x\to\infty}{\longrightarrow} 0$ de manière exponentielle.

Le second type de loi $G_{1,\alpha}$, est appelé loi de Fréchet ou encore loi à « queue lourde » dans le sens où $\bar{F}(x) \underset{x\to\infty}{\longrightarrow} 0$ de manière polynomiale (donc moins vite qu'une loi de Gumbel).

Le dernier type de loi, $G_{2,\alpha}$, est appelé loi de Weibull ou encore loi à « queue finie » dans le où il existe x_M tel que $x \geq x_M \Rightarrow \bar{F}(x) = 0$.

Ces trois lois peuvent être écrites sous la forme de la distribution généralisée des valeurs extrêmes (GEV) suivante (paramétrisation de Von Mises) :

$$G_\gamma(x) = \exp(-(1+\gamma x)^{-1/\gamma}\ si\ 1+\gamma x > 0$$

On remarque que $\lim_{\gamma \to 0} G_\gamma(x) = \exp(-e^{-x}) = G_0(x)$

Loi de Pareto Généralisée

Définition : Soit G_γ une GEV et x_M la borne sup de son support ; on appelle loi de Pareto généralisée (GPD) la loi suivante :

$$\begin{aligned} H_\gamma(x) &= 1 + \log\left(G_\gamma(x)\right) si\ 0 < x < x_M \\ &= 1 - (1+\gamma x)^{-1/\gamma} \end{aligned}$$

On remarque que si $H_\gamma(x) = 1 - (1+\gamma x)^{-\frac{1}{\gamma}}, 1+\gamma x > 0$ alors $G_\gamma(x) = \exp(-\overline{H_\gamma}(x))$

γ peut être directement interprété comme un indice de risque et il est donc naturel de chercher à l'estimer.

Pour cela, il existe différentes méthodes et le cas des estimateurs paramétriques ne sera pas traité ici.

Estimateur de l'indice de risque

Concernant les estimateurs non paramétriques, l'estimateur de Hill est très utilisé mais nécessite des lois de type Fréchet uniquement. L'autre estimateur, plus général, est celui de Pickands.

L'estimateur de Hill de γ pour k dans {1, …,n-1} est le suivant :

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \log(X_{n-i+1,n}) - \log(X_{n-k,n})$$

Cet estimateur peut être interprété comme une valeur de la pente d'un diagramme quantile-quantile à l'infini et est sensible au choix du nombre de points retenus dans la queue de distribution.

L'estimateur de Pickands, valable pour toute valeur de γ, est donné par la formule suivante :

$$\gamma^p_{k,n} = \frac{1}{\ln 2}\ln\left(\frac{X_{[k/4,n]} - X_{[k/2,n]}}{X_{[k/2,n]} - X_{k,n}}\right)$$

Où [x] est la partie entière de x.

Estimation de la fonction de dépendance de queue

Présentation des copules

La copule est un outil permettant de caractériser la dépendance entre plusieurs variables aléatoires là où en général les corrélations linéaires ne sont pas en mesure de les représenter de manière efficace.

Définition : on appelle copule bivariée C : $[0,1]^2 \rightarrow [0,1]$ une fonction telle que :

- $C(u,0) = C(0,u) = 0 \; \forall u \in [0,1]$
- $C(u,1) = C(1,u) = u \; \forall u \in [0,1]$
- $C(v_1,v_2) - C(v_1,u_2) - C(u_1,v_2) + C(u_1,u_2) \geq 0 \; \forall (u_1,u_2,v_1,v_2) \in [0,1]^4$ $tq \; u_1 \leq v_1 \; et \; u_2 \leq v_2$

La copule multivariée se définit de la même manière avec $u \in [0,1]^d$

Théorème de Sklar : Soit F une fonction de répartition bivariée de marges F_1 et F_2. Alors la copule associée à F s'écrit de la manière suivante :

$C(u_1,u_2) = C\big(F_1(x_1),F_2(x_2)\big) = F\big(F_1^{\leftarrow}(u_1),F_2^{\leftarrow}(u_2)\big) = F(x_1,x_2)$ où $F^{\leftarrow}$ est l'inverse généralisée de F.

Si, de plus, les marges F_1 et F_2 sont continues, alors C est unique.

La densité f d'une loi bivariée peut s'écrire en fonction de la densité c de la copule associée et des densités marginales f_1 et f_2 par :

$$f(x_1,x_2) = c\big(F_1(x_1),F_2(x_2)\big) \times f_1(x_1) \times f_2(x_2)$$

Où $c(u_1,u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} C(u_1,u_2)$

Bornes de Fréchet : On appelle bornes de Fréchet les deux fonctions définies sur $[0,1]^2$ suivantes : $W(u_1,u_2) = \max(u_1+u_2-1,0) \; et \; M(u_1,u_2) = \min(u_1,u_2)$

Propriétés

Pour toute copule C, on a $M(u_1,u_2) \leq C(u_1,u_2) \leq W(u_1,u_2)$.

Si X_1 et X_2 sont deux v.a. continues de copule associée C et que h_1 et h_2 sont deux fonctions strictement croissantes, alors $h_1(X_1)$ et $h_2(X_2)$ ont également C pour copule associée, ie la copule est invariante par transformation strictement croissante des v.a.

Dépendance de queue

La dépendance de queue est une notion qui permet de mesurer la probabilité des réalisations extrêmes simultanées.

La dépendance de queue d'une copule est telle que :

$$\mathbb{P}(U_1 \leq u_1 | U_2 \leq u_2) = \frac{C(u_1, u_2)}{u_2}$$
$$\mathbb{P}(U_1 \geq u_1 | U_2 \geq u_2) = \frac{1 - u_1 - u_2 + C(u_1, u_2)}{1 - u_2}$$

Alors on a les coefficients de dépendance de queue à gauche $\lambda_L = \lim_{u \to 0^+} \frac{C(u,u)}{u}$ et à droite $\lambda_U = \lim_{u \to 1^-} \frac{1-2u+C(u,u)}{1-u}$.

La copule de survie S se déduit de la copule C par la relation :

$$S(u) = u - 1 + C(1 - u), \forall u \in [0,1]^d$$

On a alors $S(u) = \bar{F}(\bar{F}_i^{\leftarrow}(u_i), i = 1, \dots, d)$

La fonction de dépendance de queue (si elle existe) est :

$$\Lambda(y) = \lim_{\alpha \to 0} \frac{1}{\alpha} S(\alpha y), y \in (0, \infty)^d$$

On remarque que $\Lambda(1) = \lambda_U = \mathbb{P}\left(Y_i \geq \bar{F}_i^{\leftarrow}(\alpha), i \neq j | Y_j \geq \bar{F}_j^{\leftarrow}(\alpha)\right)$. On dit alors que le vecteur multivarié Y est possède une dépendance de queue si $\lambda_U \in]0,1]$. Si $\lambda_U = 0$, le vecteur Y est indépendant pour la queue.

Dépendance de queue conditionnelle

Cette partie est une extension à la partie précédente dans le sens où l'on rajoute une variable indépendante X de dimension p.

On obtient alors la copule de survie conditionnelle suivante :

$$S(u|x) = \bar{F}(q(u|x)|x), u = (u_1, \dots, u_d)$$

Où $q(u|x) = (\bar{F}_i^{\leftarrow}(u_1|x), i = 1, \dots, d)^T$

On peut alors définir la fonction de dépendance de queue suivante :

$$\Lambda(y|x) = \lim_{\alpha \to 0} \frac{1}{\alpha} S(\alpha y|x), y \in (0, \infty)^d$$

Pour la suite, nous supposerons que $\Lambda(y|x) > 0$.

On note $\hat{g}_h(x) = \frac{1}{nh^p} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$ l'estimateur de la densité de X.

Marges connues

Lorsque les marges de Y sont connues, la copule de survie conditionnelle peut s'estimer par :

$$\tilde{S}_n(u|x) = \frac{1}{nh^p \hat{g}_h(x)} \sum_{j=1}^{n} \mathbb{I}\{Y_j \geq q(u|x)\} K(\frac{x - X_j}{h})$$

On obtient alors pour l'estimateur de la fonction de dépendance de queue conditionnelle :

$$\widetilde{\Lambda}_n(y|x) = \frac{1}{\alpha_n} \tilde{S}_n(\alpha_n y|x)$$

Pour obtenir la normalité asymptotique de l'estimateur de Λ(y|x), on peut montrer qu'il y a deux conditions nécessaires :

- Il existe une fonction b(.|x) telle que $\lim_{t\to\infty} b(t|x) = 0$ et une fonction $c_\Lambda(.|x)$ telles que pour tout y dans [0,∞[^{d}, $\lim_{t\to\infty} \frac{\Lambda(y|x) - tS\left(t^{-1}y|x\right)}{b(t|x)} = c_\Lambda(y|x) < \infty$
- Il existe une constante positive c_g telle que $|g(x) - g(x')| \leq c_g d(x, x')$ où d(.,.) est la distance euclidienne et $(x, x') \in (\mathbb{R}^p)^2$

Lorsque ces conditions sont vérifiées, on a la normalité asymptotique recherchée, ie :

$$(nh^p \alpha_n)^{\frac{1}{2}} \left(\widetilde{\Lambda}_n(y|x) - \Lambda(y|x)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\|K\|_2^2 \Lambda(y|x)}{g(x)}\right) \text{où } (\alpha_n) \to 0 \; et \; (nh^p \alpha_n) \to \infty$$

Marges inconnues

Lorsque les marges de Y sont inconnues, il est d'abord nécessaire d'estimer q(.|x) par :

$$\hat{q}_n(u|x) = (\hat{q}_{n,i_i}(u_1|x), i = 1, \dots, d)^T$$

La copule de survie peut alors s'estimer par :

$$\hat{S}_n(u|x) = \frac{1}{nh^p \hat{g}_h(x)} \sum_{j=1}^{n} \mathbb{I}\{Y_j \geq \hat{q}(u|x)\} K(\frac{x - X_j}{h})$$

On obtient alors pour l'estimateur de la fonction de dépendance de queue :

$$\widehat{\Lambda}_n(y|x) = \frac{1}{\alpha_n} \hat{S}_n(\alpha_n y|x)$$

Pour obtenir la normalité asymptotique de $\widehat{\Lambda}_n$, on peut montrer qu'il est nécessaire d'avoir deux conditions supplémentaires aux deux précédentes, à savoir :

- Les marges conditionnelles de Y sont de type Fréchet : $\bar{F}_i(z|x) = z^{-\frac{1}{\gamma(x)}} \times l_i(z|x)$ où γ_i est une fonction positive et l_i une fonction à variation lente, ie $\lim_{z\to\infty} \frac{l_i(kz|x)}{l_i(z|x)} = 1$ pour k>0.
- La copule conditionnelle S(.|x) est de classe C^1 telle que pour tout vecteur $v(\alpha, y) vérifiant\ v(\alpha, y) \xrightarrow[\alpha\to 0^+]{} \alpha y$, $\lim_{\alpha\to 0^+} y^T \nabla S(v(\alpha, y)|x) = \Lambda(y|x)$ avec $y \in \mathbb{R}^d$

Lorsque ces conditions sont vérifiées, on a la normalité asymptotique recherchée, ie :

$$(nh^p\alpha_n)^{\frac{1}{2}}\left(\widehat{\Lambda}_n(y|x)-\Lambda(y|x)\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{\|K\|_2^2\Lambda(y|x)}{g(x)}\right) o\grave{u}\ (\alpha_n)\to 0\ et\ (nh^p\alpha_n)\to\infty$$

Simulations

La partie simulation consiste à implémenter toutes les fonctions précédentes en langage R, tenter de les appliquer et d'interpréter les résultats.

Le but étant de trouver les α_n et h_n optimaux, une matrice des valeurs de $\widehat{\Lambda}_n(y|x)$ sera représentée graphiquement afin d'avoir α_n et h_n pour les axes et visualiser l'erreur quadratique moyenne $\sum_{i=1}^{N}\frac{1}{N}\left(\widehat{\Lambda}_n(y|x)-\Lambda(y|x)\right)^2$ aux différentes coordonnées à l'aide d'une palette de couleurs. Les valeurs de y et x seront choisies arbitrairement.

Méthodes utilisées

Pour simuler un vecteur aléatoire d'une fonction de répartition donnée F, la simulation par inversion de la fonction de répartition sera utilisée. En effet, si U suit une loi uniforme sur [0,1] alors $X=F^{\leftarrow}(U)$ est une v.a. de fonction de répartition F. Cela nécessite néanmoins que F soit strictement croissante.

Pour l'estimation de la densité, le noyau d'Epanechnikov sera utilisé pour les raisons évoquées précédemment et parce qu'il est borné. Il en va de même pour l'estimation de la fonction de répartition dans le cas où les marges sont inconnues.

Modèle logistique

Pour ce premier exemple, nous considérons trois variables Y, Z et X telles que :

- $F_y(y)=\exp\left(-\frac{1}{y}\right) et\ F_z(z)=\exp\left(-\frac{1}{z}\right)$, lois de Fréchet
- $\theta(x)=\frac{\log(2)}{\log\left(\frac{2+\exp(x)}{1+\exp(x)}\right)}$
- $F(y,z|x)=\exp\left(-\left(y^{-\theta}+z^{-\theta}\right)^{\frac{1}{\theta}}\right)$
- $F(y|z,x)=\exp\left(\frac{1}{z}-\left(y^{-\theta}+z^{-\theta}\right)^{\frac{1}{\theta}}\right)\left(y^{-\theta}+z^{-\theta}\right)^{\frac{1}{\theta}-1}z^{1-\theta}$
- $X\sim\mathcal{N}(0,1)$, loi normale centrée réduite
- Z et X sont générées de manière indépendante

Avant de se lancer dans la simulation calculons d'abord la fonction de dépendance de queue :

$$\bar{F}_y^{-1}(y) = -\frac{1}{\log(1-y)}$$

$$q(\alpha y, \alpha z|x) = (-\frac{1}{\log(1-\alpha y)}, -\frac{1}{\log(1-\alpha z)})$$

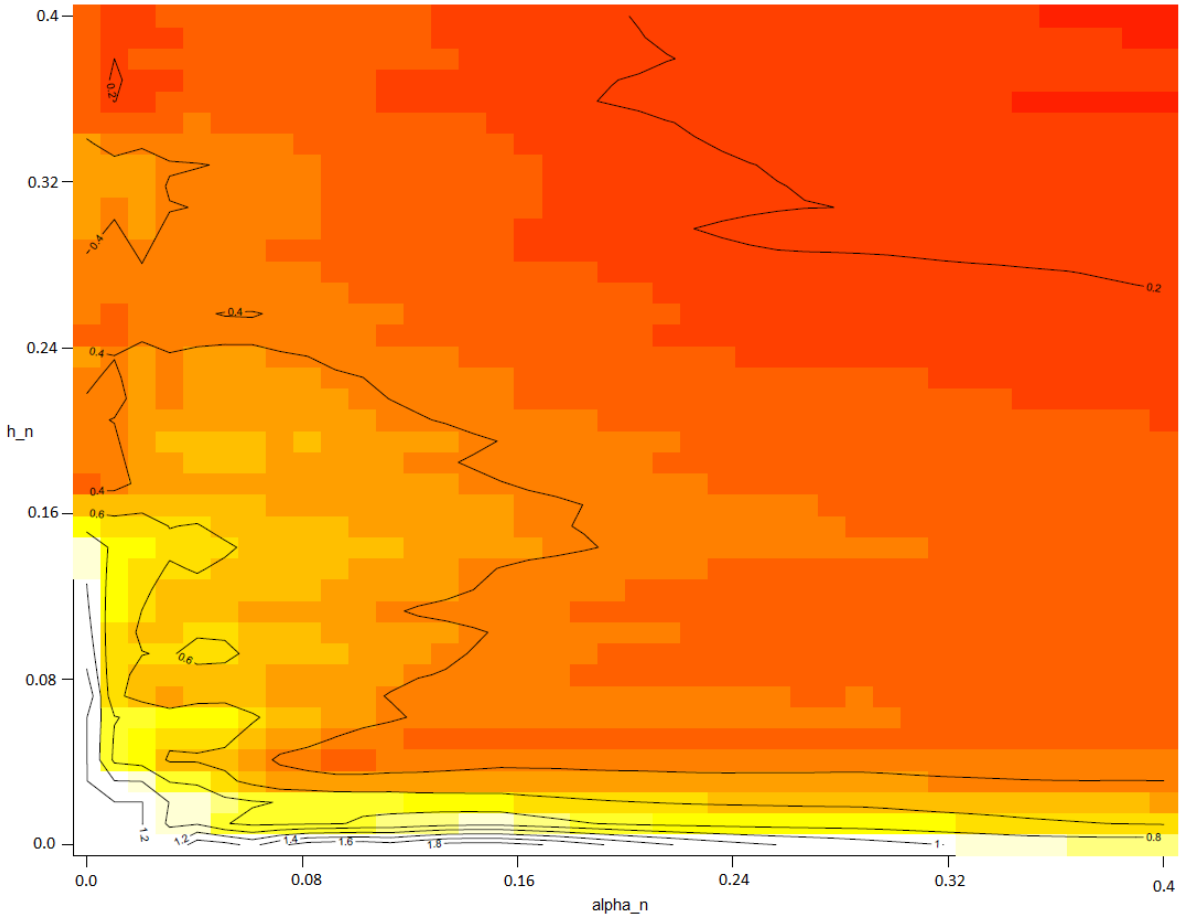
$$S(\alpha y, \alpha z|x) = \bar{F}(q(\alpha y, \alpha z|x)|\mathrm{x}) = 1 - \exp(-\big((-\log(1-\alpha y))^\theta + (-\log(1-\alpha z))^\theta\big)^{\frac{1}{\theta}})$$

Lorsque $\alpha -> 0, \log(1-\alpha t) \approx -\alpha t$ et $1 - \exp(\alpha t) \approx -\alpha t$, d'où

$$\begin{aligned} S(\alpha y, \alpha z|x) &\approx 1 - \exp\left(-\big((\alpha y)^\theta + (\alpha z)^\theta\big)^{\frac{1}{\theta}}\right) \\ &= 1 - \exp\left(-\alpha\big(y^\theta + z^\theta\big)^{\frac{1}{\theta}}\right) \\ &= \alpha\left(\big(y^\theta + z^\theta\big)^{\frac{1}{\theta}}\right) \end{aligned}$$

$$\Lambda(y, z|x) = \lim_{\alpha \to 0} \frac{1}{\alpha} S(\alpha y, \alpha z|x) = \big(y^\theta + z^\theta\big)^{\frac{1}{\theta}}$$

Les simulations sont réalisées dans les conditions suivantes : on fixe (z,y,x)=(1,1,0.5), Y, Z et X sont de taille 500 et l'EQM est calculé sur 500 triplets d'échantillons. Par ailleurs, on a $0 < \alpha_n \leq 0.4$ et $0 < h_n \leq 0.4$ avec 40 valeurs équiréparties dans ces intervalles. Malheureusement, le résultat (cf ci-dessous) n'offre pas un choix clair pour α_{opt} et h_{opt} dans les conditions de simulations utilisées.



(Le gradient de couleurs est lié à l'intervalle [0,1] dans le sens rouge -> orange -> jaune -> blanc où blanc signifie que la valeur en ce point a dépassé 1.)

En effet, il semble que l'EQM diminue au fur et à mesure que α_n et h_n augmentent, ce qui n'a pas vraiment de sens. Des résultats similaires se produisent lorsque les intervalles de α_n et h_n sont agrandis. Néanmoins, on peut noter que l'EQM reste assez faible dans les zones rouges (en-dessous d'environ 0.1) alors que la vraie valeur de $\Lambda(y = 1, z = 1|x = 0.5)$ est environ 1.098.

Le problème vient alors du code R ou du modèle qui ne respecte pas les hypothèses requises présentées dans la partie théorique. Si certaines sont simples à vérifier, comme le caractère lipschitzien de la densité ou le type des variables Y et Z (Fréchet), d'autres le sont moins.

Conclusions du mémoire

Ce mémoire sous la direction de Laurent Gardes a été une expérience enrichissante dans le sens où les différentes notions abordées ont été nouvelles pour moi mais néanmoins à portée de compréhension. J'ai ainsi eu l'occasion de lire plusieurs travaux tout en prenant soin de multiplier les sources avec un regard critique dans le but de pouvoir vérifier les informations plus facilement.

Par ailleurs, puisqu'étant principalement un travail de synthèse et de simulation, les démonstrations des différents théorèmes et propriétés ne sont pas disponibles en annexe mais sont trouvables dans les œuvres citées dans la bibliographie.

Enfin, si la partie théorique s'est faite sans trop d'encombres, je ne peux pas en dire autant sur la partie simulation. En effet, j'ai eu un moment de confusion lors de la simulation des variables dues à l'emploi de notations différentes entre les travaux. Par ailleurs, le calcul de l'erreur quadratique moyenne sous différentes conditions ne m'a pas permis de dégager une zone préférable pour les choix de α_n et h_n, le problème étant à chercher du côté du modèle ou du code R.

Je tiens cependant à préciser que je ne vois pas ce travail comme un échec mais comme une initiation à un domaine de la statistique qui m'était inconnu.

Annexe : code source commenté

Fonctions utilisées

```
#Créer une fonction inverse
inverse = function (f,m = 0,M = 100) {
      function (y) uniroot((function (x) f(x) - y), lower = m, upper =
M)[1]$root
}

#Noyau d'Epanechnikov
E=function(x){
3/4*(1-x^2)*(abs(x)<=1)
}

#Estimation d'une fonction de répartition conditionnelle utilisant le noyau
d'Epanechnikov
F_chap = function(y,Y,x,X,h){
      n = length(X)
      sum((Y<=y)*E((x-X)/h))/sum(E((x-X)/h))
}

#Estimation d'une densité utilisant le noyau d'Epanechnikov
g_chap = function(x,X,h){
      n = length(X)
      (1/(n*h))*sum(E((x-X)/h))
}

#Estimateur de S(u|x,z)
S_chap = function(u,x,Y,X,Z,h){
      n = length(X)
      (1/(n*h*g_chap(x,X,h)))*sum((Y>=q(u))*(Z>=q(u))*E((x-X)/h))
}

#Estimateur de lambda(u|x,z)
lambda_chap = function(u,x,Y,X,Z,h,alpha){
      1/alpha*S_chap(alpha*u,x,Y,X,Z,h)
}
```

Partie spécifique au modèle logistique

```
#Fonction theta(x), >=1
f_theta=function(x){
      log(2)/(log((2+exp(x))/(1+exp(x))))
}


#Fonction de répartition de z
F_z=function(z){
      exp(-1/z)
}

#Fonction de répartition de Z inversée
F_z_inv=inverse(function(z) F_z(z),0,50000)

#Fonction q(u|x)
q=inverse(function(u) (1-F_z(u)),0,50000)

#Génération de X, Y, Z
X=rnorm(500)
Z=sapply(runif(500,max=0.999),F_z_inv)
Y=c(1:500)

for (i in 1:500){
      F_y=function(y,z=Z[i],x=f_theta(X[i])){
            exp(1/z-(y^(-x)+z^(-x))^(1/x))*(y^(-x)+z^(-x))^(1/x-1)*z^(1-x)
            }
      F_y_inv=inverse(function(y) F_y(y),0.001,50000)
      Y[i] = F_y_inv(runif(1,min=0.001,max=0.999))
      }

#Représentation graphique de l'erreur quadratique moyenne
MSE = function(N){
M=matrix(0,ncol=40,nrow=40)
M2=matrix(1,ncol=40,nrow=40)
for(k in 1:N){
      X=rnorm(500)
      Z=sapply(runif(500,max=0.999),F_z_inv)
      Y=c(1:500)

      for (l in 1:500){
            F_y=function(y,z=Z[l],x=f_theta(X[l])){
                  exp(1/z-(y^(-x)+z^(-x))^(1/x))*(y^(-x)+z^(-x))^(1/x-
1)*z^(1-x)
                  }
            F_y_inv=inverse(function(y) F_y(y),0.001,50000)
            Y[l] = F_y_inv(runif(1,min=0.001,max=0.999))
            }
      for(i in 1:40){
            for (j in 1:40){
                  M2[i,j]=(lambda_chap(1,.5,Y,X,Z,i/100,j/100)-
(1^f_theta(.5)+.5^f_theta(.5))^(1/f_theta(.5)))^2
                  }
            }
      M=M+M2
}
image(M/N,zlim=c(0,1))
contour(M/N,add=T)
}
```

Bibliographie

L. Gardes, S. Girard, *Nonparametric estimation of the conditional tail dependence function.*

B. Khalifa (2008), *Estimation non-paramétrique par noyaux associés et données de panel en marketing.*

É. Youndjé (2011), *Contribution à l'estimation non-paramétrique par la méthode du noyau.*

P. Embrechts, C. Klüppelberg, T. Mikosch (1997), *Modelling extremal events for insurance and finance.*

R. Schmidt, U. Stadtmüller (2005), *Non-paramétric estimation of tail dependence.*

A. Berlinet, L. Devroye (1989), *Estimation d'une densité : un point sur la méthode du noyau.*

A. Sabourin (2012), *Introduction à la théorie des valeurs extrêmes.*

S. Girard (2007), *Introduction à la statistique des valeurs extrêmes.*

B. Raggad (2009), *Fondements de la théorie des valeurs extrêmes, ses principales applications et son apport à la gestion des risques du marché pétrolier.*

A. Gannoun, S. Girard, C. Guinot, J Saracco, *Implémentation en C d'estimateurs non paramétriques de quantiles conditionnels*