Modélisation des incertitudes
Partie 2: modèles multivariés

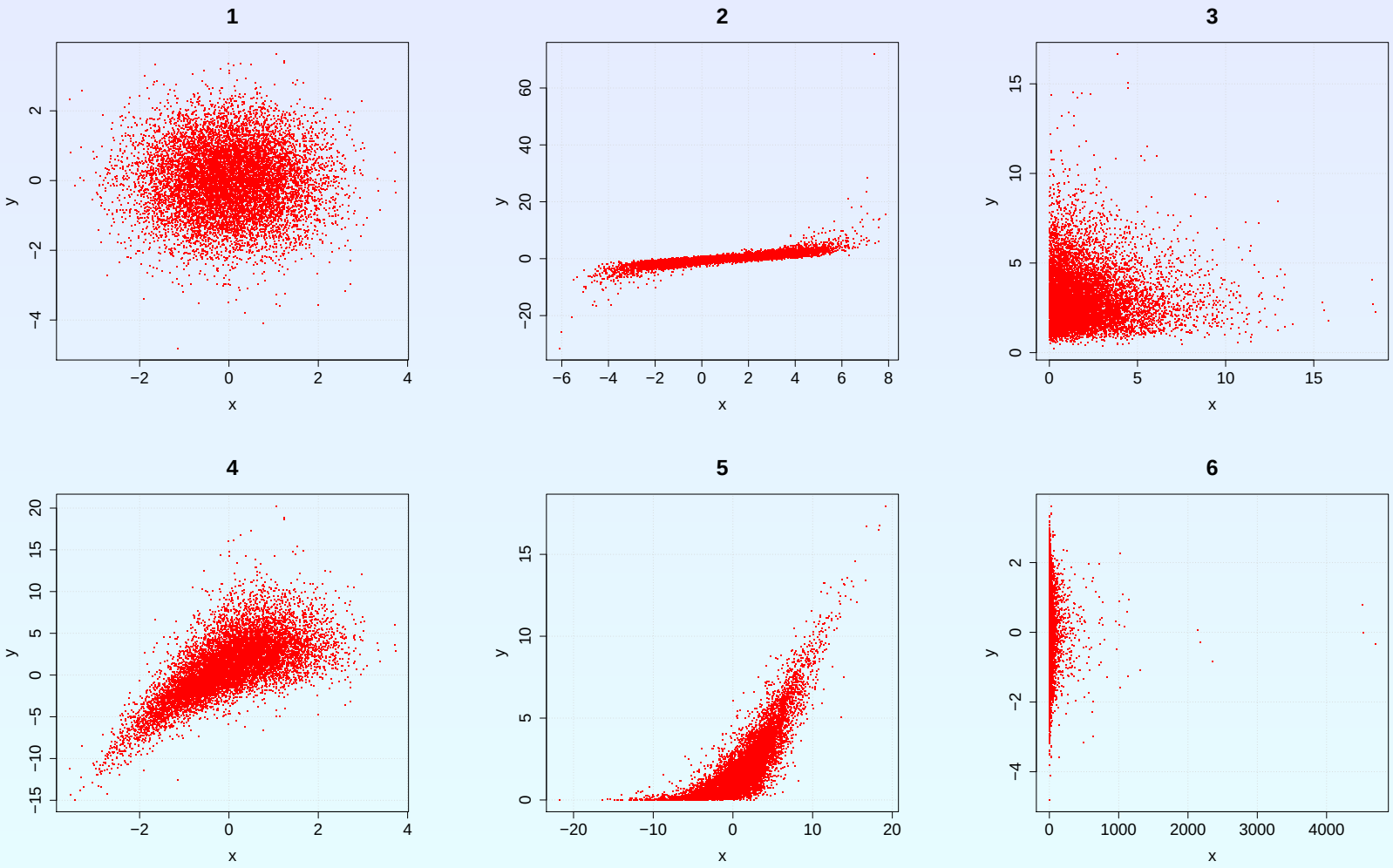
Régis LEBRUN

Airbus Central Research & Technology
regis.lebrun@airbus.com

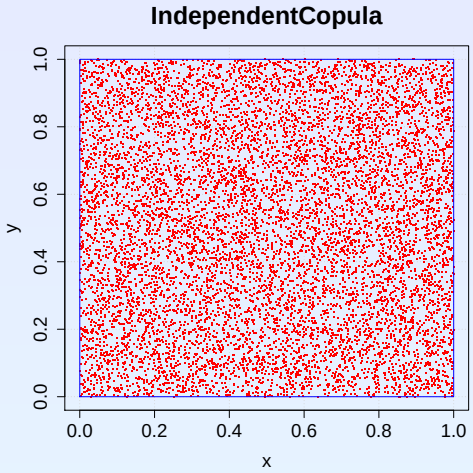
20 janvier 2020

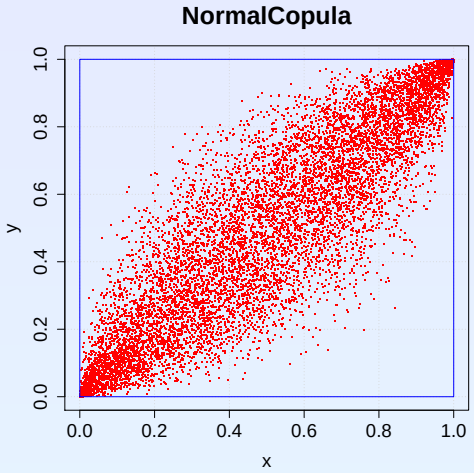
Plan de l'exposé

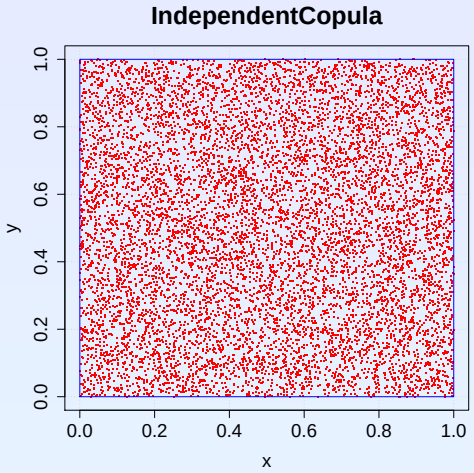
Quelques distributions bivariées. Où sont les lois à composantes indépendantes ?

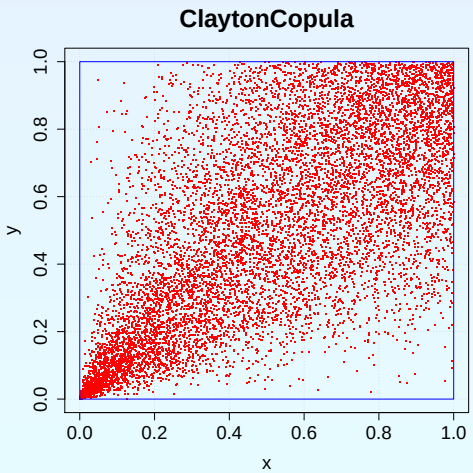


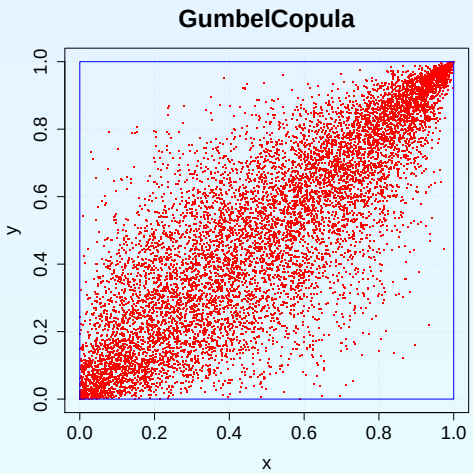
Les mêmes données dans l'espace des rangs

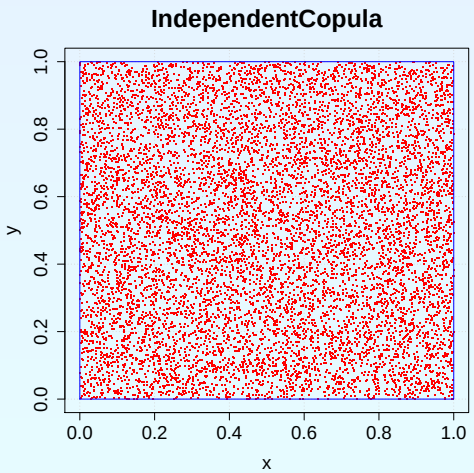












Quelques notes historiques sur les copules et la modélisation de la dépendance

1940 Hoeffding : mesure de dépendance, corrélation linéaire, fonctions de répartitions à lois marginales uniformes sur $[-1/2, 1/2]$.

1951 Fréchet : distributions multivariées à distributions marginales fixées.

1959 Sklar and Schweizer : espaces métriques probabilistes, première apparition du terme «copule».

1979 Deheuvels : tests d'indépendance, estimation non paramétrique multivariée.

1992 Darsow, Nguyen and Olsen : description des processus de Markov en termes de copules.

1999 Embrechts, Lindskog and McNeil : diffusion du concet de copule dans les mondes de la finance et de l'assurance.

2005 Mikosch : "Copulas : Tales and facts". Les copules ne sont-elles qu'un phénomène de mode ?

2009 Salmon : "Recipe for a disaster : the formula that killed Wall Street" ou comment la copule Gaussienne a tué Wall Street.

Copules : un sujet sérieux ?

L'analyse que fait Thomas Mikosch de la croissance exponentielle des travaux portant sur les copules :

2003 Google donne 10,000 réponses au mot-clé «copule»

2005 650,000 réponses...

2013 2,010,000 réponses...

- "My main concern is that this very simple concept might be something like the emperor's new clothes because it promises to solve all problems of stochastic dependence but it falls short in achieving the goal."
- "I also observed that my students are likely to be attracted to copulas than to stochastic processes. A possible reason is that one needs less than 10 minutes to understand the fundamentals of copulas, but many years of studies in order to get an idea of a genuine stochastic process."

On en reparle dans 10mn...

Les copules pour la modélisation de la dépendance I

Définition

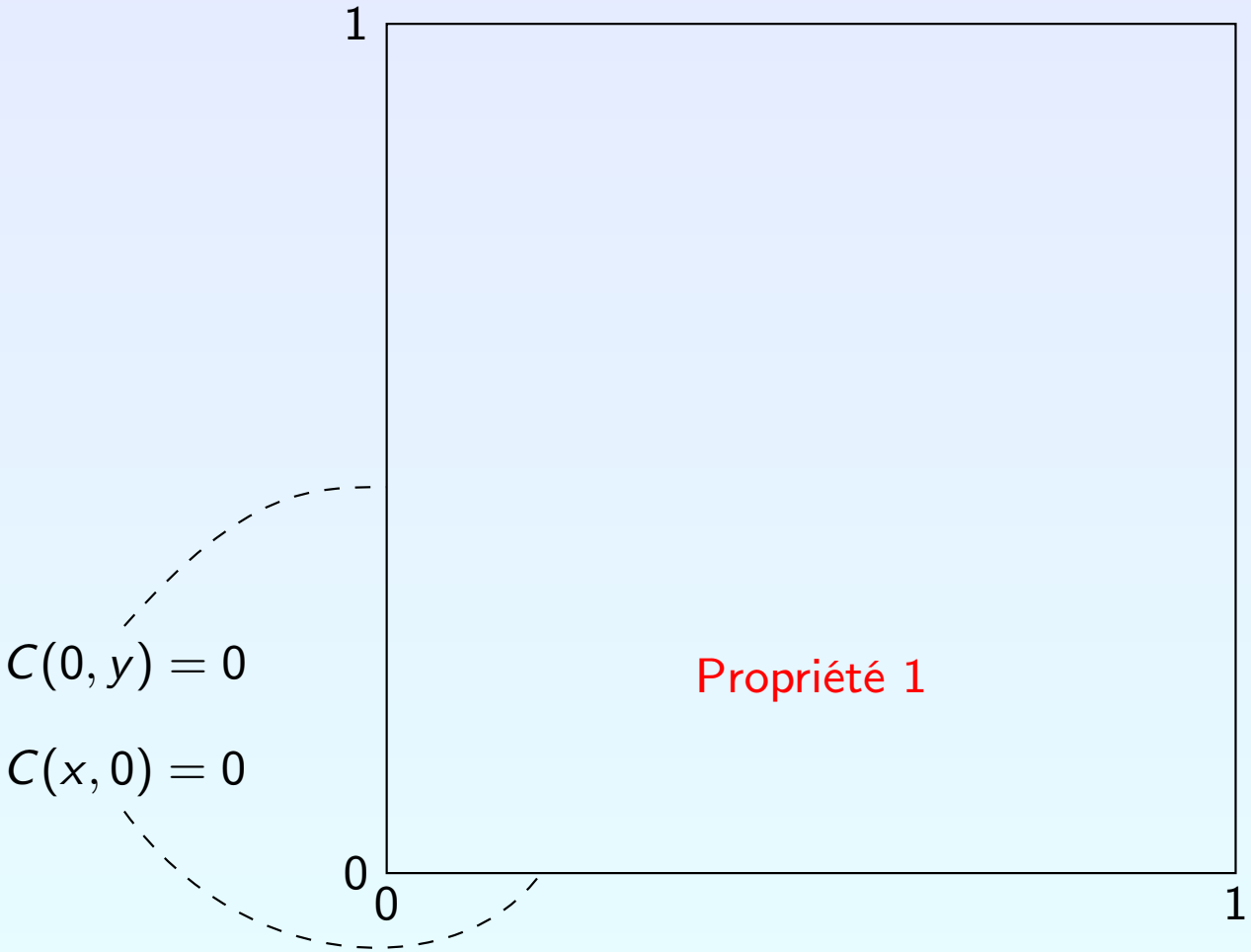
Une copule n-dimensionnelle est la restriction au cube unité $[0,1]^n$ d'une fonction de répartition multivariée dont les fonctions de répartition marginales sont uniformes sur $[0,1]$.

Propriété

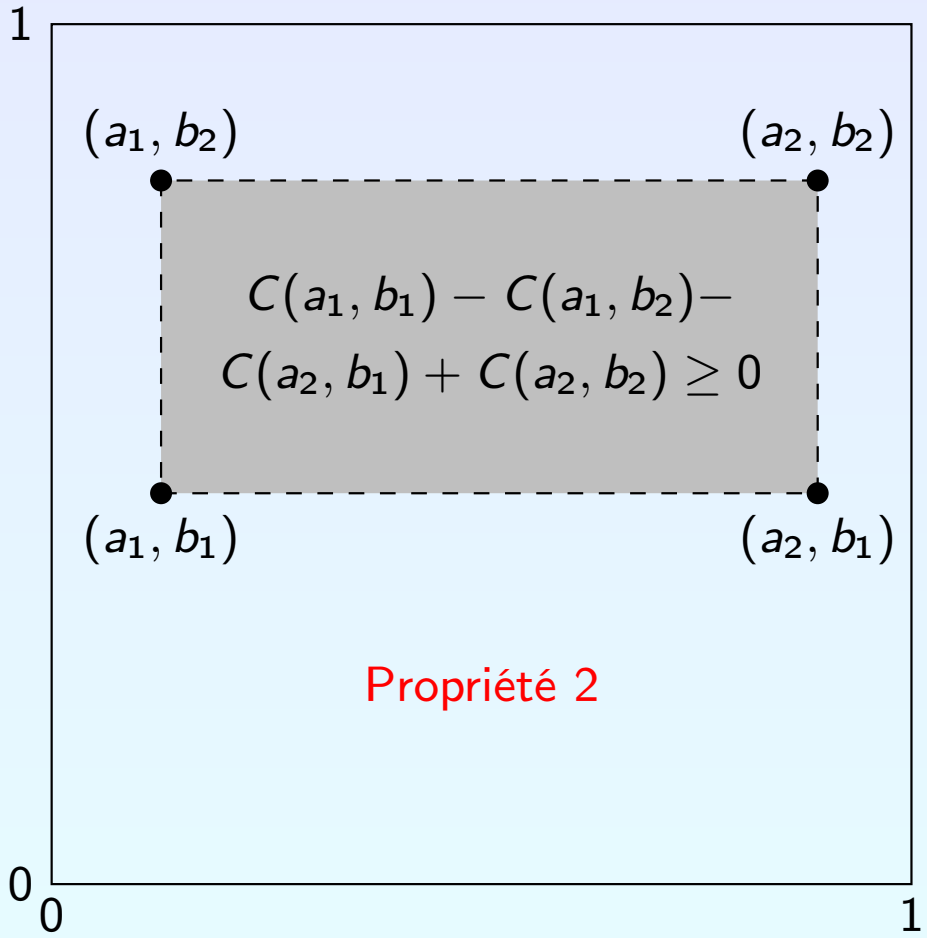
Une copule n-dimensionnelle est caractérisée par les propriétés suivantes :

1. *Pour tout* $\boldsymbol{u}$ *ayant au moins une composante nulle,* $C(\boldsymbol{u}) = 0$;
2. C *est n-croissante :* $\sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_1+\cdots+i_n} C(x_{1i_1}, \ldots, x_{ni_n}) \geq 0$ *avec* $x_{j1} = a_j$ *et* $x_{j2} = b_j$ *pour tout* $j \in \{1, \ldots, n\}$ *et* $\boldsymbol{a}, \boldsymbol{b} \in [0,1]^n$, $\boldsymbol{a} \leq \boldsymbol{b}$.
3. *Pour tout* $\boldsymbol{u}$ *ayant toutes ses composantes égales à 1 sauf éventuellement* u_k, $C(\boldsymbol{u}) = u_k$.

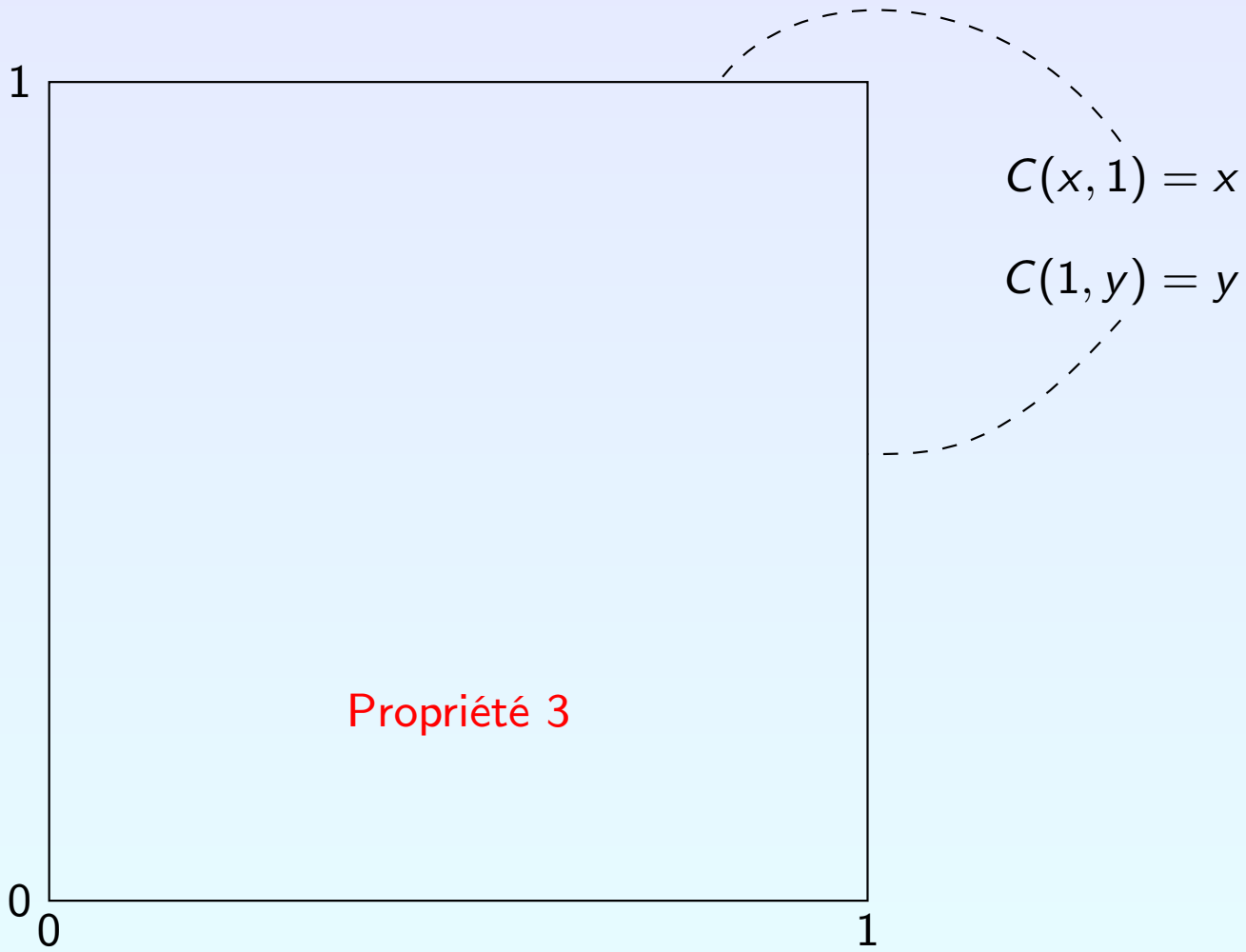
Illustration



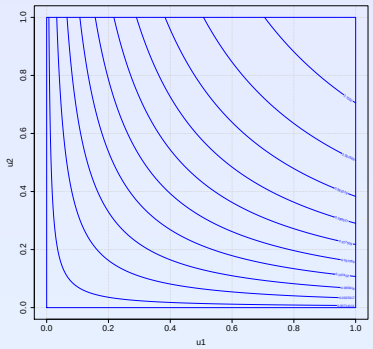
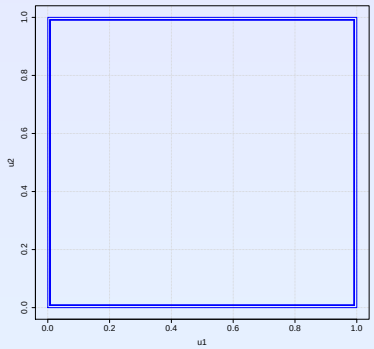
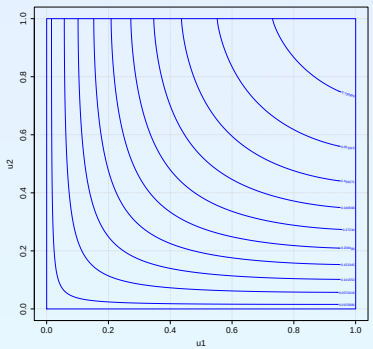
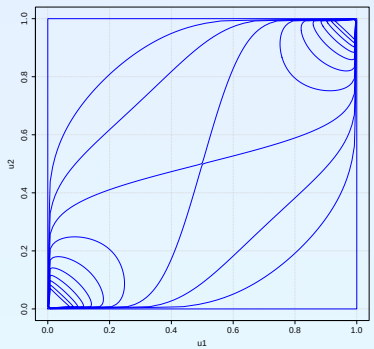
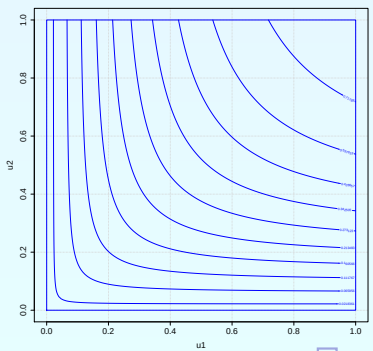
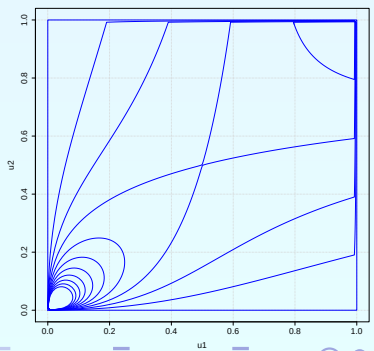
Illustration



Illustration



Exemples de copules I

Copule	C	$c = \partial^2 C / \partial u_1 \partial u_2$
Indépendante $u_1 u_2$		
Normal (ou Gaussienne) $\int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{\exp\left(-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\right)}{2\pi\sqrt{1-\rho^2}} \mathrm{d}s\,\mathrm{d}t,\ \lvert\rho < 1\rvert$		
Clayton $\left(u_1^\theta + u_2^\theta - 1\right)^{1/\theta}, \theta > 0$		

Exemples de copules II

Avec OpenTURNS :

Listing 1 – exemplesCopules.py

```
import openturns as ot

copule1 = ot.IndependentCopula(2)
copule2 = ot.NormalCopula(ot.CorrelationMatrix(2, [1.0, 0.5, 0.5, 1.0]))
copule3 = ot.ClaytonCopula(2.5)
for copule in [copule1, copule2, copule3]:
    ot.Show(copule.drawCDF())
    ot.Show(copule.drawPDF())
```

Quelques propriétés

Propriété

- *Soit C une copule n-dimensionnelle, alors*

  $$\forall \boldsymbol{u}, \boldsymbol{v} \in [0,1]^n, |C(\boldsymbol{u}) - C(\boldsymbol{v})| \leq \sum_{i=1}^{n} |u_i - v_i|$$

- *Soit $\boldsymbol{X}$ un vecteur aléatoire continu de copule C et $\alpha_1, \ldots, \alpha_n$ des fonctions strictement croissantes. Le vecteur $(\alpha_1(X_1), \ldots, \alpha_n(X_n))$ a également C pour copule.*

Bornes de Fréchet-Hoeffding

On définit les fonctions suivantes sur $[0,1]^n$:

$$W^n(\boldsymbol{u}) = \max(u_1 + \cdots + u_n - n + 1, 0) \text{ et } M^n(\boldsymbol{u}) = \min(u_1, \ldots, u_n).$$

Théorème

Pour toute copule C et tout $\boldsymbol{u} \in [0,1]^n$, on a :

$$W^n(\boldsymbol{u}) \leq C(\boldsymbol{u}) \leq M^n(\boldsymbol{u}).$$

Remarque : W^n est une copule pour $n = 2$ mais pas pour $n > 2$, alors que M^n est une copule pour tout $n \geq 2$. M^n est la copule d'un vecteur dont les composantes X_i sont presque sûrement l'image par une application strictement croissante d'une même variable aléatoire V.

Copules, mode d'emploi I

Théorème ([Sklar])

Soit F une fonction de répartition n-dimensionnelle dont les fonctions de répartition marginales sont $F_1, \ldots, F_n$. Il existe une copule C de dimension n telle que pour tout $\boldsymbol{x} \in \overline{\mathbb{R}}^n$:

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)). \tag{1}$$

Si les fonctions de répartition marginales sont continues, pour tout $\boldsymbol{u} \in [0,1]^n$ on a :

$$C(\boldsymbol{u}) = F({F_1}^{(-1)}(u_1), \ldots, {F_n}^{(-1)}(u_n))$$

Copules, mode d'emploi II

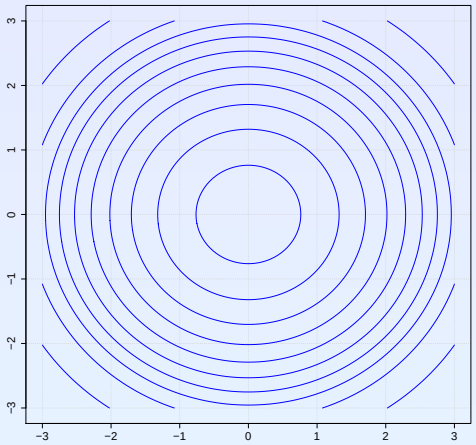
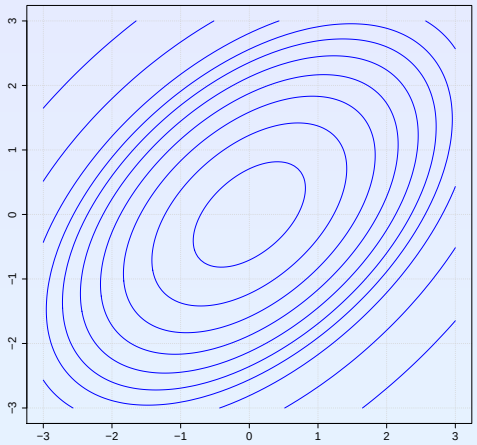
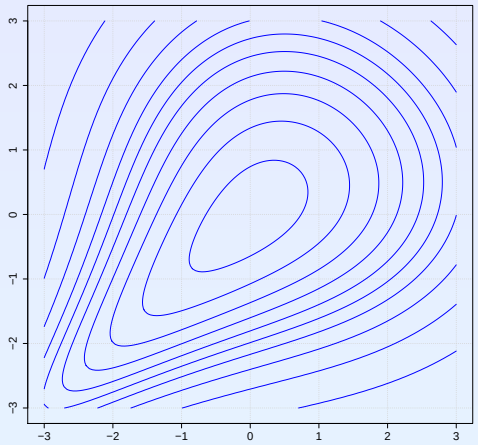
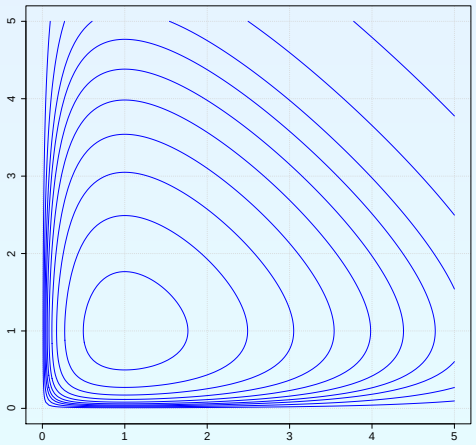
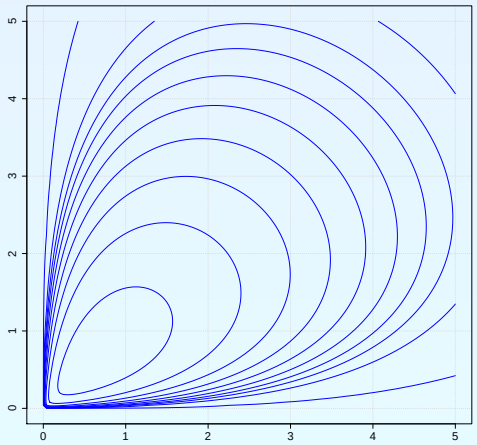
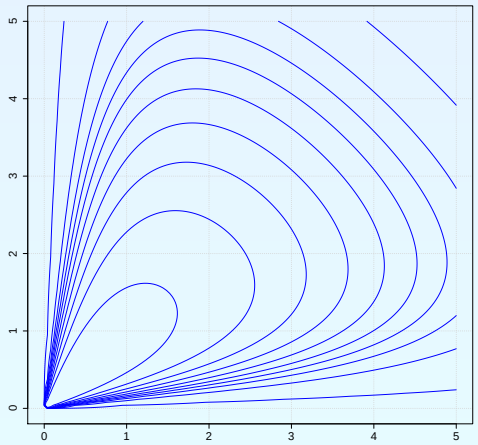
Démonstration.

- Soit $\boldsymbol{X}$ un vecteur aléatoire de dimension n et de fonction de répartition F.
- Soit V une variable aléatoire uniformément distribuée sur $[0,1]$ et indépendante de $\boldsymbol{X}$
- Pour $k = 1, \ldots, n$, soit U_k la variable aléatoire définie par $U_k = F_k(X_k)$ si F_k est continue, et par $U_k = F_k(X_k-) + V \sum_{v \in \Delta_k} \mathbb{P}(X_k = v)$ où Δ_k est l'ensemble des points de discontinuité de F_k si F_k n'est pas continue.
- Les variables aléatoires $U_1, \ldots, U_n$ sont uniformément distribuées sur $[0,1]$ et $\forall x_k \in \mathbb{R}, \{X_k \leq x_k\} = \{U_k \leq F_k(x_k)\}$ p.s.
- Soit C la fonction de répartition du vecteur aléatoire $(U_1, \ldots, U_n)$. Alors C est une copule et $\forall \boldsymbol{x} \in \mathbb{R}^n$:

$$\mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) = \mathbb{P}(U_1 \leq F_1(x_1), \ldots, U_n \leq F_n(x_n)) = C(F_1(x_1), \ldots, F_n(x_n))$$

- Si toutes les fonctions de répartition marginales F_k sont continues, leur image contient $(0,1)$ et par continuité des copules, il existe une unique copule C satisfaisant (1).

Exemples de distributions composées I

Copula	Independent	Normal	Clayton
$\mathcal{N}(0,1)$ marginals			
$\Gamma(2,1)$ marginals			

Exemples de distributions composées II

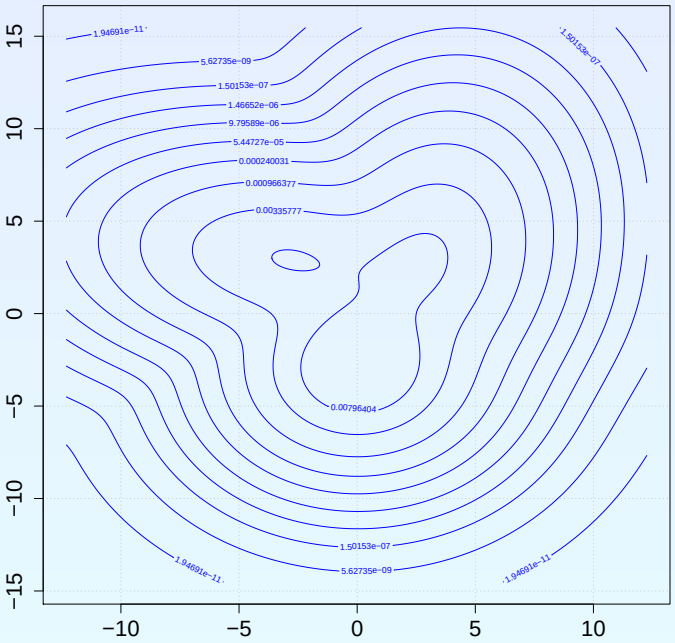
Avec OpenTURNS :

Listing 2 – loisComposees.py

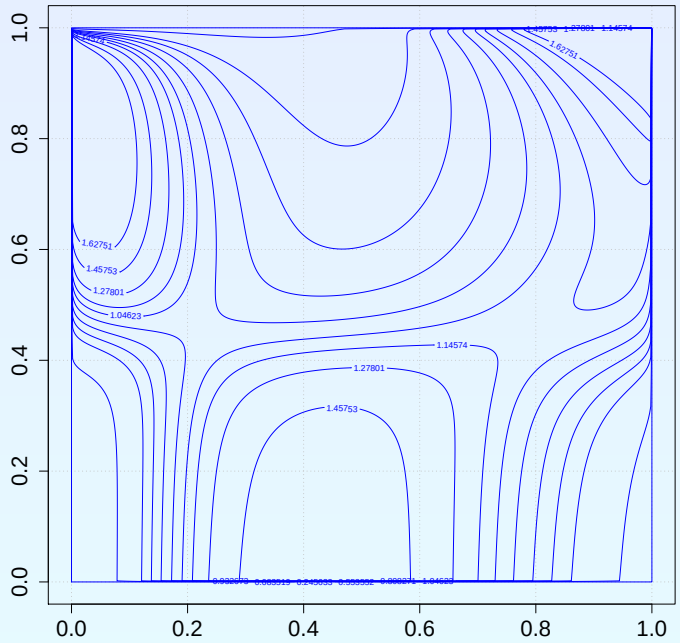
```
import openturns as ot

copule1 = ot.IndependentCopula(2)
copule2 = ot.NormalCopula(ot.CorrelationMatrix(2, [1.0, 0.5, 0.5, 1.0]))
copule3 = ot.ClaytonCopula(2.5)
liste_marginales = [[ot.Normal()]*2, [ot.Gamma(2.0, 1.0)]*2]
for copule in [copule1, copule2, copule3]:
    for marginales in liste_marginales:
        distribution = ot.ComposedDistribution(marginales, copule)
        ot.Show(distribution.drawPDF())
```

Exemple de copule extraite d'une loi multivariée I

Loi jointe



Copule



Exemple de copule extraite d'une loi multivariée II

Avec OpenTURNS :

Listing 3 – copuleSklar.py

```
import openturns as ot
import openturns.viewer as otv

d1 = ot.Normal([-3.0, 3.0], ot.CovarianceMatrix(2, [4.0, 1.0, 1.0, 9.0]))
d2 = ot.Normal([3.0, 3.0], ot.CovarianceMatrix(2, [9.0, -1.0, -1.0, 4.0]))
d3 = ot.Normal([0.0, -3.0], ot.CovarianceMatrix(2, [4.0, 0.0, 0.0, 4.0]))
distribution = ot.Mixture([d1, d2, d3], [0.3, 0.3, 0.4])
copule = ot.SklarCopula(distribution)
graph = distribution.drawPDF([512]*2)
graph.setTitle("")
graph.setXTitle("")
graph.setYTitle("")
graph.setLegendPosition("")
graph.draw("mixture.pdf", 600, 620)
graph = copule.drawPDF([512]*2)
graph.setTitle("")
graph.setXTitle("")
graph.setYTitle("")
graph.setLegendPosition("")
graph.draw("sklar.pdf", 600, 620)
```

Copules marginales, copules conditionnelles

Définition

Soit C une copule de dimension n et $k \in \{1, \ldots, n\}$. Les copules Copules marginales C_k et conditionnelles $C_{k|1,\ldots,k-1}$ de rang k sont définies par :

$$C_k(u_1, \ldots, u_k) = C(u_1, \ldots, u_k, 1, \ldots, 1)$$

$$C_{k|1,\ldots,k-1}(u_k | u_1, \ldots, u_{k-1}) = \frac{\partial^{k-1} C_k(u_1, \ldots, u_k)}{\partial u_1 \ldots u_{k-1}} \Big/ \frac{\partial^{k-1} C_{k-1}(u_1, \ldots, u_{k-1})}{\partial u_1 \ldots u_{k-1}}$$

Echantillonnage des distributions composées I

Soit $\boldsymbol{X}$ un vecteur aléatoire de fonctions de répartition $F_1, \ldots, F_n$ et de copule C. On peut échantillonner $\boldsymbol{X}$ par la procédure suivante :

1. Générer une réalisation de $\boldsymbol{u} \sim C$;
2. Une réalisation $\boldsymbol{x}$ de $\boldsymbol{X}$ est donnée par :

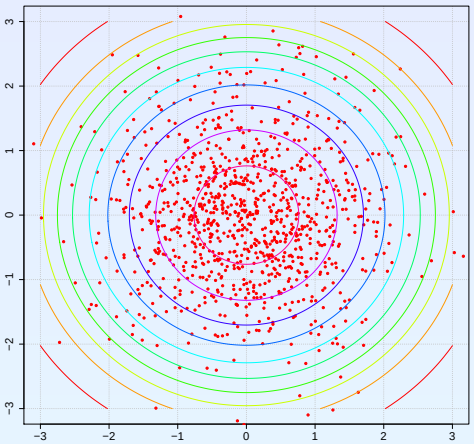
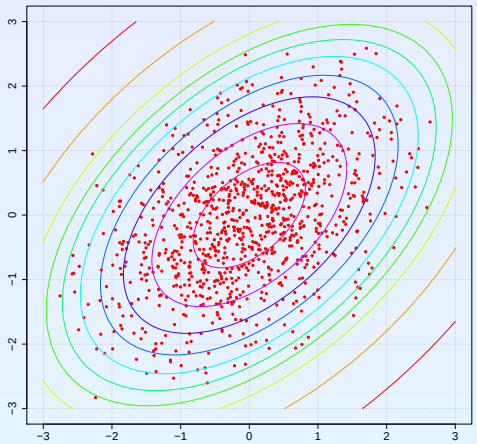
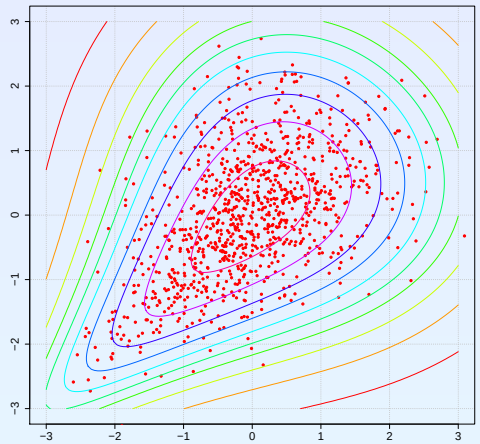
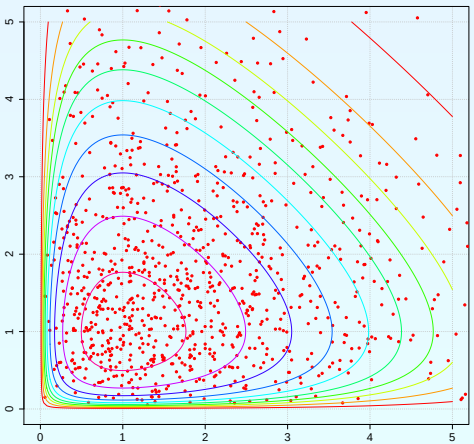
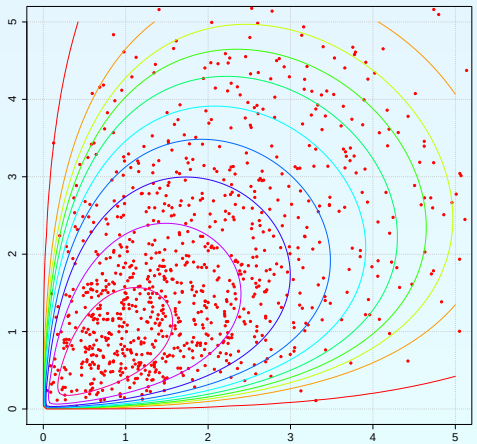
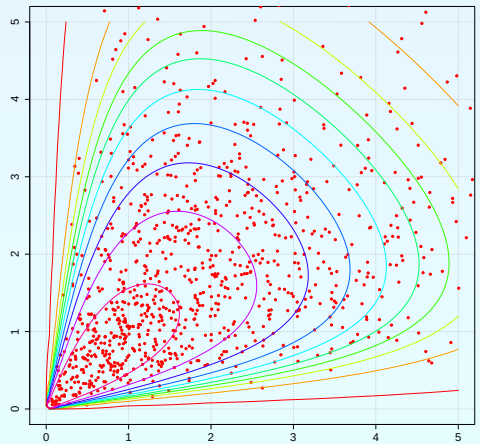
$$\boldsymbol{x} = \left(F_1^{(-1)}(u_1), \ldots, F_n^{(-1)}(u_n) \right)$$

La principale difficulté est d'échantillonner C.

1. Générer $u_1 \sim \mathcal{U}(0,1)$;
2. Pour $k \in \{2, \ldots, n\}$, générer $u_k \sim C_{k|1,\ldots,k-1}(u_1, \ldots, u_{k-1})$.
3. Le vecteur $(u_1, \ldots, u_n)$ est une réalisation de C.

Echantillonnage des distributions composées II

Remarque : pour de nombreuses copules, il existe des algorithmes de simulation spécifiques plus efficaces

Copula	Independent	Normal	Clayton
$\mathcal{N}(0,1)$ marginals			
$\Gamma(2,1)$ marginals			

Echantillonnage des distributions composées III

Avec OpenTURNS :

Listing 4 – echantillonnageComposees.py

```
import openturns as ot

size = 500
copule1 = ot.IndependentCopula(2)
copule2 = ot.NormalCopula(ot.CorrelationMatrix(2, [1.0, 0.5, 0.5, 1.0]))
copule3 = ot.ClaytonCopula(2.5)
liste_marginales = [[ot.Normal()]*2, [ot.Gamma(2.0, 1.0)]*2]
for copule in [copule1, copule2, copule3]:
    for marginales in liste_marginales:
        distribution = ot.ComposedDistribution(marginales, copule)
        sample = distribution.getSample(size)
        graph = distribution.drawPDF(sample.getMin(), sample.getMax())
        cloud = ot.Cloud(sample)
        cloud.setColor('red')
        cloud.setPointStyle('bullet')
        graph.add(cloud)
        ot.Show(graph)
```

Plus d'outils de modélisation : copules composées

Soient $C_1, \ldots, C_k$ k copules de dimensions $n_1, \ldots, n_k$ et $n = \sum_{i=1}^{k} n_i$. La fonction C definie sur $[0,1]^n$ par :

$$C(u_1, \ldots, u_n) = C_1(u_1, \ldots, u_{n_1}) \times \cdots \times C_k(u_{n-n_k+1}, \ldots, u_n)$$

est une copule de dimension n.

C'est une structure de dépendance bloc-diagonale creuse de grande dimension construite à partir de structures de dépendance pleines de faible dimension.

Encore plus d'outils de modélisation

Constructions ne changeant pas la dimension :

- Mixture de copules
- Somme ordinale de copules
- ...

Constructions changeant la dimension :

- Réseau de copules
- Regular Vine Copula
- Réseau Bayésien de copules
- ...

En pratique I

La situation change selon la dimension du problème :

- Petite dimension (2 ou 3), large choix de modèles paramétriques, estimation paramétrique ou non-paramétrique efficace. C'est typiquement le cas dans les analyses de type événement extrême climatique.
- Grande dimension apparente ($\simeq 100$) du fait de la composition de beaucoup de petites dimensions indépendantes. Dans ce cas, on bénéficie de tous les outils de la petite dimension et du mécanisme de composition de copules. C'est la situation typique de la propagation d'incertitudes dans l'industrie
- Grande dimension ($\simeq 100$) par transport d'une petite dimension stochastique : $\boldsymbol{X} = \phi(\boldsymbol{W})$ avec $\dim \boldsymbol{X} \gg \dim \boldsymbol{W} \simeq 2$. Dans ce cas il faut paramétrer le problème par $\boldsymbol{W}$ quitte à méta-modéliser ϕ.
- Grande dimension ($\gg 100$) par discrétisation d'un aléa continu : processus stochastique lu sur une grille, champ aléatoire lu sur les noeuds d'un maillage. C'est la situation typique en milieu aléatoire. Dans ce cas il faut utiliser des techniques spécifiques à l'apprentissage de processus (modèles ARMA, Gaussien) ou reparamétrer le problème par un vecteur de dimension plus raisonnable ($\simeq 100$), cf décomposition de Karhunen-Loeve.

En pratique II

- Grande dimension non réductible : on privilégie alors l'approche compromis adéquation/complexité. Etant donné une quantité d'information statistique, on choisit le modèle stochastique le plus informatif, ie réalisant un compromis optimal vraisemblance/complexité (critère BIC, modèles graphiques).

Copules et statistiques d'ordre I

Soit $\boldsymbol{X}$ un vecteur aléatoire de dimension n de fonctions de répartition connues $F_1, \ldots, F_n$. On cerche à caractériser l'ensemble $\mathcal{C}$ des copules telles que la distribution composée résultante soit telle que :

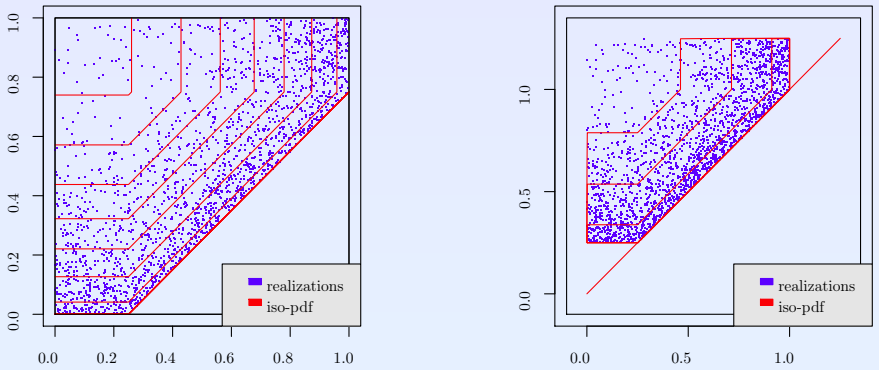
$$X_1 \leq \ldots \leq X_n \text{ a.s.}$$

Théorème

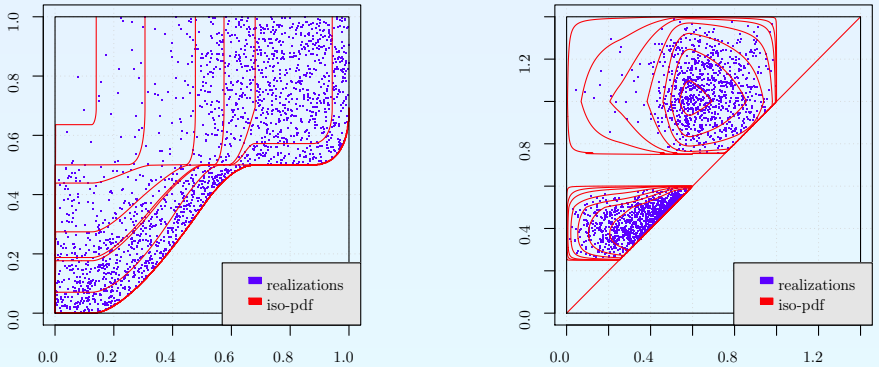
1. $\mathcal{C} \neq \emptyset$ *si et seulement si* $\forall x \in \mathbb{R}, F_n(x) \leq \ldots \leq F_1(x)$;
2. *Si* $F_1, \ldots, F_n$ *satisfont (1 et sont continues, alors* $C \in \mathcal{C}$ *si et seulement si le support de* C *est contenu dans* $\{\boldsymbol{u} \in [0,1]^n \mid F_1^{\leftarrow}(u_1) \leq \ldots \leq F_n^{\leftarrow}(u_n)\}$

Copules et statistiques d'ordre II

$$F_1 \sim \mathcal{U}(0,1),\ F_2 \sim \mathcal{U}(\tfrac{1}{4}, \tfrac{5}{4})$$



$$F_1 \sim \mathcal{T}(0, \tfrac{1}{2}, 1),\ F_2 \sim \tfrac{1}{2}\mathcal{T}(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{3}{5}) + \tfrac{1}{2}\mathcal{T}(\tfrac{3}{4}, 1, \tfrac{7}{5})$$



Examples de copules de statistiques d'ordre.

Copules et statistiques d'ordre III

Avec OpenTURNS :

Listing 5 – copuleStatistiquesOrdre.py

```
import openturns as ot

size = 1000
marginales1 = [ot.Uniform(0.0, 1.0), ot.Uniform(0.25, 1.25)]
marginales2 = [ot.Triangular(0.0, 0.5, 1.0), ot.Mixture([ot.Triangular(0.25,
    0.5, 0.6), ot.Triangular(0.75, 1.0, 1.4)])]

distribution1 = ot.MaximumEntropyOrderStatisticsDistribution(marginales1)
copule1 = ot.MaximumEntropyOrderStatisticsCopula(marginales1)
distribution2 = ot.MaximumEntropyOrderStatisticsDistribution(marginales2)
copule2 = ot.MaximumEntropyOrderStatisticsCopula(marginales2)

for d in [distribution1, copule1, distribution2, copule2]:
    graph = d.drawPDF()
    graph.setColors(['red'])
    cloud = ot.Cloud(d.getSample(size))
    cloud.setColor('blue')
    graph.add(cloud)
    ot.Show(graph)
```

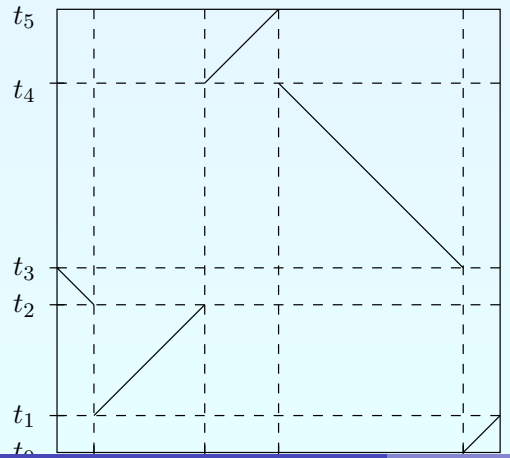
La dépendance parfaite est-elle si différente de l'indépendance ? I

Définition

Soient $X_1, \ldots, X_n$ n variables aléatoires. Elles sont dites parfaitement dépendantes s'il existe une variable aléatoire U et n fonctions bijectives $f_1, \ldots, f_n$ telles que $X_1 = f_1(U), \ldots, X_n = f_n(U)$.

Définition

Une copule C de dimension n est une permutation de la copule min si et seulement si il existe un entier $N > 0$, une partition $(0 = s_0^k < s_1 < \ldots < s_N = 1)_{k=1,\ldots,N}$ de $[0,1]$, et $N-1$ permutations σ^k de $\{1, \ldots, N\}$ telles que chaque $[s_{i-1}, s_i] \times \ldots \times [s^n_{\sigma^{N-1}(i-1)}, s^{N-1}_{\sigma^{N-1}(i)}]$ soit un hypercube dans lequel C dépose une masse $s_i - s_{i-1}$ uniformément distribuée sur sa diagonale principale.



$$s = (0, 1/12, 1/3, 1/2, 11/12, 1)$$
$$t = (0, 1/12, 1/3, 5/12, 5/6, 1)$$
$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 5 & 4 & 1 \end{pmatrix}$$

La dépendance parfaite est-elle si différente de l'indépendance ? II

Théorème

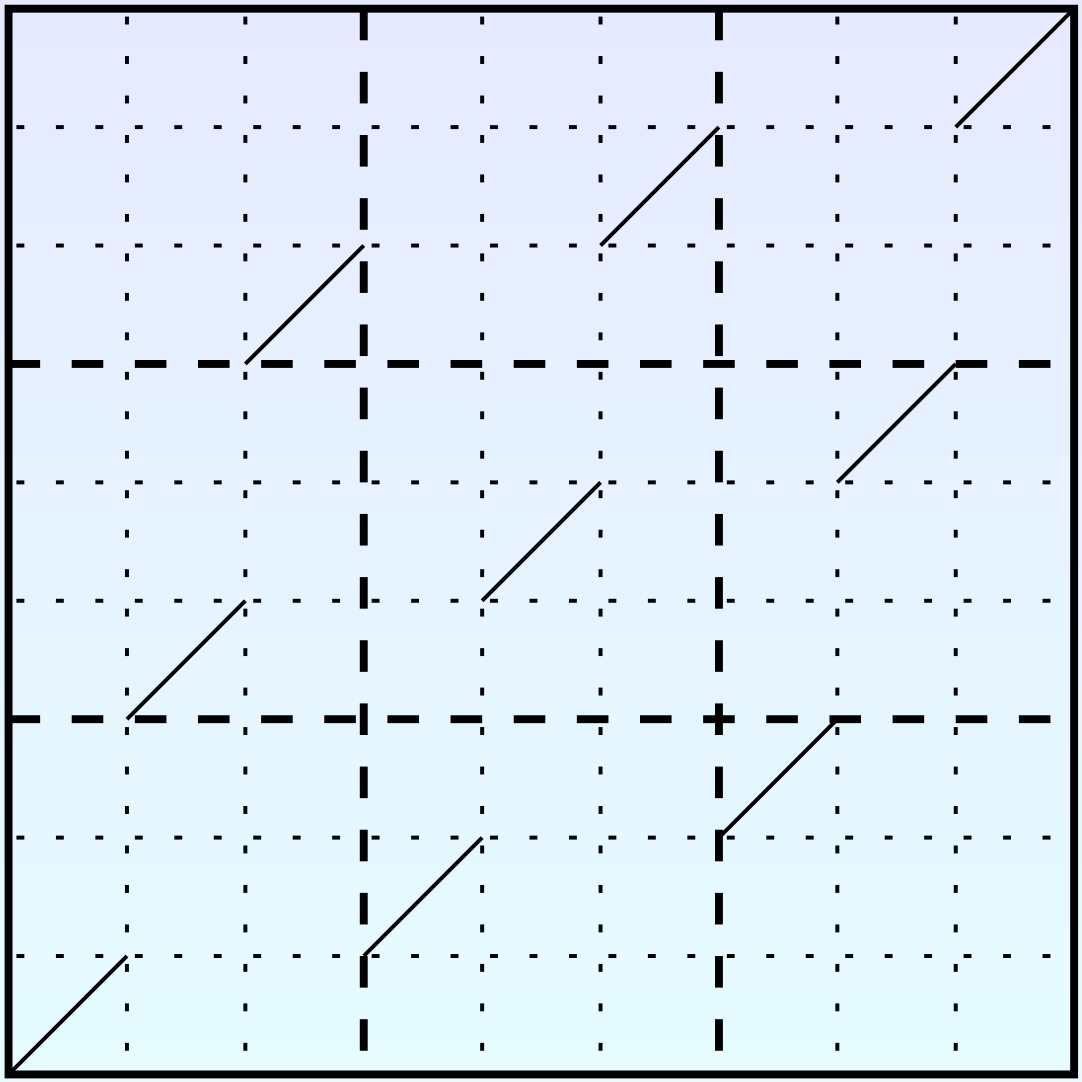
Les permutation de min forment un ensemble dense dans l'ensemble des copules muni de la norme sup.

Démonstration, cas de la copule cible indépendante Π_n.

- Soit $\epsilon > 0$, m un entier tel que $m \geq 1/\epsilon$
- On prend $M = m^n$ et on construit une permutation de min C_ϵ associée à la partition n uniforme de $[0,1]$ en M sous-intervalles et les permutations $\sigma^k(m^k(j-1)+i) = m^k(i-1)+j$ pour $i,j = 1,\ldots,m$, $k = 1,\ldots,m-1$.
- C_ϵ donne une masse de $1/M$ à chacun des M sous-hypercubes de $[0,1]^n$, et $C_\epsilon(p_1/m,\ldots,p_n/m) = p_1 \times \ldots \times p_n/m$ pour tout $p_i = 0,\ldots,m$ donc C_ϵ et Π_n coïncident en ces points. Comme Π_n et C_ϵ sont Lipschitz, on a $\|C_\epsilon - \Pi_n\|_\infty \leq n\epsilon$.

□

La dépendance parfaite est-elle si différente de l'indépendance ? III



Mesures d'association

La donnée d'une copule comme modèle de dépendance d'un vecteur aléatoire est très riche.
La notion de mesure d'association sert à résumer cette structure de dépendance dans une collection de scalaires.

Définition

Une mesure d'association r entre deux variables aléatoires X_1 et X_2 est une fonction scalaire de X_1 et X_2 telle que :

1. r est définie pour tout couple (X_1, X_2).
2. $r(X_1, X_2) \in [-1, 1], r(X_1, X_1) = 1, r(X_1, -X_1) = -1$.
3. Si X_1 et X_2 sont indépendantes, $r(X_1, X_2) = 0$.
4. Si g et h sont deux fonctions strictement croissantes, $r(X_1, X_2) = r(g(X_1), h(X_2))$.

On montre que r est une fonction de la copule de (X_1, X_2) seule

Corrélation linéaire I

Définition

La corrélation linéaire ρ entre deux variables aléatoires X_1 et X_2 telles que $\mathrm{Var}(X_1) = \sigma_1^2 < \infty$ et $\mathrm{Var}(X_2) = \sigma_2^2 < \infty$ est définie par :

$$\begin{aligned} \rho(X_1, X_2) &= \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}} \\ &= \frac{1}{\sigma_1 \sigma_2} \iint_{\mathbb{R}^2} F_{12}(x_1, x_2) - F_1(x_1) F_2(x_2) \, \mathrm{d}x_1 \mathrm{d}x_2 \end{aligned} \tag{2}$$

Propriété

- $\rho(X_1, X_2) \in [-1, 1]$ *avec* $|\rho(X_1, X_2)| = 1 \iff \exists\, a, b \in \mathbb{R}, a \neq 0, X_2 = aX_1 + b$
- X_1, X_2 *indépendantes implique* $\rho(X_1, X_2) = 0$;
- $\rho(aX_1 + b, \alpha X_2 + \beta) = \mathrm{sign}(a\alpha)\rho(X_1, X_2)$

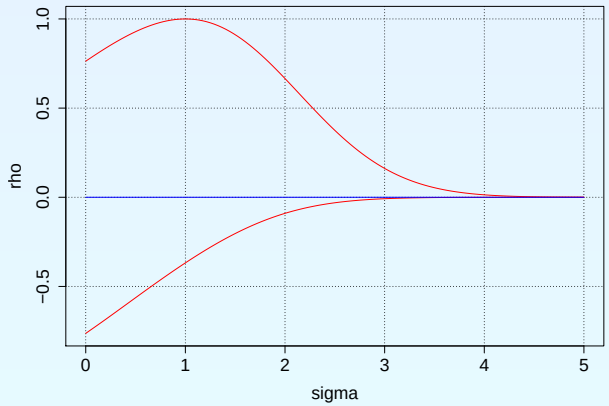
Ce n'est pas une mesure d'association ! Elle n'est pas définie pour toutes les variables aléatoires, n'est pas invariante par transformation croissante et n'est pas une fonction de la copule seule.

Corrélation linéaire II

Théorème (Fréchet-Hoeffding)

Soit (X_1, X_2) un vecteur aléatoire de lois marginales F_1, F_2 données. Les valeurs possibles de $\rho(X_1, X_2)$ forment un intervalle inclu dans $[-1, 1]$, l'inclusion étant stricte en générale.

Conséquence : il est impossible de spécifier $\rho(X_1, X_2)$ indépendamment de F_1 et F_2. Si $X_1 \hookrightarrow \mathcal{LN}(0, 1)$ et $X_2 \hookrightarrow \mathcal{LN}(0, \sigma^2)$, alors $\rho(X_1, X_2) \in [\rho_{min}, \rho_{max}] \subsetneq [-1, 1]$, avec $\rho_{min} = \frac{e^{-\sigma} - 1}{\sqrt{e-1}\sqrt{e^{\sigma^2} - 1}}$ et $\rho_{max} = \frac{e^{\sigma} - 1}{\sqrt{e-1}\sqrt{e^{\sigma^2} - 1}}$.



On note que ρ_{min} et ρ_{max} tendent vers 0 quand σ tend vers $+\infty$. Pour $\sigma = 5, \rho \in [-3\,10^{-6}, 4\,10^{-4}]$!

Rho de Spearman

Définition

Le ρ_S de Spearman entre deux variables aléatoires X_1 et X_2 est défini par :

$$\rho_s(X_1, X_2) = \rho(F_1(X_1), F_2(X_2)) = 12 \iint_{[0,1]^2} C(u, v)\, \mathrm{d}u\, \mathrm{d}v - 3$$

où C est la copule de la loi jointe de (X_1, X_2).

Propriété

- $\rho_S(X_1, X_2) \in [-1, 1]$ *avec* $|\rho_S(X_1, X_2)| = 1 \iff \exists\, \varphi$ *monotone telle que* $X_2 = \varphi(X_1)$
- X_1, X_2 *indépendantes implique* $\rho_S(X_1, X_2) = 0$;
- $\rho_S(\varphi(X_1), \psi(X_2)) = \rho_S(X_1, X_2)$ *pour toutes fonctions monotones* φ *et* ψ *de même monotonie.*

Il s'agit bien d'une mesure d'association.

Tau de Kendall

Définition

Le τ de Kendall entre deux variables aléatoires X_1 et X_2 est défini par :

$$\begin{aligned}\tau(X_1, X_2) &= \mathbb{P}[(\hat{X}_1 - \tilde{X}_1)(\hat{X}_2 - \tilde{X}_2) > 0] - \mathbb{P}[(\hat{X}_1 - \tilde{X}_1)(\hat{X}_2 - \tilde{X}_2) < 0] \\ &= 4 \iint_{[0,1]^2} C(u, v)\, \mathrm{d}C(u, v) - 1\end{aligned}$$

où $(\hat{X}_1, \hat{X}_2)$ et $(\tilde{X}_1, \tilde{X}_2)$ ont la même loi que (X_1, X_2), de copule C.

Propriété

- $\tau(X_1, X_2) \in [-1, 1]$ *avec* $|\tau(X_1, X_2)| = 1 \iff \exists\, \varphi$ *monotone telle que* $X_2 = \varphi(X_1)$
- X_1, X_2 *indépendantes implique* $\tau(X_1, X_2) = 0$;
- $\tau(\varphi(X_1), \psi(X_2)) = \tau(X_1, X_2)$ *pour toutes fonctions monotones* φ *et* ψ *de même monotonie.*

Il s'agit bien d'une mesure d'association.

Utilisation des mesures d'association

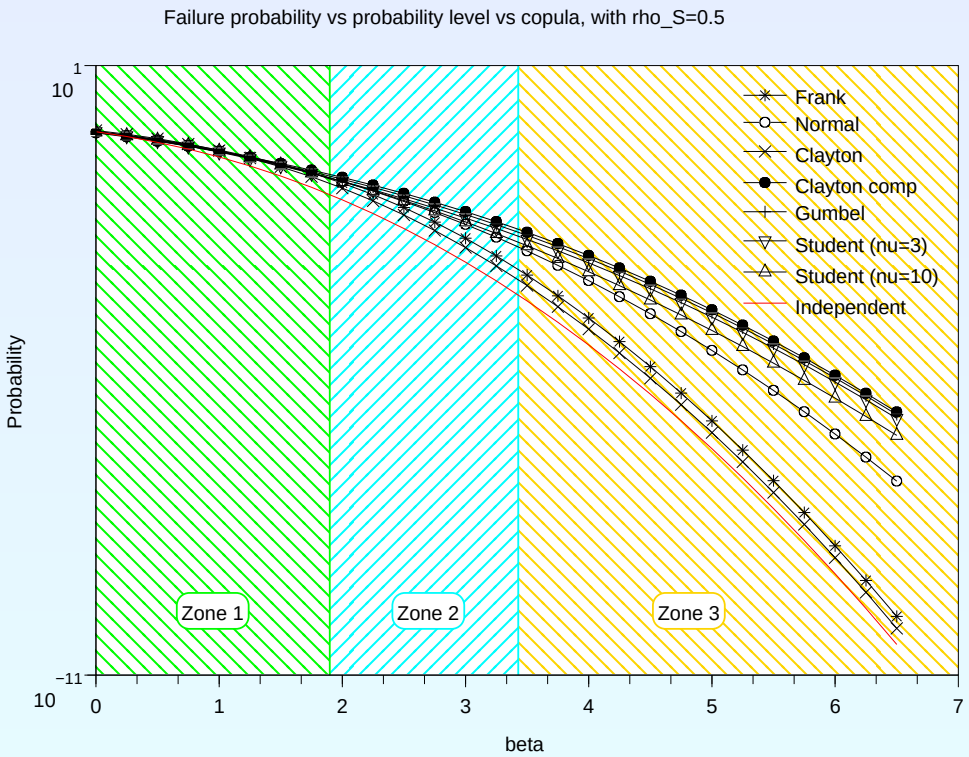
Les mesures d'association ρ_S et τ étant des fonctions de la copule seule, il est possible de les relier aux paramètres $\boldsymbol{\theta}$ de la copule $C_{\boldsymbol{\theta}}$. On construit ainsi un estimateur $\hat{\boldsymbol{\theta}}_n$ de $\boldsymbol{\theta}$ qui est robuste à l'effet des lois marginales sur l'échantillon multivarié de taille n.

Exemples :

- Copule normale $C_{\boldsymbol{R}}$: $R_{ij} = 2\sin\left(\frac{\pi}{6}\rho_{S\,ij}\right) = \sin\left(\frac{\pi}{2}\tau_{ij}\right)$
- Copule de Clayton C_θ : $\theta = \dfrac{2\tau}{1-\tau}$

Mesures d'association ou copules ?

$\mathbb{P}\left(X_1 + X_2 \geq \beta\sqrt{2}\right)$ pour $X_1, X_2 \sim \mathcal{N}(0,1)$ et différentes copules C telles que $\rho_S(X_1, X_2) = 1/2$.



β	$P_{min}(\beta)$	$P_{max}(\beta)$	ratio
1.89	$6.5\,10^{-2}$	$8.7\,10^{-2}$	1.5
3.41	$1.1\,10^{-3}$	$8.6\,10^{-3}$	10.0
6.5	$8.3\,10^{-11}$	$1.9\,10^{-6}$	$2.3\,10^{4}$

Rang et statistique dordre

La notion de rang joue un rôle central dans l'estimation des mesures d'association.

Définition

Soit $(X^k)_{k=1,\dots,N}$ un échantillon de taille N de la variable aléatoire X et $\sigma \in \mathfrak{S}_N$ une permutation aléatoire telle que $X_{\sigma(1)} \leq \dots X_{\sigma(N)}$ p.s. (une telle permutation est unique p.s si X est continue). Le rang de X^k est défini par :

$$\text{rank}(X^k) = \sigma^{-1}(k)$$

C'est la position aléatoire de X^k dans la statistique d'ordre $X_{1:N} = X_{\sigma(1)}, \dots, X_{N:N} = X_{\sigma(N)}$.

Estimation statistique I

Soit $(\boldsymbol{X}_k)_{k\in\{1,\dots,N\}}$ un échantillon de taille N d'une loi multivariée. La démarche d'estimation de la copule sous-jacente à $\boldsymbol{X}$ est :

- Identifier les fonctions de répartition marginales (estimation 1D) ;
- Transformer l'échantillon $(\boldsymbol{X}_k)_{k\in\{1,\dots,N\}}$ en l'échantillon des rangs renormalisé $(\boldsymbol{U}_k)_{k\in\{1,\dots,N\}}$.
- Estimer la copule sur la base de $(\boldsymbol{U}_k)_{k\in\{1,\dots,N\}}$

Estimation des lois marginales : toutes les techniques classiques sont possibles :

- Estimation paramétrique : $F_j^{\boldsymbol{\theta}} \in \mathcal{L}(\boldsymbol{\theta})$, on estime θ par $\hat{\boldsymbol{\theta}}_N(X_1^j,\dots,X_n^j)$ et on prend $\hat{F}_j = F_j^{\hat{\boldsymbol{\theta}}_N(X_1^j,\dots,X_n^j)}$ comme modèle marginal.
- Estimation non-paramétrique : fonction de répartition marginale empirique, reconstruction à noyaux, histogramme etc.

On distingue là encore plusieurs méthodes :

- Estimation paramétrique : $C^{\boldsymbol{\theta}} \in \mathcal{C}(\boldsymbol{\theta})$, on estime θ par $\hat{\boldsymbol{\theta}}_N(U_1^j,\dots,U_n^j)$ et on prend $\hat{C} = C^{\hat{\boldsymbol{\theta}}_N(X_1^j,\dots,X_n^j)}$ comme estimation de la copule.

Estimation statistique II

- Estimation non-paramétrique multivariée plus réciproque du théorème de Sklar : reconstruction à noyaux, histogramme etc. suivi du filtrage des marginales (légèrement) non uniformes obtenues.
- Estimation semi paramétrique dans une classe de copules à espace de paramétrage infini : copule archimédienne, copule elliptique, voir les travaux de P. Lambert, K. Kostadinov, A. Charpentier, J-D. Fermanian.

Estimation statistique : copule de Bernstein I

Définition

Soit α une fonction définie sur $[0,1]^d$ à valeurs dans $[0,1]$. On appelle copules de Bernstein C^B associée à C la copule définie par :

$$\forall (u_1,\dots,u_d)\in[0,1]^d,\quad C^B(u_1,\dots,u_d)=\sum_{i_1=0}^{m_1}\dots\sum_{i_d=0}^{m_d} C\left(\frac{i_1}{m_1},\dots,\frac{i_d}{m_d}\right)\prod_{j=1}^{d}P_{i_j,m_j}(u_j)$$

où :

- La fonction $\alpha:[0,1]^d\to[0,1]$ coïncide avec une copule sur la grille :

$$\left\{\frac{0}{m_1},\dots,\frac{m_1}{m_1}\right\}\times\dots\times\left\{\frac{0}{m_d},\dots,\frac{m_d}{m_d}\right\}$$

- P_{i_j,m_k} est le polynôme de Bernstein de paramètres $(i_j=a, m_j=b)$ donné par :

$$\forall u\in[0,1],\quad P_{a,b}(u)=\frac{b!}{a!(b-a)!}u^a(1-u)^{b-a}$$

Estimation statistique : copule de Bernstein II

Théorème

Soit $(\boldsymbol{X}_k)_{k\in\{1,\dots,N\}}$ un échantillon de taille N d'une loi multivariée de copule C et C_n la copule empirique associée. La copule de Bernstein associée à C_n converge presque sûrement vers C au sens de la norme sup, et si C est absolument continue de densité bornée c, alors la densité de la copule de Bernstein converge au sens L^2 vers c.

Test d'adéquation

Ce domaine est encore en plein essor, depuis le travail pionier de J-D. Fermanian basé sur une comparaison du modèle proposé avec une reconstruction à noyaux multivariée. Voir les travaux suivants :

- C. Genest, B. Rémillard, D. Beaudoin, Goodness-of-fit tests for copulas : A review and a power study, Insurance Mathematics & Economics, in press.
- D. Fermanian, Goodness-of-fit tests for copulas, Journal of Multivariate Analysis 95 (2005) 119-152.

Bonne nouvelle : en 2D, les tests semblent performants dès $N = 150$ observations

Estimation des mesures d'association I

Définition

Soit $\left((X_1^k, X_2^k)\right)_{k=1,\dots,N}$ un échantillon de taille N du vecteur aléatoire $\boldsymbol{X} = (X_1, X_2)$. L'estimateur du ρ de Spearman $\hat{\rho}_{S,N}(\boldsymbol{X})$ est défini par :

$$\hat{\rho}_{S,N}(\boldsymbol{X}) = \frac{\sum_{k=1}^{N} \left(\text{rank}(X_1^k) - \overline{\text{rank}}(X_1)\right)\left(\text{rank}(X_2^k) - \overline{\text{rank}}(X_2)\right)}{\sqrt{\sum_{k=1}^{N} \left(\text{rank}(X_1^k) - \overline{\text{rank}}(X_1)\right)^2 \sum_{k=1}^{N} \left(\text{rank}(X_2^k) - \overline{\text{rank}}(X_2)\right)^2}}$$

where $\overline{\text{rank}}(X_1) = \frac{1}{N}\sum_{k=1}^{N} \text{rank}(X_1^k)$ and $\overline{\text{rank}}(X_2) = \frac{1}{N}\sum_{k=1}^{N} \text{rank}(X_2^k)$.

Théorème

Soit $\boldsymbol{X}$ un vecteur aléatoire continu bidimensionnelbe. Alors :

$$\hat{\rho}_{S,N}(\boldsymbol{X}) \xrightarrow{a.s} \rho_S(\boldsymbol{X}) \text{ quand } N \to \infty$$

$$\sqrt{N}\left(\hat{\rho}_{S,N}(\boldsymbol{X}) - \rho_S(\boldsymbol{X})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_{\rho_S}^2) \text{ quand } N \to \infty$$

Estimation des mesures d'association II

où la variance asymptotique $\sigma^2_{\rho_S}$ est donnée par :

$$\begin{aligned}\sigma^2_{\rho_S} =& 144\eta_{22} + \rho_S^2\left\{\frac{9}{10} + 72\eta_{22}\right\} && \text{si } \eta_{11} = 0\\ =& 144\eta_{22} + \rho_S^2\left\{\frac{9}{10} + 72\eta_{22} - 12\frac{\eta_{13}+\eta_{31}}{\eta_{11}}\right\} && \text{si } \eta_{11} \neq 0\end{aligned}$$

où $\eta_{k\ell} = \iint_{[0,1]^2}\left(u_1 - \frac{1}{2}\right)^k\left(u_2 - \frac{1}{2}\right)^\ell c(u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2$ et c est la densité de la copule de $\boldsymbol{X}$.

Définition

Soit $\left((X_1^k, X_2^k)\right)_{k=1,\dots,N}$ un échantillo, de taille N du vecteur aléatoire $\boldsymbol{X} = (X_1, X_2)$. L'estimateur du tau de Kendall $\hat{\tau}_N(X_1, X_2)$ est défini par

$$\hat{\tau}_N(\boldsymbol{X}) = \frac{2}{N(N-1)}\sum_{1\leq i<j\leq N}\mathrm{sgn}(X_1^i - X_1^j)\,\mathrm{sgn}(X_2^i - X_2^j)$$

Estimation des mesures d'association III

Théorème

Soit $\boldsymbol{X}$ *un vecteur aléatoire bidimensionnel. On a :*

$$\hat{\tau}_N(\boldsymbol{X}) \xrightarrow{a.s} \tau(\boldsymbol{X}) \text{ quand } N \to \infty$$

$$\sqrt{N}\left(\hat{\tau}_N(\boldsymbol{X}) - \tau(\boldsymbol{X})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_\tau^2) \text{ quand } N \to \infty$$

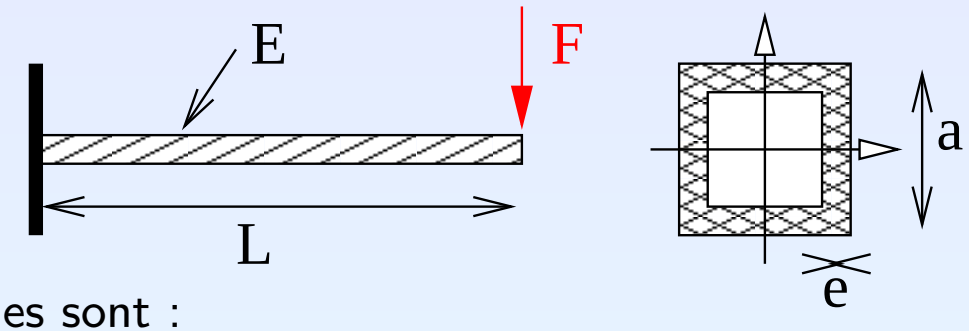
où la variance asymptotique σ_τ^2 *est donnée par :*

$$\sigma_\tau^2 = 4\mathbf{Var}\left[\mathbb{E}\left[\operatorname{sgn}(X_1 - X_1')\operatorname{sgn}(X_2 - X_2') \,|\, X_1, X_2\right]\right]$$

où $\boldsymbol{X}' = (X_1', X_2')$ *est une copie indépendante de* $\boldsymbol{X}$.

Est-ce que tout ça en vaut la peine ? I

On considère une poutre console élastique homogène isotrope dans l'hypothèse des petites perturbations :



Les données incertaines sont :

- E, F, L et I, l'inertie de flexion de la poutre.
- E et F sont indépendantes des autres variables.
- une étude statistique donne $\tau(L, I) = 0.5$.
- $E \sim \mathcal{TN}(\mu_E, \sigma_E, a_E, b_E)$, $F \sim \mathcal{U}(a_F, b_F)$, $L \sim \mathcal{U}(a_L, b_L)$, $I \sim \mathcal{B}(s_I, t_I, a_I, b_I)$

Objectif : calculer la probabilité pour que la déflection d soit supérieure à s

- On modélise la dépendance entre L et I en utilisant soit une copule normale, soit une copule de Clayton.
- Le paramétrage de ces copules est tel que $\tau(L, I) = 0.5$.
- On calcule la probabilité de l'événement $\{d > s\}$ par la méthode de Monté Carlo.

Est-ce que tout ça en vaut la peine ? II

- On obtient :
  - $\mathbb{P}(d > s) = 32.10^{-7} \pm 4.10^{-7}$ avec une confiance de 95% pour la copule de Clayton
  - $\mathbb{P}(d > s) = 104.10^{-7} \pm 4.10^{-7}$ avec une confiance de 95% pour la copule normale.

  Soit un facteur supérieur à 3 entre les deux calculs ! Il aurait sans doute été plus performant d'utiliser les données ayant conduit à $\tau(L, I) = 0.5$ à identifier directement la copule.

Challenges scientifiques : Estimation non paramétrique et test d'adéquation

Les problèmes d'estimation non paramétrique et de test d'adéquation de copules restent des problèmes difficiles :

- Les tests sont d'autant moins puissants qu'on est en grande dimension $d > 4$.
- L'estimation non paramétrique est sensible à la manière d'estimer les marginales.

Ces problématiques sont au cœur des travaux sur l'estimation de risque en finance, et les progrès sont rapides. L'enjeu est de rendre plus robuste la sélection d'une copule.
Lien avec les processus stochastiques

- La notion de copule est très performante dans la modélisation de la loi d'un vecteur aléatoire.
- Un premier lien a été fait entre la théorie des copules et celle des processus de Markov vectoriels.
- Par contre, le lien entre la copule de la loi marginale d'un processus vectoriel quelconque et la loi d'un échantillon temporel de ce processus fait encore n'a pas encore été étudié en détail.

L'enjeu est de fournir de nouveaux outils théoriques et pratiques pour paramétrer la structure de dépendance de processus, notamment pour construire des processus non gaussiens vectoriels à structure de dépendance marginale donnée.

Conclusion

- Toutes les étapes depuis l'estimation statistique à partir de données multivariées jusqu'à la simulation de Monté Carlo peuvent être réalisées via une modélisation à base de copules.
- A partir d'un ensemble de copules, il est possible d'en créer de nouvelles par assemblage de manière efficace.

Des challenges scientifiques

- Fléau de la dimension pour les aspects statistiques.
- Lien avec les processus.

Que retenir de cette présentation ?

- La notion de dépendance stochastique est exactement couverte par le concept de copule ;
- Traiter cette notion uniquement à l'aide de corrélations linéaires est (en général) une très mauvaise idée ;
- Tout modèle probabiliste multivarié possède (au moins) une copule...
- ... cependant, d'autres descriptions de la dépendance peuvent être plus adaptés au calcul ou à l'estimation statistique (processus, modèles bayésiens).


D. Kurowicka and R. Cooke.
Uncertainty Analysis with High Dimensional Dependence Modelling.
Wiley series in probability and statistics. John Wiley & Sons, 2006.


M. Sklar.
Fonctions de répartition à n dimensions et leurs marges.
Publication de l'Institut Statistique Universitaire Paris, 8 :229–231, 1959.