

Accounting for extreme-value dependence in multivariate data

Christian Genest, Université Laval, Québec

38th ASTIN Colloquium
Manchester, July 15, 2008

Outline

1. Dependence modeling through copulas
2. Rank-based inference
3. Extreme-value dependence structures
4. Testing for extremeness
5. Nonparametric estimation with known margins
6. Proposed estimators when margins are unknown
7. Efficiency comparisons
8. Final remarks

1. Dependence modeling through copulas

Consider a bivariate random vector (X, Y) with continuous margins

$$F(x) = \Pr(X \leq x), \quad G(y) = \Pr(Y \leq y).$$

Following Sklar (1959), one may write

$$H(x, y) = \Pr(X \leq x, Y \leq y) = C(F(x), G(y)),$$

for a **unique copula** C , viz.

$$C(u, v) = \Pr(U \leq u, V \leq v) = \Pr\{F(X) \leq u, G(Y) \leq v\}.$$

Copula models

A **copula model** for (X, Y) obtains when

$$H(x, y) = \Pr(X \leq x, Y \leq y) = C(F(x), G(y))$$

is assumed to hold for parametric families

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

Such models provide flexibility in the choice of X and Y margins.
The copula induces the dependence between them, e.g.,

$$X \perp Y \quad \Leftrightarrow \quad C(u, v) \equiv uv.$$

2. Rank-based inference

Copula models are not always appropriate, but the dependence in a continuous pair (X, Y) is **always** characterized by its copula.

Because copulas are invariant by increasing transformations of the margins, viz.

$$(X, Y) \mapsto (\phi(X), \psi(Y))$$

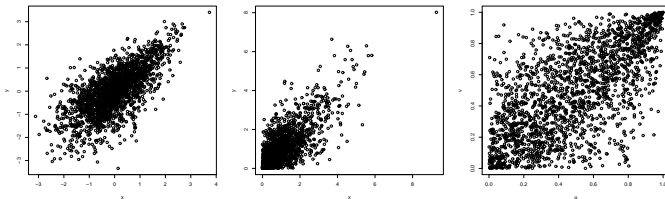
the dependence in a random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ is best represented by the ranks, viz.

$$R_i = \text{rank}(X_i) = \text{rank}(\phi(X_i)),$$

$$S_i = \text{rank}(Y_i) = \text{rank}(\psi(Y_i)).$$

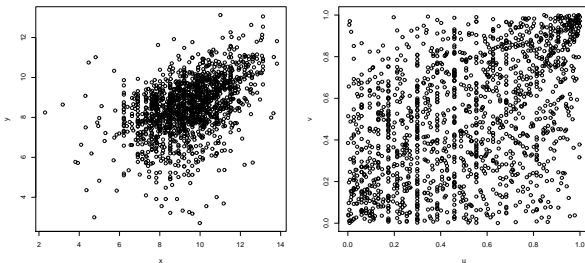
To study the dependence, get rid of the marginals!

The pairs $(R_i/n, S_i/n)$ are “pseudo-observations” from the underlying copula that characterizes the dependence structure.



Samples of size 2000 from two distributions with the same Gumbel copula ($\tau = 1/2$)

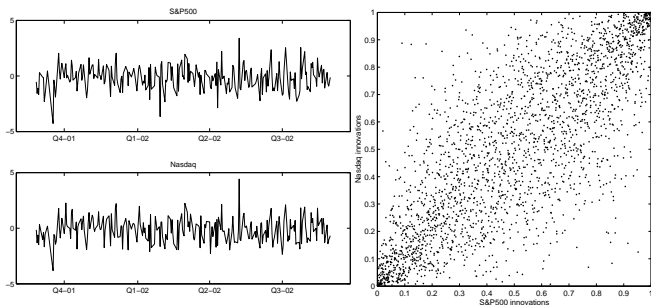
Example 1: LOSS/ALAE data



Original data (left) and pairs of normalized ranks (ranks)

Source: Frees & Valdez (1998)

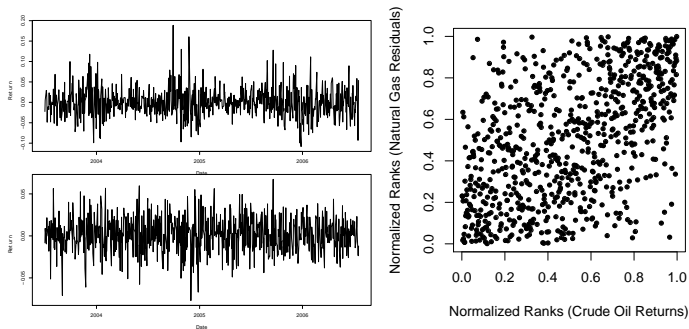
Example 2: Nasdaq versus S&P500 innovations



Original data (left) and pairs of normalized ranks (ranks)

Source: Van den Goorbergh et al. (2005)

Example 3: Prices of oil and gas



Original data (left) and pairs of normalized ranks (ranks)

Source: Genest et al. (2008)

3. Extreme-value dependence structures

Clusters of points near $(0, 0)$ and $(1, 1)$ suggest that the dependence structure may be an **extreme-value copula**, viz.

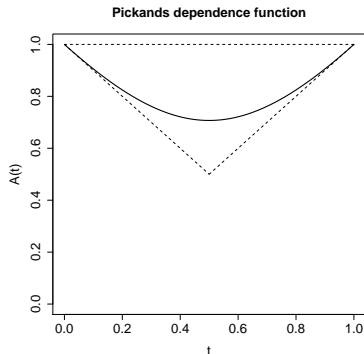
$$C(u, v) = \exp \left[\log(uv) A \left\{ \frac{\log(v)}{\log(uv)} \right\} \right].$$

where $A : [0, 1] \rightarrow [0, 1]$ is convex and

$$\max(t, 1 - t) \leq A(t) \leq 1, \quad t \in [0, 1].$$

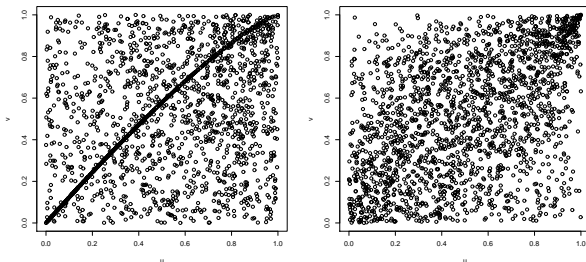
This happens in many other actuarial/financial applications, although (X, Y) does **not** have an extreme-value distribution.

Pickands dependence function



Tail dependence coefficient: $\lambda = 2\{1 - A(1/2)\} \in [0, 1]$

Illustration



Samples of size 2000 from two extreme-value copulas

Left: Marshall–Olkin (0.5, 0.4)

Right: Student extreme-value copula with 2 d.f. and $\rho = 0.5$

Issues of interest

A. How can one check that C is an extreme-value copula?

B. If C is an extreme-value copula, how can it be estimated?

N.B.: A rank-based estimate is useless for prediction purposes, but

- it can help select an appropriate family of extreme-value copulas;
- it can serve as a “golden standard” against which to test goodness-of-fit.

4. Testing for extremeness

Ghoudi et al. (1998) show how to test the hypothesis

H_0 : C is an extreme-value copula.

Their test is based on the probability integral transformation

$$(X, Y) \mapsto W = H(X, Y) = C(U, V).$$

The latter has been extensively studied by Genest & Rivest (1993, 2001), as well as Barbe et al. (1996).

Key observation

Under H_0 , the distribution of $W = H(X, Y)$ is given by

$$K(w) = \Pr(W \leq w) = w - (1 - \tau)w \log(w), \quad w \in (0, 1]$$

with

$$\tau = 4\mathbb{E}(W) - 1 = \int_0^1 \frac{t(1-t)}{A(t)} dA'(t).$$

Consequently,

$$\mu_i = \mathbb{E}(W^i) = \frac{i\tau + 1}{(i+1)^2}, \quad i \in \{1, 2, \dots\}.$$

Test statistic

If H_0 is true, one must have

$$-1 + 8E(W) - 9E(W^2) = 0.$$

This can be tested using the pseudo-observations

$$W_1 = H_n(X_1, Y_1), \dots, W_n = H_n(X_n, Y_n).$$

The test statistic proposed by Ghoudi et al. (1998) amounts to

$$S_n = -1 + \frac{8}{n} \sum_{i=1}^n W_i - \frac{9}{n} \sum_{i=1}^n W_i^2.$$

Procedure

Estimate the variance of S_n by the jackknife, viz.

$$\hat{\sigma}_n^2 = \frac{n-1}{n} \sum_{i=1}^n (S_n^{(-i)} - S_n)^2.$$

Because S_n is a U -statistic, an approximate P -value is given by

$$\Pr \left(|Z| > \frac{|S_n|}{\sigma_n} \right).$$

See Lee (1990) or Ghoudi et al. (1998) for details.

Alternative solution of Ben Ghorbal et al. (submitted)

To reduce the bias associated with the jackknife, estimate the finite- or large-sample variance of S_n .

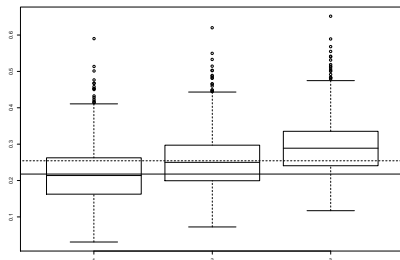
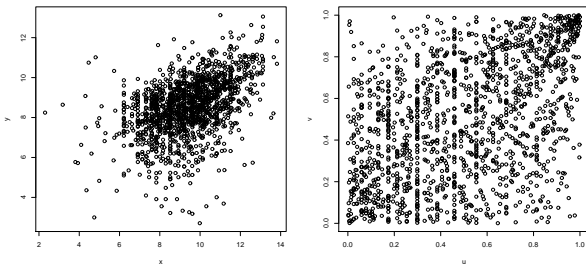


Illustration in the case of independence

Example

For the LOSS/ALAE data of Frees & Valdez (1998),

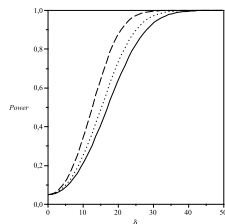
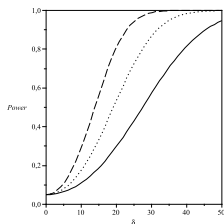
$$S_n \approx 0.059 \quad \text{and} \quad P\text{-value} \approx 95.26\%.$$



The Gumbel–Hougaard extreme-value copula fits very very well!

Limitations of the test

- 1.- It tends to be too liberal when the sample size is small (e.g., it isn't conclusive for oil/gas prices).
- 2.- The power depends on the true family of extreme-value copulas under the null hypothesis.



5. Nonparametric estimation with known margins

When the margins F and G are **known with certainty**, one can assume WLOG that the sample

$$(X_1, Y_1), \dots, (X_n, Y_n)$$

arises from some copula C (i.e., margins are uniform).

Under the hypothesis that C is an extreme-value copula, there are several possible estimates for the **Pickands dependence function**, A .

Choice of estimators

- Pickands (*Bull. Inter. Statist. Inst.*, 1981)
- Deheuvels (*Statist. Probab. Lett.*, 1991)
- Capéraà et al. (*Biometrika*, 1997)
- Hall & Tajvidi (*Bernoulli*, 2000)
- Jiménez et al. (*JMVA*, 2001)
- Segers (*Nova Science Publ.*, 2007)
- Zhang et al. (*JMVA*, 2007)

Basic fact leading to all these estimators

Recall $(U, V) = (F(X), G(Y)) \sim C$. Let

$$\xi(t) = \frac{-\log U}{1-t} \wedge \frac{-\log V}{t}, \quad t \in (0, 1).$$

Then $\xi(t) \sim \exp\{A(t)\}$, and hence

$$\begin{aligned} E\{\xi(t)\} &= 1/A(t), \\ E\{\log \xi(t)\} &= -\log A(t) - \gamma, \end{aligned}$$

where $\gamma = 0.5772\dots$ is Euler's constant.

Pickands and Capéraà–Fougères–Genest (CFG) estimators

Segers (2007) shows that they can be expressed in the form

$$1/A_n^P(t) = \frac{1}{n} \sum_{i=1}^n \xi_{i,n}(t)$$

and

$$\log A_n^{\text{CFG}}(t) = -\frac{1}{n} \sum_{i=1}^n \log \xi_{i,n}(t) - \gamma$$

as a function of

$$\xi_{i,n}(t) = \frac{-\log U_i}{1-t} \wedge \frac{-\log V_i}{t}, \quad i \in \{1, \dots, n\}.$$

Endpoint corrections to ensure that $\hat{A}(0) = \hat{A}(1) = 1$

Correction to Pickands' estimator proposed by Deheuvels (1991):

$$1/A_n^D(t) = 1/A_n^P(t) - (1-t)\{1/A_n^P(0) - 1\} - t\{1/A_n^P(1) - 1\}.$$

Simple endpoint corrections to the CFG estimator take the form

$$\log A_n^{\text{CFG}}(t) - p(t) \log A_n^{\text{CFG}}(0) - \{1 - p(t)\} \log A_n^{\text{CFG}}(1),$$

where $p(t) = 1 - t$ in Capéraà et al. (1997).

The **optimal choice** of $p(t)$ is identified by Segers (2007).

6. Proposed estimators when margins are unknown

For $i \in \{1, \dots, n\}$, let

$$U_{i,n} = R_i/n, \quad V_{i,n} = S_i/n,$$

and

$$\xi_{i,n}(t) = \frac{-\log U_{i,n}}{1-t} \wedge \frac{-\log V_{i,n}}{t}, \quad t \in (0, 1).$$

Rank-based versions of the Pickands and CFG estimators are:

$$\begin{aligned} 1/A_n^{\text{P}}(t) &= \frac{1}{n} \sum_{i=1}^n \xi_{i,n}(t), \\ \log A_n^{\text{CFG}}(t) &= -\frac{1}{n} \sum_{i=1}^n \log \xi_{i,n}(t) - \gamma. \end{aligned}$$

Reformulation in terms of Deheuvels' empirical copula

The **empirical copula** (Deheuvels, 1979) is defined by

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(U_{i,n} \leq u, V_{i,n} \leq v), \quad u, v \in [0, 1].$$

The limiting behavior of A_n^P and A_n^{CFG} depends on C_n , because

$$\begin{aligned} 1/A_n^P(t) &= \int_0^1 C_n(u^{1-t}, u^t) \frac{du}{u}, \\ \log A_n^{\text{CFG}}(t) &= \int_0^1 \{C_n(u^{1-t}, u^t) - \mathbf{1}(u > e^{-1})\} \frac{du}{u \log u} - \gamma. \end{aligned}$$

Result from Stute (1984) and Tsukahara (2005)

Let α be a C -Brownian bridge, i.e., a centered Gaussian random field on $[0, 1]^2$ with covariance function

$$\text{cov}\{\alpha(u, v), \alpha(u', v')\} = C(u \wedge u', v \wedge v') - C(u, v)C(u', v').$$

If C has continuous partial derivatives of second order, then

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

where the limit is a centered Gaussian process on $[0, 1]^2$, viz.

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

Consequences

If A is twice continuously differentiable, then as $n \rightarrow \infty$, one finds

$$\begin{aligned}\mathbb{A}_n^P &= \sqrt{n}(A_n^P - A) \rightsquigarrow \mathbb{A}^P, \\ \mathbb{A}_n^{\text{CFG}} &= \sqrt{n}(A_n^{\text{CFG}} - A) \rightsquigarrow \mathbb{A}^{\text{CFG}},\end{aligned}$$

in the space $\mathcal{C}([0, 1])$, where for all $t \in [0, 1]$,

$$\begin{aligned}\mathbb{A}^P(t) &= -A^2(t) \int_0^1 \mathbb{C}(u^{1-t}, u^t) \frac{du}{u}, \\ \mathbb{A}^{\text{CFG}}(t) &= -A(t) \int_0^1 \mathbb{C}(u^{1-t}, u^t) \frac{du}{u \log u}.\end{aligned}$$

Observations

1. The proof of this result:
 - is **not** a simple corollary of the Continuous Mapping Theorem;
 - uses strong approximations given by Tsukahara (2000);
 - requires a new result on weighted bivariate empirical processes.
2. Endpoint corrections make **no difference** asymptotically.
3. Limiting covariance functions depend on A and can be estimated consistently from the data.
4. Construction of asymptotic confidence intervals and bands is thus possible.

7. Efficiency comparisons

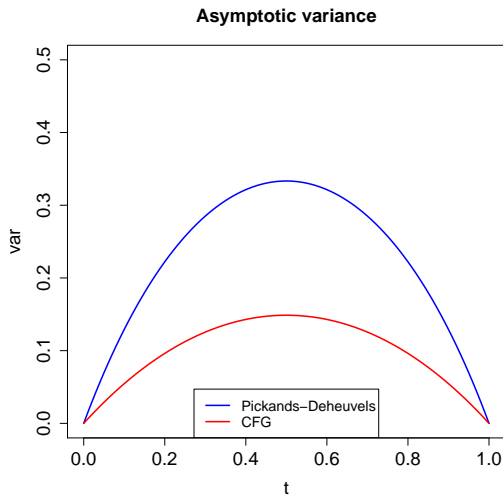
X and Y are independent if and only if $A \equiv 1$. In that case,

$$\begin{aligned}\mathbb{A}_n^{\text{P}}(t) &\rightsquigarrow \mathcal{N}(0, \sigma_{\text{P}}^2), \\ \mathbb{A}_n^{\text{CFG}}(t) &\rightsquigarrow \mathcal{N}(0, \sigma_{\text{CFG}}^2),\end{aligned}$$

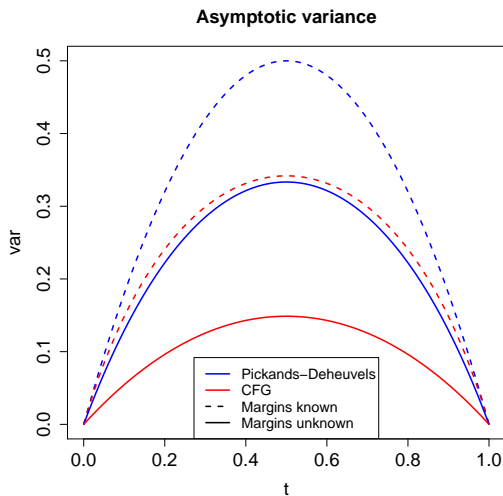
where

$$\begin{aligned}\sigma_{\text{P}}^2 &= \frac{3t(1-t)}{(2-t)(1+t)}, \\ \sigma_{\text{CFG}}^2 &= 2 \int_0^1 \int_0^v (1-v^{1-t})(1-v^t) \frac{du}{\log u} \frac{dv}{v \log v}.\end{aligned}$$

Comparison between Pickands–Deheuvels and CFG



Comparison with the case of known marginals



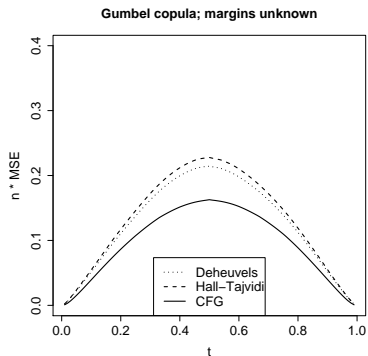
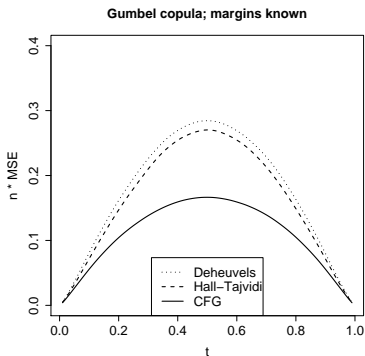
Other dependence structures

Extensive simulations imply that in terms of efficiency:

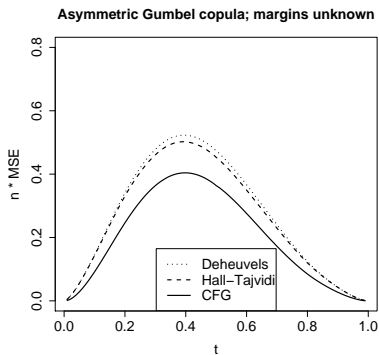
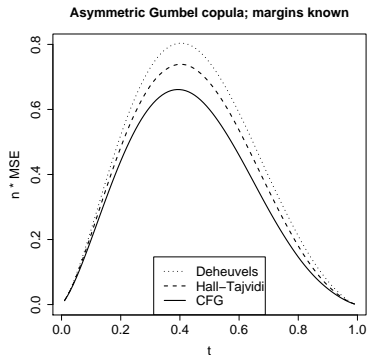
- CFG \succ Pickands–Deheuvels \approx Hall-Tajvidi;
- Rank-based \succ Known margins.

Henmi (2005) documents the paradoxical effect of nuisance parameters on the efficiency of estimators.

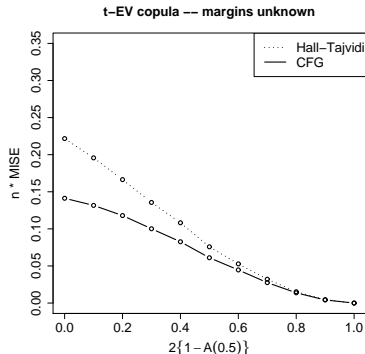
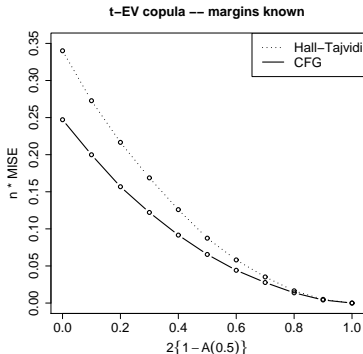
Gumbel copula with $\lambda \approx 0.25$



Asymmetric Gumbel copula with $\lambda \approx 0.25$



MISE for the Cauchy extreme-value copula



Graph of $100 \times MISE$ as a function of $\lambda = 2\{1 - A(1/2)\}$

8. Final remarks

- Rank-based versions of nonparametric estimators have been proposed for the Pickands dependence function of a bivariate extreme-value copula.
- The new estimators are asymptotically normal and unbiased.
- The CFG estimator is superior to the Pickands estimator and variants thereof.
- Even when margins are known, more efficient procedures arise if information about margins is ignored and ranks are used instead.
- Goodness-of-fit tests for extreme-value copulas are under development.

Acknowledgements

This talk is based in part on joint work with Johan Segers, Johanna Nešlehová, and Noomen Ben Ghorbal.

Funding in support of this work was provided by:

- the Natural Sciences and Engineering Research Council of Canada;
- the Fonds québécois de la recherche sur la nature et les technologies;
- the Institut de finance mathématique de Montréal.