Modeling count data with copulas: Should we?

Christian Genest
Johanna Nešlehová

Tartu, June 28, 2007

Fact of life:

Copula modeling
has become
exceedingly
popular
in recent years.

Fact of life:

Copula modeling
has become
exceedingly
popular
in recent years.

"Even I agree!"

(Thomas Mikosch)

What is a copula model for a (bivariate) distribution H?

It consists of assuming

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

for some

$$C \in (C_\theta), \quad F \in (F_\alpha), \quad G \in (G_\beta).$$

Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from H, the aim is to estimate the unknown parameters and retrieve $C = C_{\theta_0}$.

When H is continuous, this can be done consistently, but...

$$\boxed{\text{What if } X, Y \in \{0, 1, \ldots\}?}$$

1. Lack of uniqueness of the copula

If H is continuous, there is a unique function C such that

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

The copula C can be retrieved from H, viz.

$$C(u,v) = H\{F^{-1}(u), G^{-1}(v)\}, \quad u, v \in (0,1).$$

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

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

What happens in the discrete case?

If H is discrete, there are several functions A such that

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

The following is a solution but not a copula (or a distribution):

$$B(u, v) = H\{F^{-1}(u), G^{-1}(v)\}, \quad u, v \in (0, 1).$$

The following is another solution (i.e., $D \neq B$) and not a copula:

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

2. Extent of the unidentifiability issue

Given a bivariate distribution function H with discrete margins, let $\mathcal{C}_H$ be the set of copulas C for which

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

Questions:

- ✓ Can we get a sense of the size of the set $\mathcal{C}_H$?
- ✓ What are the "smallest" and "largest" elements in $\mathcal{C}_H$?

Pointwise bounds on $\mathcal{C}_H$

It is well known that in general

$$W(u,v) \;\le\; C(u,v) \;\le\; M(u,v), \quad u,v \in [0,1]$$

where W and M are the Fréchet–Hoeffding bounds.

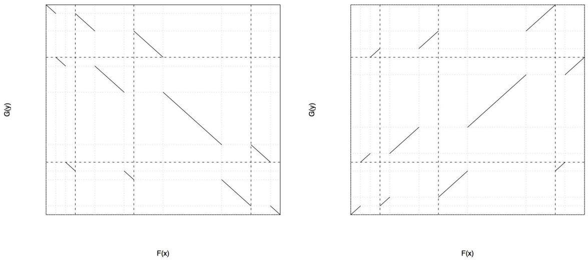
To assess the extent of unidentifiability, one needs sharp bounds

$$C_H^-(u,v) \;\le\; C(u,v) \;\le\; C_H^+(u,v), \quad u,v \in [0,1]$$

that apply to any $C \in \mathcal{C}_H$, i.e., to any copula compatible with H.

Such bounds exist; they were derived by Carley (2002).

Holly Carley's bounds: concrete example



	$X = 0$	$X = 1$	$X = 2$	$X = 3$	Total
$Y = 2$	1	2	3	0	6
$Y = 1$	1	3	6	2	12
$Y = 0$	1	1	3	1	6
	3	6	12	3	24

Carley bounds for Kendall's tau and Spearman's rho

Explicit expressions are available for Carley bounds on

$$\tau(C) = -1 + 4 \int\int C(u,v) dC(u,v), \quad \rho(C) = -3 + 12 \int\int C(u,v) dv du.$$

A sense of the unidentifiability issue is conveyed by

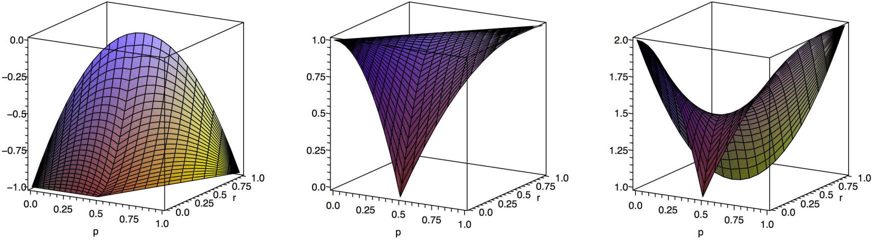
$$[\kappa(C_H^-), \kappa(C_H^+)]$$

for any measure of concordance κ (Scarsini 1984).

"I'm Holly [not Holy]"

Example: X and Y are Bernoulli

For $\Pr(X = 0) = Pr(Y = 0) = p$ and $\Pr(X = 0, Y = 0) = r$:



Plot of $\tau(C_H^-)$ and $\tau(C_H^+)$ as a function of p and r;
the difference between the two bounds is shown in the right panel.

3. Interplay between copula and dependence

In the continuous case, C characterizes dependence, e.g.,

$$
\begin{aligned}
C(u,v) &= uv \quad \Leftrightarrow \quad X \perp Y, \\
C(u,v) &= \min(u,v) \quad \Leftrightarrow \quad G(Y) = F(X), \\
C(u,v) &= \max(0, u+v-1) \quad \Leftrightarrow \quad G(Y) = 1 - F(X).
\end{aligned}
$$

Also if $\kappa(X, Y)$ is a measure of association, then

$$\kappa(X, Y) = \kappa(C).$$

In the discrete case, copula $\neq$ dependence

If $(X, Y) \sim H(x, y) = C\{F(x), G(y)\}$, then

$$C(u, v) = uv \quad \Rightarrow \quad X \perp Y$$

but

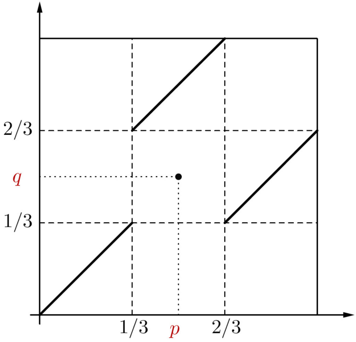
$$X \perp Y \quad \not\Rightarrow \quad C(u, v) = uv.$$

Similarly, monotone functional dependence is not equivalent to

$$H(x, y) = W\{F(x), G(y)\} \quad \text{or} \quad H(x, y) = M\{F(x), G(y)\}.$$

Example from Marshall (1996)

Take $X \sim \text{Bernoulli}(1-p)$, $Y \sim \text{Bernoulli}(1-q)$.



- $(p,q) \in [0,1/3] \times [0,1/3]$: perfect positive dependence
- $(p,q) = (1/\sqrt{3}, 1/\sqrt{3})$: independence
- $(p,q) \in [2/3,1] \times [2/3,1]$: perfect negative dependence

4. Other consequence of margin-dependence

All traditional measures of association depend on margins. ☹

As an illustration, suppose X and Y are Bernoulli with

$$\Pr(X=0)=p, \quad \Pr(Y=0)=q, \quad \Pr(X=0,Y=0)=r.$$

Then, e.g.,

$$\begin{aligned}\tau(X,Y) &= \Pr\{(X_1-X_2)(Y_1-Y_2)>0\} \\ &\quad -\Pr\{(X_1-X_2)(Y_1-Y_2)<0\} \\ &= r-pq.\end{aligned}$$

A theorem due to Marshall (1996)

"Let $\mathcal{H}$ be the class of bivariate distribution functions whose support is contained in $\mathbb{N}^2$.

Assume that κ is a dependence measure such that

$$C \in \mathcal{C}_H \quad \Rightarrow \quad \kappa(H) = \kappa(C)$$

holds for all $H \in \mathcal{H}$.

Then κ is constant." ☹

5. Consequences for inference

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

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

$$\left(\frac{R_1}{n}, \frac{S_1}{n}\right), \ldots, \left(\frac{R_n}{n}, \frac{S_n}{n}\right).$$

This amounts to estimating the margins conservatively, because

$$\hat{U}_i = F_n(X_i) = \frac{1}{n} \sum_{j=1}^{n} 1(X_j \leq X_i) = \frac{R_i}{n}, \quad i \in \{1, \ldots, n\}.$$

Most popular approaches to estimation

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

$$\ell(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log[c_\theta\{F_n(x_i), G_n(y_i)\}].$$

- Use a moment estimator of θ, 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ is an iid sample from

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

with F and G discrete.

Do the same strategies work?

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

$$X_i = X_j \quad \text{or} \quad Y_i = Y_j \quad \text{or both.}$$

- How do we account for ties?

Adjustment for ties, e.g., for inversion of τ

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): $\tau_{b,n} = (N_c - N_d)/\sqrt{N_x N_y}$

where

$$N_x = \sum_{i<j} \mathbf{1}(x_i \neq x_j) \quad \text{and} \quad N_y = \sum_{i<j} \mathbf{1}(y_i \neq y_j).$$

Modest simulation experiment

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

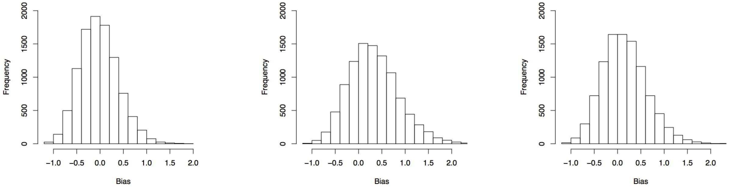
$$H_\theta(x, y) = C_\theta\{F(x), G(y)\},$$

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

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

Example: Geometric distributions



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

$$\Pr(X = 0) = 0.05\,, \quad \Pr(Y = 0) = 0.1 \quad \text{and} \quad \theta = 2.$$

What is the source of this bias?

It can be seen that τ_n is an unbiased estimator of

$$\tau(H) = \tau(C_H^{\maltese}),$$

where $C_H^{\maltese}$ is a specific element of $\mathcal{C}_H$. However, $C_H^{\maltese} \neq C_\theta$.

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. ☹

Is θ estimable at all?

In the continuous case, no problem!

In the discrete case,

- ✓ The issue is not completely settled yet.
- ✓ Rank-based methods seem hopeless. ☹
- ✓ Even with the full likelihood, an identifiability issue remains (maybe).

There are cases where maximum likelihood works! ☺☺

Let X, Y be Bernoulli with $\Pr(X = 0) = p$, $\Pr(Y = 0) = q$,

$$\Pr(X = 0, Y = 0) = C_\theta(p, q).$$

Suppose the dependence arises through an FGM family, viz.

$$C_\theta(u, v) = uv + \theta uv(1 - u)(1 - v), \quad \theta \in [-1, 1].$$

Generate 10,000 random samples of size $n = 100$.

Likelihood

Denote

$$p_{ij} = \Pr(X = i, Y = j), \quad i, j \in \{0, 1\}.$$

The log-likelihood to be maximized is

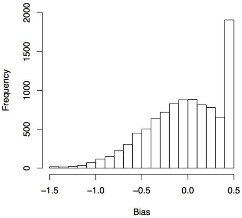
$$n_{00} \log(p_{00}) + n_{01} \log(p_{01}) + n_{10} \log(p_{10}) + n_{11} \log(p_{11}),$$

where

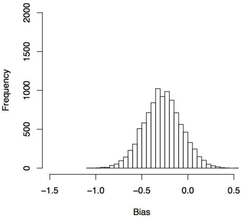
$$p_{00} = C_\theta(p, q) = pq + \theta pq(1 - p)(1 - q)$$

and $p_{01} = p - p_{00}$, $p_{10} = q - p_{00}$, $p_{11} = 1 - p - q + p_{00}$.

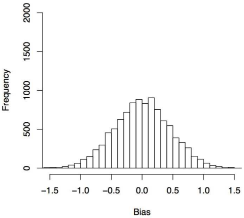
Results



$\hat{\theta}$ based on ML



$\hat{\theta}$ based on τ_n



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

Remember: focus on bias, not on normality!

6. Conclusion: Are copula models useful for discrete data?

Despite the unidentifiability issue, models of the type

$$H(x, y) = C\{F(x), G(y)\}, \quad C \in (C_\theta)$$

are still valid, even when X and Y are discrete.

Furthermore,

- H often inherits dependence properties from C.
- θ continues to govern association between X and Y.

Dependence properties of C are inherited by H

If X and Y are discrete and

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

then

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

Here, DEP could be either of the following dependence concepts:

PQD, LTD, RTI, SI, LRD.

θ is still a dependence parameter

In order for a family (C_θ) to yield meaningful models, a fundamental requirement is

$$\theta < \theta' \quad \Rightarrow \quad C_\theta(u,v) \le C_{\theta'}(u,v) \quad \text{(i.e., } C_\theta \prec_{\mathrm{PQD}} C_{\theta'}\text{)}.$$

This implies, e.g.,

$$\theta < \theta' \quad \Rightarrow \quad \tau(C_\theta) \le \tau(C_{\theta'}) \quad \text{and} \quad \rho(C_\theta) \le \rho(C_{\theta'}).$$

Given a PQD-ordered copula family (C_θ), suppose that

$$H_\theta(x,y) = C_\theta\{F(x), G(y)\}, \quad x, y \in \mathbb{R}.$$

Then whether X and Y are discrete or not, one has

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

In the discrete case, however, the reverse implication holds only for the very special copula:

$$H_\theta \prec_{\mathrm{PQD}} H_{\theta'} \quad \Leftrightarrow \quad C_\theta^{\maltese} \prec_{\mathrm{PQD}} C_{\theta'}^{\maltese}.$$

Summary

- ✓ The road to copula modeling of count data is treacherous.
- ✓ Much research remains to be done, particularly concerning inferential aspects of the problem.
- ✓ For more details, read

  C. Genest & J. Nešlehová (2007).
  A primer on copulas for count data.
  The ASTIN Bulletin, 37, in press.

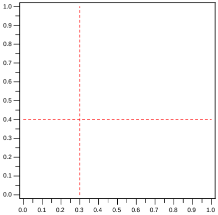
Any questions?

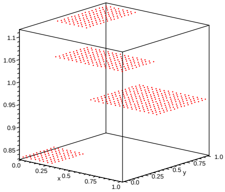
Encore: The “continuization” procedure

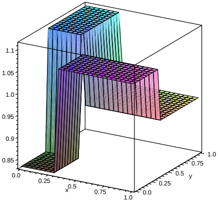
- ✓ If H is discrete, it defines a contingency table.
- ✓ Spread the mass uniformly in each cell.
- ✓ Call the resulting copula $C_H^{\maltese} \in \mathcal{C}_H$.

Illustration for Bernoulli variates X and Y:

$$\Pr(X = 0) = 0.3, \quad \Pr(Y = 0) = 0.4, \quad \Pr(X = 0, Y = 0) = 0.1$$







Good properties of $C_H^{\maltese}$

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

- $C_H^{\maltese}$ is an absolutely continuous copula.
- There exists an algebraically closed expression for it.
- $X \perp Y \Leftrightarrow C_{(X,Y)}^{\maltese}(u, v) = uv$.
- For any concordance measure, $\kappa(H) = \kappa(C_H^{\maltese})$.
- If $(\tilde{X}, \tilde{Y})$ is distributed as $C_H^{\maltese}$, then

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

In particular, DEP(X, Y) could be

- X and Y are in positive quadrant dependence
- Y is LTD or RTI in X
- Y is stochastically increasing in X
- X and Y are in positive likelihood ratio dependence

See, e.g., Denuit & Lambert (2005), Mesfioui & Tajar (2005), Nešlehová (2007).




Limitations of $C_H^{\maltese}$

$C_H^{\maltese}$ is a valiant knight but it does not solve all the problems:

- $C_H^{\maltese}$ depends on the margins.
- When $F(X) = G(Y) \not\Rightarrow C^{\maltese}_{(X,Y)} = \min(u, v)$.
- When $F(X) = \bar{G}(Y) \not\Rightarrow C^{\maltese}_{(X,Y)} = \max(0, u + v - 1)$.
- In fact, $C^{\maltese}_{(X,Y)}$ never equals M or W.
- As a consequence, one has always $|\kappa(C^{\maltese}_{(X,Y)})| < 1$.