Analyse de la dépendance avec R, une brève introduction aux copules

–R User Group Toulouse–

Tom Rohmer

19 septembre 2019

Summary

Summary

Autour de la dépendance, exemple du crash boursier du 19 octobre 1987

- Des données multivariées et sériellement dépendantes.
- Question: Changement dans la dépendance?

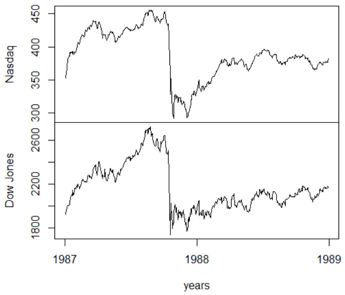


Figure: Nasdaq, Dow Jones and the "black Monday" (1987-10-19), R library QRM

Mesurer la dépendance

Considérons $(X_1, Y_1), \ldots, (X_n, Y_n)$ des copies indépendantes de (X, Y).

1. Coefficient de Pearson

$$\tau_\pi = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} \qquad \hat{\tau}_\pi = \frac{\displaystyle\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\displaystyle\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}}$$

2. Coefficient de Spearman

$$\rho = \frac{\mathrm{cov}(F(X), G(Y))}{\sqrt{\mathrm{var}(F(X))\mathrm{var}(G(Y))}} \qquad \hat{\rho} = \hat{\tau}_\pi(\underline{R}_X, \underline{R}_Y)$$

3. Coefficient de concordance (Kendall, ginni, beta de Blomqvist,etc.)

```
>cor(X,Y,method=c("pearson", "kendall", "spearman"))
```

Summary

Analyse multivariée, dépendance, copule

Théorème de Sklar (1959)

Soit $(X_1, \dots, X_d)$ un vecteur aléatoire. Notons $H(\boldsymbol{x}) = \mathrm{P}(X_1 \leq x_1, \dots X_d \leq x_d)$ sa f.d.r. et soient $F_1, \dots, F_d$ les f.d.r. marginales, supposées continues. Alors il existe une unique fonction $C : [0,1]^2 \to [0,1]$ telle que:

$$H(\boldsymbol{x}) = C\left(F_1(x_1), \dots, F_d(x_d)\right), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

[Copules]

1. Caracteriser des structures de dépendance (non nécéssairement linéaires)
2. Modéliser les intéractions entre plusieurs covariables
3. Expliquer un phénomène en fonction de ces intéractions
4. Détecter des changements dans la dépendance, dans le temps ou l'espace
5. Application sur un large spectre de données: finance, biologie, génétique.

Classical copulas

- Independence copula:

$$C^{\Pi}(\boldsymbol{u}) = \prod_{j=1}^{d} u_j;$$

- Normal copulas

$$C_{\Sigma}^{N}(\boldsymbol{u}) = \Phi_{d,\Sigma}\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\};$$

- Gumbel–Hougaard copulas:

$$C_{\theta}^{GH}(\boldsymbol{u}) = \exp\left(-\left[\sum_{j=1}^{d}\{-\log(u_j)\}^{\theta}\right]^{1/\theta}\right), \quad \theta \geq 1;$$

- Clayton copulas:

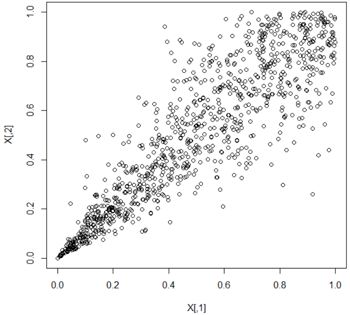
$$C_{\theta}^{Cl}(\boldsymbol{u}) = \left(\sum_{j=1}^{d} u_j^{-\theta} - d + 1\right)^{-1/\theta}, \quad \theta > 0.$$

→My Copulas here

Package R: copula

Kojadinovic, Ivan and Yan, Jun and others
Modeling multivariate distributions with continuous margins using the copula R package
Journal of Statistical Software, 2010

```
\library(copula)
 X=rCopula(100,claytonCopula(5))
head(X)
           [,1]      [,2]
 [1,] 0.7331002 0.5318563
 [2,] 0.5412551 0.5522277
 [3,] 0.7795055 0.9763939
 [4,] 0.8916388 0.7930044
 [5,] 0.4153212 0.4089746
 [6,] 0.3304539 0.4557191
```



Summary

Copulas and test for breaks detection

$$\mathcal{H}_0 : \exists\, F \text{ such that } \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \text{ have c.d.f. } F.$$

Sklar's theorem allows to rewrite $\mathcal{H}_0$ as $\mathcal{H}_{0,m} \cap \mathcal{H}_{0,c}$ where

$\mathcal{H}_{0,m} \cap \mathcal{H}_{0,c}$:

$$\begin{aligned} \mathcal{H}_{0,c} :& \quad \exists C, \text{ such that } \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \text{ have copula } C \\ \mathcal{H}_{0,m} :& \quad \exists F_1, \ldots, F_d \text{ such that } \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \text{ have m.c.d.f. } F_1, \ldots, F_d\,. \end{aligned}$$

- Construction of a test for $\mathcal{H}_0$ more powerful than its predecessors against alternatives involving a change in the copula, based on the CUSUM approach.
- F, $F_1, \ldots, F_d$ and C are unknown.

Estimation non-paramétrique de la copule

Considérons les vecteur $\boldsymbol{U}_i = (F_1(X_{i1}), \ldots, F_d(X_{id}))$. La copule du vecteur aléatoire est exactement la fonction de répartition du vecteur aléatoire U_i:

$$C(u_1, \ldots, u_d) = P(U_{i1} \leq u_1, \ldots, U_{id} \leq u_d).$$

Pour $j = 1, \ldots, d$ soit $F_{1:n,j}$ la f.d.r. empirique associée à l'échantillon $X_{1j}, \ldots, X_{nj}$. Pour $i = 1, \ldots, n$, considérons les vecteurs (pseudo-observations):

$$\hat{\boldsymbol{U}}_i^{1:n} = (F_{1:n,1}(X_{i1}), \ldots, F_{1:n,d}(X_{id})) = \frac{1}{n}(R_{i1}^{1:n}, \ldots, R_{id}^{1:n}),$$

Copule empirique, Rüschendorf(1976), Deheuvels(1979)

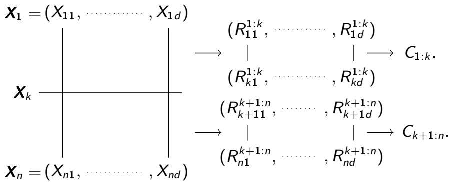
$$C_{1:n}(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\hat{\boldsymbol{U}}_i^{1:n} \leq \boldsymbol{u}), \qquad \boldsymbol{u} \in [0,1]^d.$$

$$\begin{array}{lcccc}
\boldsymbol{X}_1 = (X_{11}, \cdots\cdots, X_{1d}) & & (R_{11}^{1:n}, \cdots\cdots, R_{1d}^{1:n}) & & \\
\quad\;\; \vdots \qquad\qquad\quad \vdots & \longrightarrow & \vdots \qquad\qquad\quad \vdots & \longrightarrow & C_{1:n}. \\
\boldsymbol{X}_n = (X_{n1}, \cdots\cdots, X_{nd}) & & (R_{n1}^{1:n}, \cdots\cdots, R_{nd}^{1:n}) & &
\end{array}$$

Break detection in copula

A Cramér-von Mises statistic:

$$S_n = \max_{k \in \{1,\dots,n-1\}} \frac{1}{n} \sum_{i=1}^{n} \left(\sqrt{n} \frac{k}{n} \frac{(n-k)}{n} \{ C_{1:k}(\hat{\boldsymbol{U}}_i^{1:n}) - C_{k+1:n}(\hat{\boldsymbol{U}}_i^{1:n}) \} \right)^2 ,$$



S_n est une fonctionnelle du processus de copule empirique séquentielle

Sequential empirical copula process

$$\mathbb{C}_n(s,t,\boldsymbol{u}) = \frac{1}{\sqrt{n}} \sum_{i=\lfloor ns \rfloor + 1}^{\lfloor nt \rfloor} \left\{ \mathbf{1}(\hat{\boldsymbol{U}}_i^{\lfloor ns \rfloor + 1 : \lfloor nt \rfloor} \leq \boldsymbol{u}) - C(\boldsymbol{u}) \right\}.$$

Tests basés sur un rééchantillonage de la statistique

i.i.d. multipliers (voir par.ex. van der Vaart and Wellner(2000))

A sequence of i.i.d. multipliers $(\xi_i)_{i\in\mathbb{Z}}$ satisfies the following conditions:

- For all $i \in \mathbb{Z}$, ξ_i are independent of observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$
- $\mathbb{E}(\xi_0) = 0$, $\mathrm{var}(\xi_0) = 1$ and $\int_0^\infty \{\mathrm{P}(|\xi_0| > x)\}^{1/2} \mathrm{d}x < \infty$.

For $(s, t, \boldsymbol{u}) \in [0, 1]^{d+2}$, $s \leq t$, consider the processes

$$\check{\mathbb{B}}_n^{(m)}(s, t, \boldsymbol{u}) = \frac{1}{\sqrt{n}} \sum_{i=\lfloor ns\rfloor+1}^{\lfloor nt\rfloor} \xi_{i,n}^{(m)} \{\mathbf{1}(\hat{\boldsymbol{U}}_i^{\lfloor ns\rfloor+1:\lfloor nt\rfloor} \leq \boldsymbol{u}) - C_{\lfloor ns\rfloor+1:\lfloor nt\rfloor}(\boldsymbol{u})\}.$$

Finallement $\check{S}_n^{(m)}$ est une fonctionnelle (complexe) du processus $\check{\mathbb{B}}_n^{(m)}$.

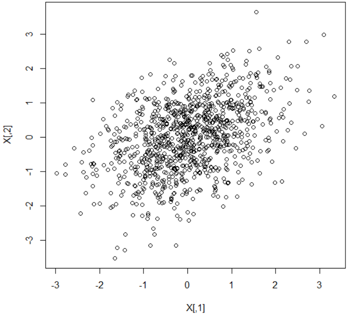
$$\hat{p}_{val} = \frac{1}{M} \sum_{m=1}^{M} 1(\check{S}_n^{(m)} > S_n)$$

Exemple 1: pas de changement dans la copule ni dans les marges

```
library(npcopTest)
set.seed(12345)
n=100
sigma = matrix(c(1,0.4,0.4,1),2,2)
X=matrix(rep(0,n*2),n,2)
for(j in 1:n)
X[j,]=t(chol(sigma))%*%rnorm(2)
```

ou bien dans library mvtnorm

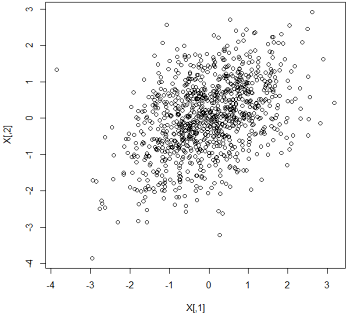
```
X = rmvnorm(100,mean=rep(0,2),sigma)
```




```
>CopTestdm(X)
        Test for change-point detection based on the multivariate empir
        c.d.f. with change in the m.c.d.f. at time(s) m=100

data:  X
Snm = 0.0092805, p-value = 0.597
```

Exemple 1 bis: changement dans la copule mais pas dans les marges

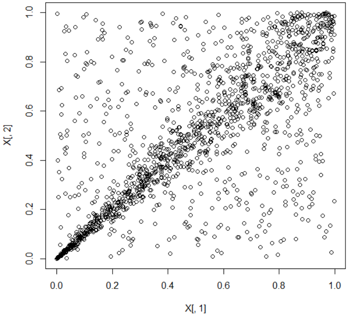
```
n=100
k=50
sigma1 = matrix(c(1,0.2,0.2,1),2,2)
sigma2 = matrix(c(1,0.6,0.6,1),2,2)
X=matrix(rep(0,n*2),n,2)
for(j in 1:k)
X[j,]=t(chol(sigma1))%*%rnorm(2)
for(j in (k+1):n)
X[j,]=t(chol(sigma2))%*%rnorm(2)
```




```
>CopTestdm(X)
        Test for change-point detection based on the multivariate empir
        c.d.f. with change in the m.c.d.f. at time(s) m=100

data:  X
Snm = 0.037066, p-value = 0.001
```

Exemple 1 bisbis: changement structurel dans la copule mais pas dans les marges

```
library(copula)
X1<-rCopula(100,claytonCopula(10))
X2<-cbind(runif(50),runif(10))
X<-rbind(X1,X2)
```




```
Test for change-point detection based on the multivariate empirical
c.d.f. with change in the m.c.d.f. at time(s) m=150

data:  X
Snm = 0.10976, p-value < 2.2e-16
```

Break detection in the copula when there exists a change in marginal distribution at time $m = \lfloor nt \rfloor$, $t \in (0, 1)$ known.

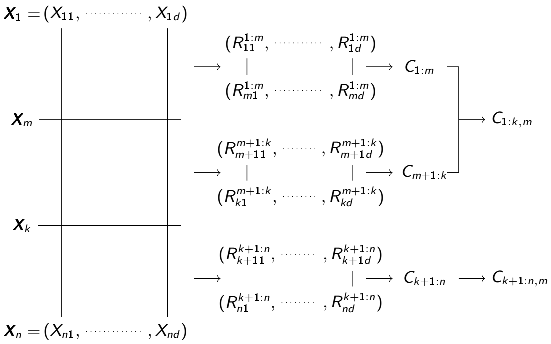
Consider the following null hypothesis

$\mathcal{H}_0 = \mathcal{H}_{1,m} \cap \mathcal{H}_{0,c}$:

$$
\begin{aligned}
\mathcal{H}_{0,c} :& \exists C, \text{ such that } \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \text{ have copula } C \\
\mathcal{H}_{1,m} :& \exists F_1, \ldots F_d \text{ and } F_1', \ldots F_d' \text{ such that } \boldsymbol{X}_1, \ldots, \boldsymbol{X}_m \\
& \text{have m.c.d.f. } F_1, \ldots F_d \text{ and } \boldsymbol{X}_{m+1}, \ldots, \boldsymbol{X}_n \text{ have m.c.d.f. } F_1', \ldots F_d'.
\end{aligned}
$$

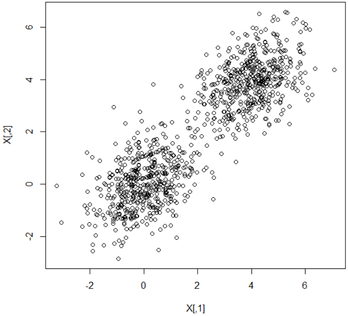
In the same way, we construct $C_{1:k,m}$ and $C_{k+1:n,m}$ from sub sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ and $\boldsymbol{X}_{k+1}, \ldots, \boldsymbol{X}_n$ for k in $\{1, \ldots, n-1\}$.

Figure: Case of $m \leq k$:



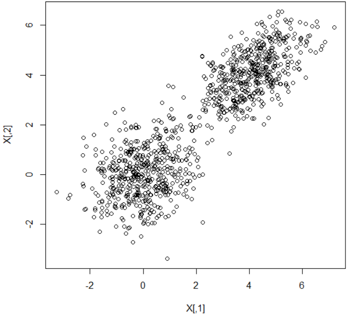
Exemple 2: pas de changement dans la copule mais changement dans les marges

```
m=50
mean1 = rep(0,2)
mean2 = rep(4,2)
X[1:m,] = X[1:m,]+mean1
X[(m+1):n,] = X[(m+1):n,]+mean2
plot(X)
```




```
>CopTestdm(X,b=0.5)
        Test for change-point detection based on the multivariate empir
        c.d.f. with change in the m.c.d.f. at time(s) m=50

data:  X
Snm = 0.0092499, p-value = 0.598
```

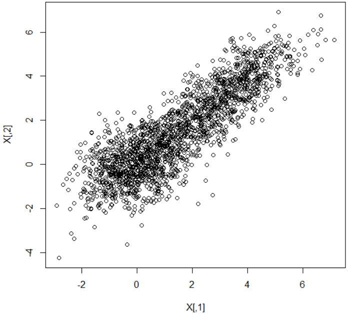
Exemple 3: Changement dans la copule et changement dans les marges




```
n=100
m=50
mean1 = rep(0,2)
mean2 = rep(4,2)
sigma1 = matrix(c(1,0.2,0.2,1),2,2)
sigma2 = matrix(c(1,0.6,0.6,1),2,2)
X=matrix(rep(0,n*2),n,2)
for(j in 1:m)
X[j,]=t(chol(sigma1))%*%rnorm(2)
for(j in (m+1):n)
X[j,]=t(chol(sigma2))%*%rnorm(2)
X[1:m,] = X[1:m,]+mean1
X[(m+1):n,] = X[(m+1):n,]+mean2
> CopTestdm(X,b=0.5)
        Test for change-point detection based on the multivariate empir
        c.d.f. with change in the m.c.d.f. at time(s) m=50

data:  X
Snm = 0.044528, p-value < 2.2e-16
```

Exemple 4: Changement dans la copule et 2 changements dans les marges

```
n=200
m1 = 100
m2 = 150
k = 50

sigma1 = matrix(c(1,0.2,0.2,1),2,2)
sigma2 = matrix(c(1,0.6,0.6,1),2,2)

X=matrix(rep(0,n*2),n,2)
for(j in 1:k)
X[j,]=t(chol(sigma1))%*%rnorm(2)
for(j in (k+1):n)
X[j,]=t(chol(sigma2))%*%rnorm(2)
```



Exemple 4: Changement dans la copule et 2 changements dans les marges

```
mean1 = rep(0,2)
mean2 = rep(2,2)
mean3 = rep(4,2)
X[1:m1,]=X[1:m1,]+mean1
X[(m1+1):m2,]=X[(m1+1):m2,]+mean2
X[(m2+1):n,]=X[(m2+1):n,]+mean3

>CopTestdm(X,b=c(0.5,0.75))

        Test for change-point detection based on the multivariate empir
        c.d.f. with change in the m.c.d.f. at time(s) m=100, 150

data:  X
Snm = 0.027974, p-value = 0.007
```