# Joint modelling of the bulk and tail of bivariate data

Lídia André

Jennifer Wadsworth and Adrian O'Hagan

23rd EYSM, 12 September 2023

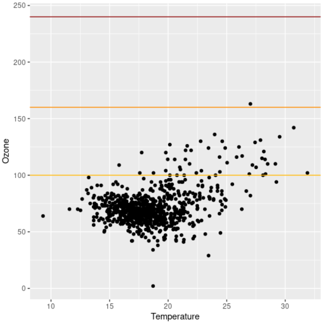
EPSRC
Engineering and Physical Sciences Research Council

STOR-i | Lancaster University

UCD DUBLIN

# Motivation

Interest not only in the extremes but also the bulk of the distribution - *e.g.* environmental applications



# Univariate Framework

There have been proposed parametric, semi-parametric and non-parametric models

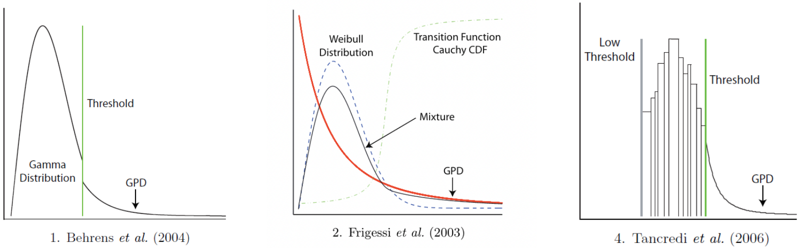


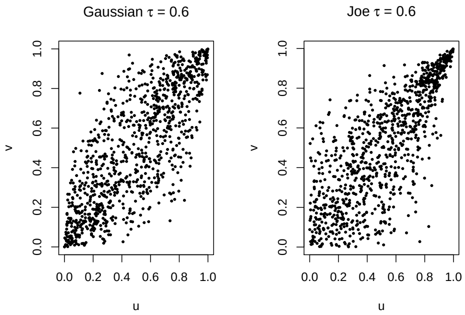

Figure 1: Taken from Scarrott and MacDonald (2012)

# Copulas

In a multivariate setting we are also concerned about the dependence between variables.

A copula $C$ is a joint distribution of a random vector $(X_1, \ldots, X_d)$

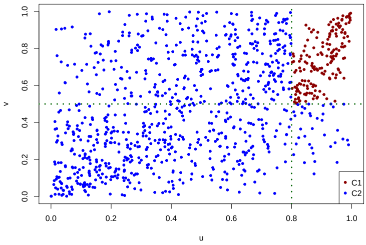
$$F(x_1, \ldots, x_d) = C\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right), \quad d \geq 2$$



# Multivariate Framework

Aulbach et al. (2012) model the full data set by fitting one copula to the body and another to the upper tail

- It sometimes doesn't offer a smooth transition between the two copulas
- It requires the choice of thresholds
- The likelihood of the model doesn't have a closed form so no inference was done



# Weighted Copula Model

For $(u^*, v^*) \in [0,1]^2$, we define the density $c^*$ as

$$c^*(u^*, v^*; \gamma) = \frac{\pi(u^*, v^*; \theta) c_t(u^*, v^*; \boldsymbol{\alpha}) + [1 - \pi(u^*, v^*; \theta)] c_b(u^*, v^*; \boldsymbol{\beta})}{K(\gamma)}$$

[1] For more details see André et al. (2023)

# Weighted Copula Model

For $(u^*, v^*) \in [0,1]^2$, we define the density $c^*$ as

$$c^*(u^*, v^*; \gamma) = \frac{\pi(u^*, v^*; \theta) c_t(u^*, v^*; \boldsymbol{\alpha}) + [1 - \pi(u^*, v^*; \theta)] c_b(u^*, v^*; \boldsymbol{\beta})}{K(\gamma)}$$

- $c_t, \; c_b \rightarrow$ copula densities tailored to the tail and body, respectively.

[1] For more details see André et al. (2023)

# Weighted Copula Model

For $(u^*, v^*) \in [0,1]^2$, we define the density $c^*$ as

$$c^*(u^*, v^*; \gamma) = \frac{\pi(u^*, v^*; \theta) c_t(u^*, v^*; \boldsymbol{\alpha}) + [1 - \pi(u^*, v^*; \theta)] c_b(u^*, v^*; \boldsymbol{\beta})}{K(\gamma)}$$

- $c_t, \ c_b \rightarrow$ copula densities tailored to the tail and body, respectively.
- $\pi(u^*, v^*; \theta) \rightarrow$ dynamic weighting function, defined in $[0,1]^2$ and increasing in $u^*$ and $v^*$

[1] For more details see André et al. (2023)

# Weighted Copula Model

For $(u^*, v^*) \in [0,1]^2$, we define the density $c^*$ as

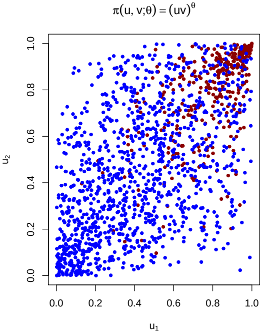
$$c^*(u^*, v^*; \gamma) = \frac{\pi(u^*, v^*; \theta) c_t(u^*, v^*; \boldsymbol{\alpha}) + [1 - \pi(u^*, v^*; \theta)] c_b(u^*, v^*; \boldsymbol{\beta})}{K(\gamma)}$$

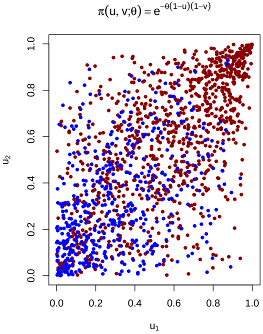
- $c_t, \; c_b \rightarrow$ copula densities tailored to the tail and body, respectively.
- $\pi(u^*, v^*; \theta) \rightarrow$ dynamic weighting function, defined in $[0,1]^2$ and increasing in $u^*$ and $v^*$
- $\boldsymbol{\gamma} = (\theta, \boldsymbol{\alpha}, \boldsymbol{\beta}) \rightarrow$ vector of model parameters
- $K(\boldsymbol{\gamma}) \rightarrow$ normalising constant [1]

[1] For more details see André et al. (2023)

# Weighted Copula Model

- Doesn't require a choice of threshold
- Offers a smooth transition between the body and tail copulas
- However, it is also hard to perform inference on it



# Inference

The inference on the model was achieved by fitting the copula of the density $c^*$ via numerical integration as follows

$$c(u, v; \gamma) = \frac{c^* \left( F_{U^*}^{-1}(u), F_{V^*}^{-1}(v); \gamma \right)}{f_{U^*} \left( F_{U^*}^{-1}(u) \right) f_{V^*} \left( F_{V^*}^{-1}(v) \right)}$$

where

$$F_{U^*}(u^*) = P[U^* \leq u^*] = \int_0^{u^*} \int_0^1 c^*(u, v) \mathrm{d}v \mathrm{d}u$$

$$f_{U^*}(u^*) = \int_0^1 c^*(u^*, v) \, \mathrm{d}v, \qquad v \in (0, 1)$$

# Extremal Dependence Properties

It is important to know if extreme values of the variables are likely to occur together (**asymptotic dependence**) or not (**asymptotic independence**)

$$\chi = \lim_{r \to 1} P\left[F_Y(y) > r \mid F_X(x) > r\right],$$

$$P\left[F_Y(y) > r \mid F_X(x) > r\right] \sim \mathcal{L}(1-r)(1-r)^{\frac{1}{\eta}-1} \quad \text{as } r \to 1$$

- Asymptotic Dependence (AD): $\chi > 0$ and $\eta = 1$
- Asymptotic Independence (AI): $\chi = 0$ and $\eta \neq 1$

# Extremal Dependence Properties

Depending on the weighting function used, $c_b$ has an influence in $\chi$ in some cases:

- If $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$ and $c_t$ is AD, $\chi$ is dominated by $\chi_t$ with an influence of $\chi_b$
- If $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$ and $c_t$ is AI, $\chi$ is that from $c_t$
- If $\pi(u^*, v^*; \theta) = \exp\{-\theta(1 - u^*)(1 - v^*)\}$, $\chi$ is that from $c_t$ (independently of the nature of $c_t$)

# Extremal Dependence Properties

When $c_b$ is a Frank copula (AI) with parameter $\beta \in \mathbb{R}$, $c_t$ is a Gumbel copula (AD) with parameter $\alpha > 1$, and $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$, $\theta > 0$,

$$\chi = \frac{2 - 2^{1/\alpha}}{1 + \beta\left(1 - \exp\{-\beta\}\right)^{-1} \int_0^1 (1 - (v^*)^\theta) e^{-\beta(1-v^*)} \mathrm{d}v^*}$$

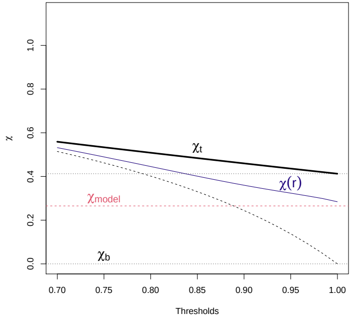
and $\eta = 1$

If $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$,

$$\chi = 2 - 2^{1/\alpha} \quad \text{and} \quad \eta = 1$$

($\chi_b = 0$, $\eta_b = 0.5$, $\chi_t = 2 - 2^{1/\alpha}$ and $\eta_t = 1$)

# Extremal Dependence Properties



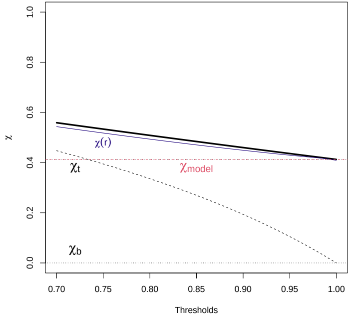




Figure 2: Weight functions: $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$ (left) and $\pi(u^*, v^*; \theta) = \exp\{-\theta(1 - u^*)(1 - v^*)\}$ (right) with $\gamma = (1.5, 2, 3.488889)$

# Extremal Dependence Properties

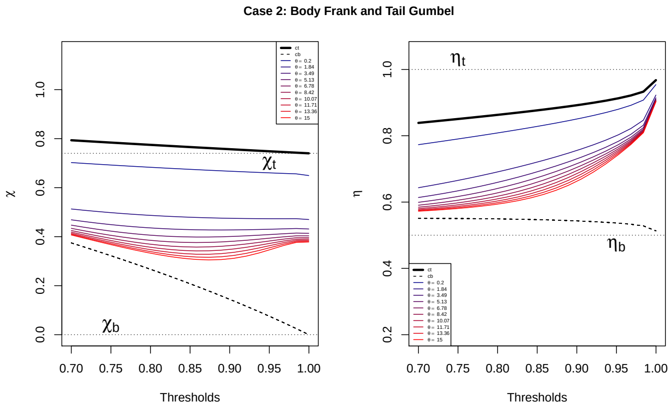



Figure 3: Weight function: $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$.

# Extremal Dependence Properties

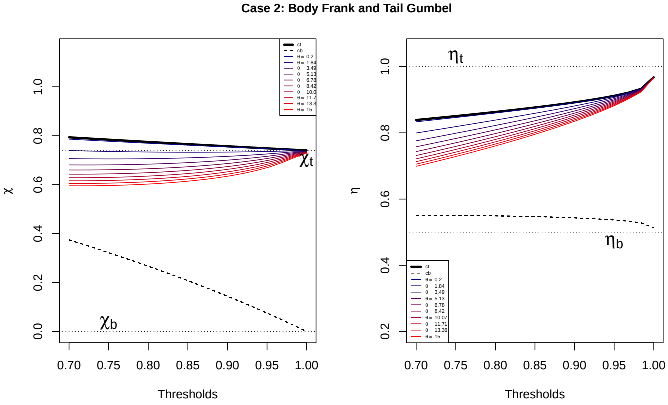


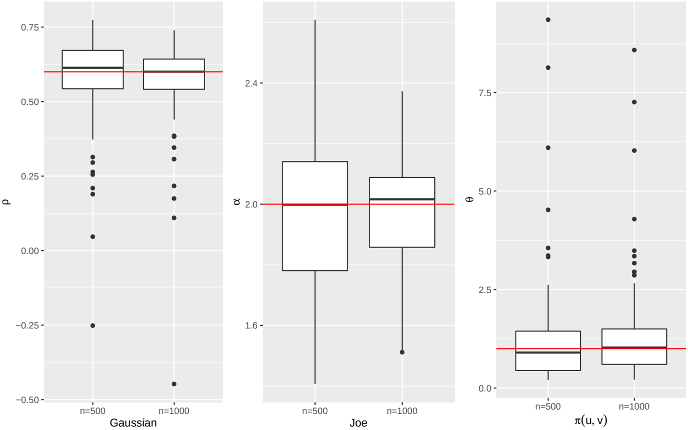

Figure 4: Weight function: $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$.

# Parameter Estimation

Simulation setup:

- $c_t$ : Gaussian copula with $\rho = 0.6$
- $c_b$ : Joe copula with $\alpha = 2$
- $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$ with $\theta = 1$
- $n = 500$ and $n = 1000$
- 100 repetitions

# Parameter Estimation

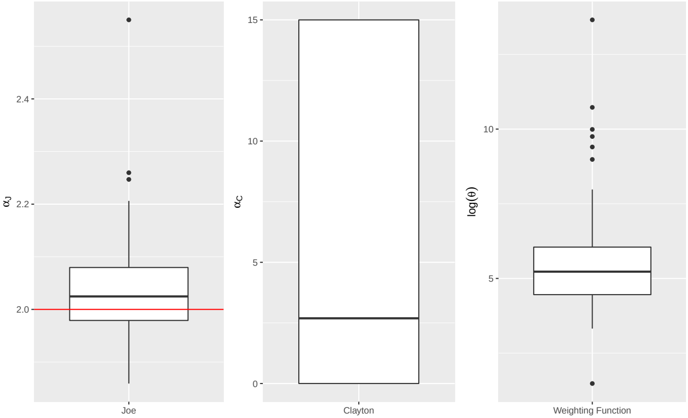


# Model Misspecification

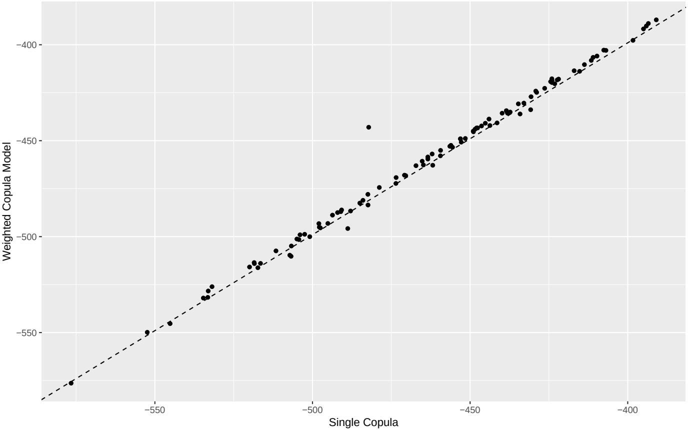
Simulation setup:

- True data from a Joe copula with $\alpha = 2$
- $c_t$ : Clayton copula
- $c_b$ : Joe copula (true)
- $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$
- $n = 1000$
- 100 repetitions

# Model Misspecification



# Model Misspecification

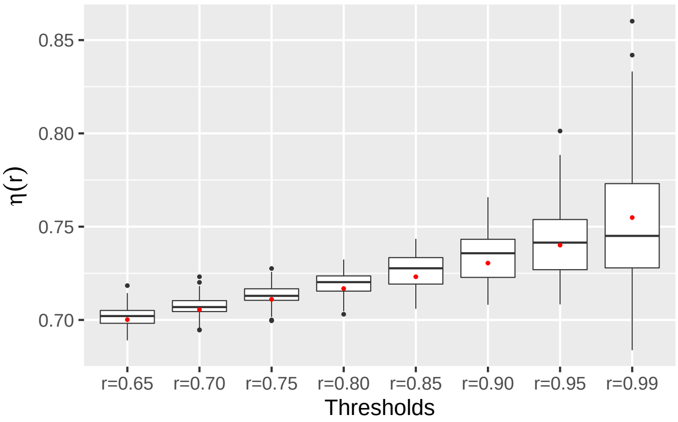


# Model Misspecification

Simulation setup:

- True data from a Gaussian copula with $\rho = 0.65$
- Models considered:
  1. $c_t$ : Joe copula; $c_b$ : Frank copula
  2. $c_t$ : Hüsler-Reiss copula; $c_b$ : Clayton copula
  3. $c_t$ : **Inverted Gumbel copula;** $c_b$ : **Student t copula** $\rightarrow$ best average AIC
  4. $c_t$ : Coles-Tawn copula; $c_b$ : Galambos copula
- $\pi(u^*, v^*; \theta) = (u^* v^*)^{\theta}$
- $n = 1000$
- Each model was fitted 50 times

# Model Misspecification



# Ozone and Temperature Data

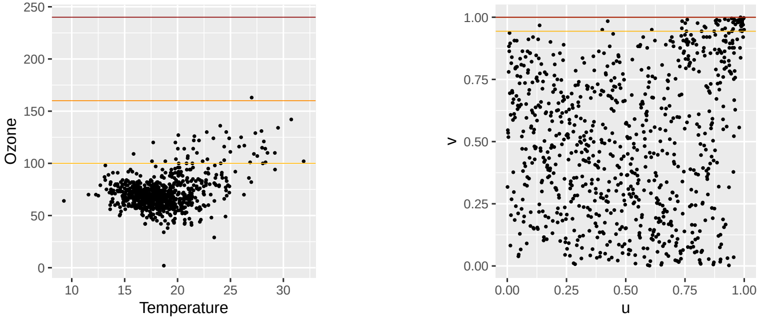
- Temperature may influence the levels of Ozone concentration in the air
- The legal thresholds for $O_3$ levels in the UK might then be found in the body and not just in the tails of the data

UK legal thresholds:

| Levels | Low | Moderate | High | Very High |
|---|---|---|---|---|
| $O_3\,(\mu g/m^3)$ | $[0, 100]$ | $[101, 160]$ | $[161, 240]$ | $> 240$ |

We applied our model to the summers between 2011 and 2019 of Blackpool, UK

# Ozone and Temperature Data

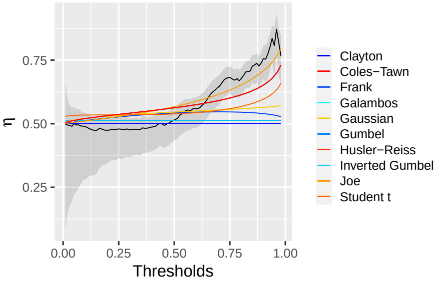



Apart from the upper tail, the variables seem to be negative correlated

# Ozone and Temperature Data

Fitting a single copula

| Copula | AIC |
|---|---|
| Clayton | 2.0 |
| Gaussian | -28.6 |
| Frank | -15.8 |
| Joe | **-143.6** |
| Gumbel | -97.4 |
| Student t | -52.8 |
| Inverted Gumbel | 0.1 |
| Hüsler-Reiss | -99.1 |
| Coles-Tawn | -99.0 |
| Galambos | -95.9 |




None of the single copulas showed negative correlation

# Ozone and Temperature Data

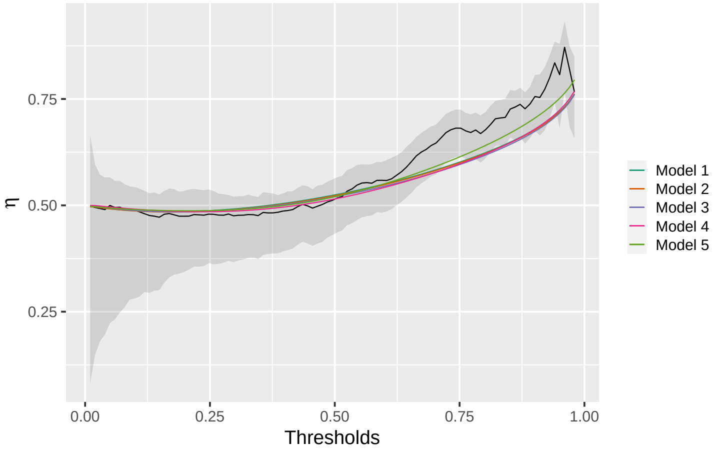
Fitting the weighted copula model with

$$\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$$

| Model | | Parameters | | | AIC |
|---|---|---|---|---|---|
| $c_b$ | $c_t$ | $\hat{\beta}$ | $\hat{\boldsymbol{\alpha}}$ | $\hat{\theta}$ | |
| Gaussian | Hüsler-Reiss | -0.40 | 1.24 | 0.35 | -176.1 |
| Gaussian | Galambos | -0.41 | 0.79 | 0.34 | -172.1 |
| Gaussian | Coles-Tawn | -0.33 | 0.35, 2.86 | 0.43 | -158.4 |
| Frank | Coles-Tawn | -2.52 | 0.33, 4.80 | 0.37 | -163.2 |
| Frank | Joe | -4.11 | 1.61 | 0.18 | **-184.9** |

The models with the best AIC all show negative correlation in the copulas tailored to the body

# Ozone and Temperature Data



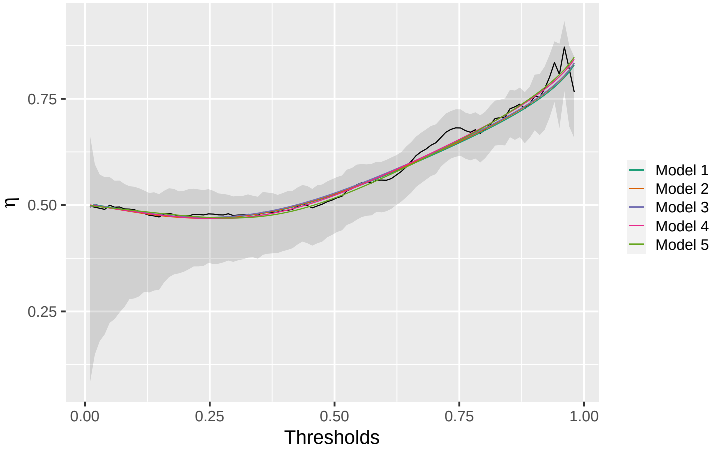
# Ozone and Temperature Data

Fitting the weighted copula model with

$$\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$$

| Model | | Parameters | | | AIC |
|---|---|---|---|---|---|
| $c_b$ | $c_t$ | $\hat{\beta}$ | $\hat{\boldsymbol{\alpha}}$ | $\hat{\theta}$ | |
| Gaussian | Hüsler-Reiss | -0.74 | 1.33 | 3.32 | **-240.1** |
| Gaussian | Galambos | -0.72 | 0.90 | 3.55 | -237.2 |
| Gaussian | Coles-Tawn | -0.74 | 0.85, 0.79 | 3.25 | -234.8 |
| Frank | Coles-Tawn | -4.51 | 0.87, 1.02 | 4.33 | -235.7 |
| Frank | Joe | -6.49 | 1.72 | 2.45 | -232.9 |

# Ozone and Temperature Data



# Ozone and Temperature Data

Other diagnostics

| Models | Kendall's $\tau$ | $P[T \geq 24, O_3 \geq 100]$ | $P[O_3 \geq 100 \mid 22 \leq T \leq 23]$ |
|---|---|---|---|
| Empirical | 0.0812 | 0.0302 | 0.1330 |
| (95% CI) | (0.0173 , 0.1867) | (0.0147 , 0.0544) | (0.0227 , 0.1944) |
| Model 1 | 0.0690 | 0.0246 | 0.1441 |
| Model 2 | 0.0663 | 0.0250 | 0.1412 |
| Model 3 | 0.0770 | 0.0251 | 0.1429 |
| Model 4 | 0.0779 | 0.0262 | 0.1392 |
| Model 5 | 0.0718 | 0.0267 | 0.1366 |

# Conclusions

- Our model provides a better fit than just fitting a single copula to the data
- It is flexible - it is able to capture different structures within the same data set
- However, it is computationally expensive
- Further Steps:
  - Account for non-stationarity - incorporate covariates

# Questions?

Thank you all for listening!

# References I

André, L. M., Wadsworth, J. L., and O'Hagan, A. (2023). Joint modelling of the body and tail of bivariate data. *Computational Statistics & Data Analysis*.

Aulbach, S., Bayer, V., and Falk, M. (2012). A Multivariate Piecing-Together Approach with an Application to Operational Loss Data. *Bernoulli*, 18:455–475.

Scarrott, C. and MacDonald, A. (2012). A Review of Extreme Value Threshold Estimation and Uncertainty Quantification. *Revstat Statistical Journal*, 10:33–60.

## $\chi$ when $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$

$$c_1 = 2 - 2^{1/\alpha} = \chi_{Gumbel},$$

$$c_2 = (2^{1/\alpha} - 1 - C_\alpha)(\theta - 1),$$

$$c_3 = \beta\theta \left(1 - \exp\{-\beta\}\right)^{-1},$$

$$c_4 = 1,$$

$$c_5 = -\theta/2 + o\left((1-r)^2\right), \quad \text{as } r \to 1$$

$$c_6 = \beta \left(1 - \exp\{-\beta\}\right)^{-1} \int_0^1 (1 - (v^*)^\theta) e^{-\beta(1-v^*)} \mathrm{d}v^*,$$

$$c_7 = -\frac{1}{2} \int_0^1 B_{v^*,\beta,\theta} \mathrm{d}v^*$$

with

$$B_{v^*,\beta,\theta} = \frac{2\beta^2(1-(v^*)^\theta)(1-\exp\{-\beta v^*\})\exp\{-2\beta(1-v^*)\}}{(1-\exp\{-\beta\})^2} - \frac{\beta\theta(v^*)^\theta \exp\{-\beta(1-v^*)\}}{1-\exp\{-\beta\}} - \frac{\beta^2(1-(v^*)^\theta)\exp\{-\beta(1-v^*)\}}{1-\exp\{-\beta\}}$$

## $\chi$ when $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$

$$
\begin{aligned}
\chi &= \lim_{r \to 1} P\left[F_Y(y) > r \mid F_X(x) > r\right] \\
&= \lim_{r \to 1} \frac{c_1(1-r) + c_2(1-r)^2 + c_3(1-r)^3 + o\left((1-r)^3\right)}{c_4(1-r) + c_5(1-r)^2 + c_6(1-r) + c_7(1-r)^2 + o\left((1-r)^2\right)} \\
&= \lim_{r \to 1} \left( \frac{c_1}{c_4 + c_6} + \left[ \frac{c_2 - c_1(c_5 + c_7)}{(c_4 + c_6)^2} \right] (1-r) + \mathcal{O}\left((1-r)^2\right) \right) \\
&= \frac{c_1}{c_4 + c_6} = \frac{2 - 2^{1/\alpha}}{1 + \beta\left(1 - \exp\{-\beta\}\right)^{-1} \int_0^1 (1 - (v^*)^\theta) e^{-\beta(1-v^*)} \mathrm{d}v^*}
\end{aligned}
$$

## $\chi$ when $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$

$$
\begin{aligned}
c_1 &= 2 - 2^{1/\alpha} = \chi_{Gumbel}, \\
c_2 &= 1, \\
c_3 &= \frac{1}{\alpha}, \\
c_4 &= -\frac{1}{2}\int_0^1 A_{v^*,\beta,\theta} \mathrm{d}v^*
\end{aligned}
$$

with

$$
\begin{aligned}
A_{v^*,\beta,\theta} = &- \frac{2\beta^2(1-\exp\{-\beta\})\exp\{-2\beta(1-v^*)\}}{(1-\exp\{-\beta\})^2} - \frac{\beta\theta(1-v^*)\exp\{-\beta(1-v^*)\}}{1-\exp\{-\beta\}} \\
&+ \frac{\beta^2\exp\{-\beta(1-v^*)\}}{1-\exp\{-\beta\}}
\end{aligned}
$$

## $\chi$ when $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$

$$
\begin{aligned}
\chi &= \lim_{r\to 1} P\left[F_Y(y) > r \mid F_X(x) > r\right] \\
&= \lim_{r\to 1} \frac{c_1(1-r) + o\left((1-r)^2\right)}{c_2(1-r) + c_3(1-r)^2 + c_4(1-r)^2 + o\left((1-r)^2\right)} \\
&= \lim_{r\to 1} \left(\frac{c_1}{c_2} - \frac{c_3 + c_4}{c_2^2}(1-r) + \mathcal{O}\left((1-r)^2\right)\right) = \frac{c_1}{c_2} = 2 - 2^{1/\alpha}
\end{aligned}
$$