Copula modeling for discrete data

Christian Genest & Johanna G. Nešlehová

in collaboration with Bruno Rémillard

McGill University and HEC Montréal

ROBUST, September 11, 2016

Main question

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are from a discrete distribution H.

More specifically, assume

$$X, Y \in \{0, 1, \ldots\}.$$

$\boxed{\text{Can we still do copula modeling?}}$

Lack of uniqueness of the copula

In the continuous case, there is a unique function $C : [0,1]^2 \to [0,1]$ such that

$$H(x,y) = C\{F(x), G(y)\}, \quad x, y \in \mathbb{R}.$$

In the discrete case, there are several functions $A : [0,1]^2 \to [0,1]$ such that

$$H(x,y) = A\{F(x), G(y)\}, \quad x, y \in \mathbb{R}.$$

This class of functions is denoted $\mathcal{A}$, but note that not all its members are copulas!

Lack of uniqueness of the copula

In the continuous case, C is the distribution function of the pair $(U, V) = (F(X), G(Y))$, i.e.,

$$C(u, v) = \Pr(U \leq u, V \leq v), \quad u, v \in (0, 1).$$

In the discrete case,

$$D(u, v) = \Pr(U \leq u, V \leq v), \quad u, v \in (0, 1).$$

is a distribution function too, but it is not a copula!

As a consolation, $D \in \mathcal{A}$.

Many paradoxical results follow...

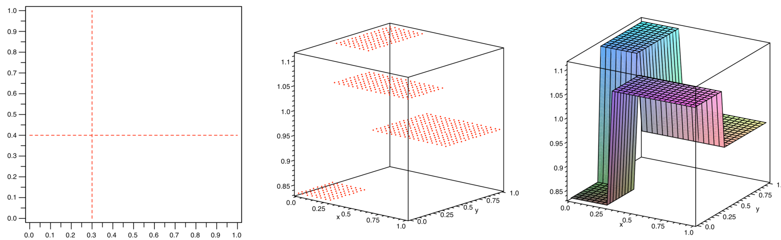
- ✓ As soon as F or G are discrete, the set $\mathcal{C}_H$ of admissible copulas is infinitely large (though its bounds can be identified).
- ✓ Members of $\mathcal{C}_H$ can embody completely different types of dependence.
- ✓ Measures of association, dependence concepts, and orderings become margin dependent.

Genest & Nešlehová (2007), *ASTIN Bulletin*.

Saving the connection between copulas and dependence?

1. H defines a contingency table.
2. Spread the mass uniformly in each cell.
3. Call the resulting copula $C^{\maltese} \in \mathcal{C}_H$ the bilinear extension copula.

Illustration for Bernoulli variates X and Y:



Our main discovery

$C^{\maltese}$ is the best possible candidate if you want to think of the copula associated with a discrete H, because...

- ✓ $C^{\maltese}$ is an absolutely continuous copula.
- ✓ $X \perp Y \Leftrightarrow C^{\maltese}(u, v) = uv$.
- ✓ For any concordance measure, $\kappa(H) = \kappa(C^{\maltese})$.
- ✓ If $(\tilde{X}, \tilde{Y})$ is distributed as $C^{\maltese}$, then

$$\mathrm{DEP}(X, Y) \Leftrightarrow \mathrm{DEP}(\tilde{X}, \tilde{Y}).$$

Here, DEP can refer to PQD, LTD, RTI, SI, LRD.

Are copula models for discrete data of interest?

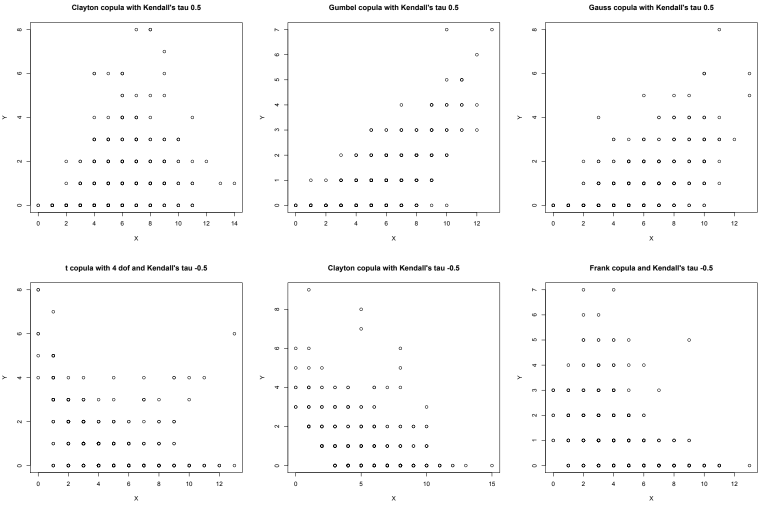
A copula model for H, viz.

$$H(x, y) = C\{F(x), G(y)\}$$

with $F \in (F_\alpha)$, $G \in (G_\beta)$ and $C \in (C_\theta)$

is a perfectly valid construction, even if F and G are discrete.

Yes, they are!



$X \sim \mathcal{P}(5)$ and $Y \sim \mathcal{G}(0.6)$ with various copulas and positive (top) and negative (bottom) association.

Theoretical back-up

- ✓ H often inherits dependence properties from C because

$$\mathrm{DEP}(U, V) \Rightarrow \mathrm{DEP}(X, Y),$$

where $(U, V) \sim C$ and DEP means PQD, LTD, RTI, SI, LRD.

- ✓ θ can continue to govern association between X and Y, viz.

$$C_\theta \prec_{\mathrm{PQD}} C_{\theta'} \quad \Rightarrow \quad H_\theta \prec_{\mathrm{PQD}} H_{\theta'}.$$

Can we do inference?

Assume $(X_1, Y_1), \ldots, (X_n, Y_n)$ is an iid sample from

$$H_\theta(x, y) = C_\theta\{F(x), G(y)\}$$

with F and G discrete.

$\boxed{\text{How can one fit and validate such a model?}}$

The quest for the answer is the subject of our ongoing research.

Naïve suggestion

Draw 10,000 samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ of size $n = 100$ from

$$H_\theta(x, y) = C_\theta\{F(x), G(y)\},$$

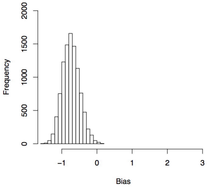
where C_θ is a Clayton copula and F, G are discrete distributions.

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

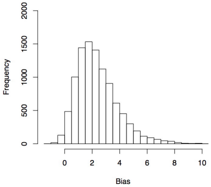
$$\hat{\theta} = 2\,\frac{\hat{\tau}}{1 - \hat{\tau}}.$$

Illustration: Poisson margins

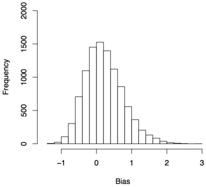
Take $\theta = 2$ and Poisson margins with $\mathrm{E}(X) = 1$ and $\mathrm{E}(Y) = 2$.



$\hat{\theta}$ based on τ_n



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



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

What is going on?

It can be seen that τ_n is an *unbiased* estimator of

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

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

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

$$X_i = F^{-1}(U_i) \quad \text{and} \quad Y_i = G^{-1}(V_i) \quad \not\Rightarrow \quad (F(X_i), G(Y_i)) \sim C_\theta.$$

In short, the discretization of (U_i, V_i) is irreversible. ☹

Can't we just randomize?

1. Take a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from H whose margins are count distributions.

2. Add an independent noise to each component of the pair (X_i, Y_i), viz.

$$\tilde{X}_i = X_i + U_i - 1, \quad \tilde{Y}_i = Y_i + V_i - 1,$$

   where $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ are independent samples from the standard uniform distribution on $(0, 1)$.

3. The randomized sample $(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_n, \tilde{Y}_n)$ then stems from a distribution whose margins are continuous.

Two additional estimators

4. Compute $\tau_n(\tilde{X}, \tilde{Y})$, the sample version of Kendall's tau based on the randomized sample
$$(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_n, \tilde{Y}_n).$$
This gives a moment-estimate of θ, viz. $\hat{\theta} = g^{-1}(\bar{\tau}_n)$.

5. Alternatively, one can compute the pseudo-likelihood estimate of θ based on the randomized sample.

6. To eliminate the uncertainty induced by randomization, repeat the previous steps N times and compute the average of the values of the estimate of θ.

… that do not work either.

Average and st. deviation of six estimates of θ in the Illustration:

	Estimate of θ based on					
	$\tau_n(U,V)$	$\tau_n(X,Y)$	$\tau_{a,n}(X,Y)$	$\tau_{b,n}(X,Y)$	$\tau_n(\tilde{X},\tilde{Y})$	MLE$(\tilde{X},\tilde{Y})$
Av.	2.039	1.262	4.358	2.213	1.269	1.144
S.d.	0.446	0.243	1.495	0.537	0.285	0.779

Randomization is bound to fail, because the copula of $(\tilde{X},\tilde{Y})$ is $C^{\maltese}$. However, remember that

$$C^{\maltese} \neq C_\theta.$$

The empirical bilinear extension copula

- ✓ Compute the empirical cdf H_n corresponding to the sample

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

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

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

$$c_n^{\maltese}(u, v) = n \times \frac{n_{ij}}{n_{i\bullet} n_{\bullet j}}$$

for all $u \in (F_n(i-1), F_n(i)]$, $v \in (G_n(j-1), G_n(j)]$, $i, j \in \mathbb{N}$.

- ✓ Observe that $C_n^{\maltese}$ is rank-based.

Wait a minute...

A sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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 mid-rank coefficient for testing monotone trend

$$\rho_n^* = \frac{12}{n^3} \left\{ \sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S}) \right\}$$

- ✓ Kendall's coefficient for testing monotone trend

$$\tau_n^* = \frac{2}{n^2} \{\#(\text{concordant pairs}) - \#(\text{discordant pairs})\}$$

Surprise!

It can be seen that

$$\chi^2 = n \int_0^1 \int_0^1 \{c_n^{\maltese}(u,v) - 1\}^2 \mathrm{d}u \, \mathrm{d}v,$$

$$\rho_n^* = 12 \int_0^1 \int_0^1 \{C_n^{\maltese}(u,v) - uv\} \mathrm{d}u \, \mathrm{d}v,$$

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

Here's an idea

In the continuous case, many inferential procedures derive from the limiting behavior of the empirical copula process

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

In the discrete case, one can investigate the asymptotic behavior of the empirical Maltese copula process

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

hoping that it would be as useful as in the continuous case.

Known margins in the continuous case

When F and G are known and continuous, C can be estimated by the empirical distribution function B_n of the sample

$$(F(X_i), G(Y_i)), \quad 1 \leq i \leq n.$$

It is well-known that in this case,

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

converges weakly in $\mathcal{C}[0,1]^2$ 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 and G 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

$$(F(X_i), G(Y_i)), \quad 1 \leq i \leq n.$$

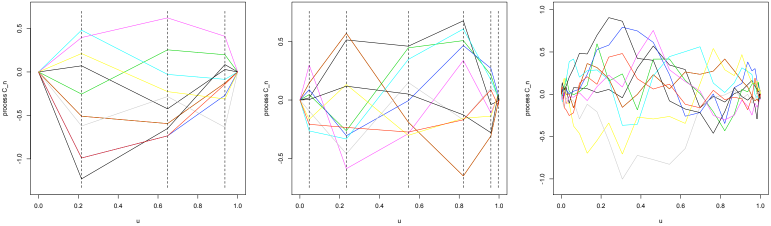
Theorem

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 ten 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 $\mathcal{C}[0,1]^2$ to a centered Gaussian process $\mathbb{C}$,

$$\mathbb{C}(u,v) = \mathbb{B}_C(u,v) - \frac{\partial}{\partial u} C(u,v)\,\mathbb{B}_C(u,1) - \frac{\partial}{\partial v} C(u,v)\,\mathbb{B}_C(1,v).$$

Bad news in the discrete case

Suppose that H is a bivariate Bernoulli distribution with

$$F(0) = p, \quad G(0) = q, \quad H(0,0) = r,$$

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 not be 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 illustrate, consider a similar process in the univariate case.

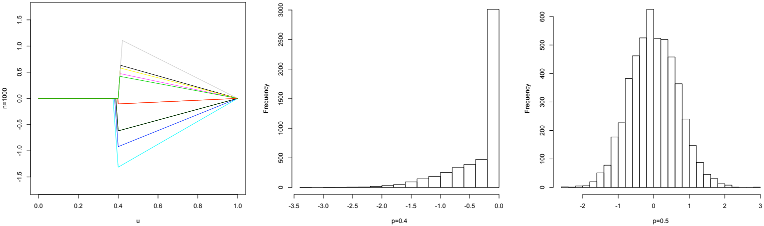
Take F Bernoulli with $F(0) \in (0,1)$ and let F_n be its empirical counterpart based on a sample of size n from F.

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

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

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 $\mathbb{E}_n$ for the Bernoulli distribution with $p = 0.4$ and sample size 1000 (left). Histograms of $5,000$ realizations of $\mathbb{E}_n(u)$ when $u = 0.4$ (middle) and $u = 0.5$ (right) based on samples of size $n = 10,000$.

All's well that ends well!

Consider the set

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

Theorem

Let K be an arbitrary compact subset of $\mathcal{O}$. Then $\mathbb{C}_n^{\maltese}$ converges weakly on $\mathcal{C}(K)$ as $n \to \infty$ to $\mathbb{C}^{\maltese}$ given for every $u, v \in \mathcal{O}$ by

$$\mathbb{B}_C^{\maltese}(u,v) - \frac{\partial}{\partial u} C^{\maltese}(u,v)\,\mathbb{B}_C^{\maltese}(u,1) - \frac{\partial}{\partial v} C^{\maltese}(u,v)\,\mathbb{B}_C^{\maltese}(1,v),$$

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.

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

Its consistent estimator is given by

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

where R_i and S_i are the componentwise mid-ranks. Consequently,

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

It works!

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

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

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

Theorem

As $n \to \infty$,

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

References

C. Genest & J. Nešlehová (2007). A primer on copulas for count data. *The ASTIN Bulletin*, **37**, 475–515.

C. Genest & J.G. Nešlehová & B. Rémillard (2014). On the empirical multilinear copula process for count data. *Bernoulli*, 20, 1344–1371.

C. Genest & J.G. Nešlehová & B. Rémillard (2014). Asymptotics of the empirical multilinear copula process. *Submitted.*

Research funded by

CANADA RESEARCH CHAIRS

Fonds de recherche sur la nature et les technologies Québec

CANSSI INCASS

NSERC CRSNG

McGill