© Théo Jalabert

Inférence statistique des copules (première partie)

Esterina Masiello

M2 Actuariat 2020 – 2021

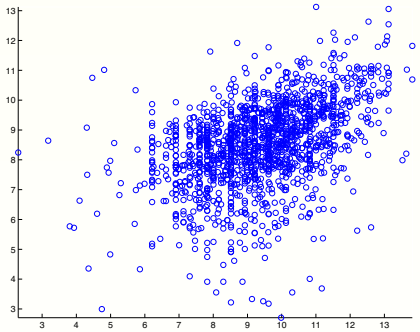
Cours en complément d'ERM.
→ Projet

Le problème est le suivant...

On est intéressé par la relation existante entre deux variables aléatoires réelles X et Y.

Par exemple, X : montant du sinistre (indemnity payment (LOSS))

Y : montant des dépenses d'enquêtes et de procédures (adjustment expenses (ALAE))



Représentation graphique des données LOSS/ALAE ($n = 1500$) de Frees and Valdez (1998) (données originales (échelle logarithmique)).

© Théo Jalabert

Une des possibles solutions : la copule !

Soit $F(x) = \Pr(X \leq x)$ et $G(y) = \Pr(Y \leq y)$ les fonctions de répartition de X et Y, respectivement.

Selon Sklar (1959), la fonction de répartition jointe du couple (X, Y) peut être écrite sous la forme

Copule décrit la structure de dépendance entre 2 variables.

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

pour une certaine fonction de répartition C dont les marges sont uniformes sur $[0, 1]$.
De plus, si F et G sont continues, la copule C est alors unique et on peut écrire:

$$\forall (u, v) \in [0, 1]^2 \qquad C(u, v) = H(F^{-1}(u), G^{-1}(v))$$

En fait, il s'agit de la loi du couple $(U, V) = (F(X), G(Y))$.

© Théo Jalabert

Un des enjeux avec les copules

Un des principaux enjeux avec les copules est choisir la copule qui s'adapte le mieux aux données d'intérêt.

Une des questions le plus fréquemment posé par les praticiens ou les jeunes chercheurs s'intéressant à ces thématiques pour la première fois est :

"Quelle copule faudrait-il utiliser pour un problème donné ?"

La question n'a pas de réponse évidente !

© Théo Jalabert

Méthodes d'estimation pour des modèles de copules

1) Estimation par l'approche paramétrique → je connais tout sauf les paramètres
→ Hyp sur loi X et Y puis modèle de copule mais on connait pas les paramètres

2) Estimation par l'approche semiparamétrique → Je connais le modèle de copules mais on ne fait plus d'hyp sur les marges

3) Estimations fondées sur la méthode des moments → On va essayer d'exploiter certains indicateurs de dépendance (ex: Tau de Kendall...) pour estimer le param de copule

4) Estimation par l'approche nonparamétrique → On ne connait plus rien, pas d'hypothèse on veut estimer la copule.

© Théo Jalabert

Encore un peu de notation

La notation suivante sera employée dans toute cette partie :

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ est un échantillon aléatoire de loi H.
- $(R_1, S_1), \ldots, (R_n, S_n)$ sont les paires de rangs associées.

  R_i est la place occupée par X_i dans l'échantillon ordonné en ordre croissant. Ex: $x_i : (7, 3, 10, 1)$ $(1, 3, 7, 10)$ $\rightarrow r_i : (3, 2, 4, 1)$

  place de 7 dans l'échantillon ordonné.

  place de 3 dans l'échantillon ordonné.
- C est la copule sous-jacente à la loi H.
- $\mathcal{C} = \{C_\theta : \theta \in \Theta\}$ est une famille paramétrique de copules.

© Théo Jalabert

1) Estimation par l'approche paramétrique

Supposons que certains modèles paramétriques soient envisagés pour F, G et C, e.g.,

$$F \in (F_\delta), \qquad G \in (G_\eta), \qquad C \in \mathcal{C} = \{C_\theta, \theta \in \Theta\}$$

Comment peut-on estimer le paramètre de dépendance θ ?

Des méthodes existent pour estimer les paramètres inconnus:

a) Maximum de vraisemblance

b) Méthode IFM (Inference function for margins)

© Théo Jalabert

Exemple

Pour fixer les idées, prenons

- F normale avec paramètre (bivarié) $\delta = (\mu, \sigma^2)$;
- G gamma avec paramètre $\eta = (\alpha, \lambda)$;
- $C \in \mathcal{C} = (C_\theta)$, la famille de copules FGM, définie pour chaque $\theta \in [-1, 1]$ par

$$C_\theta(u, v) = uv + \theta uv(1-u)(1-v), \qquad u, v \in (0, 1)$$

© Théo Jalabert

a) Maximum de vraisemblance

Notons les densités associées à F, G et C par

$$f_\delta, \qquad g_\eta, \qquad c_\theta$$

La contribution de chaque observation (x, y) à la vraisemblance est alors égale à

$$f_\delta(x) g_\eta(y) c_\theta\{F_\delta(x), G_\eta(y)\}$$

loi jointe

$H(x,y) = C_\theta(F_\delta(x), G_\eta(y))$ Théorème de Sklar

$\mathcal{L}((x_i, y_i)_{i=1,\dots,m}; \theta, \delta, \eta) = \prod_{i=1}^{m} h(x_i, y_i)$

densité de la copule

La densité jointe du couple (X, Y) s'écrit $h(x,y) = c_\theta(F_\delta(x), G_\eta(y)) f_\delta(x) g_\eta(y)$

© Théo Jalabert

a) Maximum de vraisemblance

L'estimation par maximum de vraisemblance de δ, η et θ est déterminée en maximisant

$$\begin{aligned} l(\delta,\eta,\theta) &= \sum_{i=1}^{n} \log\{f_\delta(x_i)\} + \log\{g_\eta(y_i)\} \\ &\quad + \sum_{i=1}^{n} \log[c_\theta\{F_\delta(x_i), G_\eta(y_i)\}] \end{aligned}$$

Noter que seul le dernier terme fait intervenir θ.

Inconventients : l'optimisation d'une fonction de plusieurs variables est exigeante au plan du calcul ! De plus, il y a un risque non négligeable de ne pas identifier correctement le maximum global.

Avantage : méthode que l'on connait bien et qui donne des estimateurs asympt., sans biais et de variance minimale.

MV - Exemple d'application en R

Nous générons un échantillon d'une loi bivariée avec marges gamma et copule gaussienne : → package Copula.

permet de construire le modèle de copule

```
myMvd <- mvdc(copula = ellipCopula(family = "normal",
param = 0.5),
+ margins = c("gamma", "gamma"), paramMargins =
list(list(shape = 2,
+ scale = 1), list(shape = 3, scale = 2)))
n <- 200
dat <- rmvdc(myMvd, n)
```

génère un échantillon suivant le modèle de Copule.

Les paramètres à estimer sont donc les paramètres des marges $\delta = (2, 1)$ et $\eta = (3, 2)$ et le paramètre $\theta = 0.5$ de la copule.

MV - Exemple d'application en R

La log-vraisemblance en correspondance de la vraie valeur du paramètre est :

```
loglikMvdc(c(2,1,3,2,0.5),dat,myMvd)
[1] -781.1641
```

Afin d'obtenir une estimation des paramètres par MV, nous utilisons la fonction `fitMvdc`, qui demande des valeurs initiales (celles de la méthodes des moments par exemple).

© Théo Jalabert

MV - Exemple d'application en R

Valeur init des paramètres à estimer

```
mm <- apply(dat, 2, mean)
vv <- apply(dat, 2, var)
b1.0 <- c(mm[1]2̂/vv[1], vv[1]/mm[1])
b2.0 <- c(mm[2]2̂/vv[2], vv[2]/mm[2])
a.0 <- sin(cor(dat[, 1], dat[, 2], method =
"kendall") * pi/2)

start <- c(b1.0, b2.0, a.0)
fit <- fitMvdc(dat, myMvd, start = start,
+ optim.control = list(trace = TRUE, maxit = 2000))
```

© Théo Jalabert

MV - Exemple d'application en R

```
fit
The ML estimation is based on 200 observations.
Margin 1 :
Estimate Std.  Error
m1.shape 1.830479 0.1689802
m1.scale 1.037388 0.1100475
Margin 2 :
Estimate Std.  Error
m2.shape 3.515646 0.3362403
m2.scale 1.628037 0.1672812
Copula:
Estimate Std.  Error
rho.1 0.419909 0.05820196
The maximized loglikelihood is -777.327
The convergence code is 0
```

attendu pour paramètres (2, 1, 3, 2, 0.5)
donc pas si mal.

b) Inference Functions for Margins (Joe, 1997)

Approche en deux temps de Joe :

1) On estime d'abord δ et η séparément à partir des log-vraisemblances marginales

$$\sum_{i=1}^{n} \log\{f_\delta(x_i)\} \qquad \sum_{i=1}^{n} \log\{g_\eta(y_i)\}$$

2) Ensuite on prend pour estimation de θ la valeur $\hat{\theta}$ qui maximise

$$\sum_{i=1}^{n} \log\{c_\theta(F_{\hat{\delta}}(X_i), G_{\hat{\eta}}(Y_i))\}$$

IFM - Exemple d'application en R

La méthode peut être appliquée en R grâce à la fonction fitCopula

```
loglik.marg <- function(b, x) sum(dgamma(x, shape =
b[1], scale = b[2],log = TRUE))
ctrl <- list(fnscale = -1)
b1hat <- optim(b1.0, fn = loglik.marg, x = dat[, 1],
control = ctrl)$par
b2hat <- optim(b2.0, fn = loglik.marg, x = dat[, 2],
control = ctrl)$par
udat <- cbind(pgamma(dat[, 1], shape = b1hat[1],
scale = b1hat[2]),
+ pgamma(dat[, 2], shape = b2hat[1], scale =
b2hat[2]))
fit.ifl <- fitCopula(udat, myMvd@copula, start = a.0)
```

IFM - Exemple d'application en R

L'estimation obtenue est la suivante :

```
c(b1hat, b2hat, fit.ifl@est)
[1] 1.8301906 1.0375602 3.5167486 1.6284508 0.4199346

fit.ifl
The ML estimation is based on 200 observations.
Estimate Std.  Error z value Pr(>|z|)
rho.1 0.4199346 0.05368175 7.822671 5.107026e-15
The maximized loglikelihood is 19.41620
The convergence code is 0
```

© Théo Jalabert

Forces et faiblesses

Principal avantage : Facilité d'implémentation.

Principal inconvenient :
Un mauvais choix de marges risque de "contaminer" l'estimation du paramètre de dépendance θ, même s'il est vrai que $C \in (C_\theta)$.

Joe (2005, JMVA) a montré que l'estimateur IFM est asymptotiquement normal et, sous des conditions de régularité raisonnables, il est aussi efficace que l'estimateur du MV dans le voisinage de l'indépendance ($\theta \approx 0$)

2) Estimation par l'approche semiparamétrique

On suppose que les marges sont inconnues et on les estime par la f.d.r. empirique :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(X_i \leq x)} \quad G_n(y) = \frac{1}{n}\sum_{i=1}^{n} I_{(Y_i \leq y)}$$

La méthode CML (Canonical Maximum Likelihood) (Oakes, 1994) consiste à maximer la log pseudo-vraisemblance, donnée par

$$\sum_{i=1}^{n} \log\{c_\theta(F_n(X_i), G_n(Y_i))\}$$

L'estimateur du maximum de pseudo-vraisemblance est fondée sur les rangs, puisque

$$F_n(X_i) = \frac{R_i}{n}, \qquad G_n(Y_i) = \frac{S_i}{n}$$

pour tout $i = 1, \ldots, n$.

© Théo Jalabert

Méthode CML - Exemple

Nous disposons des données suivantes : Exemple uniquement illustratif.

i	1	2	3	4	5	6
X_i	-2.224	-1.538	-0.807	0.024	0.052	1.324
Y_i	0.431	1.035	0.586	1.465	1.115	-0.847

i	1	2	3	4	5	6
R_i	1	2	3	4	5	6
S_i	2	4	3	6	5	1

© Théo Jalabert

Logarithme de la pseudo-vraisemblance

Sous l'hypothèse $C \in \mathcal{C} = \{C_\theta\}$, l'équation à résoudre est la suivante :

$$\begin{aligned} \dot{l}(\theta) &= \frac{\partial}{\partial\theta} l(\theta) \\ &= \sum_{i=1}^{n} \dot{c}_\theta \left(\underbrace{\frac{R_i}{n+1}}_{F_n(X_i)}, \underbrace{\frac{S_i}{n+1}}_{G_n(Y_i)} \right) / c_\theta \left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right) \\ &= 0 \end{aligned}$$

© Théo Jalabert

Cas de la copule FGM...

Nous avons :

$$c_\theta(u,v) = 1 + \theta(1-2u)(1-2v)$$

et donc

$$\frac{\dot{c}_\theta(u,v)}{c_\theta(u,v)} = \frac{(1-2u)(1-2v)}{1+\theta(1-2u)(1-2v)}$$

pour tout $(u,v) \in (0,1)$

La COPULE FARLIE-GUMBEL-MORGNSTER (FGM) s'écrit :

$C_\theta(u,v) = uv + \theta uv(1-u)(1-v)$

$c_\theta(u,v) = \frac{\partial^2 C_\theta(u,v)}{\partial u \partial v}$

Cas de la copule FGM...

La fonction de log pseudo-vraisemblance est donnée par :

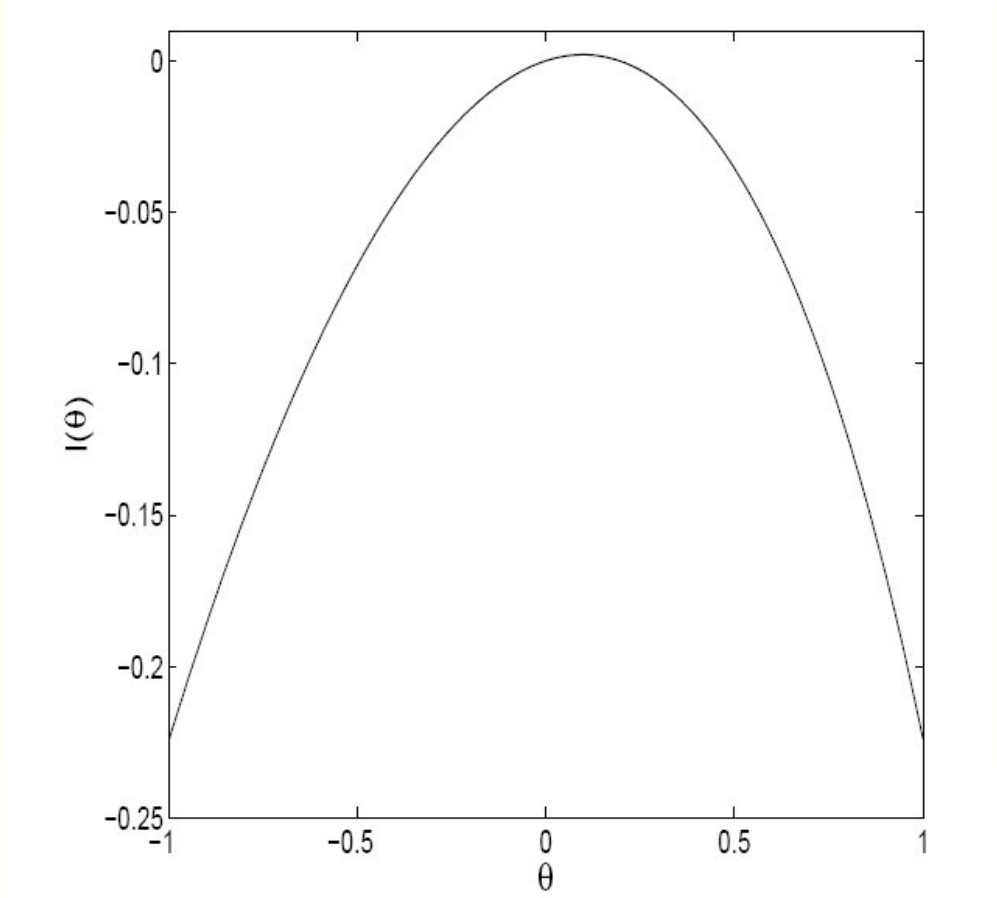
$$l(\theta) = \sum_{i=1}^{n} \log \left\{ 1 + \theta \left(1 - \frac{2R_i}{n+1} \right) \left(1 - \frac{2S_i}{n+1} \right) \right\}$$

La fonction pseudo-score qui correspond est donnée par:

$$\begin{aligned} \dot{l}(\theta) &= \sum_{i=1} n \frac{(1 - \frac{2R_i}{n+1})(1 - \frac{2S_i}{n+1})}{1 + \theta(1 - \frac{2R_i}{n+1})(1 - \frac{2S_i}{n+1})} \\ &= \sum_{i=1} n \frac{(n+1-2R_i)(n+1-2S_i)}{(n+1)^2 + \theta(n+1-2R_i)(n+1-2S_i)} \end{aligned}$$

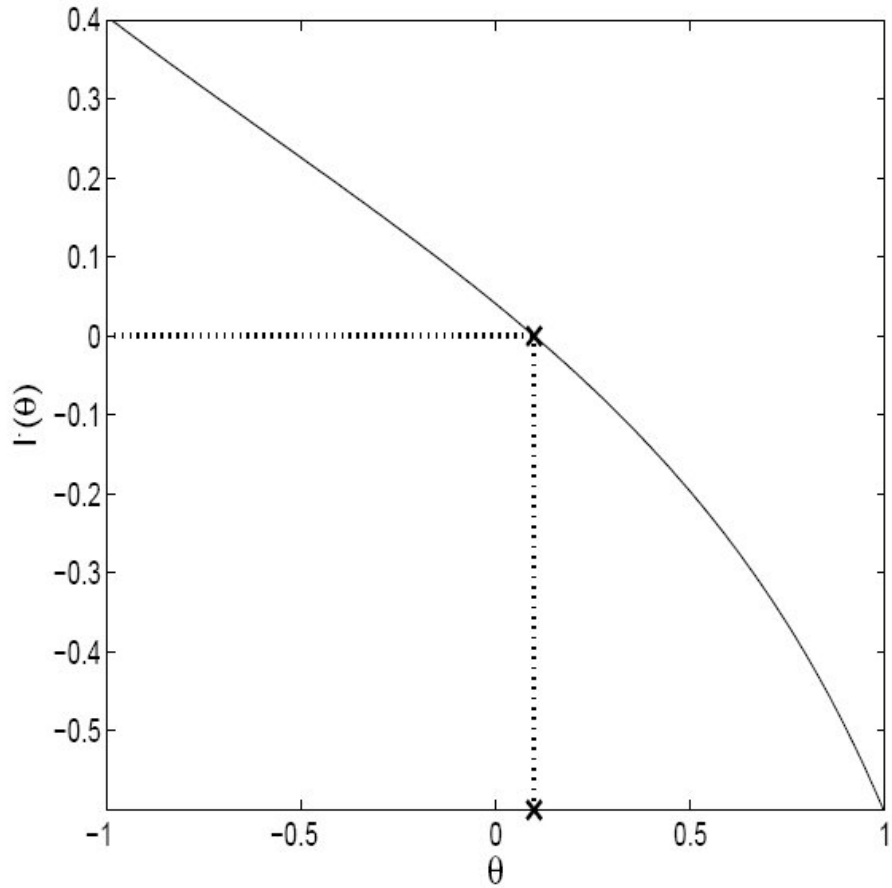
Log pseudo-vraisemblance FGM

→ $\hat{\theta}$ maximise cette vraisemblance



Fonction pseudo-score de la copule FGM

→ $\hat{\theta}$ est celui qui annule cette fct°



© Théo Jalabert

Estimation ponctuelle

Après substitution, on trouve $\hat{\theta}_n = 0.0989$ comme racine de l'équation (solution numerique):

$$\begin{aligned} \dot{l}(\theta) &= \frac{15}{49+15\theta} - \frac{3}{49-3\theta} + \frac{1}{49+\theta} \\ &+ \frac{5}{49+5\theta} + \frac{9}{49+9\theta} - \frac{25}{49-25\theta} = 0 \end{aligned}$$

CML - Exemple d'application en R

```
eu <- cbind((rank(dat[, 1]) - 0.5)/n, (rank(dat[, 2])
- 0.5)/n)
fit.cml <- fitCopula(eu, myMvd@copula, start = a.0)

fit.cml
The ML estimation is based on 200 observations.
Estimate Std.  Error z value Pr(>|z|)
rho.1 0.4304444 0.05311327 8.104272 4.440892e-16
The maximized loglikelihood is 20.29229
The convergence code is 0
```

valeur de la stat du test $H_0: \rho=0$ vs $H_1: \rho\neq 0$

p.value (très faible $\Rightarrow$ rejet de H_0)

Remarquez que l'estimation obtenue pour le paramètre θ de la copule est très différente des estimations obtenues par les méthodes MV et IFM !

© Théo Jalabert

Méthode CML (Canonical Maximum Likelihood) (Oakes, 1994)

Principal avantage :
la méthode est aussi facile à implémenter que celle de Joe. Vous n'avez qu'à travailler avec les rangs !

Principal inconvenient :
l'emploi d'estimations non paramétriques pour les marges induit une perte d'information. Si l'on connaissait la forme des marges avec certitude, la méthode de Joe mènerait à des estimations plus précises (i.e., de variance plus petite).

Mais en réalité, quand connaît-on vraiment les marges ?

© Théo Jalabert

3) Méthode des moments

θ paramètre de dépendance de la copule.

Si une certaine mesure de dépendance k peut être écrite comme :

$$k = h(\theta)$$

pour une certaine fonction h appropriée, alors

$$\hat{\theta} = h^{-1}(\hat{k})$$

sera l'estimateur des moments de θ.

Exemple : inversion du tau de Kendall. En effet, des expressions très simples pour $\tau = \Psi(\theta)$ existent pour la plupart des copules dans les modèles méta-elliptiques et les modèles archimédiens. Puisque τ_n est une mesure de dépendance non paramétrique, on peut estimer θ par :

$$\hat{\theta}_n = \Psi^{-1}(\tau_n)$$

© Théo Jalabert

METHODE DES MOMENTS

Supposons que le paramètre de dépendance $\theta \in \mathbb{R}$.
Si la famille de copules C_θ est ordonnée par DÉPENDANCE POSITIVE PAR QUADRANT (DPQ) i.e
$\theta_1 < \theta_2 \Rightarrow C_{\theta_1} \leq C_{\theta_2}$ alors $\tau = g(\theta)$ et $\rho = h(\theta)$ sont des fonct° ↗ de τ et ρ respt.
avec τ le τ de Kendall
ρ le ρ de Spearman

Le concept de DPQ est basé sur la comparaison des masses de proba accordées aux différents quadrants du plan p/rt à la masse correspondante sous l'hypothèse d'indépendance.
On dit aussi qu'une copule est en DPQ ssi $C(u,v) \geq uv$

On estime τ et ρ non paramétriquement par τ_n et ρ_n et ensuite on obtient une estimation de θ par l'approche basée sur la méthode des moments :

$$\hat{\theta} = g^{-1}(\tau_n) \text{ et/ou } \hat{\theta} = h^{-1}(\rho_n)$$

Rappel : MESURES DE CORRELATION DES RANGS ET CONCORDANCE

On considère une paire de v.a continues (X,Y)

Définition :
Le τ de Kendall associé au couple (X,Y) est défini par :

$$\tau(X,Y) = \mathbb{P}((X-X')(Y-Y')>0) - \mathbb{P}((X-X')(Y-Y')<0)$$

où (X',Y') est $\perp\!\!\!\perp$ de (X,Y) et possède les m̂ marges que (X,Y)

τ s'interprète comme la différence entre la proba de concordance et la proba de discordance.

* $\underbrace{-1}_{\text{antimonotone}} \leq \tau(X,Y) \leq \underbrace{1}_{\text{comonotone}}$ et $\tau(X,Y) = 0$ en cas d'indépendance

Définition :
Le ρ de Spearman associé au couple (X,Y) est défini comme :

$$\rho(X,Y) = 3\mathbb{P}((X-X^\perp)(Y-Y^\perp)>0) - 3\mathbb{P}((X-X^\perp)(Y-Y^\perp)<0)$$

où $(X^\perp, Y^\perp)$ est indépendant de (X,Y) et possède les mêmes marges que (X,Y) et les composantes sont indep. ($C(u,v)=uv$)

* $-1 \leq \rho(X,Y) \leq 1$

On peut écrire $\rho = 12 \int_{[0,1]^2} uv \, dC_\theta(u,v) - 3 = 12 \int_{[0,1]^2} uv \, c_\theta(u,v) \, du \, dv - 3$

© Théo Jalabert

Méthode des moments - Exemple

Comme exemple, nous montrons qu'il est possible d'obtenir une estimation de θ par inversion du coefficient ρ de Spearman.
En fait, dans le modèle FGM

$$\rho = \frac{\theta}{3}$$

Une estimation du paramètre θ peut être donc obtenue par

$$\theta_n = 3\rho_n$$

où

$$\rho_n = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} R_i S_i - 3\frac{n+1}{n-1}$$

Avec les données de notre exemple, nous obtenons :

$$\theta_n = 3/35 \approx 0.086$$ → dépendance très faible.

Méthode des moments - cont.

Comment calculer un intervalle de confiance?
Borkowf (2002, CSDA) a montré que

$$\rho_n \to^L N\left(\rho, \frac{\sigma_C^2}{n}\right)$$

où la variance asymptotique σ_C^2 depend de la copule C sous-jacente. Par la méthode Delta

$$\theta_n \to^{\mathrm{L}} N\left(\theta, \frac{[\sigma_C h'(\rho_n)]^2}{n}\right)$$

© Théo Jalabert

Rappel : Méthode delta

Soit T_n un estimateur du paramètre θ et supposons vouloir estimer une fonct° $g(\theta)$.
Alors, sous certaines conditions, si $\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$, on a que :

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 (g'(\theta))^2)$$

© Théo Jalabert

Méthode des moments - cont.

Un estimateur convergent de σ_C^2 est donné par

$$\sigma_n^2 = 144(-9A_n^2 + B_n + 2C_n + 2D_n + 2E_n)$$

avec

$$A_n = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{n+1}\frac{S_i}{n+1}$$

$$B_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_i}{n+1}\right)^2\left(\frac{S_i}{n+1}\right)^2$$

$$C_n = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{R_i}{n+1}\frac{S_i}{n+1}1_{\{R_k\leq R_i, S_k\leq S_j\}} + \frac{1}{4} - A_n$$

© Théo Jalabert

Méthode des moments - cont.

$$D_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{S_i}{n+1} \frac{S_j}{n+1} \max\left(\frac{R_i}{n+1}, \frac{R_j}{n+1} \right)$$

$$E_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{R_i}{n+1} \frac{R_j}{n+1} \max\left(\frac{S_i}{n+1}, \frac{S_j}{n+1} \right)$$

© Théo Jalabert

Le τ de Kendall peut être estimé par $\tau_m = \frac{4}{m(m-1)} \sum_{i<j} \mathbb{1}_{\{(X_i-X_j)(Y_i-Y_j)>0\}} - 1$

$\sqrt{m}\left(\frac{\tau_m - \tau}{4S}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ avec $S^2 = \frac{1}{m}\sum_{i=1}^{m} (W_i + \tilde{W}_i - 2\bar{W})^2$

où $\tilde{W}_i = \frac{1}{m}\mathrm{Card}(\{j : X_i \le X_j, Y_i \le Y_j\})$ → aussi noté $\frac{1}{m}\#\{j : X_i \le X_j, Y_i \le Y_j\}$

$W_i = \frac{1}{m}\mathrm{Card}(\{j : X_i > X_j, Y_i > Y_j\})$

$\bar{W} = \frac{1}{m}\sum_{i=1}^{m} W_i$

Grâce à la méthode delta, on obtient que :

$$\tilde{\theta} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\theta, \frac{1}{m}(4S\, g'(\tau_m))^2\right) \qquad \text{où } \tilde{\theta} = g(\tau_m)$$

Pour la copule FGM, la relation qui lit θ et τ s'écrit $\tau = \frac{2\theta}{9}$ (cela vient de $\tau = 4\int_{[0,1]^2} C_\theta(u,v)\,dC_\theta(u,v) - 1$)

© Théo Jalabert

4) Estimation par l'approche nonparamétrique

Question plus générale : comment peut-on estimer (ou choisir)

$$C(u,v) = \text{distribution de} \quad (U,V) = (F(X), G(Y))$$

alors que dans la pratique, on a des observations

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim H(x,y) = C\{F(x), G(y)\}$$

et que ni F ni G ne sont connus ?

© Théo Jalabert

Soit $(X,Y) \sim H$ où $X \sim F$ et $Y \sim G$

$$\xrightarrow{\text{Sklar}} H(x,y) = C(F(x), G(y))$$

$F(x) = u$ et $G(y) = v$

$$\rightarrow \underbrace{H(F^{-1}(u), G^{-1}(v))}_{} = C(u,v)$$

$$= \mathbb{P}(X \leq F^{-1}(u), Y \leq G^{-1}(v))$$

$$= \mathbb{P}(F(X) \leq u, G(Y) \leq v) = \mathbb{P}(U \leq u, V \leq v)$$ donc la copule C est la loi de $(U,V) = (F(X), G(Y))$

$$\rightarrow \begin{cases} (U,V) \sim C \\ (X,Y) \sim H \end{cases}$$

$(U,V) = (F(X), G(Y))$

$$\rightarrow \boxed{(U_i, V_i) = (F_m(X_i), G_m(Y_i)) = \left(\frac{R_i}{m}, \frac{S_i}{m}\right)}$$

© Théo Jalabert

Solution

Pour obtenir des observations $(U_1, V_1), \ldots, (U_n, V_n) \sim C(u, v)$, on estime les marges F et G par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(X_i \leq x)}, \qquad G_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1_{(Y_i \leq y)}$$

Puis on pose

$$U_i = F_n(X_i), \qquad V_i = G_n(Y_i)$$

$1 \leq i \leq n$.

Point crucial

La procédure décrite ci-dessus revient à poser

$$U_i = \frac{R_i}{n}$$

$$V_i = \frac{S_i}{n}$$

où R_i est le rang de X_i parmi $X_1, \ldots, X_n$ et S_i est le rang de Y_i parmi $Y_1, \ldots, Y_n$.

Comment sélectionner une famille de copules appropriée...

Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon de loi inconnue H.
Soient F_n et G_n les f.d.r. empiriques associées à F et G.

a) Une premiere étape vers la sélection d'une famille de copules consiste à dessiner les points

RANK-RANK PLOT := Plot des

$$\left(\frac{R_i}{n}, \frac{S_i}{n}\right) = (F_n(X_i), G_n(Y_i)), \quad i \in \{1, \ldots, n\}$$

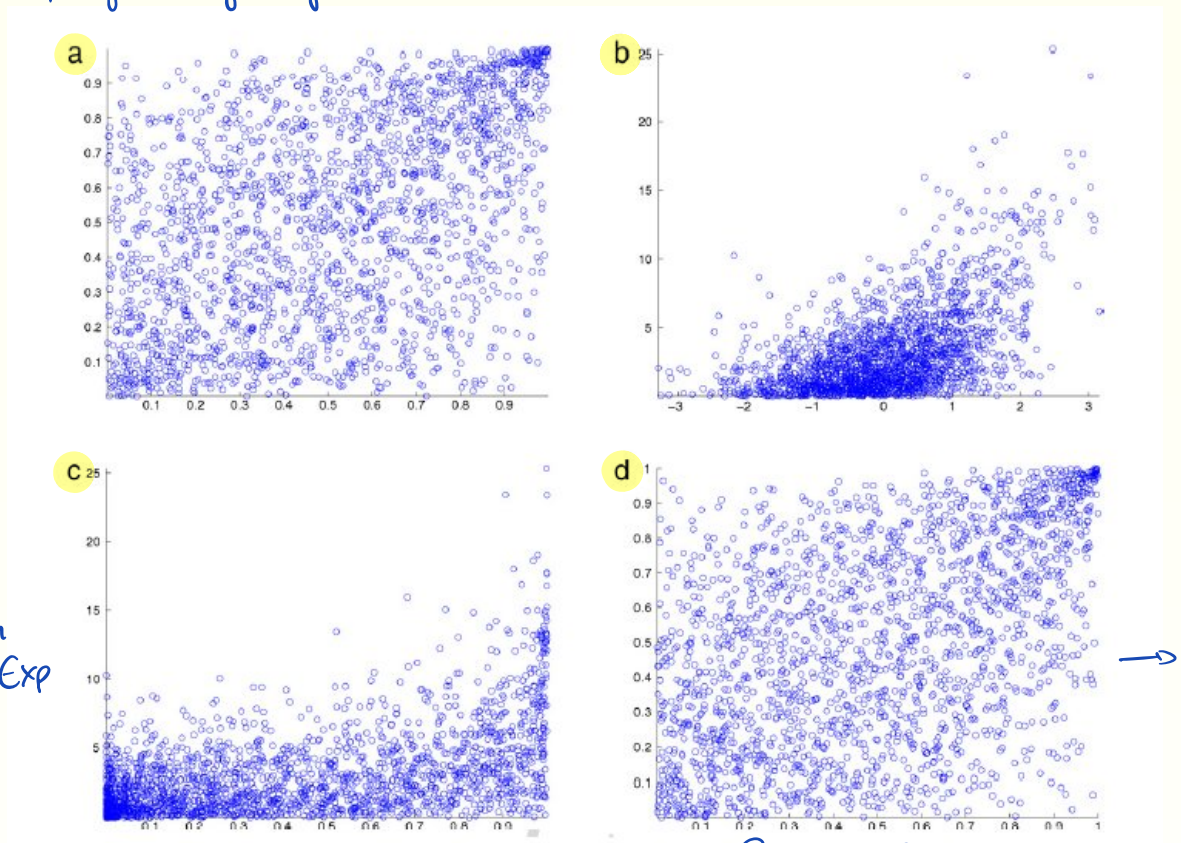
où R_i est le rang de X_i parmi $X_1, \ldots, X_n$ et S_i est le rang de Y_i parmi $Y_1, \ldots, Y_n$.
La motivation qui se cache derrière cette approche graphique est la suivante : les pseudo observations $(R_i/n, S_i/n)$ sont des substituts proches des paires inobservables $(U_i, V_i) = (F(X_i), G(Y_i))$ qui forment un échantillon provenant de C.

© Théo Jalabert

Structure de dep = Copule de Gumbel
Ce sont les marges qui changent

marges Unif Unif

marge Gauss Exp



Lorsqu'on veut avoir image structure de dep il faut pas faire le plot des X, Y car les marges peuvent cacher la structure de dep => rank-rank plot

Bin et Exp

Rank Rank Plot

→ Copule asym
=> non aux copules ellipt., gauss...

Scatter et rank-rank plots pour un échantillon de taille $n = 2000$ d'une copule de Gumbel-Hougaard ($\tau = 1/3$) avec marges (a) $F = G = \mathcal{U}(0,1)$; (b) $F = \mathcal{N}(0,1)$, $G = \mathcal{E}(3)$; (c) $F = \mathcal{B}(0.5, 0.7)$, $G = \mathcal{E}(3)$; (d) le rank-rank plot associé aux trois échantillons est le même.

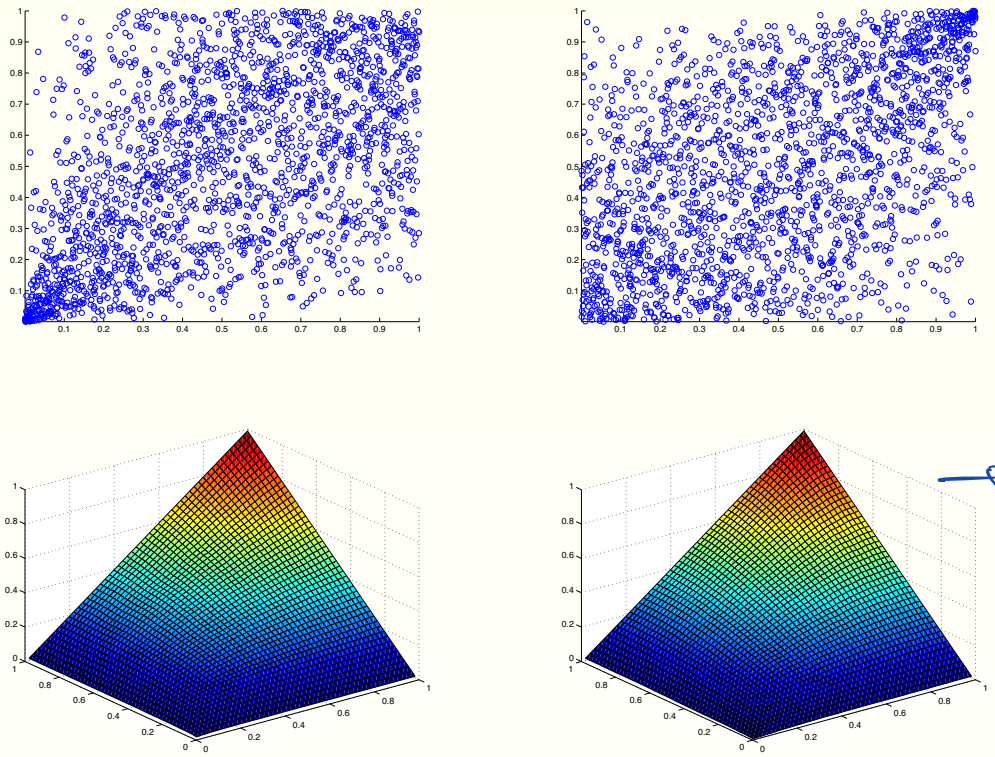
b) Copule empirique

Deheuvels (1979) appelle "copule empirique" la fonction

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} 1_{\left\{\frac{R_i}{n}\leq u, \frac{S_i}{n}\leq v\right\}}, \qquad u,v \in (0,1)$$

Cette loi de probabilité place une masse de $1/n$ en chaque paire de rangs normalisés.

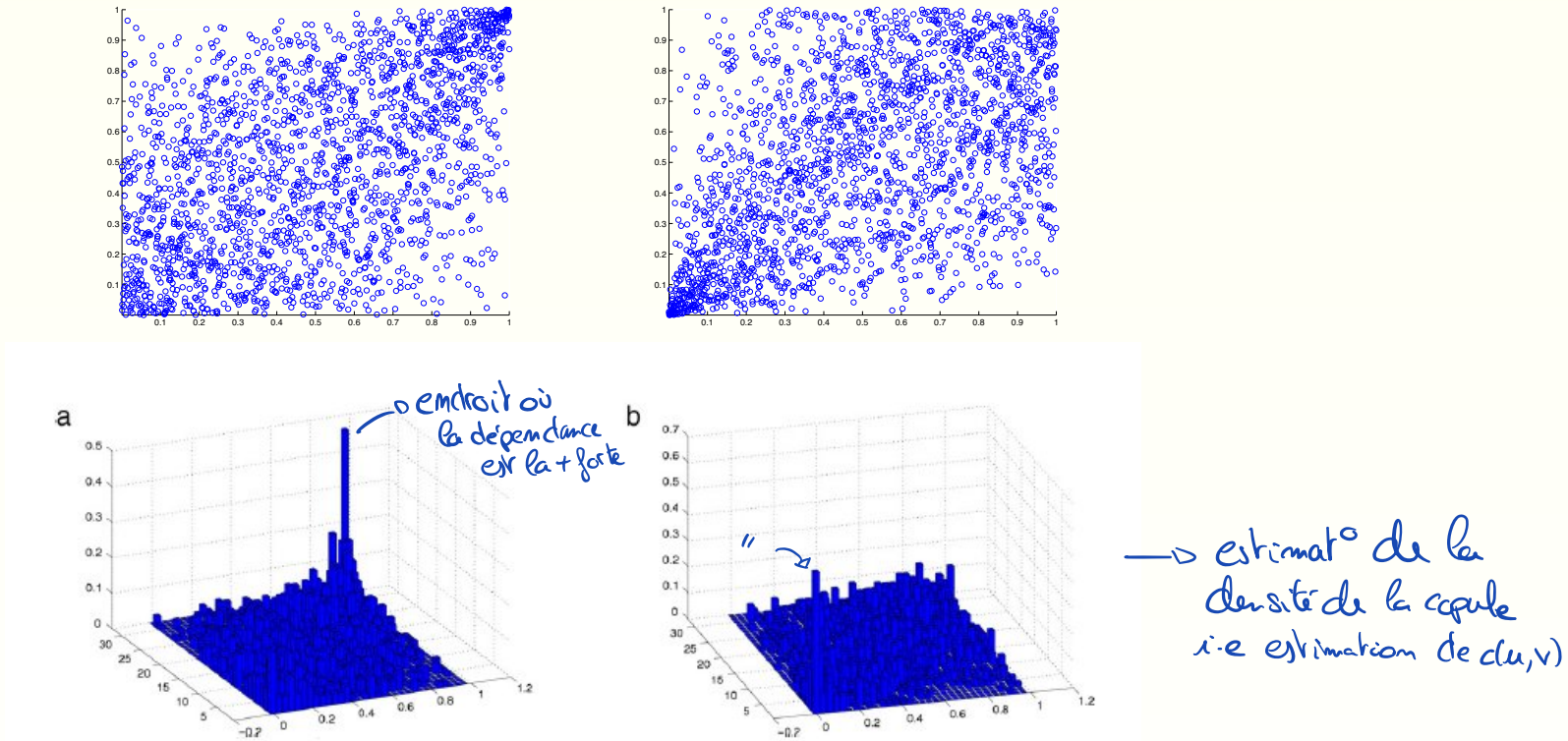
Il s'agit de la f.d.r. empirique C_n des couples $(R_i/n, S_i/n)$.



→ Copule empirique très utile d'un pt de vue théorique mais d'un pt de vue pratique ce n'est pas très utile.

Rank-rank plots et copule empirique correspondante pour deux échantillons de taille $n = 2000$ d'une copule de Clayton $(\theta = 1)$ et d'une copule de Gumbel-Hougaard $(\theta = 1.5)$. Dans les deux cas, les marges sont $F = \mathcal{B}(0.5, 0.7)$ et $G = \mathcal{E}(3)$.

c) Histogramme 3D



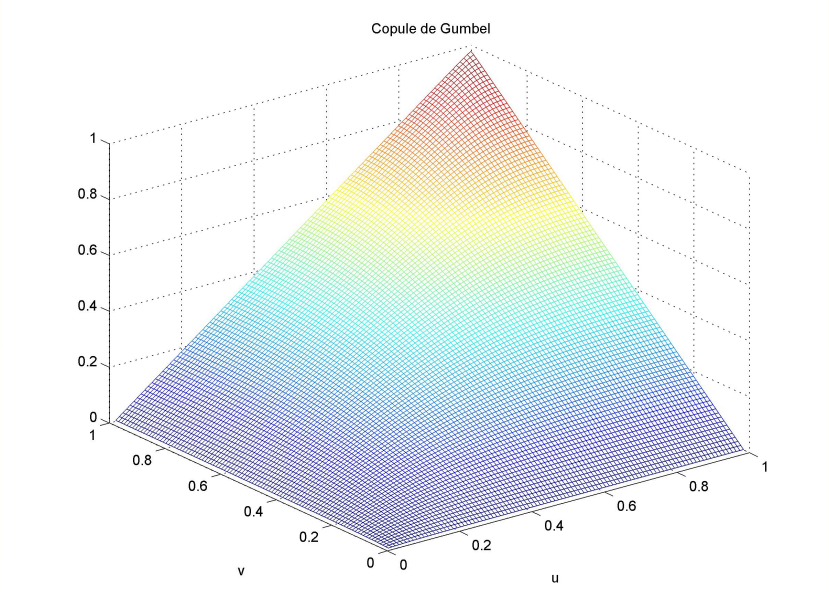
Rank-rank plots et histogramme 3D correspondant pour deux échantillons de taille $n = 2000$ d'une copule de Gumbel-Hougaard $(\theta = 1.5)$ et d'une copule de Clayton $(\theta = 1)$.

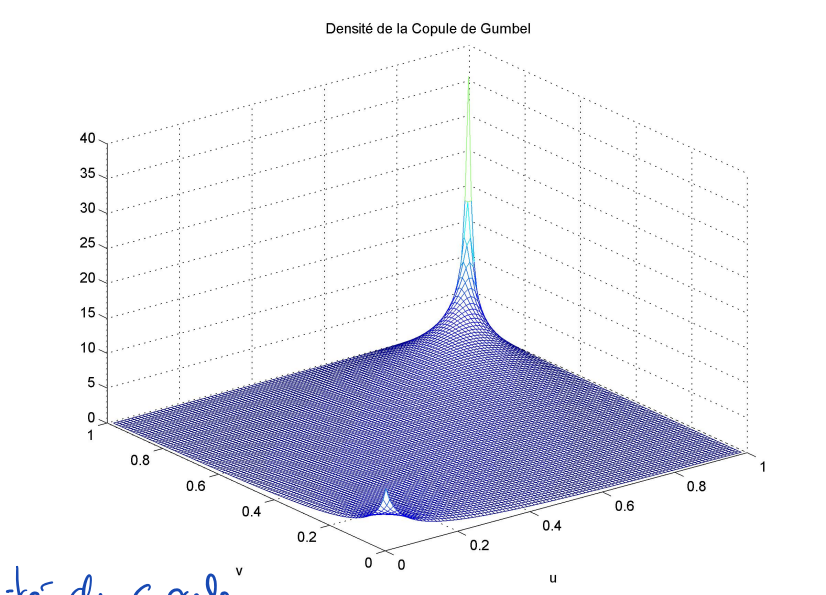
On cherche à estimer la densité d'une copule...

C doit être Δ^2

Si C est deux fois dérivable, on peut définir sa densité comme :

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





densité de copule
permet de décrire exactem^t la structure de dépendance de l'échantillon

© Théo Jalabert

Ce qui existe en estimation de c...

- un estimateur à noyau;
- un estimateur basé sur un critère de distance minimale (Biau and Wegcamp (2005));
- un estimateur basé sur les ondelettes.

Notions de base sur l'estimation nonparamétrique d'une densité

L'idée de base pour obtenir un estimateur de la densité au point x consiste à compter le nombre d'observations dans un voisinage de x (e.g. en $[x-h, x+h)$ pour $h > 0$).
Pour cela, considérons l'"histogramme mobile" ou ''"estimateur naif" proposé par Rosenblatt (1956):

$$\hat{f}(x) = \frac{1}{2nh} \sum_{i=1}^{n} 1_{\{X_i \in [x-h, x+h)\}}$$

Des extensions de cet estimateur peuvent être facilement obtenue en utilisant d'autres définitions de voisinage de x,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

où K est le noyau (e.g. $K(\omega) = 1_{\{|\omega| \leq 1\}}/2$).

L'estimateur à noyau de c

Un estimateur à noyau de la densité de copule basé sur l'échantillon $(F_n(X_1), G_n(Y_1)), \ldots, (F_n(X_n), G_n(Y_n))$ est donné par:

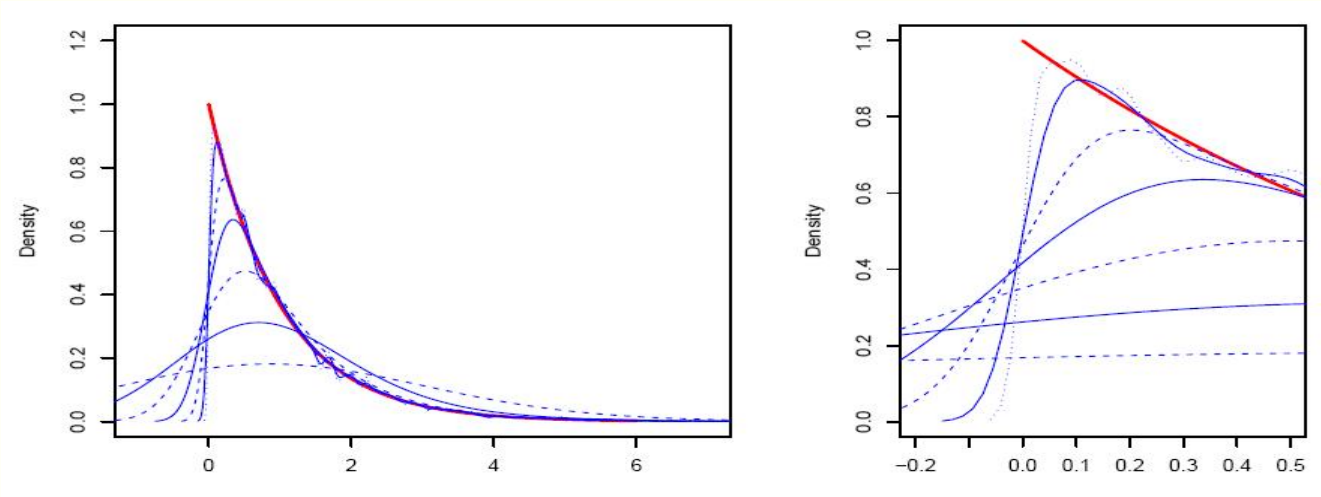
$$\hat{c}(u,v) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{u - F_n(X_i)}{h}, \frac{v - G_n(Y_i)}{h}\right)$$

(Behnen et al. (1985), Gijbels et Mielniczuk (1990), Fermanian (2005))

Quelques problèmes théoriques à etudier:

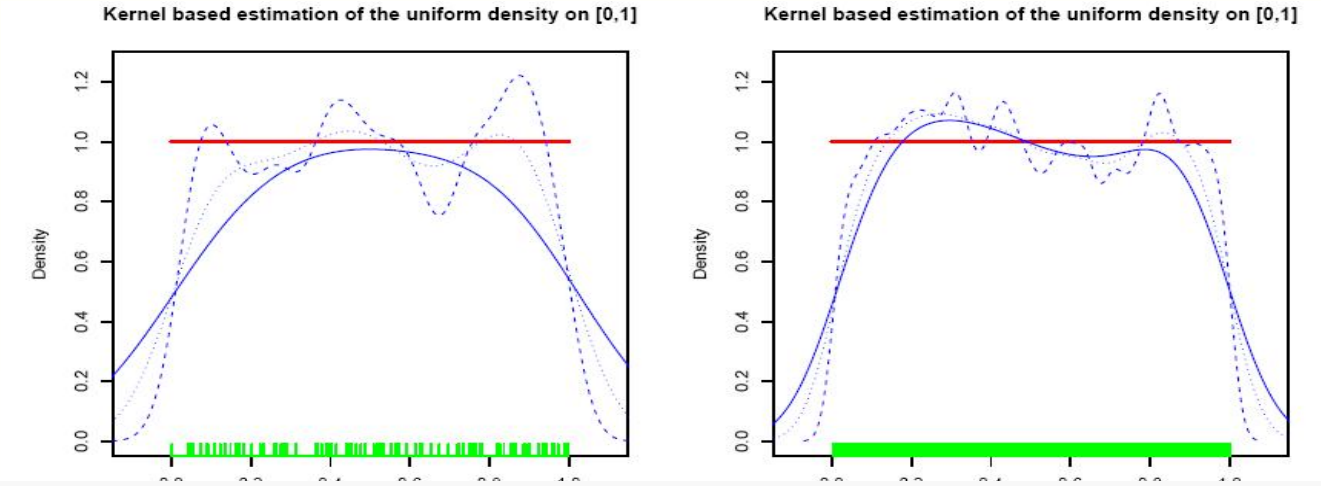
- problèmes de bord;
- pas de vitesse de convergence.

Dans le cas de densité unidimensionnelle...

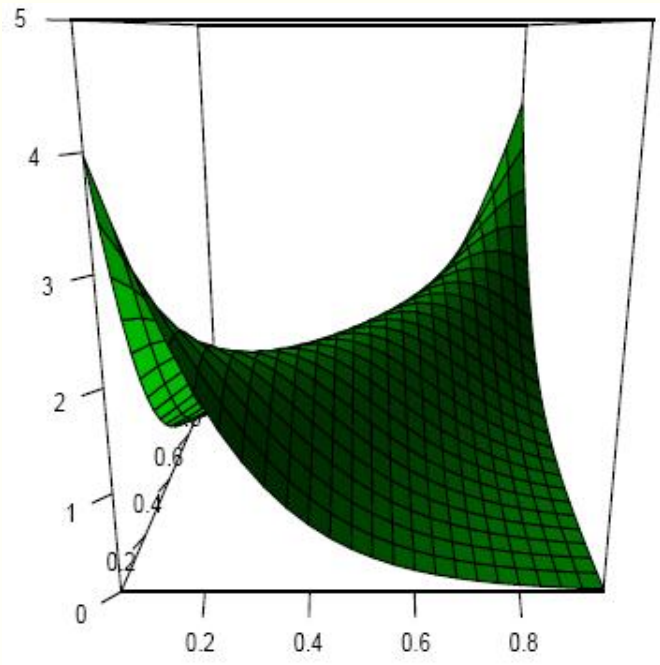


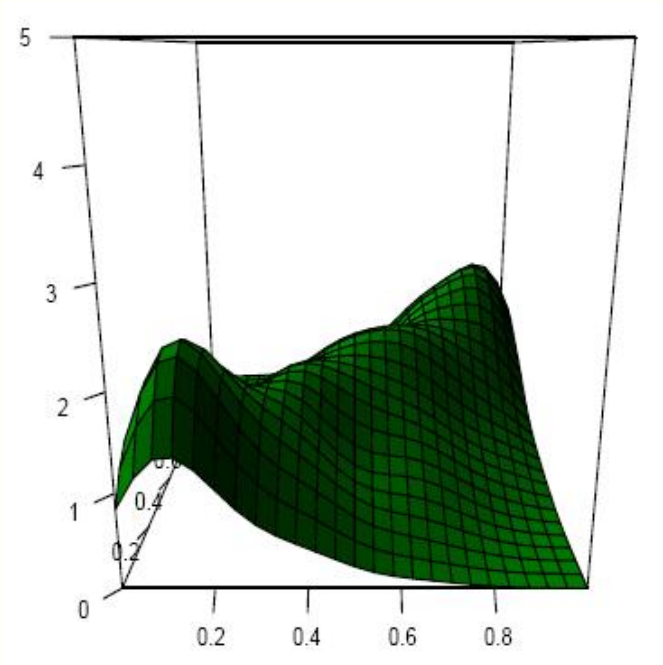
Ce qui change ce sont les valeurs des fenêtres $[x-h, x+h]$

Estimation de la densité d'une exponentielle (haut) et d'une uniforme (bas) à travers le noyau ($n = 10000$).



Et pour la copule...





Copule de Frank (gauche) et estimation de la densité de la copule de Frank à travers le noyau gaussian (droite).

- Article bibliographique: Charpentier et al. (2007)

Les problèmes de bord

Lorsque l'on s'intéresse à la dépendance de queue, sous-estimer la densité de copule dans le voisinage de $(1, 1)$ peut entraîner des problèmes graves, par exemple des problèmes importants de under-pricing!

Donc, comment bien estimer sur les bords ?

© Théo Jalabert

Obtenir une meilleure estimation sur les bords...

Plusieures techniques ont été proposées afin d'obtenir une meilleure estimation sur les bords :

- boundary kernel (Muller (1991));
- mirror image modification (Deheuvels and Hominal (1989), Schuster (1985));
- transformed kernel (Devroye and Gyorfi (1981), Wand, Marron and Ruppert (1991)).

Dans le cas particulier de densités sur $[0, 1]$:

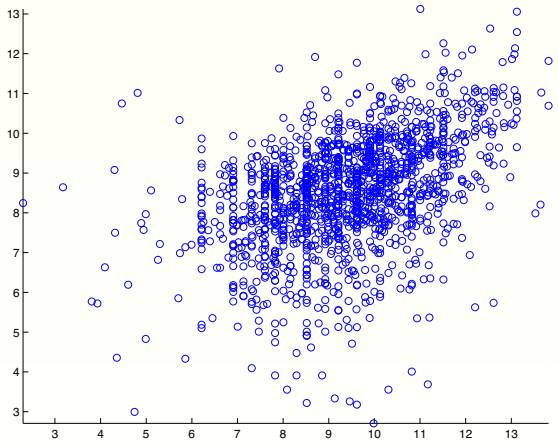
- Beta kernel (Brown and Chen (1999), Chen (1999, 2000))

Et aussi :

- average of histograms inspiré par Dersko (1998).

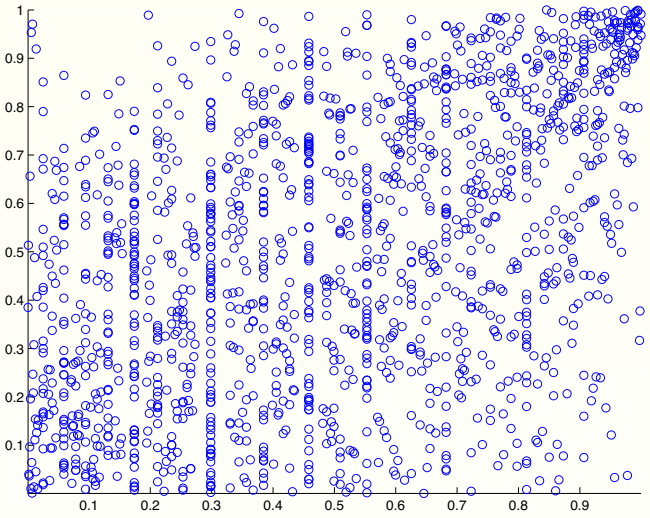
Le problème était...

X : indemnité (LOSS)
Y : frais de traitement de la réclamation (y compris les frais légaux, le cas échéant) (ALAE)



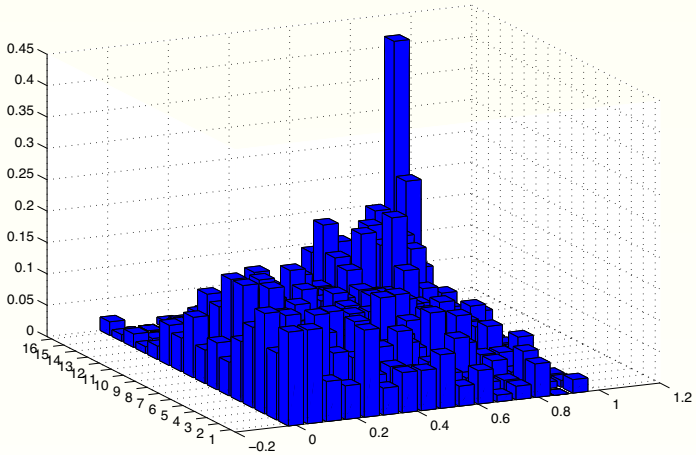
Représentation graphique des donnés LOSS/ALAE ($n = 1500$) de Frees and Valdez (1998)(données originales (échelle logarithmique))

Premiere étape



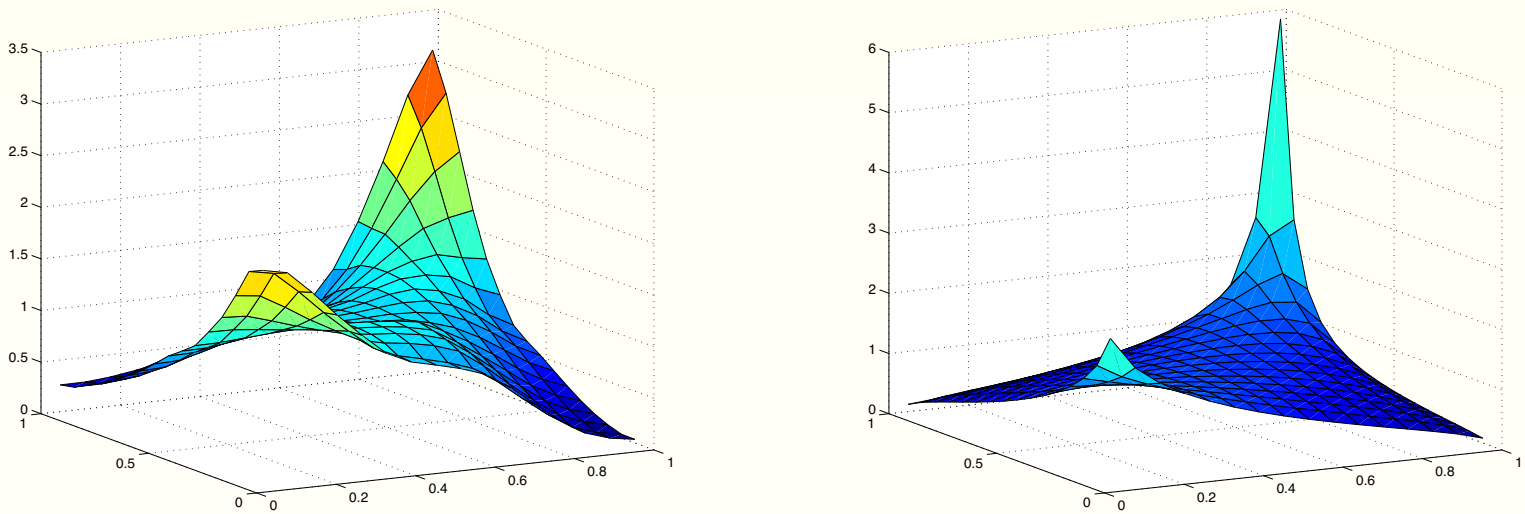
Rank-rank plot des données LOSS/ALAE ($n = 1500$) de Frees and Valdez (1998).

Deuxieme étape



histogramme 3D sur une grille 16×16 des données LOSS/ALAE ($n = 1500$) de Frees and Valdez (1998).

© Théo Jalabert



Densité de la copule estimée par ondelettes c^* (gauche) et densité d'une copule de Gumbel de paramètre $\hat{\theta} = 1.4$ (droite).