# Wie kann man Abhängigkeiten zwischen diskreten und gemischten Risiken aufdecken?

Johanna G. Nešlehová

in collaboration with Chr. Genest, O. A. Murphy and B. Rémillard

McGill University, Montréal, Canada

9. April 2013

# Outline

A. Prelude

B. Inferential issues

C. The empirical checkerboard copula

D. Tests of independence

E. The empirical checkerboard copula process

F. References

# The horseshoe crab data

Data on $n = 173$ female crabs having a male attached in their nest:

$X_1$ : Colour of the female (4 categories)
$X_2$ : Spine condition (3 categories)
$X_3$ : Carapace width (cm)
$X_4$ : Number of satellites, i.e., other males around the female
$X_5$ : Weight (kg)

Source:
Brockmann (1996)
*Ethology* **102**: 1–21.

# Let's build a copula model

Choose $d$ univariate distributions from (parametric) classes and assume that

$$F_1 \in \mathcal{F}_1, \quad \ldots, \quad F_d \in \mathcal{F}_d.$$

Choose a copula

$$C \in \{C_\theta : \theta \in \Theta\}$$

and construct a joint distribution of $(X_1, \ldots, X_d)$ via

$$H(x_1, \ldots, x_d) = C_\theta\{F_1(x_1), \ldots, F_d(x_d)\}.$$

# Issues of interest

(1) How can one test whether $X_1, \ldots, X_d$ are dependent?

(2) If there is dependence, how can it be modeled and estimated?

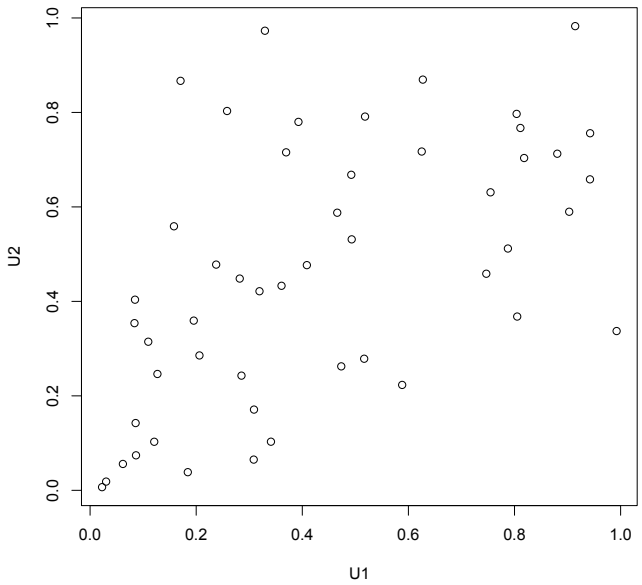
(3) How can the model be validated?

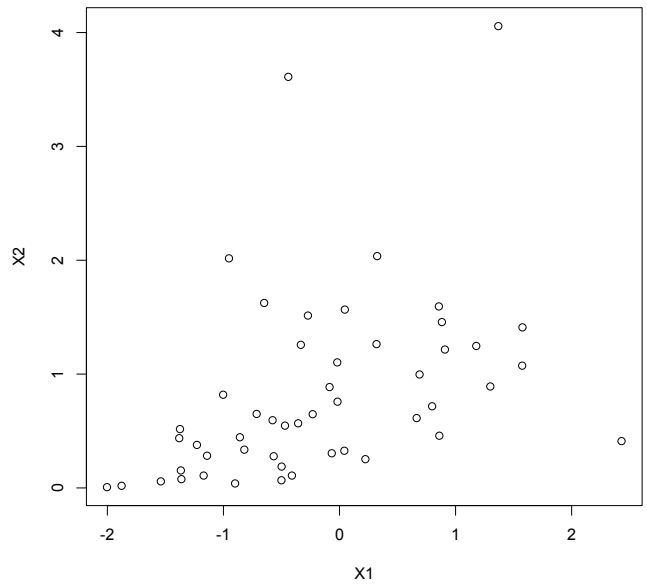
When at least one margin has atoms, these issues become complex and not fully understood.

This talk will be concerned with (1).

# Illustration of a copula model in the continuous case

Take $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{E}(1)$ and $C$ to be Clayton(2).



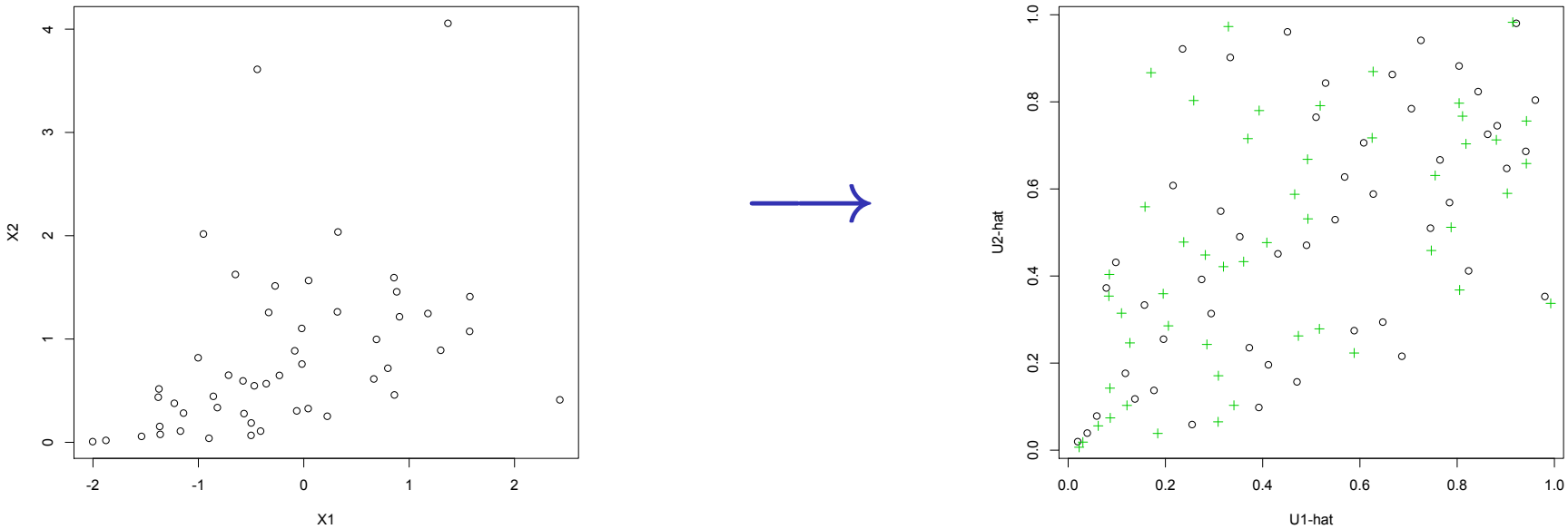

$$(U_1, U_2) \longrightarrow (F_1^{-1}(U_1), F_2^{-1}(U_2))$$

$$(F_1(X_1), F_2(X_2)) \longleftarrow (X_1, X_2)$$

# In reality, the margins are often unknown ...

No problem! Just estimate them and set

$$\hat{U}_{ij} = F_{nj}(X_{ij}) = \frac{1}{n}\sum_{k=1}^{n} 1(X_{kj} \leq X_{ij}) = \frac{R_{ij}}{n}, \quad i \in \{1, \ldots, n\}, \quad j \in \{1, 2\}.$$



# Road to inference in the continuous case

In the continuous case, the copula is unique and invariant by strictly increasing transformations of the margins.

Inference on $\theta$ can thus be based on the maximally invariant statistics, i.e., the normalized ranks

$$\left(\frac{R_{11}}{n}, \frac{R_{12}}{n}\right), \ldots, \left(\frac{R_{n1}}{n}, \frac{R_{n2}}{n}\right).$$

# Most popular approaches to estimation

- ✓ Maximize the log pseudo-likelihood as per Genest et al. (1995):

$$\ell(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln[c_\theta\{F_{n1}(x_{i1}), F_{n2}(x_{i2})\}].$$

- ✓ Use a moment estimator of $\theta$, e.g.,

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

where $\tau : \Theta \to [-1, 1] : \theta \mapsto \tau(C_\theta)$ is one-to-one and

$$\tau_n = (N_c - N_d) \Big/ \binom{n}{2}.$$

# What happens in the discrete case?

Assume $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$ is a random sample from

$$H_\theta(x, y) = C_\theta\{F_1(x_1), F_2(x_2)\}$$

with $F_1$ and $F_2$ discrete.

Do the same strategies work?

- ✓ Ties occur in the data, e.g., for some $i \neq j$,

$$X_{i1} = X_{j1} \quad \text{or} \quad X_{i2} = X_{j2} \quad \text{or both.}$$

- ✓ How do we account for ties?

# Adjustment for ties, e.g., for inversion of $\tau$

Different options can be envisaged:

Option 1 (split ties): $\tau_n = (N_c - N_d) \Big/ \binom{n}{2}$

Option 2 (ignore ties): $\tau_{a,n} = (N_c - N_d)/(N_c + N_d)$

Option 3 (adjust for ties): $\tau_{b,n} = (N_c - N_d)/\sqrt{N_1 N_2}$

where

$$N_1 = \sum_{i<j} \mathbf{1}(X_{i1} \neq X_{j1}) \quad \text{and} \quad N_2 = \sum_{i<j} \mathbf{1}(X_{i2} \neq X_{j2}).$$

# A simulation experiment

Draw $10,000$ samples $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$ of size $n = 100$ from

$$H_\theta(x_1, x_2) = C_\theta\{F_1(x_1), F_2(x_2)\}\,,$$

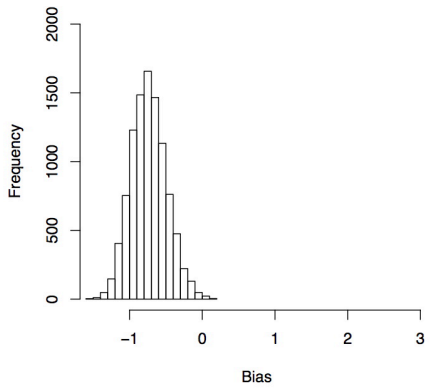
where $C_\theta$ is a Clayton copula and $F_1$, $F_2$ are discrete distributions.

Given that $\tau = \theta/(\theta + 2)$, pick $\hat{\tau} \in \{\tau_n, \tau_{a,n}, \tau_{b,n}\}$ and set

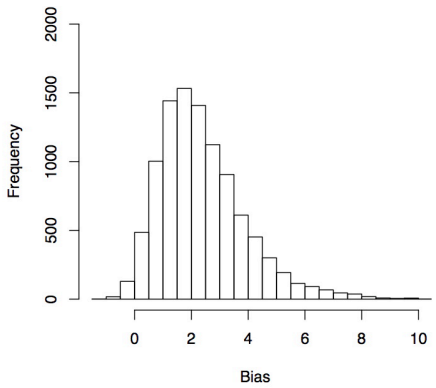
$$\hat{\theta} = 2\,\frac{\hat{\tau}}{1 - \hat{\tau}}\,.$$

# Illustration: Poisson margins

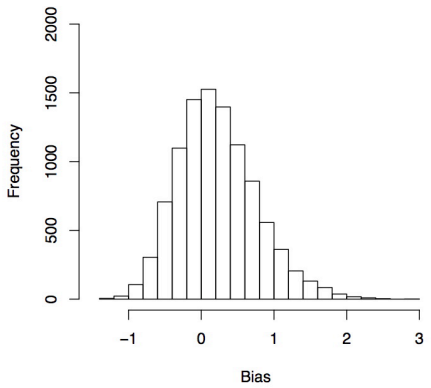
Take $\theta = 2$ and Poisson margins, e.g., $X_1 \sim \mathcal{P}(1)$, $X_2 \sim \mathcal{P}(2)$.




$\hat{\theta}$ based on $\tau_n$




$\hat{\theta}$ based on $\tau_{a,n}$



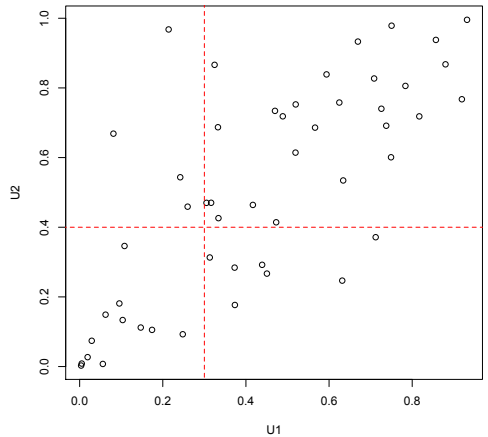

$\hat{\theta}$ based on $\tau_{b,n}$

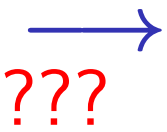
# The source of the bias

In general, $\tau_{a,n}$ and $\tau_{b,n}$ are biased estimators of $\tau(C_\theta)$ because

$$X_{i1} = F_1^{-1}(U_{i1}) \text{ and } X_{i2} = F_2^{-1}(U_{i2}) \not\Rightarrow (F_1(X_{i1}), F_2(X_{i2})) \sim C_\theta.$$

In other words, the discretization of $(U_{i1}, U_{i2})$ is irreversible. ☹



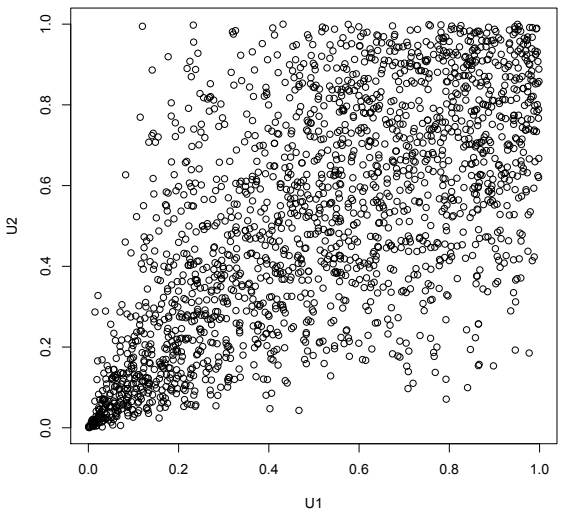

| | $X_1 = 0$ | $X_1 = 1$ |
|---|---|---|
| $X_2 = 1$ | 4 | 27 |
| $X_2 = 0$ | 12 | 7 |

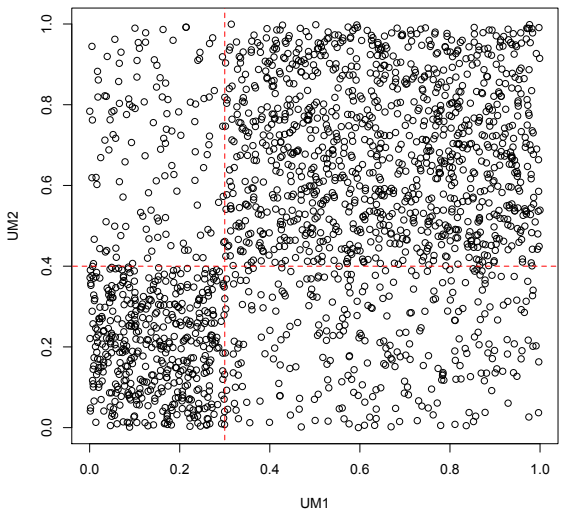
# What exactly is going on?

It can be seen that $\tau_n$ is an unbiased estimator of

$$\tau(H) = \tau(C^{\maltese}).$$

But $C^{\maltese} \neq C_\theta$ for most copula families except at independence.



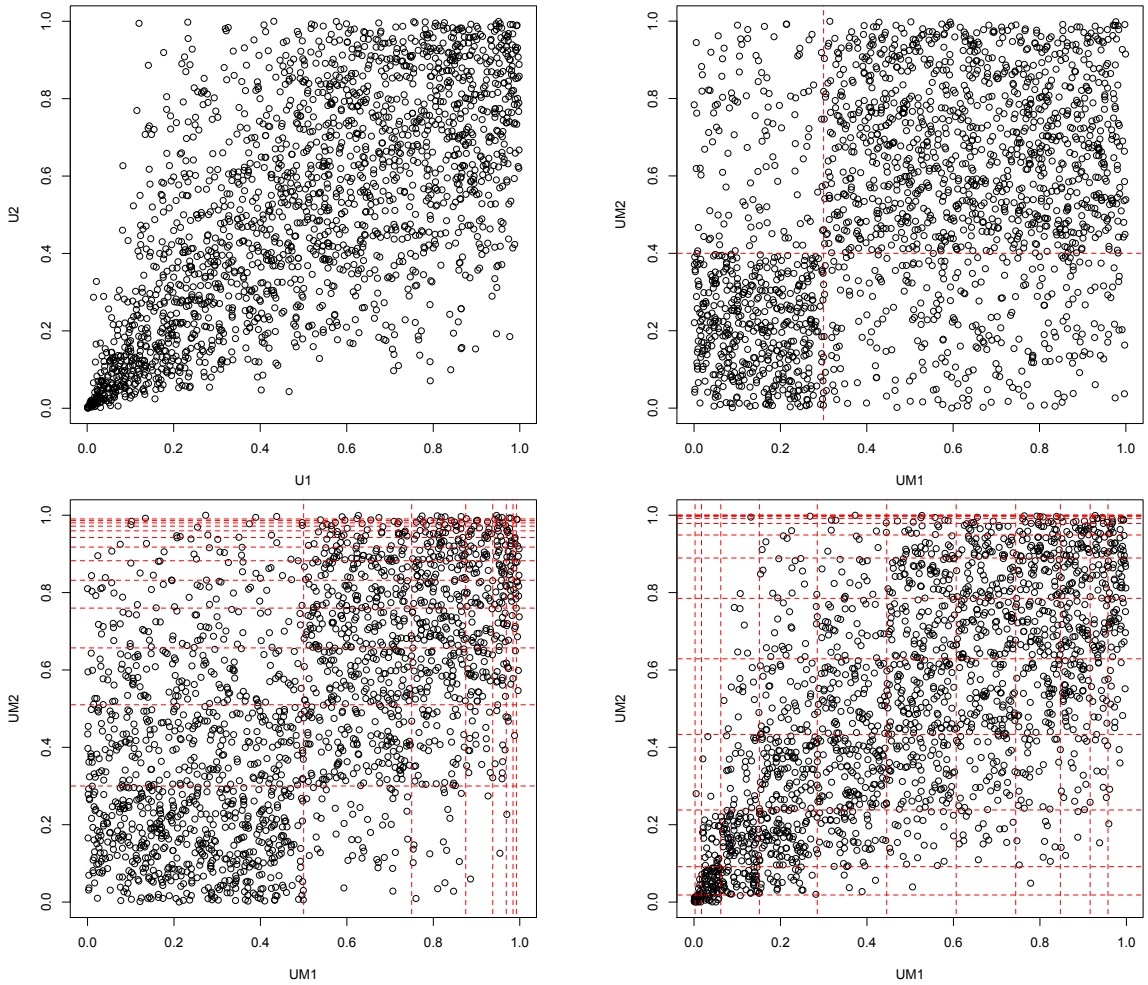
# Not everything is lost

- ✓ While the data cannot be used to estimate $C_\theta$, they provide an estimate of $C^{\maltese}$.
- ✓ $C^{\maltese}$ reflects many dependence properties of $H$. For example, $X_1$ and $X_2$ are independent if and only if $C^{\maltese} = \Pi$.
- ✓ Observe that $C^{\maltese}$ depends on $C_\theta$ as well as on the margins.
- ✓ For simplicity, assume that

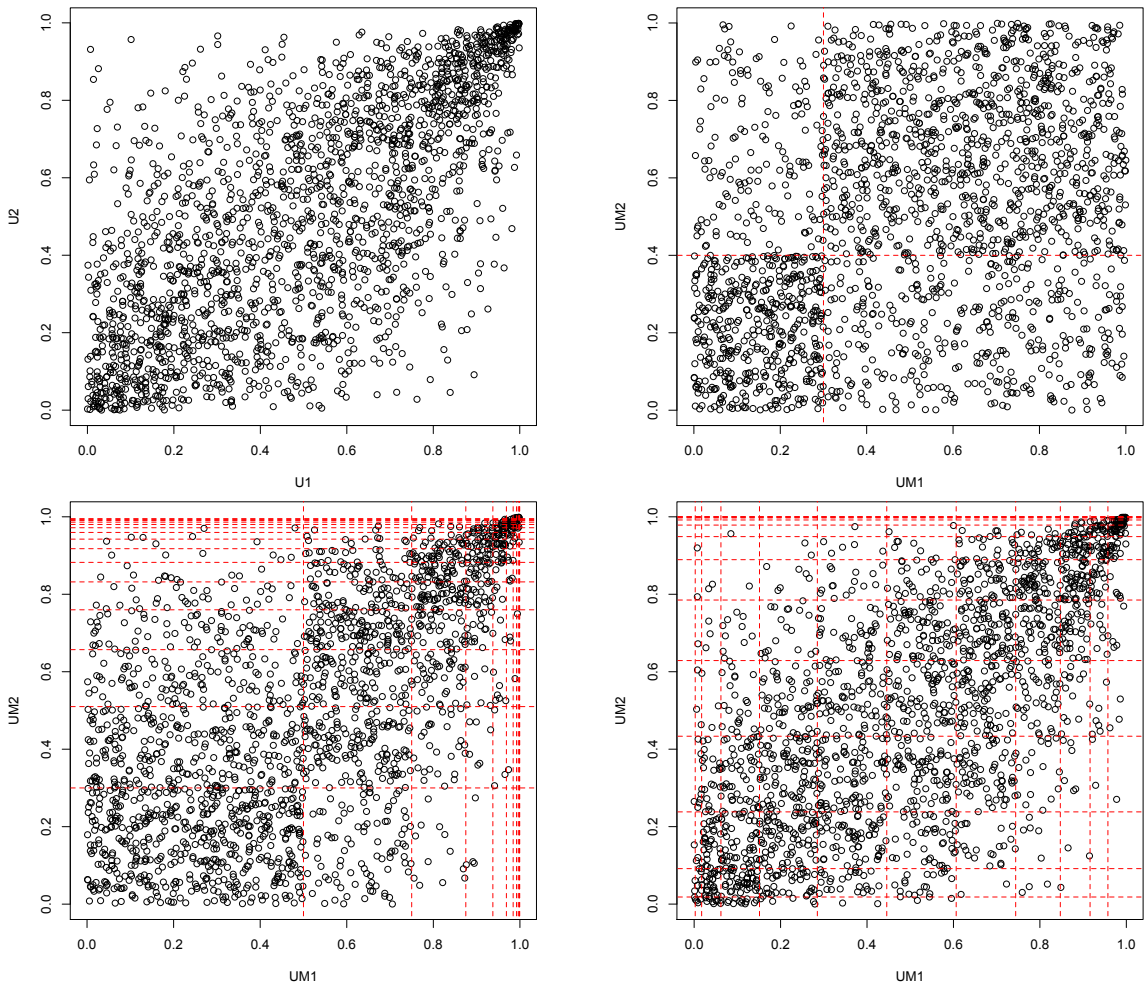
  (1) Only $d = 2$ variables are observed.

  (2) The variables are counts, i.e., supported on $\{0, 1, 2, \dots\}$.

# Clayton(2) with Bernoulli, Geometric, Poisson margins



# Gumbel(2) with Bernoulli, Geometric, Poisson margins



# Empirical checkerboard copula

- ✓ Compute the empirical cdf $H_n$ corresponding to the sample

$$(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2}).$$

- ✓ Denote its bilinear extension copula by $C_n^{✠}$.

- ✓ $C_n^{✠}$ is explicit; its density is given by

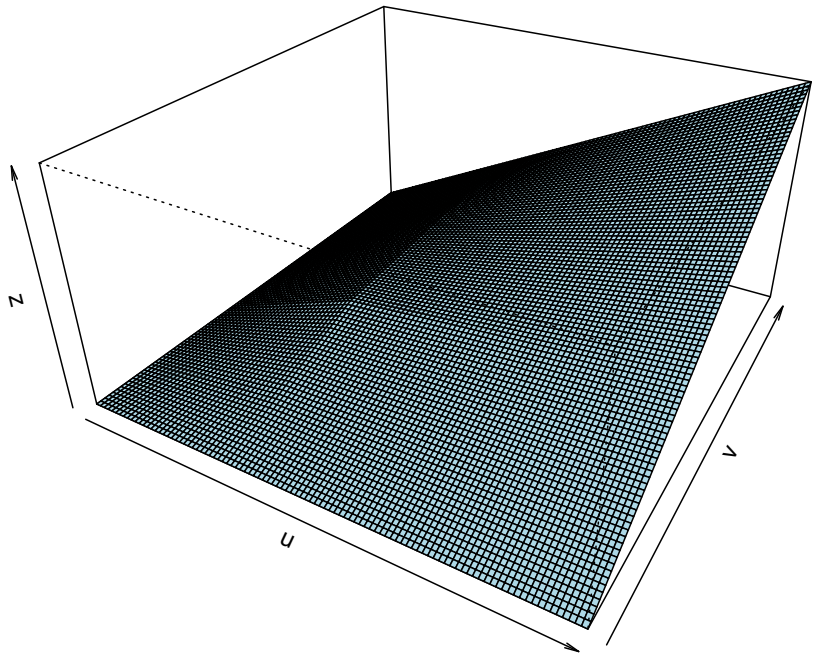
$$c_n^{✠}(u_1, u_2) = n \times \frac{n_{ij}}{n_{i\bullet} n_{\bullet j}}$$

whenever for some $i, j \in \{0, 1, 2, \ldots\}$,

$$F_{n1}(i-1) < u_1 \leq F_{n1}(i) \quad \text{and} \quad F_{n2}(j-1) < u_2 \leq F_{n2}(j).$$

# Illustration with Bernoulli data

| | $X_1 = 0$ | $X_1 = 1$ |
|---|---|---|
| $X_2 = 1$ | 4 | 27 |
| $X_2 = 0$ | 12 | 7 |



# Wait a minute...

The sample $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$ defines a contingency table.

- ✓ Pearson's chi-squared statistic for testing independence:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i\bullet} n_{\bullet j}/n)^2}{n_{i\bullet} n_{\bullet j}/n}.$$

- ✓ Spearman's midrank coefficient for testing monotone trend:

$$\rho_n^* = \frac{12}{n^3} \left\{ \sum_{i=1}^{n} (R_{i1}^* - \bar{R})(R_{i2}^* - \bar{R}) \right\}.$$

# Surprise!

It can be seen that

$$\chi^2 = n \int_0^1 \int_0^1 \{c_n^{\maltese}(u_1, u_2) - 1\}^2 \mathrm{d}\Pi(u_1, u_2),$$

$$\rho_n^* = 12 \int_0^1 \int_0^1 \{C_n^{\maltese}(u_1, u_2) - u_1 u_2\} \mathrm{d}\Pi(u_1, u_2),$$

$$\tau_n^* = -1 + 4 \int_0^1 \int_0^1 C_n^{\maltese}(u_1, u_2)\, \mathrm{d}C_n^{\maltese}(u_1, u_2).$$

# New and consistent tests of independence

$X_1$ and $X_2$ are independent if and only if for all $u_1, u_2 \in [0,1]$,

$$C^{\Pi}(u_1, u_2) = u_1 \times u_2.$$

The hypothesis $H_0$ of independence can thus be tested using

$$S_{n1} = n \int_0^1 \cdots \int_0^1 \{C_n^{\Pi}(u_1, u_2) - u_1 u_2\}^2 \mathrm{d}u_1 \mathrm{d}u_2$$

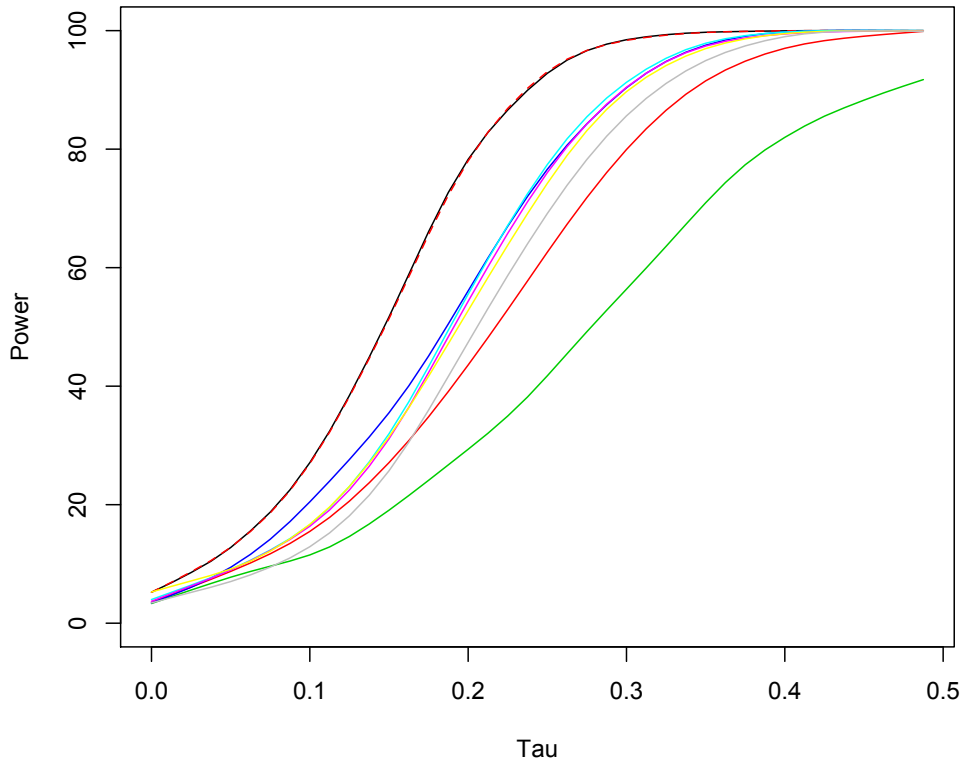
$$S_{n2} = n \int_0^1 \cdots \int_0^1 \{C_n^{\Pi}(u_1, u_2) - u_1 u_2\}^2 \mathrm{d}C_n^{\Pi}(u_1, u_2),$$

$$T_n = \sqrt{n} \times \sup_{u_1, u_2 \in [0,1]} |C_n^{\Pi}(u_1, u_2) - u_1 u_2|$$

# A little power study for the start

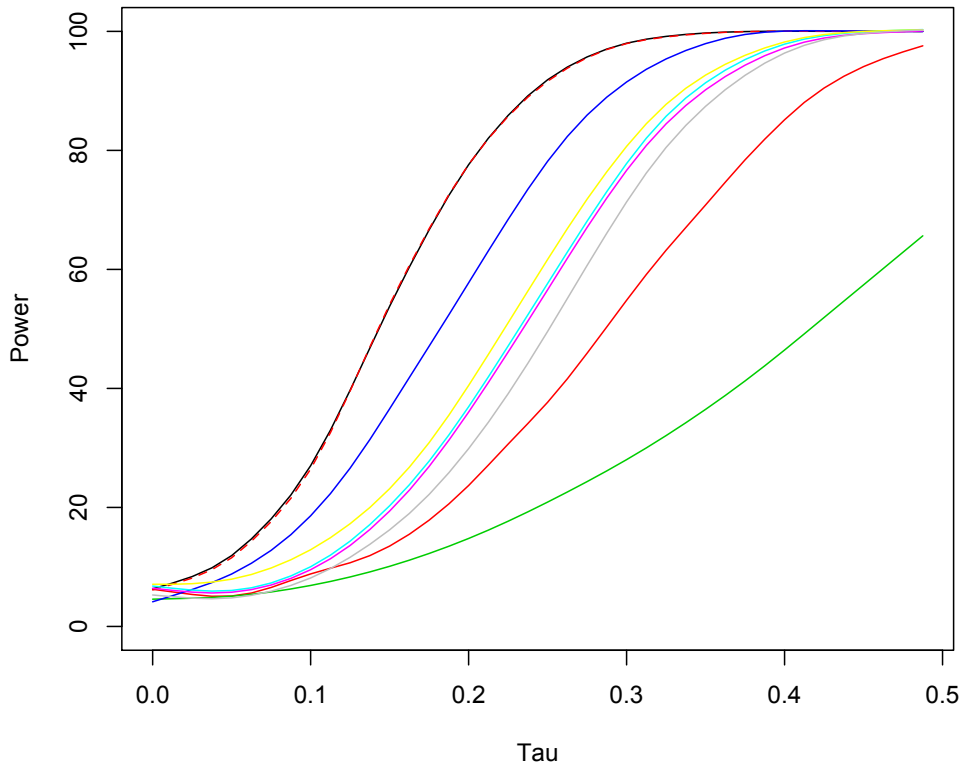
- ✓ Take the level to be $\alpha = 5\%$.
- ✓ Take the sample size to be $n = 100$.
- ✓ Take copula alternatives (FGM, Clayton, Gauss, Frank, Gumbel).
- ✓ Compare $S_{n1}$ to $S_{n2}$ (dashed), $T_n$, Pearson's $\chi^2$, Pearson's $\chi^2$ with MC, Likelihood ratio statistic and Zelterman's statistic.
- ✓ Compute approximate critical values for $S_{n1}$, $S_{n2}$ and $T_n$ using MC (assuming known margins).
- ✓ Make 10,000 repetitions.

# Examples



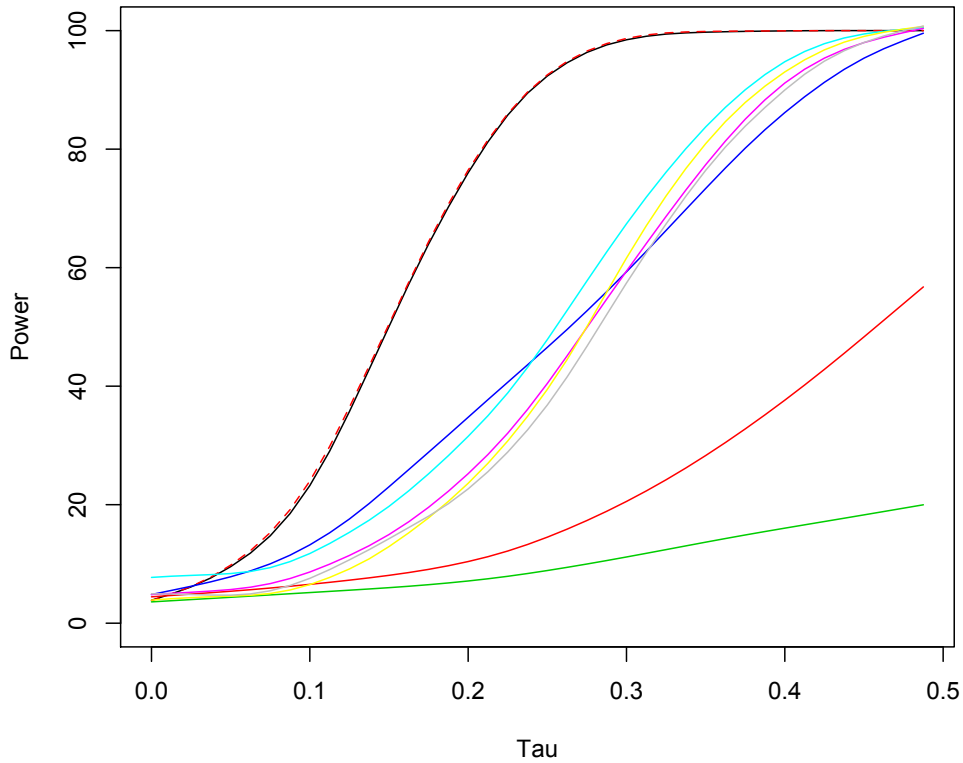

$X_1 \sim \text{Bin}(4, 0.5)$, $X_2 \sim \text{Bin}(4, 0.5)$ and $C$ is Clayton

# Examples (cont'd)



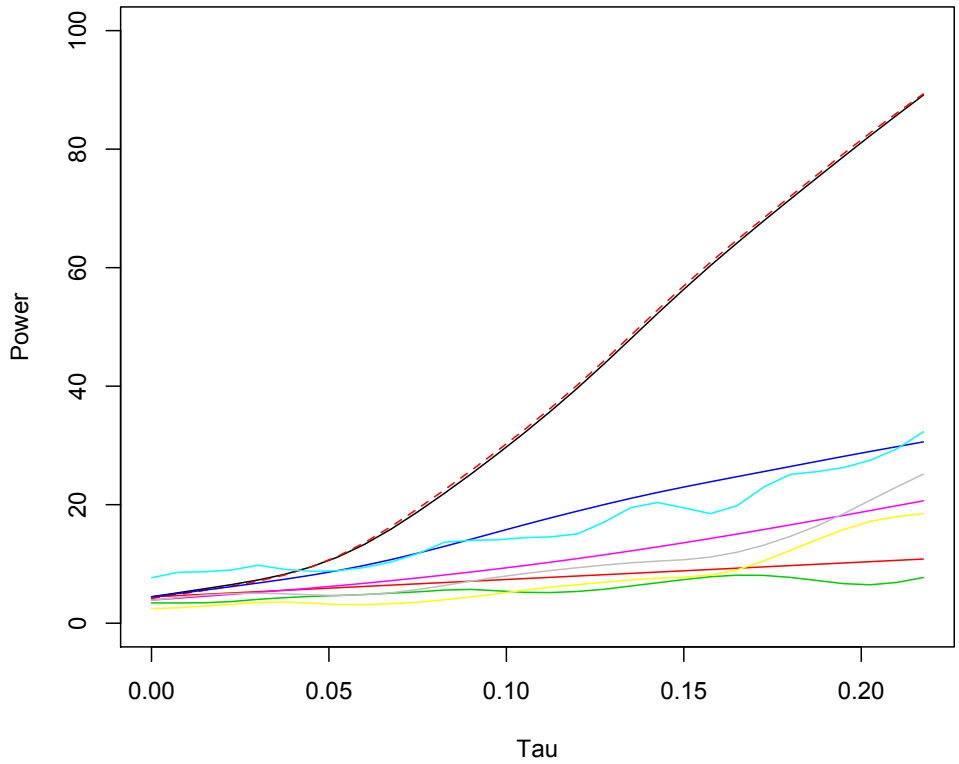

$X_1 \sim \text{Bin}(4, 0.5)$, $X_2 \sim \text{Bin}(4, 0.5)$ and $C$ is Gaussian

# Examples (cont'd)



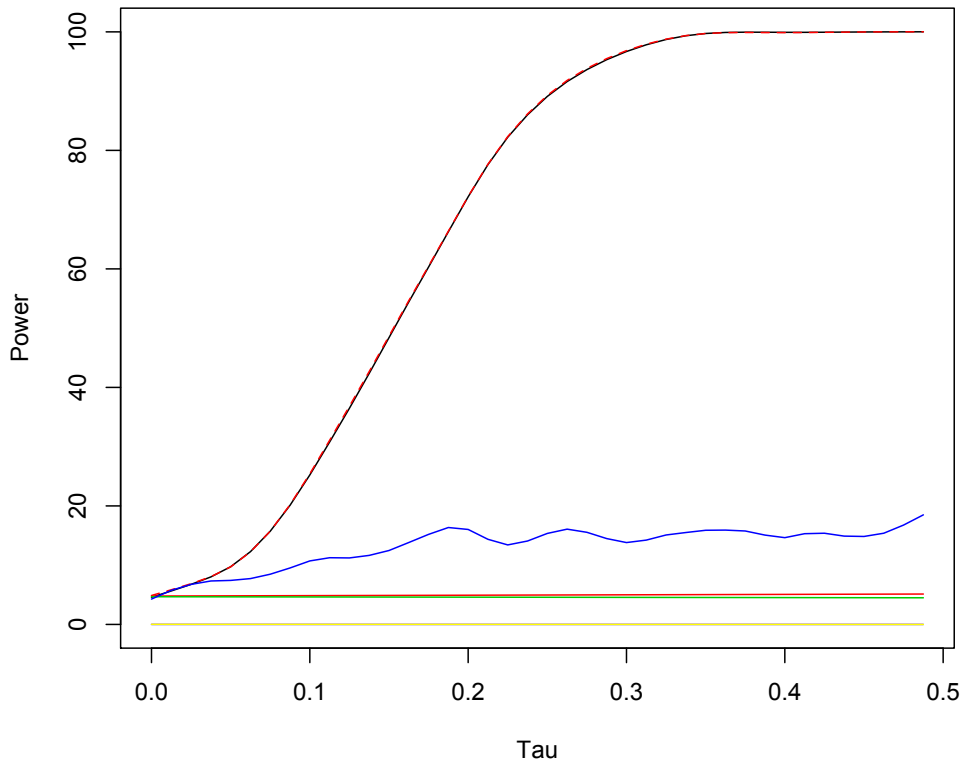

$X_1 \sim \mathrm{Bin}(4, 0.5)$, $X_2 \sim \mathrm{NBin}(5, 5/7)$ and $C$ is Gaussian

# Examples (cont'd)




$X_1 \sim \mathcal{P}(2)$, $X_2 \sim \mathcal{P}(2)$ and $C$ is FGM

# Examples (cont'd)




$X_1 \sim \mathcal{P}(1/2)$, $X_2 \sim \mathcal{N}(0,1)$ and $C$ is Frank

# What to do when the margins are unknown?

To compute (approximate) $p$-values of the new tests when the margins are unknown, one has to investigate the asymptotic behaviour of the empirical checkerboard copula process

$$\mathbb{C}_n^{\maltese} = \sqrt{n}\,(C_n^{\maltese} - C^{\maltese}).$$

This process is also the stepping stone to estimation and more advanced techniques, including goodness-of-fit testing.

# Known margins in the continuous case

When $F_1$ and $F_2$ are known and continuous, $C$ can be estimated by the empirical distribution function $B_n$ of the sample

$$(F_1(X_{i1}), F_2(X_{i2})), \quad i \in \{1, \ldots, n\}.$$

It is well known that in this case,

$$\mathbb{B}_n = \sqrt{n}(B_n - C)$$

converges weakly in $\ell^\infty[0,1]^2$ to a $C$-Brownian sheet $\mathbb{B}_C$, i.e., to a centered Gaussian process with covariance function

$$\mathrm{cov}\{\mathbb{B}_C(u,v), \mathbb{B}_C(w,z)\} = C(u \wedge w, v \wedge z) - C(u,v)C(w,z).$$

# Known margins in the discrete case

When $F_1$ and $F_2$ are known and supported on $\mathbb{N}$, $C^{\maltese}$ can be estimated by bilinear interpolation $B_n^{\maltese}$ of the empirical distribution function of the sample

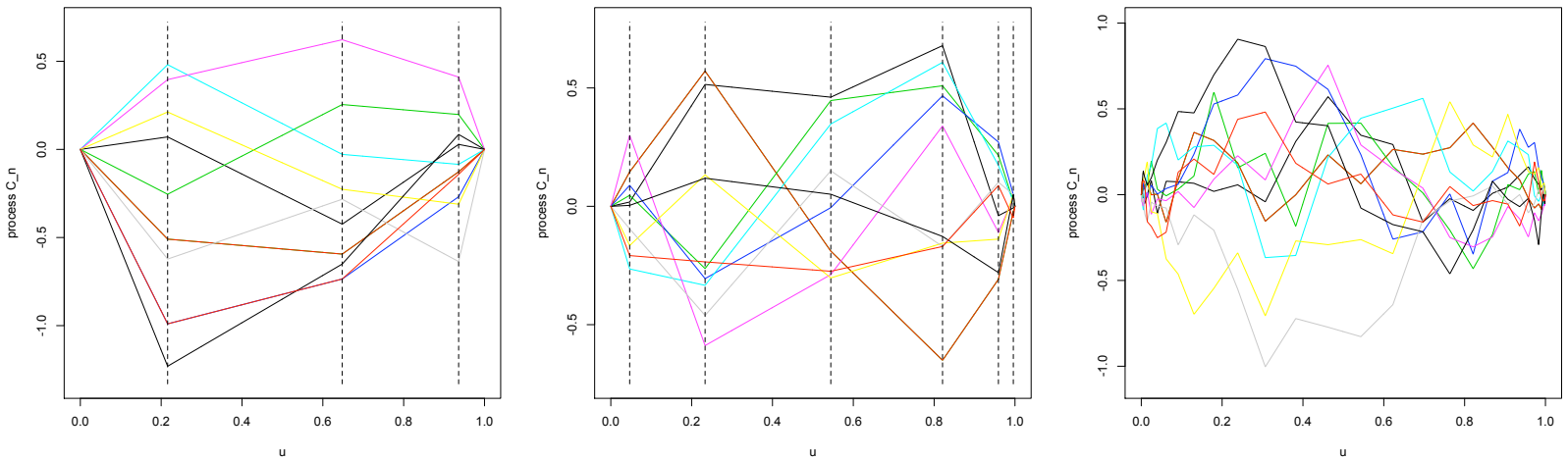
$$(F_1(X_{i1}), F_2(X_{i2})), \quad i \in \{1, \ldots, n\}.$$

Theorem. *As $n \to \infty$, the process $\mathbb{B}_n^{\maltese} = \sqrt{n}\,(B_n^{\maltese} - C^{\maltese})$ converges weakly in $\mathcal{C}[0,1]^2$ to a centered Gaussian process $\mathbb{B}_C^{\maltese}$.*

Here, $\mathbb{B}_C^{\maltese}$ is no longer a $C^{\maltese}$-Brownian sheet, but a "bilinear interpolation" thereof.

# Illustration

The limiting process $\mathbb{B}_C^{\maltese}$ is illustrated below in the univariate case when $F$ is Binomial with $p = 0.4$ and $n = 3, 10$ and $100$.




Displayed are 10 realizations of $\mathbb{B}_n^{\maltese}$ when $n = 5000$.

# Unknown margins in the continuous case

Under suitable regularity conditions, the process

$$\mathbb{C}_n = \sqrt{n}\,(C_n - C)$$

converges weakly in $\ell^\infty[0,1]^2$ to a centered Gaussian process $\mathbb{C}$ defined, for all $u_1, u_2 \in [0,1]$, by

$$\mathbb{C}(u_1,u_2) = \mathbb{B}_C(u_1,u_2) - C_1(u_1,u_2)\,\mathbb{B}_C(u_1,1) - C_2(u_1,u_2)\,\mathbb{B}_C(1,u_2).$$

# Bad news in the discrete case

Suppose that $H$ is a bivariate Bernoulli distribution with

$$F_1(0) = p, \quad F_2(0) = q, \quad H(0,0) = r,$$

where $p, q \in (0,1)$ and $r = C(p,q)$ for some copula $C$.

It can be established that the finite-dimensional margins of $\mathbb{C}_n^{\maltese}$ converge in law, although the limit may be non-Gaussian.

However, the sequence $\mathbb{C}_n^{\maltese}$ is not asymptotically equicontinuous in probability unless $r = pq$.

In other words, $\mathbb{C}_n^{\maltese}$ does not converge in $\mathcal{C}[0,1]^2$ unless $r = pq$. ☹

# What went wrong?

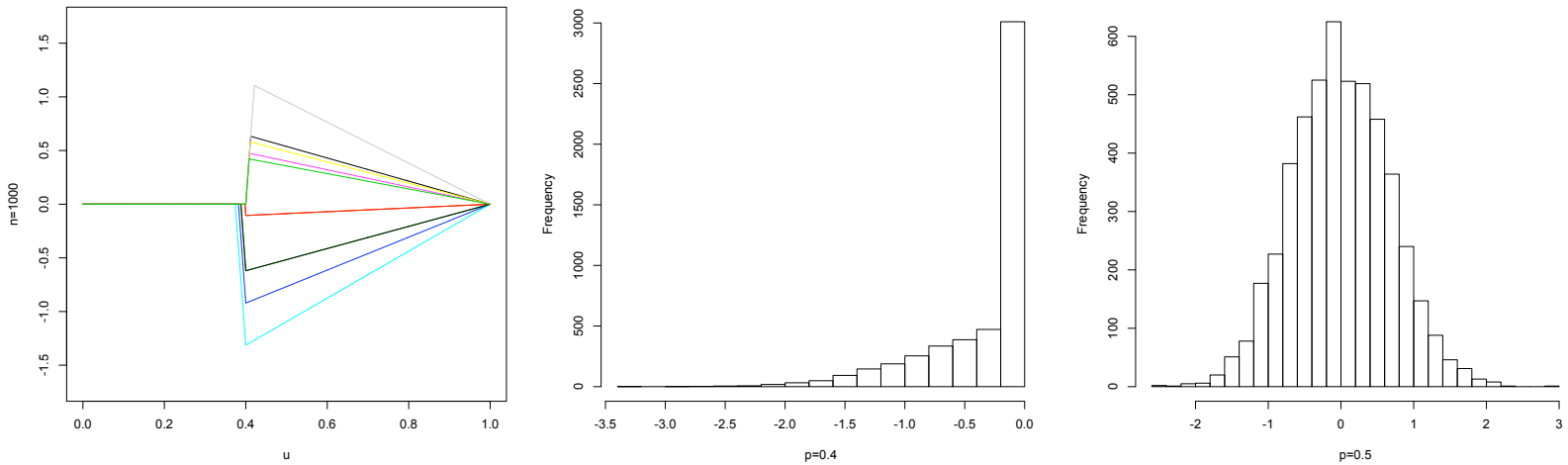
To see why, take $F$ Bernoulli with $F(0) \in (0,1)$ and let $F_n$ be its empirical counterpart based on a sample of size $n$ from $F$.

Set $F(0) = p$, $F_n(0) = p_n$ and consider

$$E_n(u) = \begin{cases} u, & u \in [0, p_n], \\ \frac{(1-u)}{1-p_n}\, p_n, & u \in [p_n, 1], \end{cases} \qquad E(u) = \begin{cases} u, & u \in [0, p], \\ \frac{(1-u)}{1-p}\, p, & u \in [p, 1]. \end{cases}$$

Then $\mathbb{E}_n = \sqrt{n}\,(E_n - E)$ does not converge in law in $\mathcal{C}[0,1]$ even though its finite-dimensional margins converge.

# Illustration




Ten realizations of the process $\mathbb{E}_n$ with $p = 0.4$ and sample size 1000 (left). Histograms of 5000 realizations of $\mathbb{E}_n(u)$ when $u = 0.4$ (middle) and $u = 0.5$ (right) based on samples of size $n = 10{,}000$.

# It works nonetheless!

Consider the set

$$\mathcal{O} = \bigcup_{(k,\ell)\in\mathbb{N}^2} \big(F_1(k-1), F_1(k)\big) \times \big(F_2(\ell-1), F_2(\ell)\big).$$

Theorem. *For arbitrary compact $K \subset \mathcal{O}$, $\mathbb{C}_n^{\maltese}$ converges weakly on $\mathcal{C}(K)$ as $n \to \infty$ to $\mathbb{C}^{\maltese}$ given for every $u_1, u_2 \in \mathcal{O}$ by*

$$\mathbb{B}_C^{\maltese}(u_1,u_2) - C_1^{\maltese}(u_1,u_2)\,\mathbb{B}_C^{\maltese}(u_1,1) - C_2^{\maltese}(u_1,u_2)\,\mathbb{B}_C^{\maltese}(1,u_2),$$

*where $\mathbb{B}_C^{\maltese}$ is the weak limit of $\mathbb{B}_n^{\maltese}$.*

# Example: Spearman's rho

Consider the non-normalized version of Spearman's rho, viz.

$$\rho = \rho(H) = \rho(C^{\maltese}) = 12 \int_0^1 \int_0^1 \{C^{\maltese}(u_1, u_2) - u_1 u_2\} \mathrm{d}\Pi(u_1, u_2).$$

Its consistent estimator is given by

$$\rho_n^* = \frac{12}{n^3} \sum_{i=1}^{n} (R_{i1}^* - \bar{R})(R_{i2}^* - \bar{R}) = \rho(C_n^{\maltese}),$$

where $R_{i1}^*$ and $R_{i2}^*$ are the componentwise midranks. Then

$$\sqrt{n}\,\{\rho_n^* - \rho(H)\} = 12 \int_0^1 \int_0^1 \mathbb{C}_n^{\maltese}(u_1, u_2)\, \mathrm{d}\Pi(u_1, u_2).$$

# Example (cont'd)

Because $[0,1]^2 \setminus \mathcal{O}$ has Lebesgue measure zero,

$$12 \int_0^1 \int_0^1 \mathbb{C}_n^{\maltese}(u_1, u_2)\, \mathrm{d}\Pi(u_1, u_2) = 12 \int_{\mathcal{O}} \mathbb{C}_n^{\maltese}(u_1, u_2)\, \mathrm{d}\Pi(u_1, u_2).$$

Furthermore, $\mathcal{O}$ can be approximated arbitrarily closely by compact sets. This lies at the heart of the following result:

Theorem. *As* $n \to \infty$,

$$\sqrt{n}\,\{\rho_n^* - \rho(H)\} \rightsquigarrow 12 \int_{\mathcal{O}} \mathbb{C}^{\maltese}(u_1, u_2)\, \mathrm{d}\Pi(u_1, u_2).$$

# Example: the new tests of independence

- ✓ Tests based on $S_{n1}$, $S_{n2}$ and $T_n$ are consistent.
- ✓ The tests can be applied to any types of margins. If no ties are present, one gets the tests of Genest & Rémillard (2004).
- ✓ Under $H_0$,

$$S_{n1} \quad \text{and} \quad S_{n2} \rightsquigarrow \int_0^1 \cdots \int_0^1 \mathbb{C}^{\maltese}(u_1, u_2) \mathrm{d}\Pi(u_1, u_2),$$

$$T_n \rightsquigarrow \sup_{u_1, u_2 \in [0,1]} |\mathbb{C}^{\maltese}(u_1, u_2)|.$$

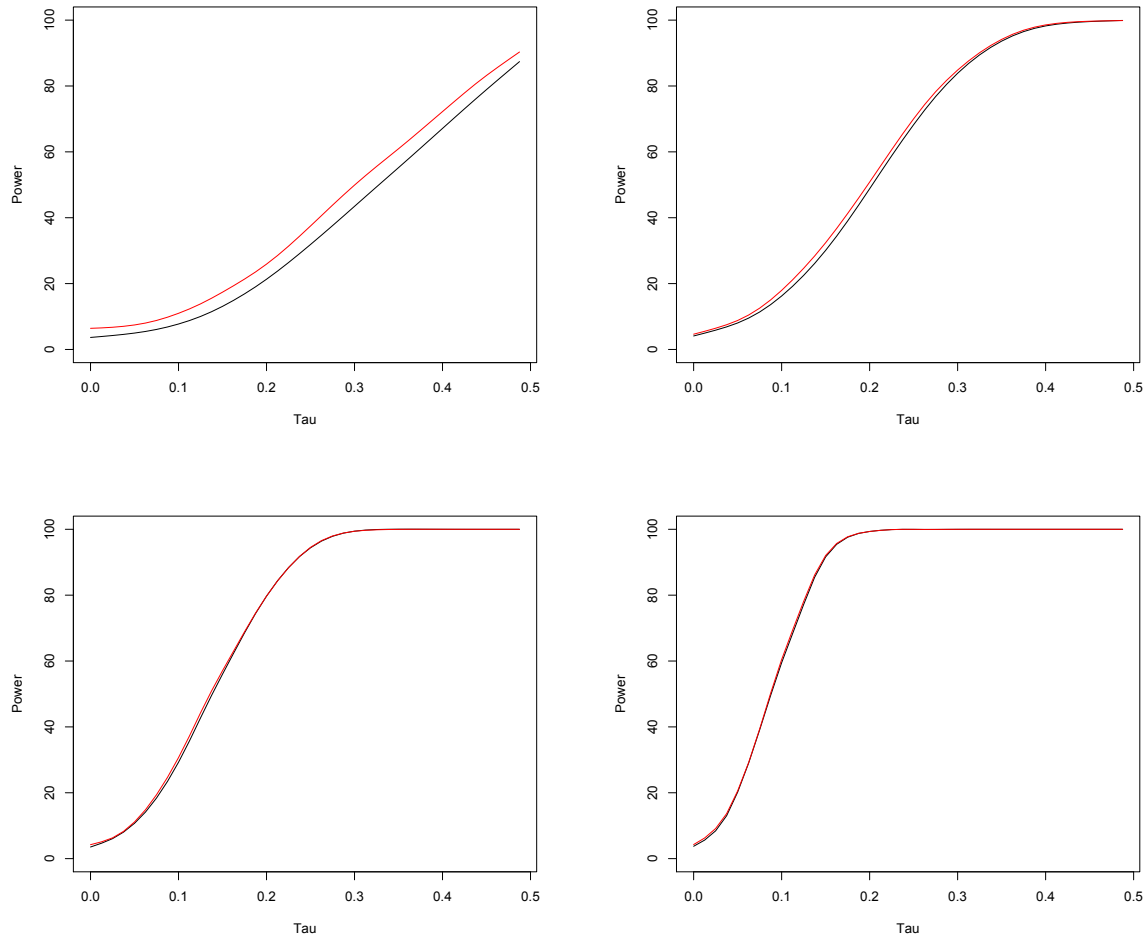
- ☹ The limit of $S_{n1}$, $S_{n2}$ and $T_n$ depends on the margins.

# How to use the limiting distribution?

Approximative $p$-values for the new tests can be computed using the multiplier bootstrap (van den Vaart & Wellner, 1996).

- ✓ This procedure has first been adapted to the copula context by Rémillard & Scaillet (2009).
- ✓ The core idea is to generate replicates from the limiting process $\mathbb{C}^{✠}$ using a sample of $M$ i.i.d. random variables, called the multipliers, with mean 0 and variance 1.
- ✓ The multiplier bootstrap is quick and can be proved to yield consistent procedures.

# A little simulation to convince you!




$\mathcal{P}(2)$ margins linked with the Frank copula; $n = 20, 50, 100, 250$.

# Application to the horseshoe crabs dataset

| Variables | $p$-value | | | | |
|---|---|---|---|---|---|
| | $S_{n2}^*$ | $\chi^2$ | $\chi^2_{MC}$ | $G^2$ | $D^2$ |
| Colour and Spine Condition | 0.000 | $4.83 \times 10^{-6}$ | $5.00 \times 10^{-4}$ | $4.28 \times 10^{-6}$ | $1.11 \times 10^{-16}$ |
| Satellites and Spine Condition | 0.031 | 0.565 | 0.535 | 0.382 | 0.252 |
| Satellites and Colour | 0.011 | 0.360 | 0.383 | 0.268 | 0.276 |

- ✓ All tests give the same conclusion for the categorical variables Colour and Spine Condition.
- ✓ The new test gives a different conclusion when one of the variables does not have a finite support.

# References

C. Genest, J. Nešlehová & B. Rémillard (2013). On the empirical multilinear copula process for count data. *Bernoulli*, in press.

C. Genest, J. Nešlehová & B. Rémillard (2013). On the estimation of Spearman's rho and related tests of independence for possibly discontinuous multivariate data. *J. Multivariate Anal.*, in press.

O.A. Murphy, C. Genest & J. Nešlehová (2013). Copula-based tests of independence for discrete or mixed data. *In preparation*.

# Any more questions?

# The multiplier bootstrap procedure

Let $n_1$ and $n_2$ be the number of distinct values of $X_{i1}$'s and $X_{i2}$'s in the sample, and denote these values by

$$\xi_1, \ldots, \xi_{n_1} \quad \text{and} \quad \eta_1, \ldots, \eta_{n_2}.$$

Step 1 For $m \in \{1, \ldots, M\}$, generate an independent random sample of size $n$ from any univariate distribution with mean zero and variance one. Let $\gamma_1^{(m)}, \ldots, \gamma_n^{(m)}$ represent the $m$th sample and $\bar{\gamma}^{(m)}$ represent the sample mean.

**Step 2** For each sample, compute

$$S_{n2}^{*(m)} = \int_0^1 \int_0^1 \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\gamma_i^{(m)} - \bar{\gamma}^{(m)}) \{\omega_{n1,i}(u) - u\} \{\omega_{n2,i}(v) - v\} \right]^2 dv\, du,$$

where

$$\omega_{n1,i}(u) = \left\{ \mathbf{1}(X_{i1} \le \xi_{j-1}) \frac{F_{n1}(\xi_j) - u}{\Delta F_{n1}(\xi_j)} + \mathbf{1}(X_{i1} \le \xi_j) \frac{u - F_{n1}(\xi_{j-1})}{\Delta F_{n1}(\xi_j)} \right\},$$

$$\omega_{n2,i}(v) = \left\{ \mathbf{1}(X_{i2} \le \eta_{k-1}) \frac{F_{n2}(\eta_k) - v}{\Delta F_{n2}(\eta_k)} + \mathbf{1}(X_{i2} \le \eta_k) \frac{v - F_{n2}(\eta_{k-1})}{\Delta F_{n2}(\eta_k)} \right\}.$$

**Step 3** The $p$-value for the test is given by

$$\frac{1}{M} \sum_{m=1}^{M} \mathbf{1}(S_{n2}^{*(m)} > S_{n2}^{*}).$$