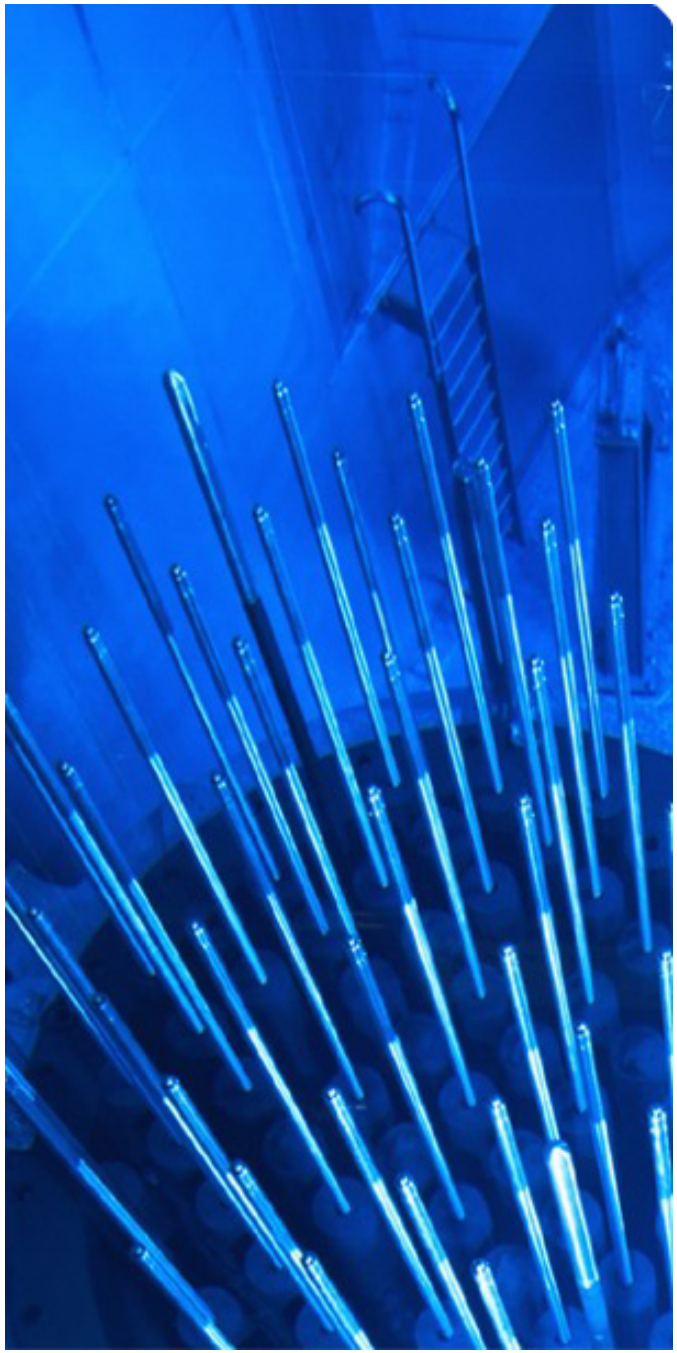
Dependence Modeling using Copulas

Anne Dutfoy

EDF R&D PERICLES
7, Boulevard Monge
91170 Palaiseau
anne.dutfoy@edf.fr

Formation ITECH
"Incertitudes Avancées"
Nov 2021

CHANGER L'ÉNERGIE ENSEMBLE



Sommaire

1. Introduction
2. Multivariate distributions and Copulas
3. Lien avec les mesures d'association
4. Estimation
5. Dependence in high dimension
6. Conclusion

Dependence Modeling using Copulas

1 Introduction

2 Multivariate distributions and Copulas

3 Lien avec les mesures d'association

4 Estimation

5 Dependence in high dimension

6 Conclusion

Motivation

Uncertainty propagation

Given :

- A random vector $\underline{X}$ taking values into $\mathbb{R}^d$ (uncertainties)
- A measurable function $f : \mathbb{R}^d \to \mathbb{R}^p$

One wants to gain information on the distribution of $\underline{Y} = f(\underline{X})$ (hence the term of propagation) :

- Some moments $\mathbb{E}(h(\underline{Y}))$ for various measurable functions h
- As a special case, the probability of some events $\mathbb{P}(\underline{Y} \in B) = \mathbb{E}(1_{Y \in B})$

Probabilistic modeling

The main objective is to build the distribution of $\underline{X}$:

- from multivariate data
- or from univariate data only
- or from expert knowledge

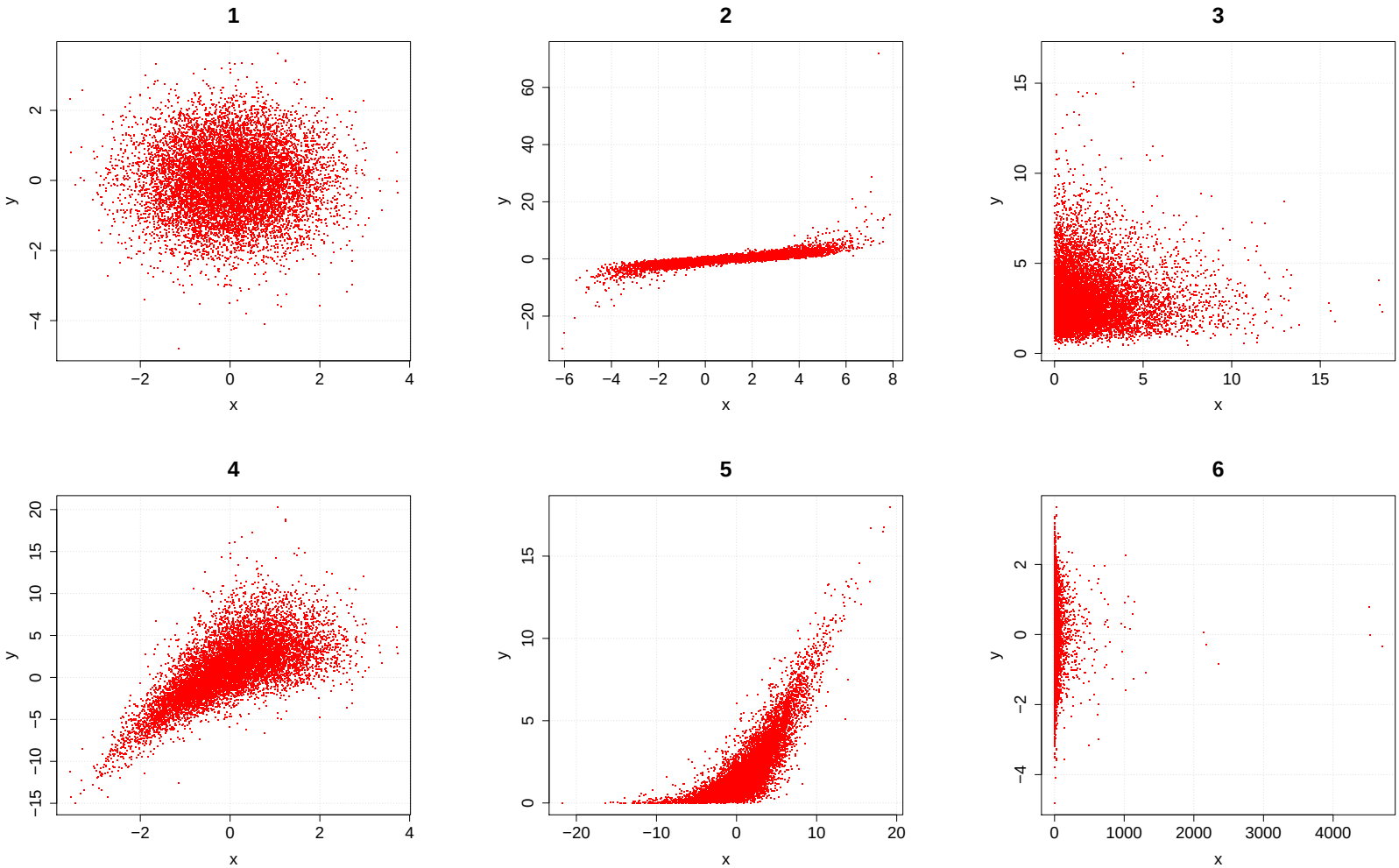
From my personal experience :

- in many applications, we have a rather good knowledge of the marginal distributions of $\underline{X}$
- in contrast, the interaction between the components of $\underline{X}$ is rather unknown

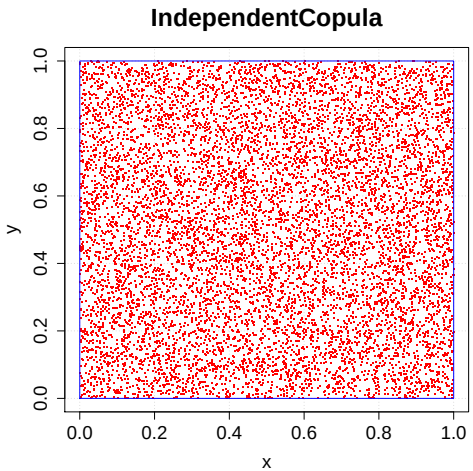
What is Dependence modeling ?

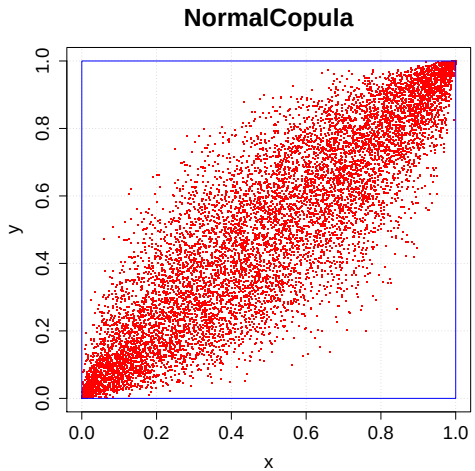
Dependence modeling is the description of this interaction, ie the description of the joint distribution function once the effect of the marginal distributions has been removed.

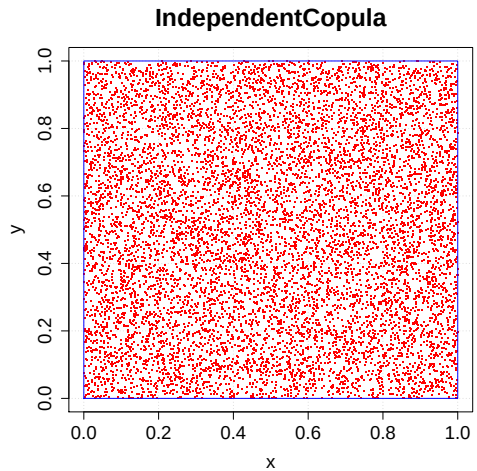
Some bidimensional distributions. Which ones have independent components ?

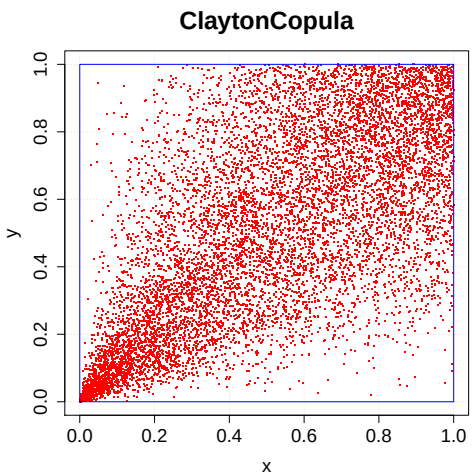


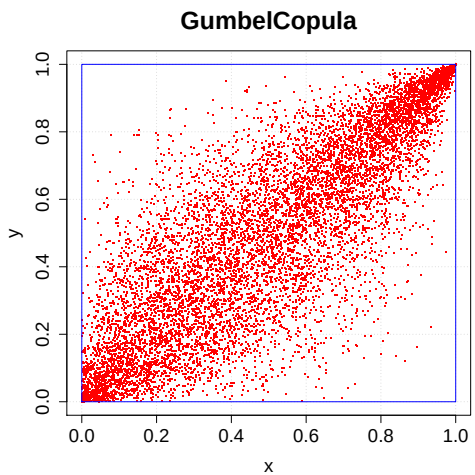
The same data, considering ranks

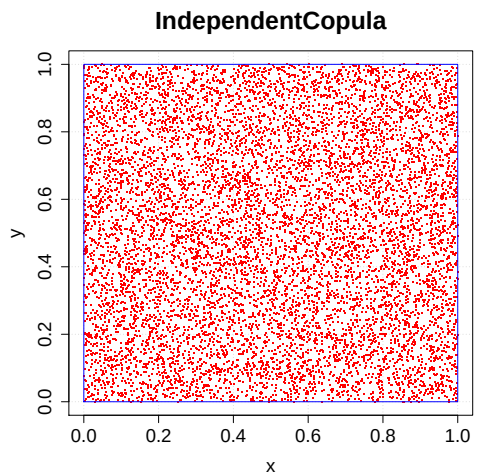












A short historical review on copulas and dependence modeling

1940 Hoeffding : measure of dependence, linear correlation multivariate distributions with uniform marginals on $[-1/2, 1/2]$.

1951 Fréchet : multivariate distributions with fixed marginal distributions.

1959 Sklar and Schweizer : probabilistic metric spaces, first occurence of the term copula.

1979 Deheuvels : independence tests, non parametric multivariate estimation.

1992 Darsow, Nguyen and Olsen : description of Markov processes in terms of copulas.

1999 Embrechts, Lindskog and McNeil : dissemination of copula methodology in financial and insurance applications.

2005 Mikosch : "Copulas : Tales and facts". Are copulas something else than a fashionable subject ?

2009 Salmon : "Recipe for a disaster : the formula that killed Wall Street"

Dependence Modeling using Copulas

1 Introduction

2 Multivariate distributions and Copulas

3 Lien avec les mesures d'association

4 Estimation

5 Dependence in high dimension

6 Conclusion

Multivariate distributions I

La copule est un ingrédient *obligatoire* de la fonction de répartition d'une loi multivariée :

Théorème de Sklar (1953)

Let F be a d-dimensional distribution function whose marginal distribution functions are $F_1, \ldots, F_d$. There exists a copula C of dimension d such that for $\underline{x} \in \mathbb{R}^d$, we have :

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

In the case of continuous marginal distributions, for all $\underline{u} \in [0,1]^d$, we have :

$$\begin{aligned} C(\underline{u}) &= F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)) && (2)\\ &= \mathbb{P}\left(X_1 \leq F_1^{-1}(u_1), \ldots, X_d \leq F_d^{-1}(u_d)\right) && (3)\\ &= \mathbb{P}\left(F_1(X_1) \leq u_1, \ldots, F_d(X_d) \leq u_d\right) && (4)\\ &= \mathbb{P}\left(U_1 \leq u_1, \ldots, U_d \leq u_d\right) && (5) \end{aligned}$$

Cette décomposition Marginales - Copule permet de créer une zoologie très importante de lois multivariées !

Copulas for dependence modeling I

Definition : Copula

A *d*-dimensional copula is the restriction to the unit cube $[0,1]^d$ of a multivariate distribution function with uniform univariate marginals on $[0,1]$.

Propriétés

Soit C une copule de dimension d, alors

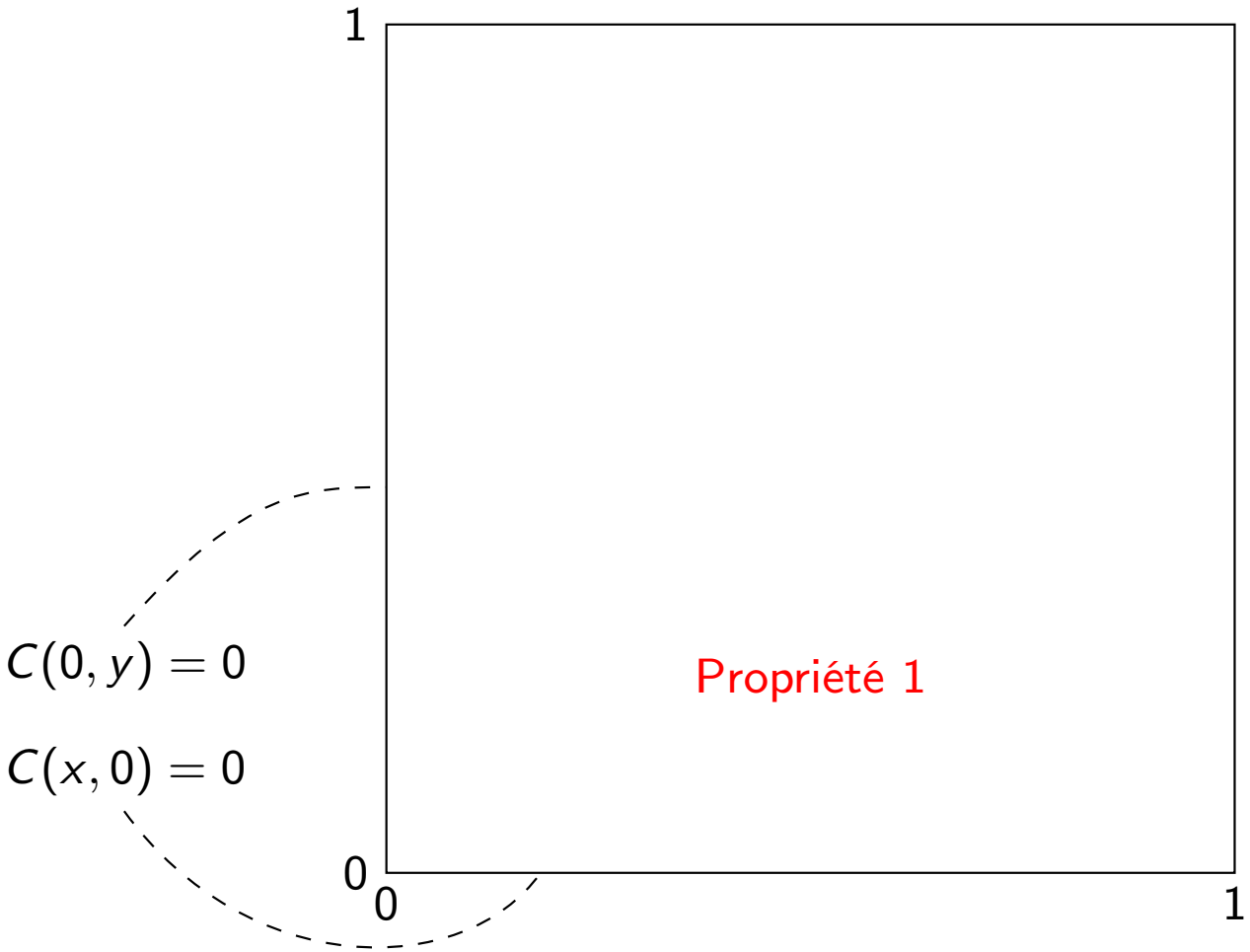
$$\forall \underline{u}, \underline{v} \in [0,1]^d, |C(\underline{u}) - C(\underline{v})| \leq \sum_{i=1}^{d} |u_i - v_i|$$

.

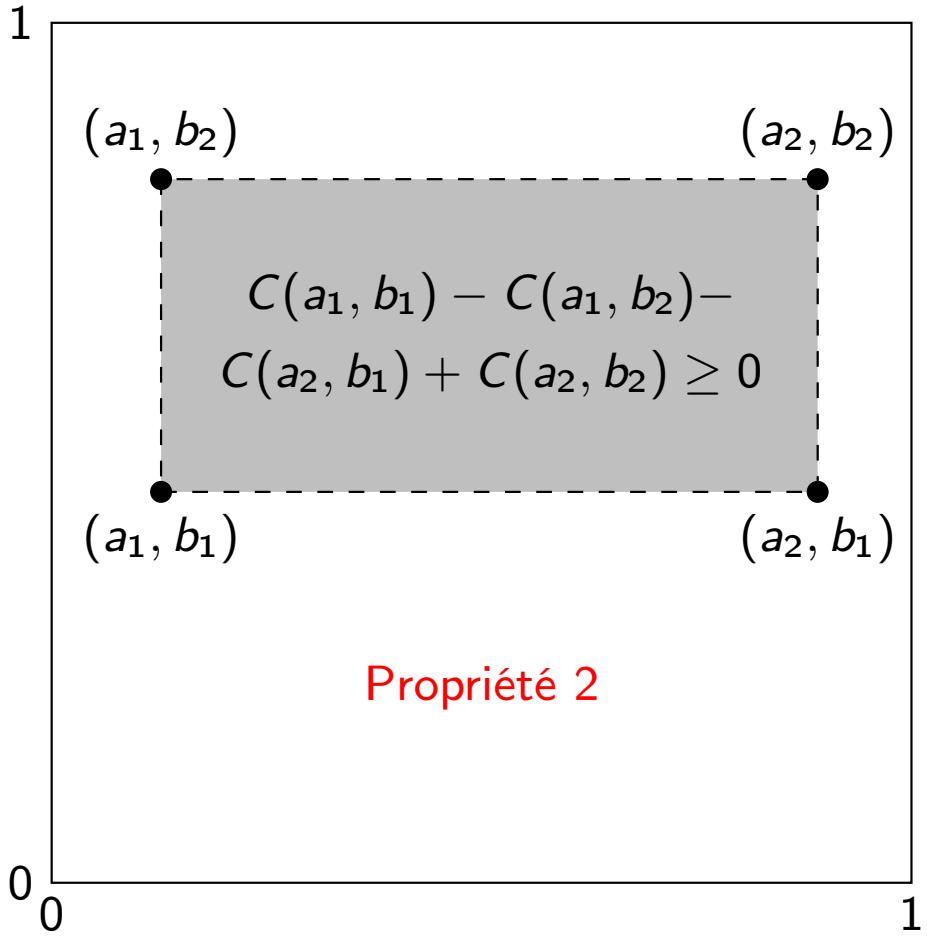
C vérifie en outre **propriétés suivantes** :

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

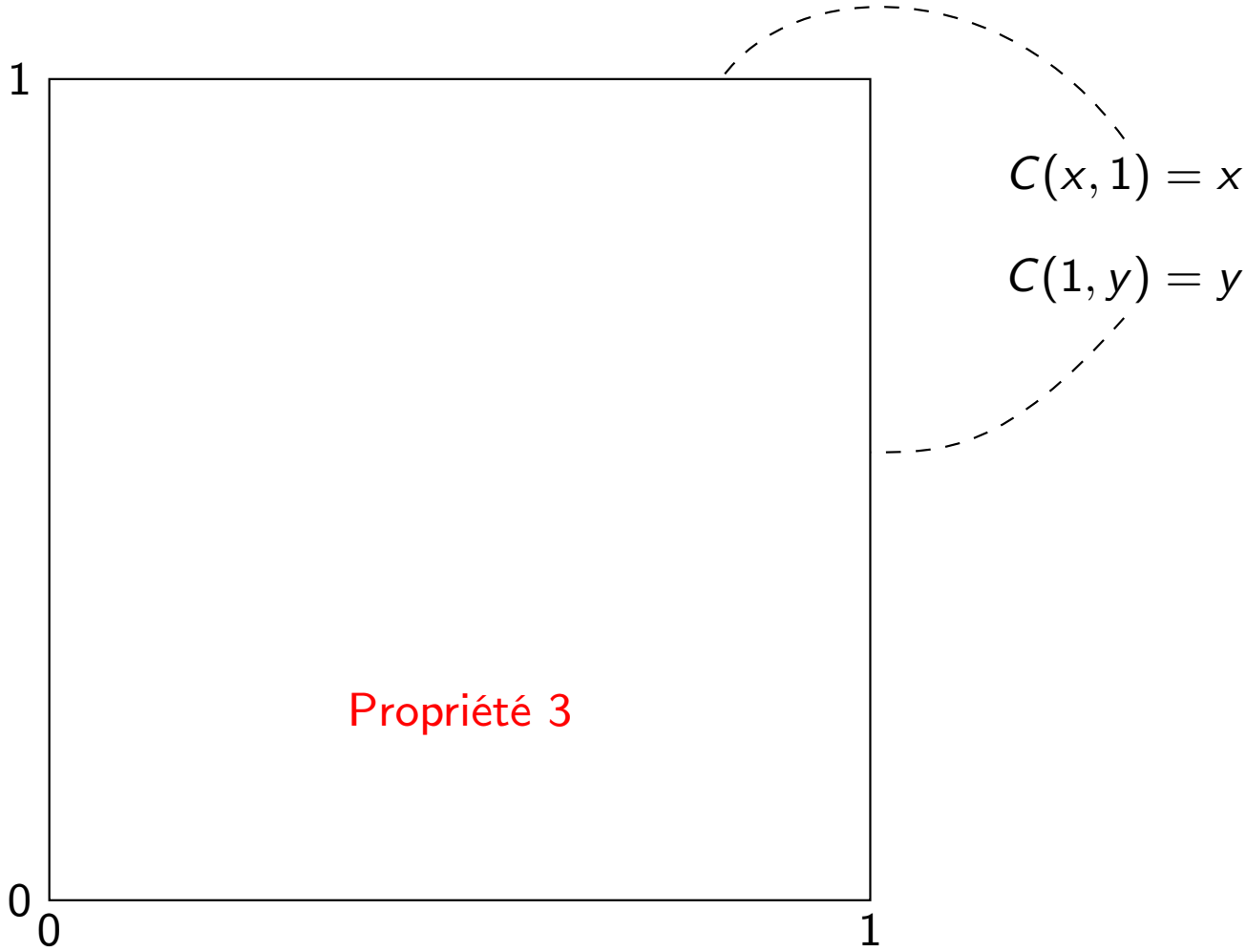
Illustration



Illustration



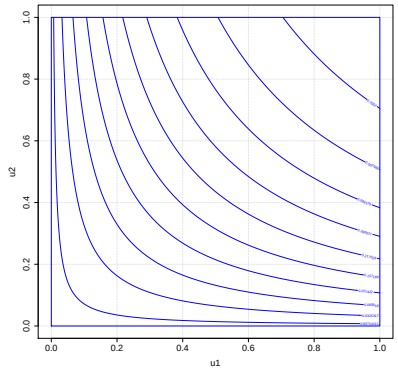
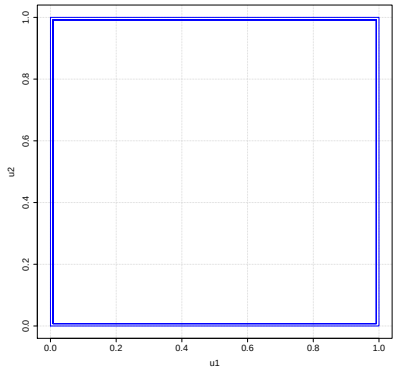
Illustration



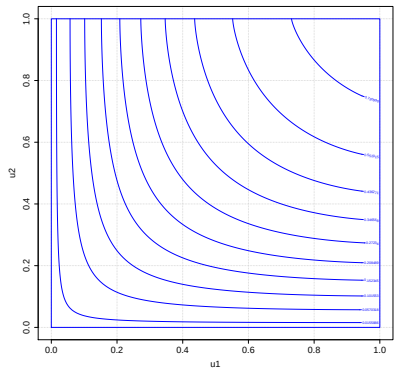
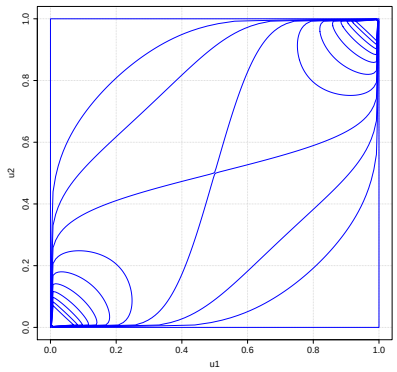
Examples

Independent

$$C(u_1, u_2) = u_1 u_2$$

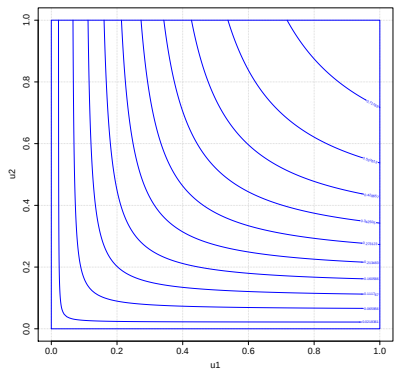
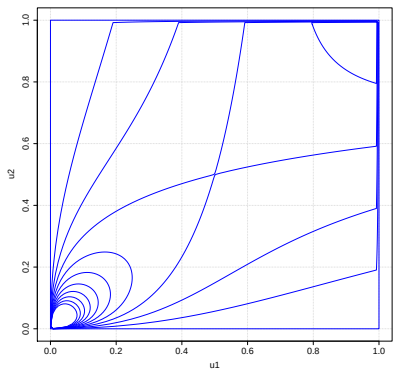



Normal

Clayton

$$C(u_1, u_2) = \left(u_1^{\theta} + u_2^{\theta} - 1\right)^{1/\theta}$$

Additional properties of copulas

Theorem : Fréchet-Hoeffding bounds

Let C be a d-dimensional copula. Then for all $\underline{u} \in [0,1]^d$ we have :

$$W_d(\underline{u}) := \max(0, u_1 + \ldots + u_d - d + 1) \leq C(\underline{u}) \leq M_d(\underline{u}) := \min(u_1, \ldots, u_d) \quad (6)$$

The copula M_d is called the **Min copula** and corresponds to random vectors $\underline{X}$ for which all the components are almost surely strictly increasing functions of a common random variable : $\underline{X} = (f_1(U), \ldots, f_d(U))$.
W_d i copula only if $d = 2$.

Theorem

Let $\underline{X}$ be a d-dimensional random vector with copula C and α_1,...,α_d be d strictly increasing functions from $\mathbb{R}$ to $\mathbb{R}$, then C is also a copula for the random vector $(\alpha_1(X_1), \ldots, \alpha_d(X_d))$.

Conclusion : Maping the data into the rank space does not change the dependence structure : $\alpha_i = \hat{F}_i$.

Familles de lois et de copules I

On peut caractériser les copules de dimension 2 par leur dépendance de queue qui quantifie la dépendance dans les valeurs extrêmes hautes ou basses.

Définition : Dépendance de queue haute

La dépendance de queue haute est le coefficient scalaire $\lambda_U \in [0,1]$:

$$\lambda_U = \lim_{u \to 1} \mathbb{P}\left(X > F_X^{-1}(u) | Y > F_Y^{-1}(u)\right) = \lim_{u \to 1} \frac{1 - 2u + C(u,u)}{1-u}$$

Propriétés : . Si $\lambda_U = 0$, alors les marginales sont indépendantes dans les valeurs extrêmes et la probabilité du cumul des aléas extrêmes $\mathbb{P}\left(X > F_X^{-1}(u) \cap Y > F_Y^{-1}(u)\right)$ est négligeable devant la probabilité marginale $\mathbb{P}\left(Y > F_Y^{-1}(u)\right)$.

Définition : Dépendance de queue basse

La dépendance de queue basse est le coefficient scalaire $\lambda_L \in [0,1]$:

$$\lambda_L = \lim_{u \to 0} \mathbb{P}\left(X < F_X^{-1}(u) | Y < F_Y^{-1}(u)\right) = \lim_{u \to 1} \frac{C(u,u)}{u}$$

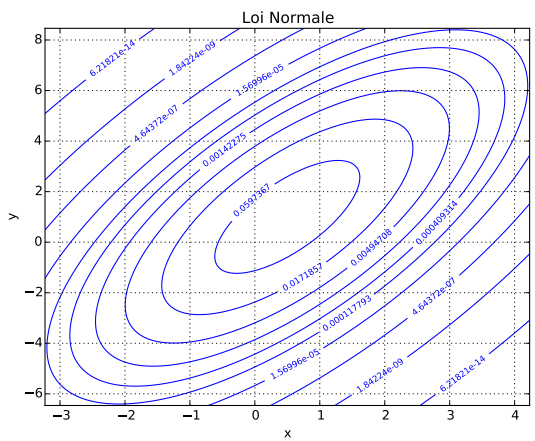
Familles de lois et de copules I

Un modèle multivarié très connu et très utilisé est celui des **vecteurs gaussiens** caractérisés par le **vecteur moyenne μ (position)** et la **matrice de covariance (dispersion)**.
Pourquoi ? Algèbre particulière : stablilité par transformation affine :

- marginales (dim 1 ou plus) : vecteurs gaussiens,
- combinaisons linéaires encore vecteurs gaussiens
- on connaît les paramètres (moyenne et covariance) de la transformation affine

$\Longrightarrow$ **On cherche les distributions ayant les mêmes propriétés algébriques : lois elliptiques**.

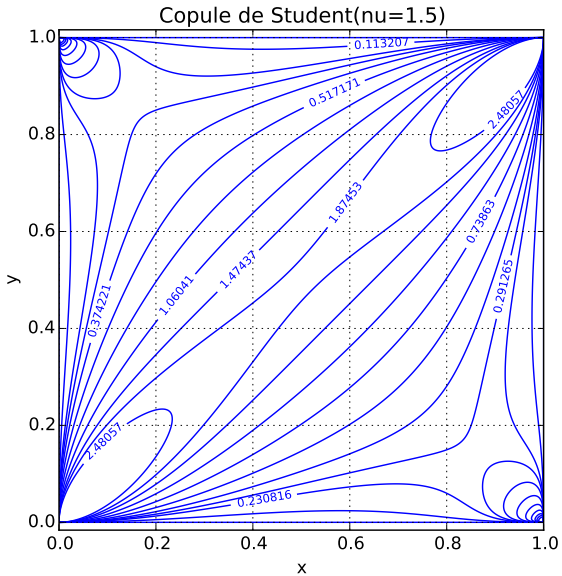
Lois elliptiques : **stables par transformation affine**.
Les lois sont caractérisées par un vecteur moyenne et une matrice de covariance (position et dispersion) + quelques paramètres spécifiques (de forme : ν de la loi de Student, ...) qui restent constants par transformation affine.

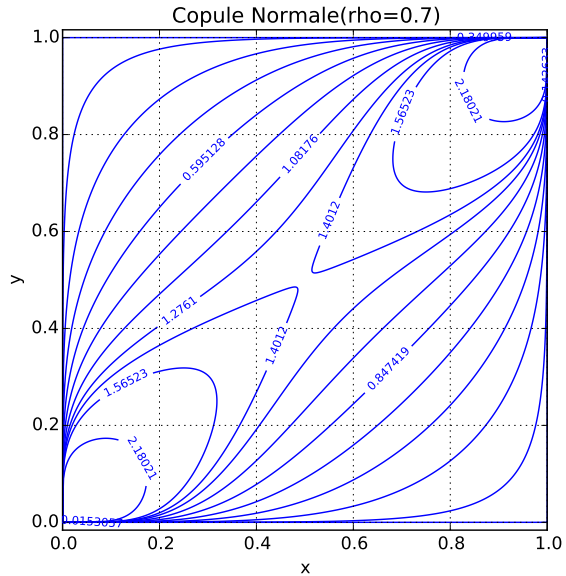


$\underline{X} = \underline{\mu} + R\,\underline{\underline{M}}\,\underline{U}$ où R est une variable aléatoire ≥ 0, $\underline{U}$ est uniformément distribuée sur la sphère unité de dimension $d_U \leq d$, $(R, \underline{U})$ indépendantes, $\underline{\underline{M}} \in \mathcal{M}(d, d_U)$.
Les lois elliptiques sont symétriques par rapport à μ $\Longrightarrow$ même comportement dans les extrêmes hauts et bas.

Familles de lois et de copules II

Copules elliptiques : **copule des lois elliptiques** (inverse de Sklar).
L'intérêt est de pouvoir utiliser la structure de dépendance des lois elliptiques et de la combiner à des marginales quelconques.
On conserve donc la symétrie par rapport au centre du graphe $[0,1]^2$: $\lambda_L = \lambda_U$.





Ex :
copule de **Student** : $\lambda_L = \lambda_U > 0$, copule **Normale** : $\lambda_L = \lambda_U = 0$.
$\Longrightarrow$ **On cherche les copules permettant de disymétriser les extrêmes hauts et bas**.

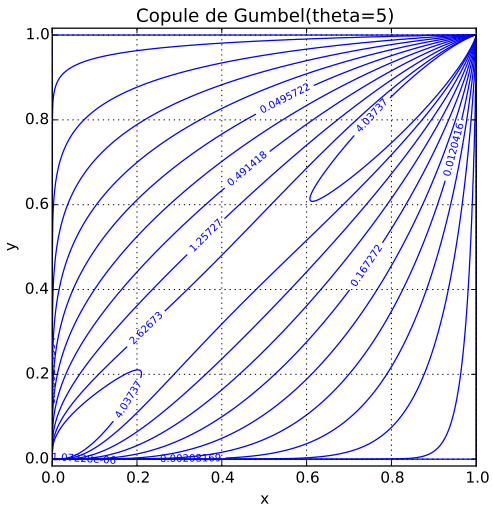
Familles de lois et de copules III

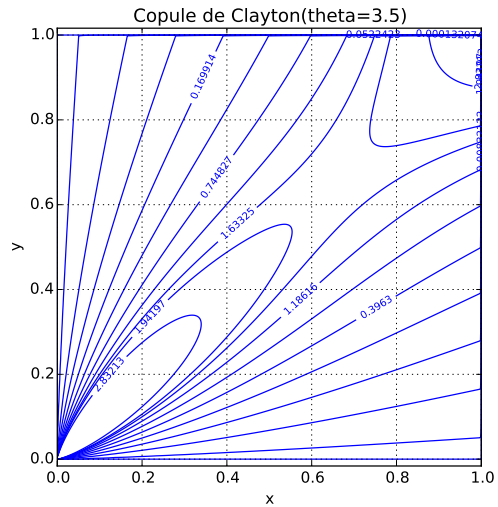
Copules archimédiennes : **disymétrie dans les extrêmes hauts et bas**.

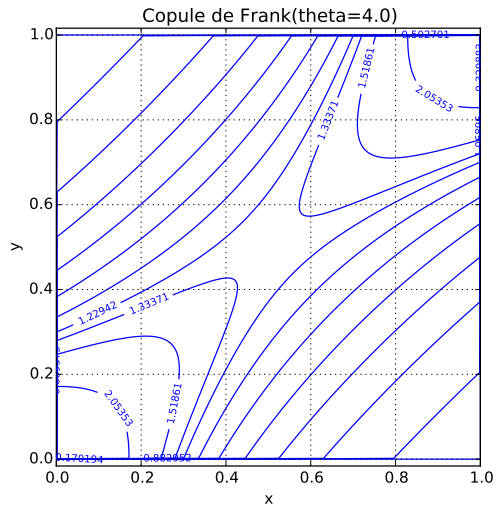
Elles sont définies à partir d'un générateur φ : $C(u,v) = \varphi^{-1}(\varphi(u) + \varphi(v))$ où φ est convexe décroissante $[0,1] \mapsto [0,+\infty]$.

Les propriétés probabilistes sur C sont déterminées par ϕ.

Les copules archimédiennes sont symétriques par rapport à la première diagonale : u et v sont interchangeables : $C(u,v) = C(v,u)$.



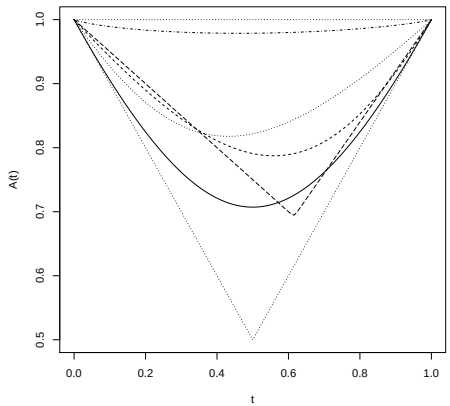




Ex :
copule de **Gumbel** : $\lambda_L = 0$, $\lambda_U > 0$ si $\theta > 1$; copule de **Clayton** : $\lambda_U = 0$, $\lambda_L > 0$ si $\theta > 0$; copule de **Frank** : $\lambda_U = \lambda_L = 0$. Cette copule présente une symétrie centrale et charge surtout le centre.

Familles de lois et de copules IV

Autre famille de copule : **copules de valeurs extrêmes**.
C'est la copule de la limite d'un maximum multidimensionnel renormalisé.
En dimension 2, elle est déterminée par la **fonction de dépendance de queue de Pickands** A, convexe et telle que $(1-t) \vee t \leq A(t) \leq 1, \quad t \in [0,1]$:



$$C(u_1,u_2) = \exp\left\{\log(u_1u_2)A\left(\frac{\log(u_2)}{\log(u_1u_2)}\right)\right\} \quad \forall \underline{u} \in [0,1]^2$$

Indépendance $\Longleftrightarrow A(1/2)=1 \Longleftrightarrow C(u_1,u_2)=u_1u_2$
Dépendance $\Longleftrightarrow A(1/2)=1/2$

Copule indépendante :

$$C(u_1,\ldots,u_d) = \Pi u_i \quad \text{et} \quad F(x_1,\ldots,x_d) = \Pi_i F_i(x_i)$$

Cette copule entraîne que toute manipulation de F revient à manipuler d fonctions de dimension 1.

Familles de lois et de copules V

Démonstration :

$$C(u_1, u_2) = \exp\left\{\log(u_1 u_2) A\left(\frac{\log(u_2)}{\log(u_1 u_2)}\right)\right\} \quad \forall \underline{u} \in [0,1]^2$$

Si $A(1/2) = 1$ alors $A(t) = 1$ pour tout t et

$$C(u_1, u_2) = exp\left\{\log(u_1 u_2) \times 1\right\} = u_1 u_2 \tag{7}$$

Si $A(1/2) = 1/2$ alors $A(t) = (1-t) \vee t$ pour tout t et donc

$$C(u_1, u_2) = exp\left\{\log(u_1 u_2) \max\left[1 - \left(\frac{\log u_2}{\log u_1 u_2}\right), \left(\frac{\log u_2}{\log u_1 u_2}\right)\right]\right\} \tag{8}$$

$$= exp\left\{\log(u_1 u_2) \max\left[\left(\frac{\log u_1}{\log u_1 u_2}\right), \left(\frac{\log u_2}{\log u_1 u_2}\right)\right]\right\} \tag{9}$$

$$= exp\left\{\min\left[\log u_1, \log u_2\right]\right\} \tag{10}$$

$$= exp\left\{\log \min\left[u_1, u_2\right]\right\} \tag{11}$$

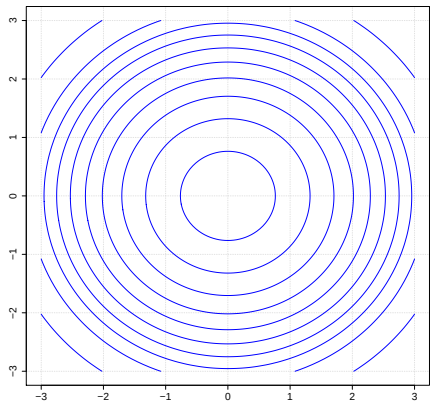
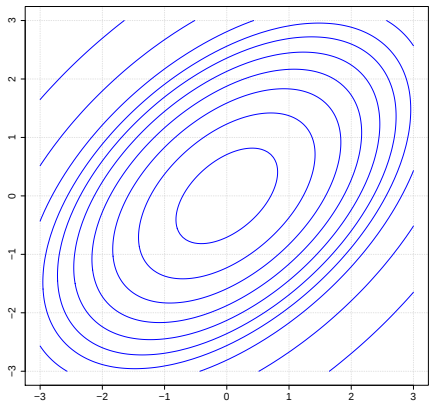
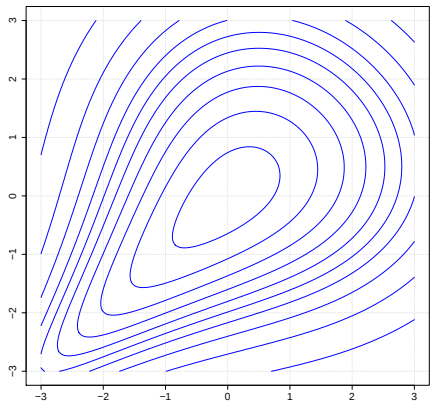
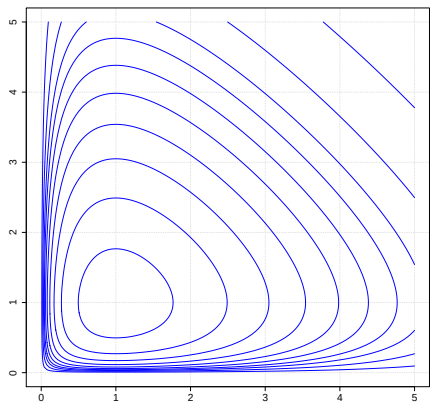
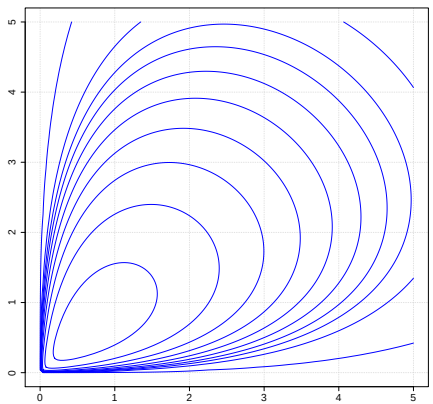
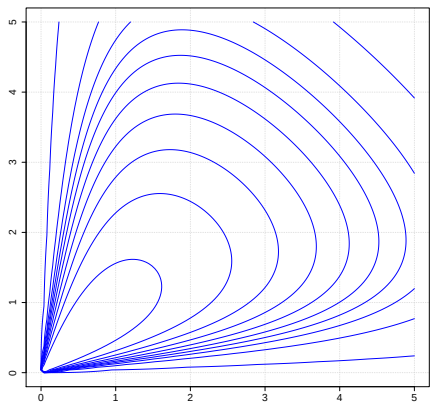
$$= \min\left[u_1, u_2\right] \tag{12}$$

$$= M_2(u_1, u_2) \tag{13}$$

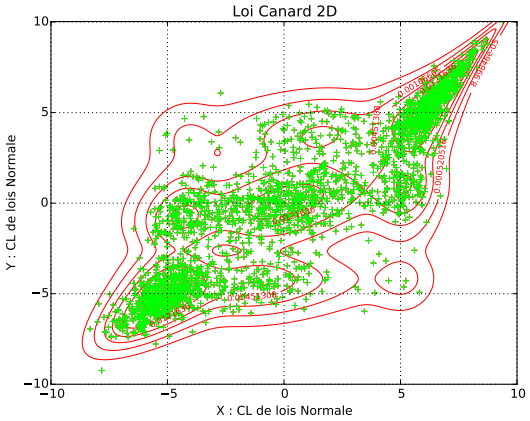
C'est la copule Min en dimension 2.

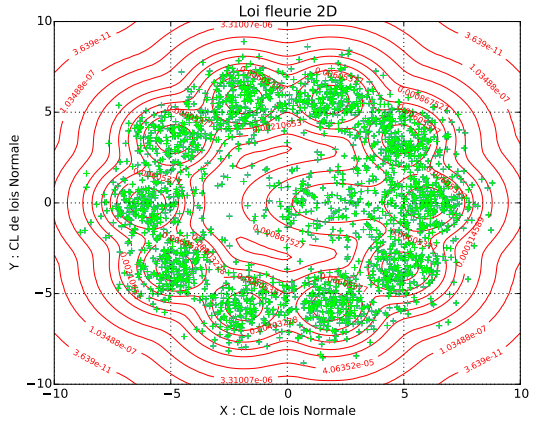
Examples of composed distribution I

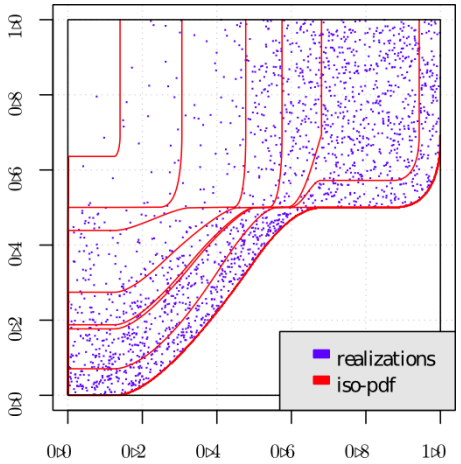
Zoologie très diversifiée : $F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$.

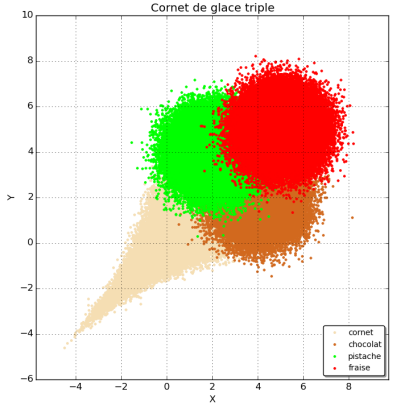
	Independent	Normal	Clayton
$\mathcal{N}(0,1)$ marginals			
$\Gamma(2,1)$ marginals			

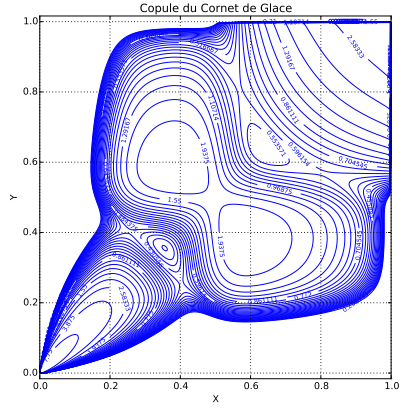
Examples of composed distribution II











Loi *Fleur* : mixture de gaussiennes + copule gausienne
Loi *Canard* : gaussiennes + copule Gumbel
Loi *Cornet de glace* : gaussiennes + copule Clayton

Sampling a copula I

Définition : Copules marginales et conditionnelles

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

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

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

Sampling a multivariate distribution needs to sample a copula :

1. Generate $u_1 \sim \mathcal{U}(0,1)$;
2. For $k \in \{2, \dots, n\}$, generate $u_k \sim C_{k|1,\dots,k-1}(u_1, \dots, u_{k-1})$.
3. The resulting point $(u_1, \dots, u_d)$ is a realization of C.

Remark : for many copulas, more efficient specialized algorithms exist

Sampling a composed distribution

Let $\underline{X}$ be a random vector with marginal distribution functions $F_1, \ldots, F_d$ and copula C. Its distribution function writes :

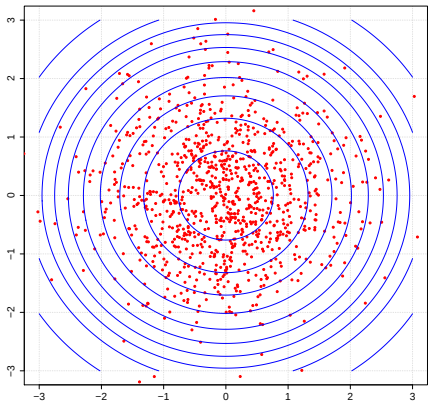
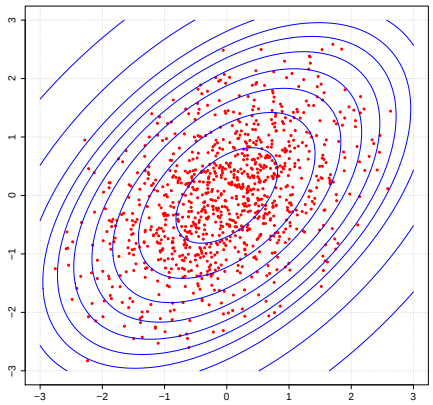
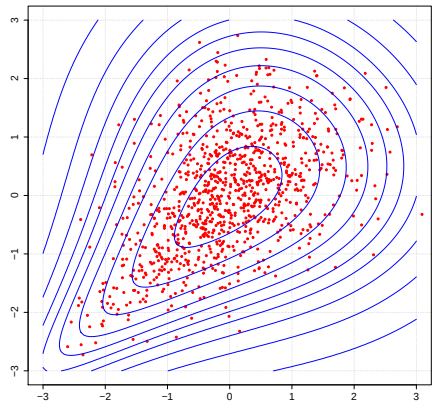
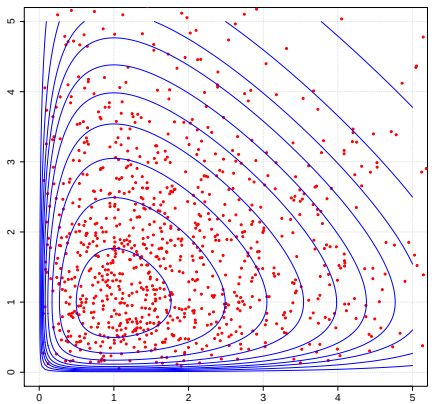
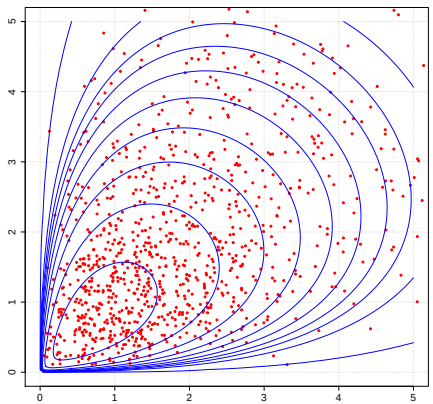
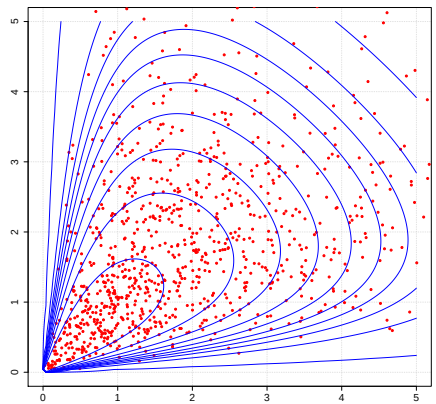
$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)).$$

One can sample $\underline{X}$ by the following two-steps procedure :

1. Generate $\underline{u} \sim C$;
2. A realization $\underline{x}$ of $\underline{X}$ is given by :

$$\underline{x} = \left(F_1^{(-1)}(u_1), \ldots, F_d^{(-1)}(u_d)\right) \tag{14}$$

Sampling of composed distributions

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

More modeling tools : ordinal sum

Définition : Somme ordinale de copules

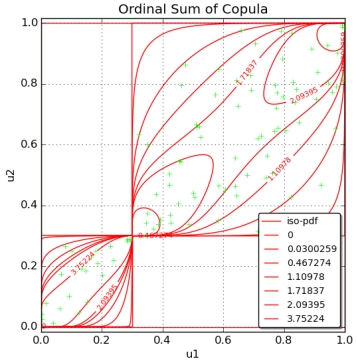
Somme ordinale de copules : Création d'une copule par sous domaine du pavé $[0,1]^d$ autour de la diagonale : chaque copule C_i est ramenée au support $[\alpha_i, \alpha_{i+1}]$ pour $0 \leq \alpha_1 \leq \cdots \leq \alpha_{n-1}$:

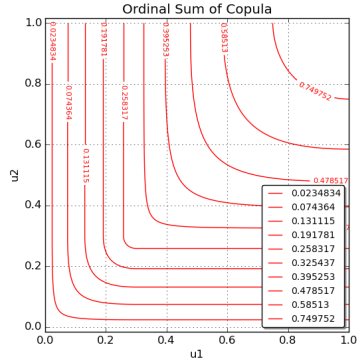
$$C(\underline{u}) = \begin{cases} \alpha_i + C_i\left(\dfrac{u_1 - \alpha_i}{\alpha_{i+1} - \alpha_i}, \ldots, \dfrac{u_d - \alpha_i}{\alpha_{i+1} - \alpha_i}\right) & \text{if } \underline{u} \in [\alpha_i, \alpha_{i+1}[^d \\ M_d(\underline{u}) & \text{sinon} \end{cases}$$

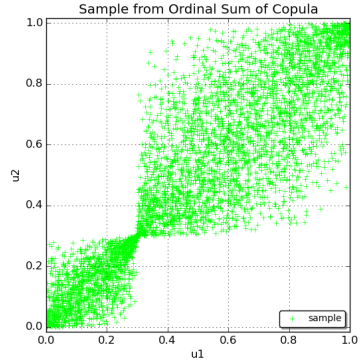
où $M_d(\underline{u}) = \min_{k=1\ldots d} u_k$ est la copule Min.

Somme ordinale des copules Gumbel et Normal :

```
-------------------------------------------------------------
>>> myCop1 = ot.GumbelCopula(2)
>>> myCop2 = ot.NormalCopula(2)]
>>> alpha = 0.3
>>> myOrdSumCop = ot.OrdinalSumCopula([myCop1, myCop2], alpha)
>>> graph = myOrdSumCop.drawPDF()
-------------------------------------------------------------
```







More modeling tools : Loi de statistique d'ordre d'entropie maximale

Définition : Loi de statistique d'ordre d'entropie maximale

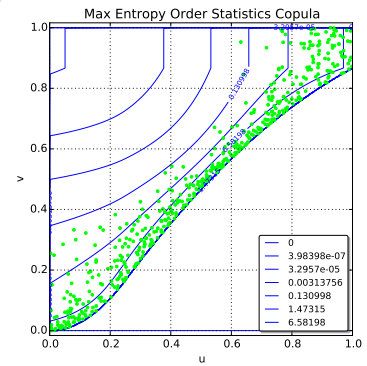
La **loi de statistique d'ordre d'entropie maximale** permet de modéliser la loi de $(X_1, \ldots, X_d)$ de lois marginales $(\boldsymbol{F_1}, \ldots, \boldsymbol{F_d})$ (avec $F_1 \geq \ldots F_d$) et tel que $\boldsymbol{X_1} \leq \cdots \leq \boldsymbol{X_d}$ presque sûrement. Cette loi est l'unique d'**entropie maximale**.

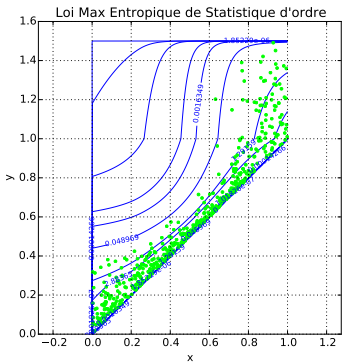
$$f_X(x) = f_1(x_1) \prod_{k=2}^{d} \phi_k(x_k) \exp\left(- \int_{x_{k-\mathbf{1}}}^{x_k} \phi_k(s) ds \right) 1_{x_{\mathbf{1}} \leq \cdots \leq x_d}$$

avec $\phi_k(x_k) = \frac{f_k(x_k)}{F_{k-\mathbf{1}}(x_k) - F_k(x_k)}$.

Loi de statistique d'ordre d'entropie maximale :

```
>>> myDist = [ot.Uniform(0.0, 1.0),  ot.Triangular(0.0, 0.25, 1.5)]
>>> myOrderStatDist = ot.MaximumEntropyOrderStatisticsDistribution(myDist)
>>> graph = myOrderStatDist.drawPDF()
```





Copule de statistique d'ordre d'entropie maximale :

```
>>> myOrderStatCop = ot.MaximumEntropyOrderStatisticsCopula(myDist)
ou
>>> myOrderStatCop = myOrderStatDist.getCopula()

>>> graph = myOrderStatCop.drawPDF()
```

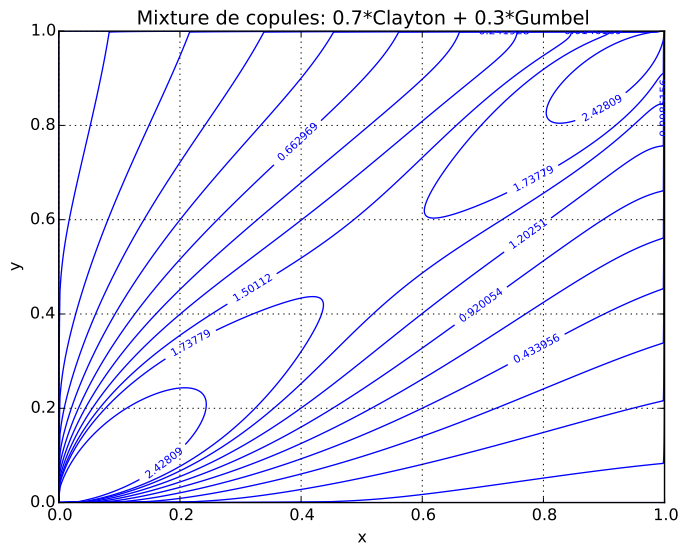
More modeling tools : Mixture de copules

Définition : Mixture de copules

Pour tout $0 < \alpha_i < 1$ et $\sum \alpha_i = 1$ et C_i une copule de dimension d. On appelle **Mixture de copules** la copule $C = \sum \alpha_i C_i$ de dimension d.

L'ensemble des copules étant convexe, C est une copule.
L'intérêt des mixtures de copules est de pouvoir mixer les comportements de plusieurs copules.
Ex : Mélanger une copule ayant $\lambda_L > 0$ avec une copule ayant une ayant $\lambda_U > 0$ et $\lambda_L \neq \lambda_U$. Les coefficients de dépendance de queue se calculent par : $\boldsymbol{\lambda_L = \sum \alpha_i \lambda_{i,L}}$ et $\boldsymbol{\lambda_U = \sum \alpha_i \lambda_{i,U}}$.

Exemple d'un mélange d'une **Clayton** $(\lambda_U = 0, \lambda_L > 0)$ et d'une **Gumbel** $(\lambda_U > 0, \lambda_L = 0)$



Dependence Modeling using Copulas

1 Introduction

2 Multivariate distributions and Copulas

3 Lien avec les mesures d'association

4 Estimation

5 Dependence in high dimension

6 Conclusion

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.

Mesure d'association

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

Défnition : Corrélation linéaire

La **corrélation linéaire** ρ entre deux variables aléatoires X_1 et X_2 telles que $\mathbb{V}(X_i) = \sigma_i^2$ soient finies, est définie par :

$$\begin{aligned} \rho(X_1, X_2) &:= \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1 \sigma_2} \\ &= \frac{1}{\sigma_1 \sigma_2} \iint_{\mathbb{R}^2} F_{12}(x_1, x_2) - F_1(x_1) F_2(x_2) \, \mathrm{d}x_1 \mathrm{d}x_2 \end{aligned} \tag{15}$$

Propriétés

- $\rho(X_1, X_2) \in [-1, 1]$ avec $|\rho(X_1, X_2)| = 1 \Longleftrightarrow \exists\, a, b \in \mathbb{R}, a \neq 0, X_2 = aX_1 + b$
- 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

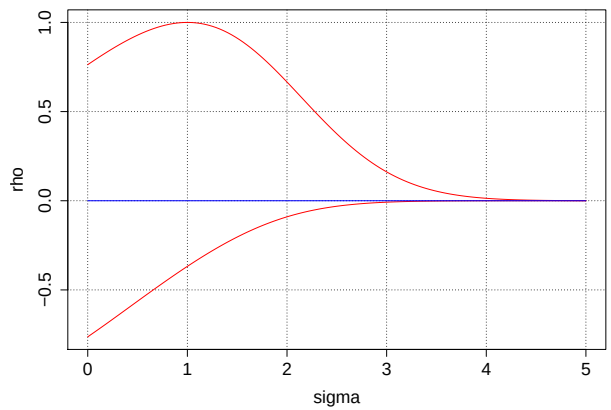
Valeurs possibles de $\rho(X_1, X_2)$

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 inclus 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

$$\rho_{min} = \frac{e^{-\sigma} - 1}{\sqrt{e-1}\sqrt{e^{\sigma^2} - 1}} \text{ et } \rho_{max} = \frac{e^{\sigma} - 1}{\sqrt{e-1}\sqrt{e^{\sigma^2} - 1}}.$$



On note que ρ_{min} et ρ_{max} tendent vers 0 quand σ tend vers $+\infty$. Pour $\sigma = 5, \rho \in [-3.10^{-6}, 4.10^{-4}]$!

Conclusion : **La corrélation linéaire ne se choisit pas indépendamment des marginales . . .**

Rho de Spearman

Définition : Rho de Spearman

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).

Propriétés

- $\rho_S(X_1, X_2) \in [-1, 1]$ avec $|\rho_S(X_1, X_2)| = 1 \Longleftrightarrow \exists\, \varphi$ 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 : Tau de Kendall

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

Propriétés

- $\tau(X_1, X_2) \in [-1, 1]$ 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, elles permettent de paramétrer la copule : on utilise pour cela les relations entre ρ_S et τ et les paramètres $\underline{\theta}$ de la copule $C_{\underline{\theta}}$:

Exemples :

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

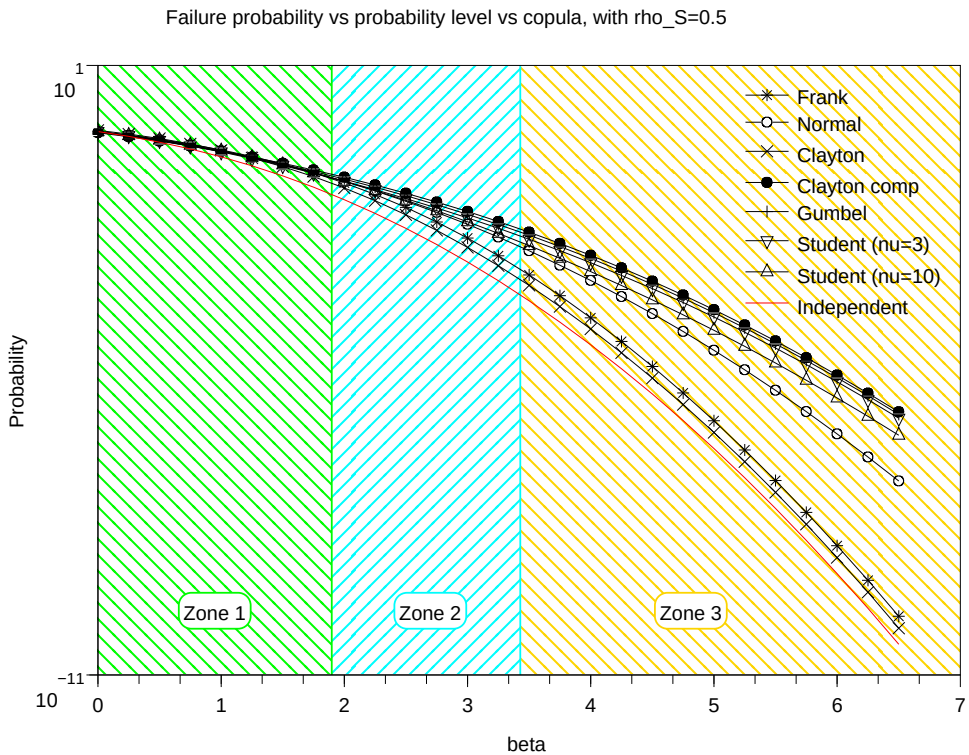
On peut ainsi construit un estimateur $\hat{\underline{\theta}}_n$ de $\underline{\theta}$ à partir des estimateurs $\hat{\rho}_S$ et $\hat{\tau}$.

Intérêt : Cette estimation paramétrique de la copule est robuste à l'estimation des lois marginales : $\hat{\underline{\theta}}_n$ est estimé directement sur l'échantillon $\underline{X}_i$ et non sur l'échantillon des rangs obtenus par l'application des F_i sur les données.

Mesures d'association ou copules ?

Spécifier une mesure d'association : est-ce suffisant ?

$\mathbb{P}\left(X_1 + X_2 \geq \beta\sqrt{2}\right)$ 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	$8.3\,10^{-11}$	$1.9\,10^{-6}$	$2.3\,10^{4}$

Conclusion : **Une mesure d'association sert à paramétrer un *modèle* de copule.**

Dependence Modeling using Copulas

1 Introduction

2 Multivariate distributions and Copulas

3 Lien avec les mesures d'association

4 Estimation

5 Dependence in high dimension

6 Conclusion

Rang et statistique dordre

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

Définition : Rang d'une donnée

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 :

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

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 $(\underline{X}_k)_{k\in\{1,\dots,n\}}$ un échantillon de taille n d'une loi multivariée $\mathcal{L}_{\underline{X}}$.

Estimation de la loi multivariée $\mathcal{L}_{\underline{X}}$:

- estimation directe de $\mathcal{L}_{\underline{X}}$: paramétrique ou non paramétrique. On peut en extraire les marginales et la copule (Sklar) ;
- estimation via la décomposition en lois conditionnelles : $F_{12}(x_1, x_2) = F_{1|2}(x_1, x_2)F_2(x_2)$
- estimation via la décomposition en marginales F_j et copule C :
  1. Identifier les fonctions de répartition marginales ;
  2. Transformer l'échantillon $(\underline{X}_k)_{k\in\{1,\dots,n\}}$ en l'échantillon des rangs renormalisé $(\underline{U}_k)_{k\in\{1,\dots,n\}}$;
  3. Estimer la copule sur la base de $(\underline{U}_k)_{k\in\{1,\dots,n\}}$.
- ...

Estimation des lois marginales :

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

Estimation statistique II

Estimation de la copule :

- Estimation paramétrique : $C^{\underline{\theta}} \in \mathcal{C}(\underline{\theta})$, on estime θ par $\underline{\hat{\theta}}_N(U_1^j, \ldots, U_d^j)$ et on prend $\hat{C} = C^{\underline{\hat{\theta}}_N(X_1^j, \ldots, X_d^j)}$ comme estimation de la copule.
- Estimation non-paramétrique : attention aux propriétés d'une copule (marginales uniformes) : les techniques de reconstruction à noyaux ne garantissent pas l'uniformité des marginales. Utiliser plutôt la copule de Bernstein.

Quelle stratégie ? : Tout paramétrique ? Tout non paramétrique ? Mixte paramétrique / non paramétrique ? (=semi-paramétrique)
Voir les travaux de P. Lambert, K. Kostadinov, A. Charpentier, J-D. Fermanian ([Fermanian], [Scaillet]) : **ne pas utiliser les marginales estimées pour créer l'échantillon des rangs ; utiliser les rangs empiriques.**
En pratique :

- estimation des modèles marginaux par max de vraisemblance ;
- estimtion de la copule dans l'espce des rangs empiriques.

Estimation non paramétrique d'une copule I

Définition : Interpolant de Bernstein

Soit α une fonction définie sur $[0,1]^d$ à valeurs dans $[0,1]$. On construit l'interpolant de Bernstein de α associé à la grille $G = \left\{\frac{0}{m_1}, \ldots, \frac{m_1}{m_1}\right\} \times \ldots \times \left\{\frac{0}{m_d}, \ldots, \frac{m_d}{m_d}\right\}$ la fonction définie par :

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

où $P_{a,b}$ est le **polynôme de Bernstein** de paramètres (a,b) défini par :

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

Copule de Bernstein

Si $\alpha : [0,1]^d \to [0,1]$ coïncide avec une copule C sur la grille G, alors C^B est une copule : c'est la copule de Bernstein associé à C.

Estimation non paramétrique d'une copule II

Définition : *Copule empirique C_n*

Soit $(\underline{X}^i)_{1\leq i\leq n}$ un échantillon de taille n, associé à l'échantillon des rangs renormalisés $(\underline{U}^i)_{1\leq i\leq n}$. On appelle (de manière impropre !) *copule empirique C_n* la fonction de répartition de la loi discrète portée par $(\underline{U}^i)_{1\leq i\leq n}$:

$$C_n(\underline{u}) = \frac{1}{n}\mathrm{card}\left\{\underline{U}^i \in [\underline{0}, \underline{u}]\right\}, \quad \forall \underline{u} \in [\underline{0}, \underline{1}]$$

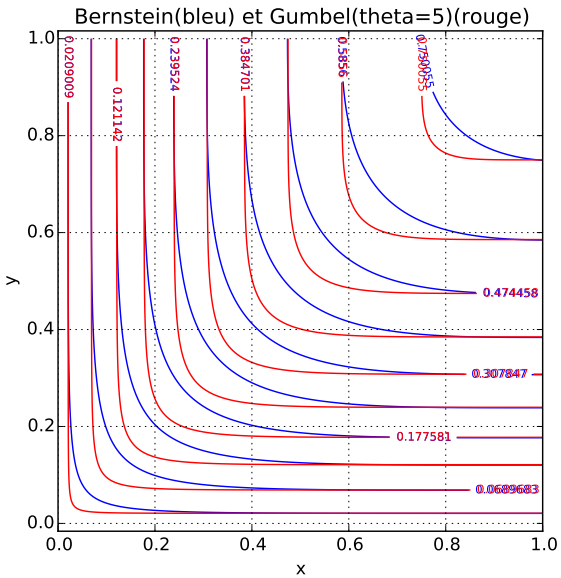
Attention : **C_n n'est pas une copule** car ses marginales 1d ne sont pas uniformes *continues* sur $[0, 1]$!

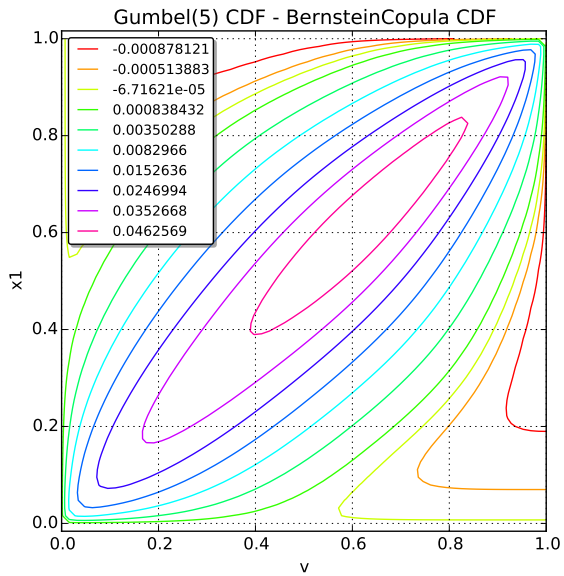
Theorème

Soit $(\underline{X}_k)_{k\in\{1,\dots,n\}}$ 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 (l'interpolant de Bernstein de 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.

Estimation non paramétrique d'une copule III

Exemple : Reconstruction d'une copule de Gumbel($\theta = 5$) à partir d'une échantillon de 10^3 points.
Erreur sur la CDF : (Gumbel - Bernstein) : max : $4.6\,10^{-3}$ et min : $-9.\,10^{-4}$





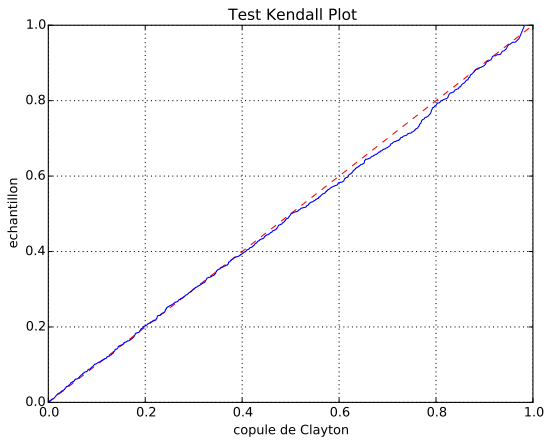
Test d'adéquation

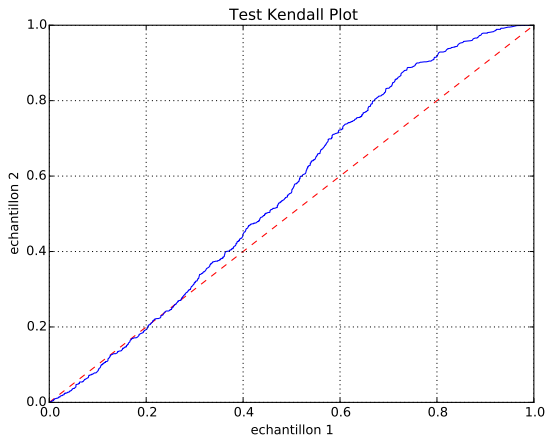
Ce domaine est encore en plein essor, depuis le travail pionnier de J-D. Fermanian basé sur une comparaison du modèle proposé avec une reconstruction à noyaux multivariée : ([Genest2009], [Fermanian]).

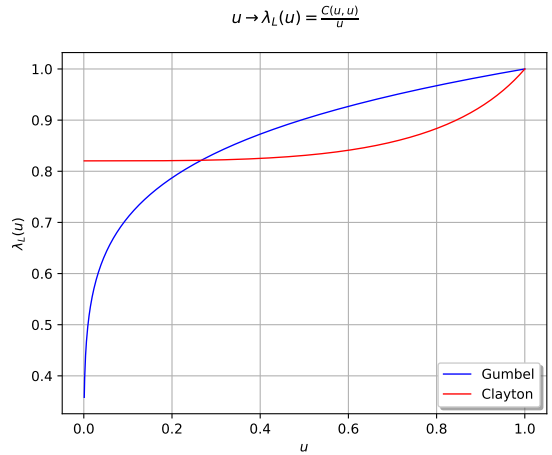
Bonne nouvelle : en 2D, les tests semblent performants dès $N = 150$ observations

Test de Kendall (dim 2) : 2 utilisations : **comparaison de la copule d'un échantillon à un modèle + comparaison de la copule de 2 échantillons**.

Le test de Kendall compare la CDF (=*fct de Kendall*) de $Z = F_{\underline{X}}(\underline{X})$ empirique et issue du modèle proposé (on a $Z = C_{\underline{X}}(F_1(X_1), F_2(X_2))$: ne dépend que de $C_{\underline{X}}$).







Ech. 1 : copule de Clayton($\theta = 3.5$). Ech. 2 : copule de Gumbel ($\theta = 5$) avec $N = 10^3$.

Conclusion :

- Mauvaise adéquation en queue haute : $\lambda_U(Clayton) = 0$ et $\lambda_U(Gumbel) = 2 - 2^{1/\theta}$
- En queue basse : c'est mieux même si $\lambda_L(Clayton) = 2^{-1/\theta}$ et $\lambda_L(Gumbel) = 0$ (décroissance lente de $\lambda_L(Gumbel)(u)$ quand $u \to 0$)

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 qui permettent de passer dans l'espace des rangs.

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.

Estimation des mesures d'association I

Estimateur du ρ de Spearman

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

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

où $\overline{rank}(X_1) = \frac{1}{N}\sum_{k=1}^{N} rank(X_1^k)$ et $\overline{rank}(X_2) = \frac{1}{N}\sum_{k=1}^{N} rank(X_2^k)$.

Théorème

Soit $\underline{X}$ un vecteur aléatoire continu bidimensionnel. Alors :

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

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

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 $\eta_{k\ell} = \displaystyle\int\int_{[0,1]^2} \left(u_1 - \frac{1}{2}\right)^k \left(u_2 - \frac{1}{2}\right)^\ell c(u_1, u_2)\, du_1\, du_2$ et c est la densité de la copule de $\underline{X}$.

Estimation des mesures d'association III

Estimateur du tau de Kendall

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

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

Théorème

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

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

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

où la variance asymptotique σ_τ^2 est donnée par :

$$\sigma_\tau^2 = 4\mathbb{V}\left(\mathbb{E}\left(sign(X_1 - X_1')\, sign(X_2 - X_2') \,|\, X_1, X_2\right)\right)$$

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

Dependence Modeling using Copulas

1 Introduction

2 Multivariate distributions and Copulas

3 Lien avec les mesures d'association

4 Estimation

5 Dependence in high dimension

6 Conclusion

How to increase the dimension ? I

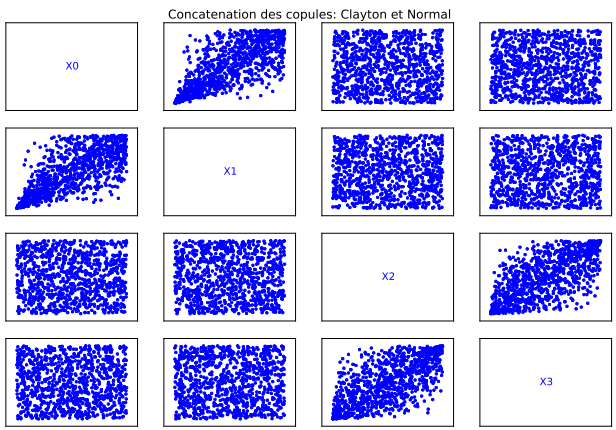
It is possible to build copulas of high dimension from copulas of low dimensions :

Definition : Composed copulas

Let $C_1, \ldots, C_k$ be k copulas of dimensions $n_1, \ldots, n_k$ and $d = \sum_{i=1}^{k} n_i$. The function C defined on $[0,1]^d$ by :

$$\boldsymbol{C(u_1, \ldots, u_d) = C_1(u_1, \ldots, u_{n_1}) \times \cdots \times C_k(u_{d-n_k+1}, \ldots, u_d)} \quad (16)$$

is a copula of dimension d. It is a sparse block-diagonal dependence structure based on several low dimensional dense dependence structures.



Exemple : Concaténation dune copule de Clayton ($\theta = 2.3$) et d'une copule Normale de dimension 2 avec $\rho = 0.7$

En pratique I

Quelle dimension ?

En pratique, il est nécessaire de bien analyser la réelle 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** 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 apparente** par **transport d'une petite dimension stochastique** : $\underline{X} = \phi(\underline{W})$ avec $\dim \underline{X} \gg \dim \underline{W} \simeq 2$. Dans ce cas il faut **paramétrer le problème par** $\underline{W}$ quitte à méta-modéliser ϕ.

En pratique II

Quelle dimension ?

- **Grande dimension réelle** ($\gg 100$) 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 (caractéristique d'un matériau indexé par l'espace). 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.
- **Grande dimension non réductible** : on privilégie alors l'approche compromis adéquation/complexité pour l'estimation de la loi multivariée. 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).

Dependence Modeling using Copulas

1 Introduction

2 Multivariate distributions and Copulas

3 Lien avec les mesures d'association

4 Estimation

5 Dependence in high dimension

6 Conclusion

Conclusion I

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 adaptées au calcul ou à l'estimation statistique (processus, modèles bayésiens)
- Toutes les étapes depuis l'estimation statistique à partir de données multivariées jusqu'à la simulation de Monte Carlo peuvent être réalisées via une modélisation à base de copules.
- A partir d'un ensemble de copules, il est possible de créer de nouvelles copules par assemblage de manière efficace.

Conclusion II

Des challenges scientifiques

- Modélisation de la dépendance en grande dimension.
- Combattre le fléau de la dimension pour les aspects statistiques.
- Lien avec les processus à temps discret : si le processus est Markovien, il est possible de traduire en termes d'opérations sur les copules la propriété de Markov.
- ...

References I

D. Fermanian.
Goodness-of-fit tests for copulas.
Journal of Multivariate Analysis, 95 :119–152, 2005.

C. Genest, B. Rémillard, and D. Beaudoin.
Goodness-of-fit tests for copulas : A review and a power study.
Insurance Mathematics and Economics, 44 :199–213, 2009.

O. Scaillet.
Kernel-based goodness-of-fit tests for copulas with fixed smoothing parameters.
Journal of Multivariate Analysis, 98 :533–543, 2007.