C. GOF TESTS FOR COPULA

C.I. Pitfalls

Dangers in adapting standard procedures developed for classical parametric distributions to the case of copulas

C.II. GOF tests

Tests based on the integrated square difference between a kernel estimate of the copula density and the estimated parametric copula density.

C.III. Bootstrap procedures

Nonparametric, parametric and semiparametric bootstrap procedures

C.IV. Simulations from copula models

C.V. Monte Carlo results

C.VI. Conclusions

C.I. Pitfalls

One of the main issue with copulas is to choose the "best" one, namely the copula that provides the best fit with the data set at hand.

The choice among possible copula specifications can be done rigorously via so-called goodness-of-fit (GOF) tests.

Most tests developed in the standard case of a cdf are based on some comparison between the empirical cdf (or another nonparametric estimate) and the estimated parametric model.

Since the copula is the cumulative distribution function of $(u_1,...,u_n)$, one might think of designing testing procedures with the empirical copula substituted for the empirical cdf, and applies the same testing procedure as in the standard case.

The difficulty comes from the fact that the univariate cumulative distribution functions $(F_1,...,F_n)$ which are needed to map the observations $Y_t = (Y_{1t},...,Y_{nt})$ towards the unit cube via $(F_1(Y_{1t}),...,F_n(Y_{nt}))$ are *unknown*.

This means that we need to use estimates $(\hat{F}_1,...,\hat{F}_n)$ of $(F_1,...,F_n)$, and work with the pseudo-observations $(\hat{u}_{1t},...,\hat{u}_{nt}) = (\hat{F}_1(Y_{1t}),...,\hat{F}_n(Y_{nt}))$ instead of the unavailable "observations" $(u_{1t},...,u_{nt}) = (F_1(Y_{1t}),...,F_n(Y_{nt}))$.

$(T \times \hat{u}_{1t},...,T \times \hat{u}_{nt})$ are the ranks.

The use of the first step estimator $(\hat{F}_1,...,\hat{F}_n)$ affects the distributional properties, and destroys the asymptotic properties of the standard test.

Example:

multidimensional chi-square tests based on some disjoint subsets $(A_1,...,A_p)$ in $\Re^n$ and the test statistic

$$\chi^2 = T\sum_{k=1}^{p}\frac{\left(P_T\left(Y\in A_k\right)-P_\theta\left(Y\in A_k\right)\right)^2}{P_\theta\left(Y\in A_k\right)}$$

tends in law towards a chi-square distribution in the standard case (even if θ is estimated) .

This chi-square testing procedure will not work anymore, after replacing the unknown marginal cumulative distribution functions by their empirical counterparts, and apply the test on the pseudo-observations.

C.II. GOF tests

In order to get well-defined testing procedures, we can use a GOF test based on the integrated square difference between a kernel estimate of the copula density and the estimated parametric copula density.

Recall that the copula density c of the copula C is such that

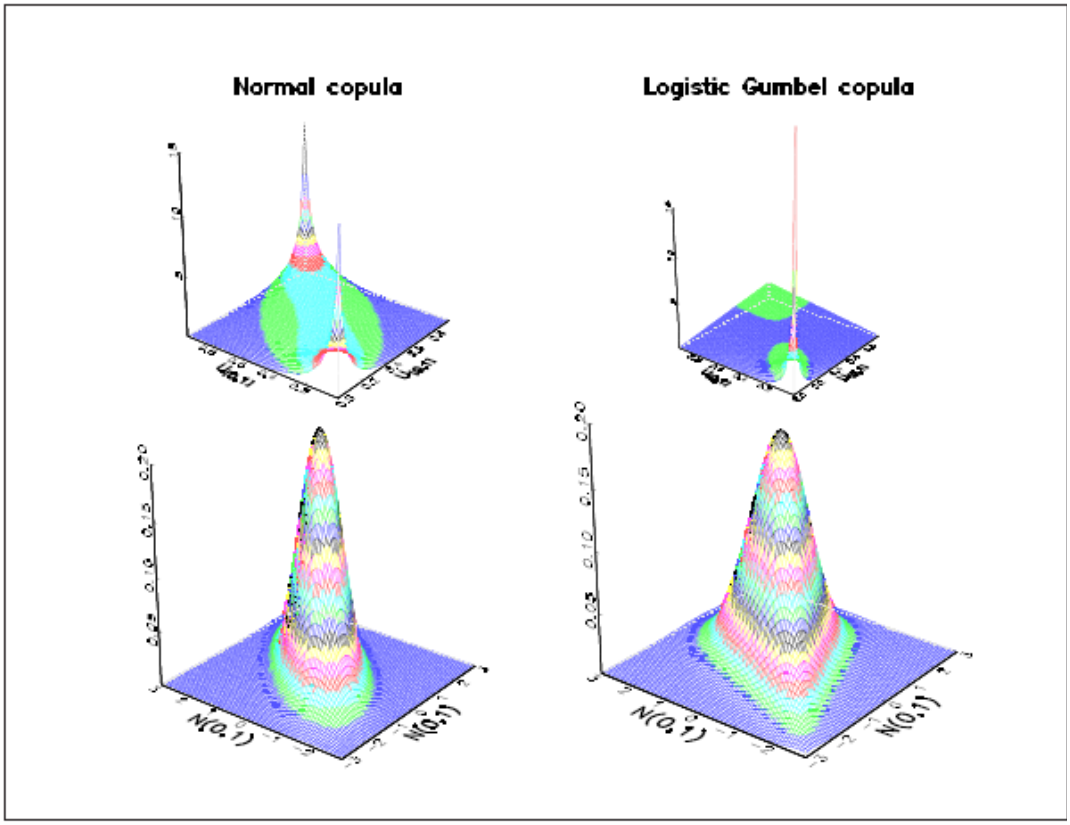
$$f(y_1, y_2) = c(F_1(y_1), F_2(y_2)) f(y_1) f(y_2)$$

or equivalently

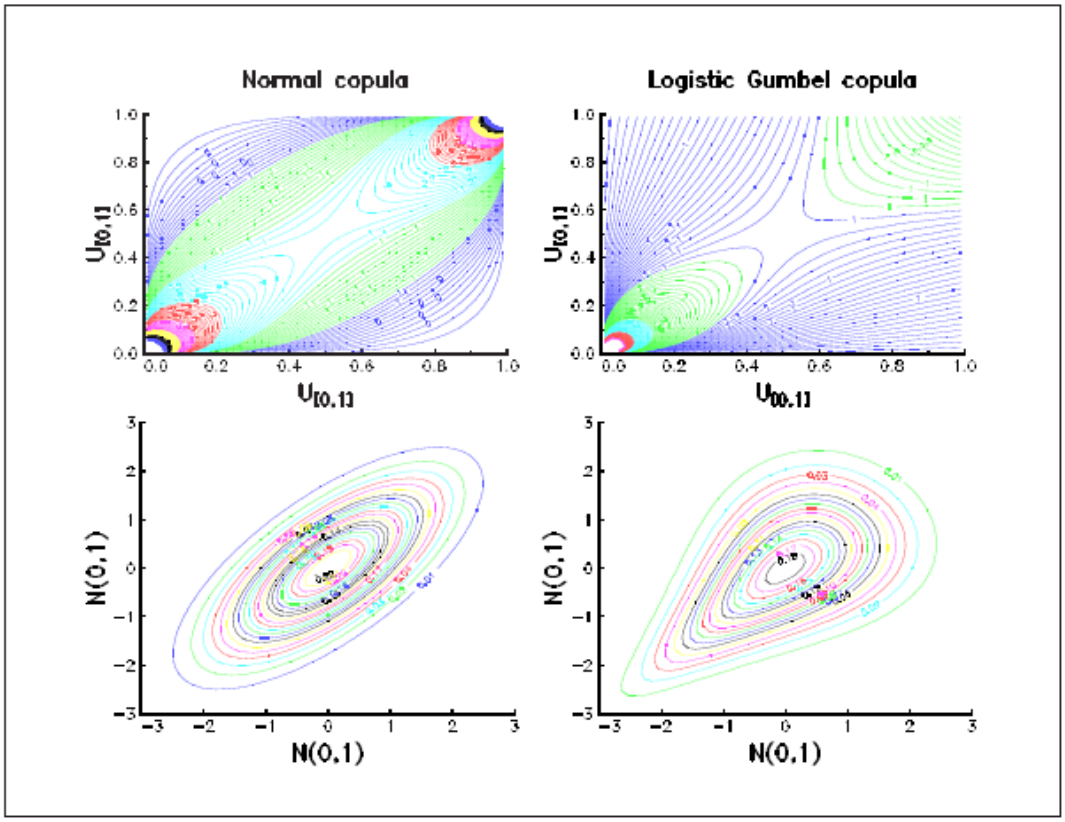
$$c(u_1, u_2) = \partial^2 C(u_1, u_2)$$

The density is defined on the unit square.

It is often more informative to visualize the copula density than the copula itself.



Graphique 1.6. Densité des copules Logistic Gumbel et Normale



Graphique 1.7. Courbes de niveau des densités des copules Logistic Gumbel et Normale

In order to estimate the copula density we can rely on a nonparametric approach based on kernel smoothing of the pseudo-observations :

$$(\hat{u}_{1t},...,\hat{u}_{nt}) = (\hat{F}_1(Y_{1t}),...,\hat{F}_n(Y_{nt})).$$

The kernel estimator of the copula density at point $u = (u_1,...,u_n)$ is simply :

$$\hat{c}(u) = \frac{1}{T}\sum_{t=1}^{T} K(\hat{u}_t - u; h)$$

where $$K(x;h) = \prod_{j=1}^{n} K_j(x_j / h_j).$$

We can then build a GOF test statistic for the parametric family $C(u;\theta)$ with density $c(u;\theta)$ based on the integrated square difference between a kernel estimate of the copula density and the estimated parametric copula density.

We get:

$$\hat{J}(w) = \int \left[\hat{c}(u) - K * c(u;\hat{\theta})\right]^2 w(u)du,$$

where * denotes convolution and w is a weight function.

The use of

$$K * c(u;\hat{\theta}) = \int K(y;h)c(u-y;\hat{\theta})dy$$

instead of $c(u;\hat{\theta})$ itself allows reducing the asymptotic bias of the test statistic.

When the bandwidth goes to zero it is possible to show that the test statistic is asymptotically normally distributed.

However the performance of the rejection rules based on the asymptotic distribution may be poor in several cases.

When the bandwidth is kept fixed (is not assumed to vanish when the sample size goes to infinity) it is possible to show that the test statistic still yields a consistent test when the weight function is set equal to one.

This means that no matter the choice of the bandwidth the limit of $\hat{J}(1)$ is such that $J \geq 0$ and $J = 0$ if the copula function is well-specified.

However the asymptotic distribution is not available in a tractable way, and we need to rely on simulation based methods to compute the rejection sets (*p*-values).

These methods are known under the name of resampling procedures since we resample from the original data.

One of these methods is called the bootstrap.

C.III. Bootstrap procedures

Asymptotic properties of estimators are valid when sample size is large,

i.e. when $T \to \infty$.

In small samples (finite distance), asymptotic properties may provide poor approximations of the real distribution of estimators.

In particular confidence intervals based on asymptotic normality may be too wide or too narrow and exhibit a wrong coverage probability.

We may use simulations to get a better approximation when samples are small.

The idea is to draw from the original data new observations in order to generate new fictitious samples which mimic the behavior of the original observed sample.

Nonparametric bootstrap procedures:

1) start from initial data $Y_1,...,Y_T$, and compute the empirical cdf.

2) derive ST independent drawings by sampling randomly in the initial data with replacement $Y_1^s,...,Y_T^s$, $s = 1,...,S$, each sample is called a bootstrap sample

(draw ST realizations of a uniform $[0,1]$ variable, and invert the empirical cdf).

3) for each simulated sample $s = 1,...,S$ of length T, compute the estimate $\hat{\theta}^s$, for example the empirical mean

$$\hat{\theta}^s = \hat{m}_s = \frac{1}{T}\sum_{t=1}^{T} Y_t^s .$$

4) the empirical distribution of the estimates $\hat{\theta}^s$, $s = 1,...,S$ will constitute a good approximation (consistent estimator when the number S of simulated sample goes to infinity) for the true distribution of the estimator in small sample

Pseudo random generators are available in most softwares.

Since we draw new observations from the empirical (nonparametric) cdf, this is called a *nonparametric* bootstrap.

Parametric bootstrap procedures:

If we postulate a parametric model for the data we may also draw from the parametric distribution once its parameter has been estimated.

This is called a *parametric* bootstrap.

Example:

If we estimate the parameter β of a regression under the assumption of normal error terms, we can generate a parametric bootstrap sample as:

$$Y_t^s = X_t \hat{\beta} + e_t^s, \qquad t = 1,\ldots,T,$$

where the error terms e_t^s are drawn from a $N(0, s^2)$, with s^2 being the estimated variance of the innovations.

Semiparametric bootstrap procedures:

The *semiparametric* bootstrap is a mixture of the two previous procedures.

One part of the model is parametric, and the other part is nonparametric.

Example:

If we estimate the parameter β of a regression, we can generate a parametric bootstrap sample as:

$$Y_t^s = X_t\hat{\beta} + e_t^s, \qquad t = 1,\ldots,T,$$

where the error terms e_t^s are drawn with replacement from the estimated residuals $e_t = Y_t - X_t\hat{\beta}$, $t = 1,\ldots,T$.

C.IV. Simulations from copula models

The problem we address here is the simulations of a random vector $Y = (Y_1,...,Y_d)$, whose distribution is characterized by

$$F(y_1,...,y_d) = C(F_1(y_1),...,F_d(y_d))$$

The problem consists :

first in simulating the random vector $U = (U_1,...,U_d)$, whose distribution is the copula C,

and then in using the transformation

$$X = (F_1^{-1}(U_1),...,F_d^{-1}(U_d))$$

We will thus concentrate in the following on methods which allow simulating $U = (U_1,...,U_d)$ from the associated copula.

a) Distribution method:

This is the mirror method of the one presented above.

We have

$$C(u_1,...,u_d) = F(F_1^{-1}(u_1),...,F_d^{-1}(u_d))$$

Hence if we can easily draw X from F, we can apply the transformation

$$U = (F_1(X_1),...,F_d(X_d)).$$

Examples :

1. Normal copula with correlation matrix ρ

Using the cholesky decomposition $\rho = PP'$, we can easily draw $X = P\eta$ where η is the stack of random $N(0,1)$, and compute $U_i = \Phi(X_i)$ where Φ is the cdf of a *N(0,1)*.

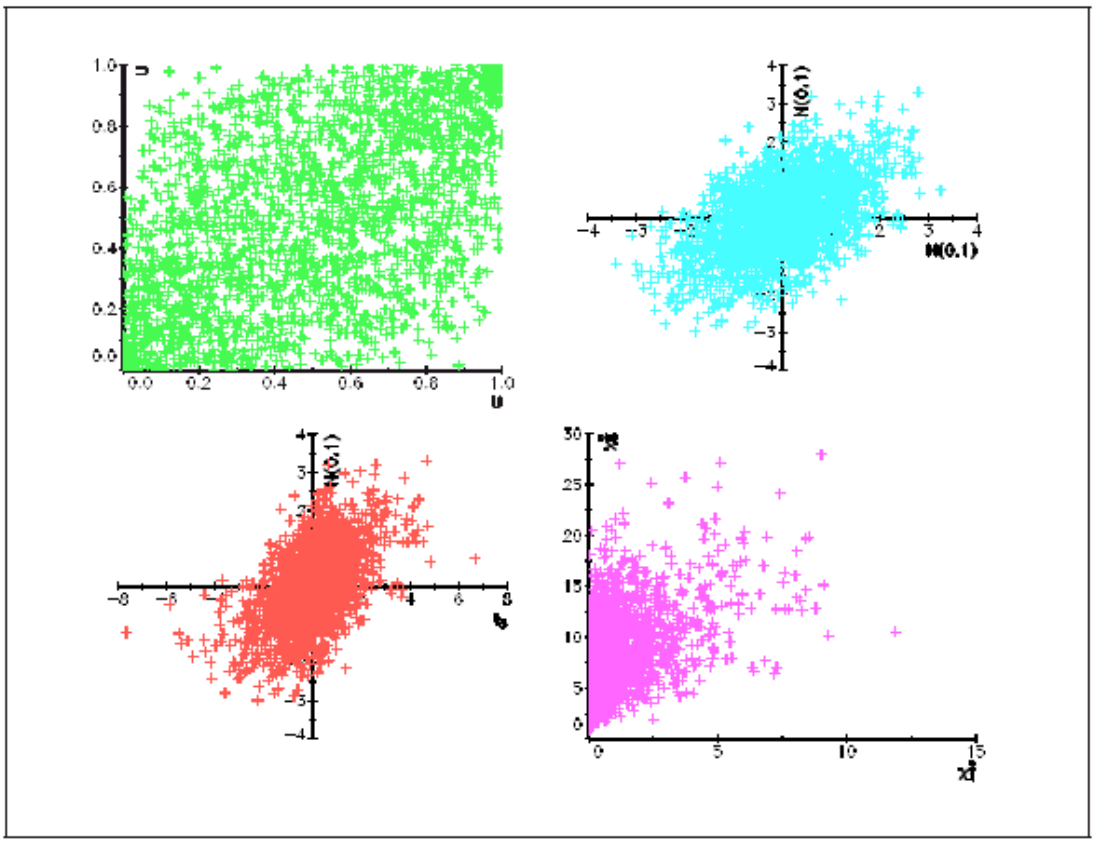
Remark: Box Muller method for normal

if u_1 and u_2 are independent random variates from $U[0,1]$, then the variates

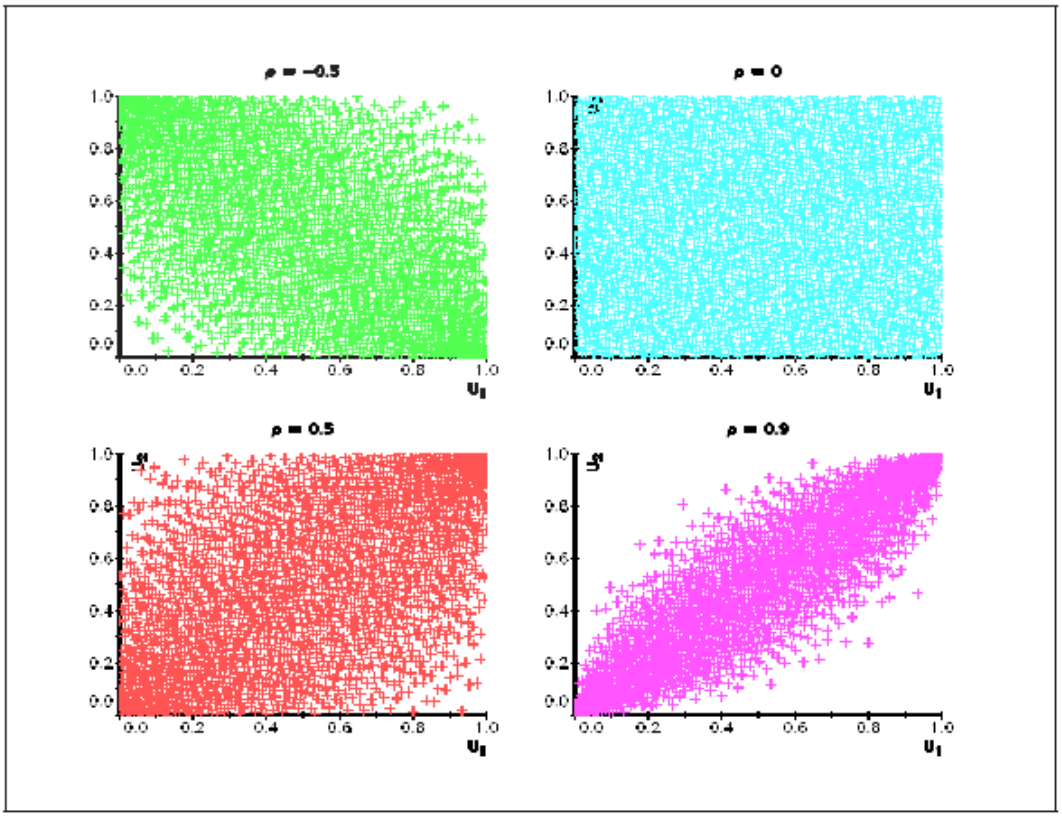
$$\eta_1 = (-2\ln(u_1))^{1/2}\cos(2\pi u_2),$$

$$\eta_2 = (-2\ln(u_1))^{1/2}\sin(2\pi u_2),$$

are independent random variates from $N(0,1)$.



Graphique 6.1. Simulation de 4 distributions (générateur LCG)



Graphique 6.3. Simulation de la copule Normale

2. Student copula with correlation matrix ρ and number ν of degrees of freedom

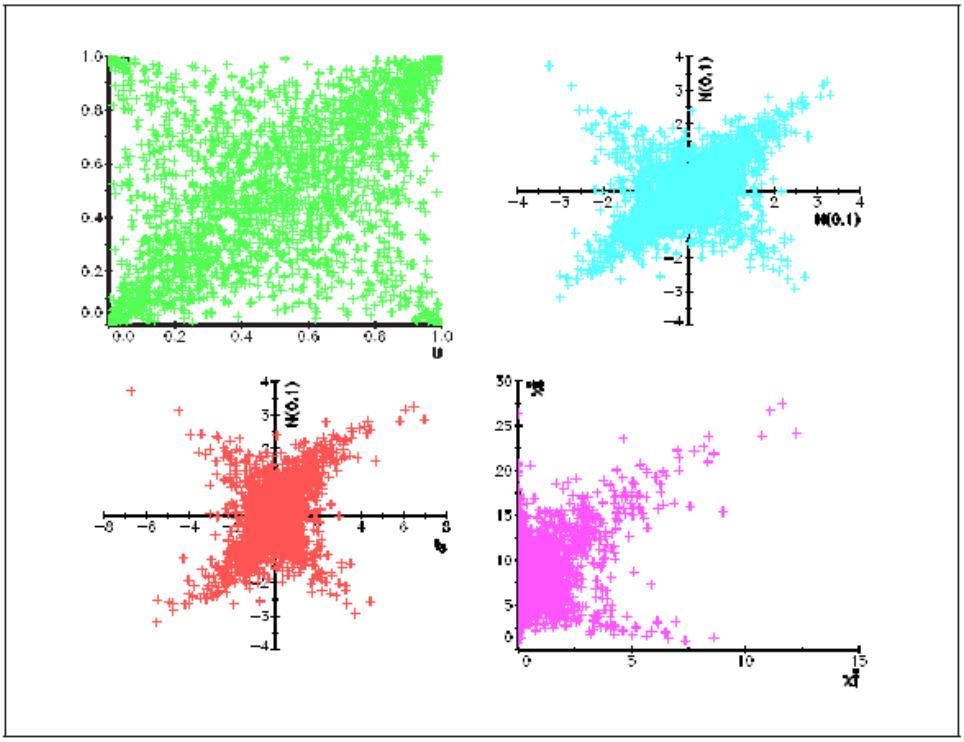
Let X be a random vector whose distribution is a multivariate $t_{\rho,\nu}$.

Then we have : $$X = \frac{X^*}{\sqrt{\chi_\nu^2 / \nu}}$$

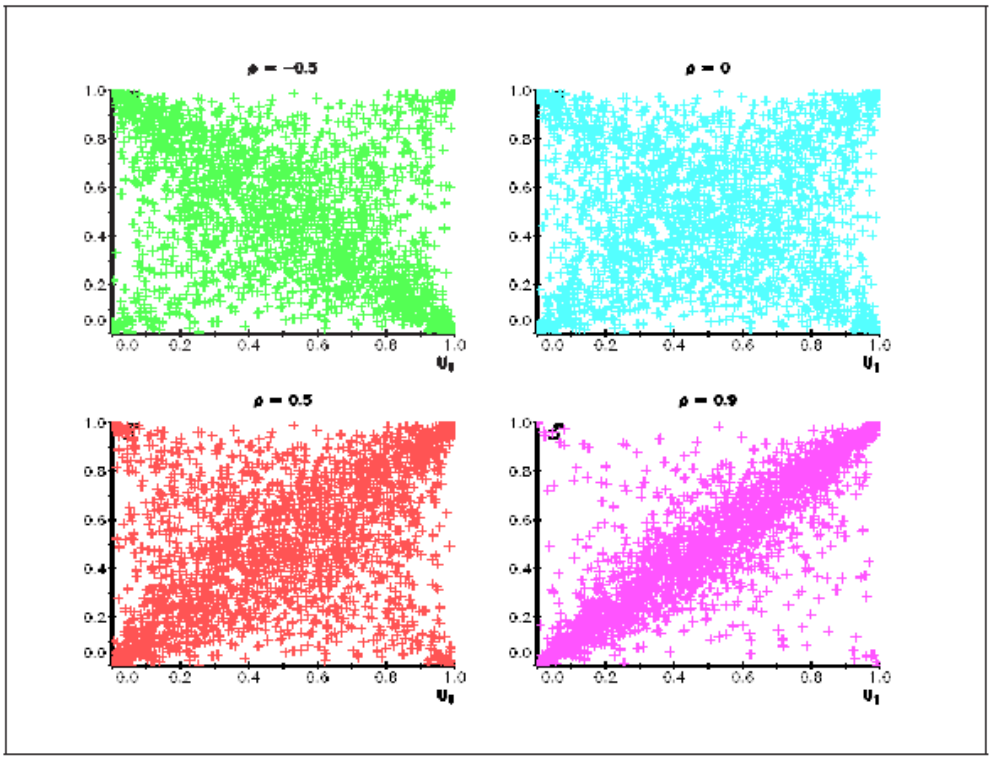
with X^* a Gaussian vector with correlation matrix ρ and χ_ν^2 a chi-square random variable with number ν of degrees of freedom.

From X, we can compute $U_i = t_\nu(X_i)$ where t_ν is the cdf of a student with number ν of degrees of freedom.

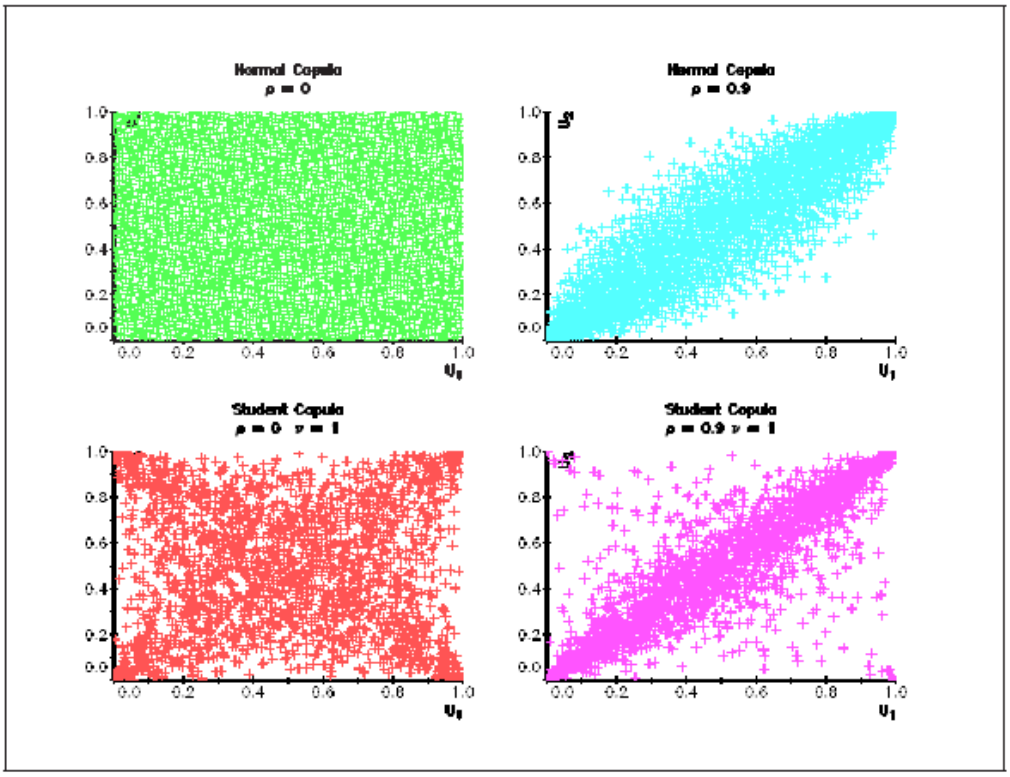
The student copula yields further dependence in the tails, and does not correspond to the independent copula even if the correlation is zero.



Graphique 6.5. Simulation de 4 distributions avec une copule Student ($\rho = 0.5$, $\nu = 1$)



Graphique 6.4. Simulation de la copule Student ($\nu = 1$)



Graphique 6.6. Comparaison des copules Normale et Student

b) Conditional method:

Let us consider the bivariate case. Let $U = (U_1, U_2)$ be a random vector whose distribution is C.

We know that

$$P[U_1 \leq u_1] = C(u_1, 1) = u_1$$

and that

$$P[U_2 \leq u_2 | U_1 = u_1]$$
$$= \partial_1 C(u_1, u_2) = C_{2|1}(u_1, u_2)$$

Since $C(U_1, 1)$ and $C_{2|1}(u_1, U_2)$ are two uniform random variables, we get the following algorithm:

a. Simulate two uniforms v_1 and v_2.

b. Take $u_1 = v_1$.

c. Take $u_2 = C_{2|1}^{-1}(u_1, v_2)$.

Example :

Frank copula :

$$C(u_1,u_2;\theta) = -\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{(e^{-\theta}-1)}\right)$$

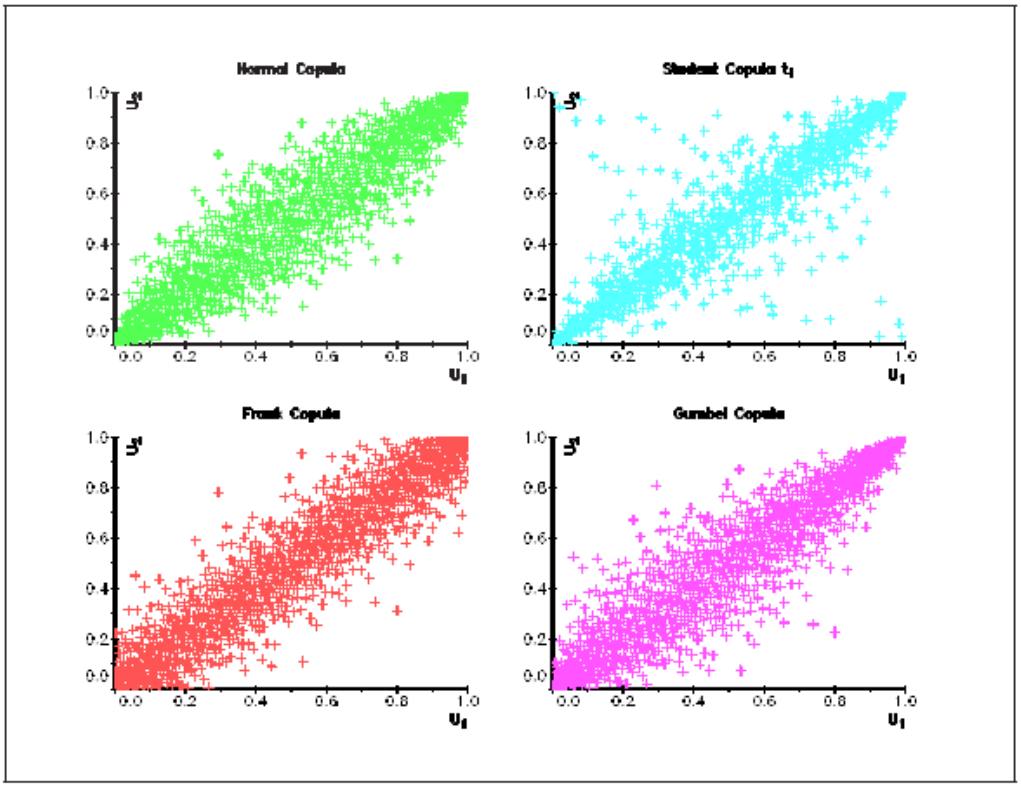
we have

$$C_{2|1}^{-1}(u_1,v_2) = -\frac{1}{\theta}\ln\left(1+\frac{v_2(e^{-\theta}-1)}{v_2+(1-v_2)e^{-\theta u_1}}\right)$$

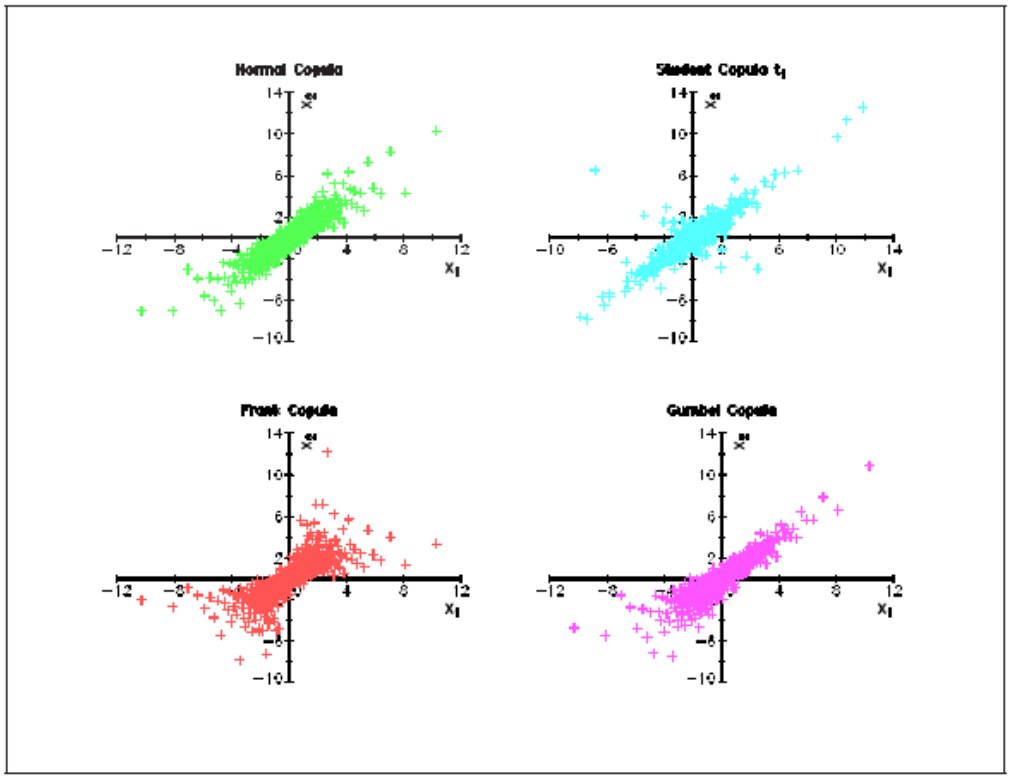
Remark:

When $C_{2|1}^{-1}(u_1,v_2)$ is not available in closed-form we may use a numerical method to solve $C_{2|1}(u_1,u_2)=v_2$.

This is the case for the Gumbel copula.



Graphique 6.7. Simulation des copules Normale, Student, Frank et Gumbel (marges uniformes)



Graphique 6.8. Simulation des copules Normale, Student, Frank et Gumbel (marges t_3)

C.V. Monte Carlo results

In order to implement a testing procedure based on

$$\hat{J}(1) = \int \left[\hat{c}(u) - K * c(u;\hat{\theta}) \right]^2 du$$

we can rely on a semiparametric bootstrap.

First we draw from the estimated copula $C(u;\hat{\theta})$ in order to impose the dependence structure of the null hypothesis (well-specified copula), and then we use the inverse of the empirical margins $\hat{F}_j$ to get the bootstrap sample.

The first step is parametric; the second step is nonparametric.

Monte Carlo results:

200 MC simulations, 500 bootstrap samples, 50 and 200 observations.

Choice of bandwidth:
δh with $\delta = .1,.25,.5,1,1.5$, and h given by the rule of thumb.

Size (type I error):
True copula: Frank with parameter values: $\theta = 1,2,3$ $(\tau = .11,.21,.31)$
True margins: exponential

Power (type II error):
we contaminate the sample with 50% of the observations (mixture) coming from a student copula with 4 degrees of freedom and a correlation parameter of 0.95.

Comparison with asymptotic testing procedures and bootstrap methods based on a standardized test statistics (divide by the standard deviation).

TABLE I: Impact of bandwidth choice on size

	$n = 50$					$n = 200$				
F: $\theta = 1$	.1	.25	.5	1	1.5	.1	.25	.5	1	1.5
Asym.	.44	.00	.00	.00	.00	.51	.06	.00	.00	.00
As. Boot.	.00	.02	.04	.03	.00	.05	.04	.06	.05	.05
Boot.	.00	.02	.04	.04	.03	.05	.04	.06	.05	.05
F: $\theta = 2$	.1	.25	.5	1	1.5	.1	.25	.5	1	1.5
Asym.	.50	.00	.00	.00	.00	.51	.06	.01	.00	.00
As. Boot.	.01	.04	.04	.02	.00	.06	.05	.05	.05	.04
Boot.	.01	.04	.05	.03	.03	.06	.05	.06	.05	.05
F: $\theta = 3$	.1	.25	.5	1	1.5	.1	.25	.5	1	1.5
Asym.	.51	.01	.00	.00	.00	.48	.04	.01	.00	.00
As. Boot.	.01	.02	.03	.02	.00	.05	.04	.05	.05	.05
Boot.	.01	.02	.04	.05	.02	.04	.03	.05	.05	.05

TABLE II: Impact of bandwidth choice on power

	$n = 50$					$n = 200$				
F: $\theta = 1$	.1	.25	.5	1	1.5	.1	.25	.5	1	1.5
Asym.	.11	.05	.02	.00	.00	.65	.81	.80	.03	.00
As. Boot.	.09	.17	.27	.07	.06	.13	.77	.95	.77	.19
Boot.	.10	.18	.27	.21	.20	.13	.77	.95	.73	.26
F: $\theta = 2$	.1	.25	.5	1	1.5	.1	.25	.5	1	1.5
Asym.	.11	.01	.00	.00	.00	.48	.74	.56	.01	.00
As. Boot.	.00	.12	.17	.04	.03	.21	.90	.90	.51	.11
Boot.	.00	.10	.14	.16	.11	.22	.90	.90	.49	.17
F: $\theta = 3$	.1	.25	.5	1	1.5	.1	.25	.5	1	1.5
Asym.	.10	.01	.00	.00	.00	.54	.52	.27	.00	.00
As. Boot.	.00	.06	.11	.02	.00	.13	.69	.75	.25	.06
Boot.	.00	.06	.11	.09	.08	.13	.67	.72	.24	.13

C.VI. Conclusions

When working with copulas, observations are pseudo-observations since we transform into the unit cube via the margins.

In fact we work with ranks of the observations instead of the observations themselves.

Adaptation of standard testing procedures should be conducted with care since asymptotic distribution are affected by the first step transformation.

Bootstrap procedures may alleviate the burden caused by the complexity induced by the use of pseudo-observations.