Joint Modelling of the Body and Tail of Multivariate Data

Lídia André [1] Jennifer Wadsworth [2] Adrian O'Hagan [3]

[1]STOR-i CDT, Lancaster University, UK
[2]Department of Mathematics and Statistics, Lancaster University, UK
[3]University College Dublin, School of Mathematics and Statistics, Ireland

STOR-i excellence with impact

Lancaster University | UKRI | Engineering and Physical Sciences Research Council

1. Motivation

When **both** extreme and non-extreme data are of interest, modelling the **whole** distribution accurately is important.

However, usually the model fits the **body** well or Extreme Value Theory is employed to model the **tails**.

In a **univariate** framework, there are statistical methods available but little work has been done in a **multivariate** setting.

2. Copulas

In a multivariate setting, measuring and modelling the dependence between variables may be of interest.

Copulas are joint distributions that describe the dependence between variables independently from their marginal structure:

$$H(x, y) = C\left(F_X(x), F_Y(y)\right),$$

where $H(\cdot)$ is the joint distribution of (X, Y), $F(\cdot)$ is the marginal distribution and $C(\cdot, \cdot)$ represents the copula.

An **advantage** of copulas is that they allow us to assess the dependence structure in both the joint body and tail of a data set.

3. Extremal Dependence

When the focus is on extreme values, assessing if the variables are extreme **together** or not may also be of interest. They are **asymptotically dependent** if joint extremes occur at a similar frequency to marginal extremes or **asymptotically independent** otherwise.

Dependence Measures

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

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

where $\mathcal{L}$ is a slowly-varying function and $\eta \in (0, 1]$ is the coefficient of tail dependence [1].

(X, Y) are asymptotically **dependent** if $\chi > 0$ and $\eta = 1$ and asymptotically **independent** otherwise.

4. Proposed Model

Our model blends **two different copulas** over the whole range of the support. One copula, c_t, is tailored to the extremes and the other, c_b, is tailored to the body.

We combine these two densities to define a new density by means of a **dynamic** weighting function, in this case $\pi(u, v) = (uv)^\theta$ with $\theta > 0$:

$$c^*(u, v) = \frac{\pi(u, v)c_t(u, v) + [1 - \pi(u, v)]c_b(u, v)}{K},$$

where K is a normalising constant, $u = F_X(x)$ and $v = F_Y(y)$. We use numerical methods to fit the copula of the density c^* to data. This model has the advantage of not requiring an arbitrary choice of threshold.

5. Application to Ozone Pollution Data

From a public health perspective, we not only want to learn about the probability of exceeding harmful yet locally moderate pollutant levels, but also the probability of exceeding extreme and potentially more dangerous levels. Additionally, ozone concentration in the air seems to be dependent on temperature [2].

We applied our method to model the dependence between ozone air concentration and temperature from the summers of 2011 to 2019 in Weybourne, UK.

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

Table 1: Daily Air Quality Index (DAQI) for O_3 concentrations in the UK.

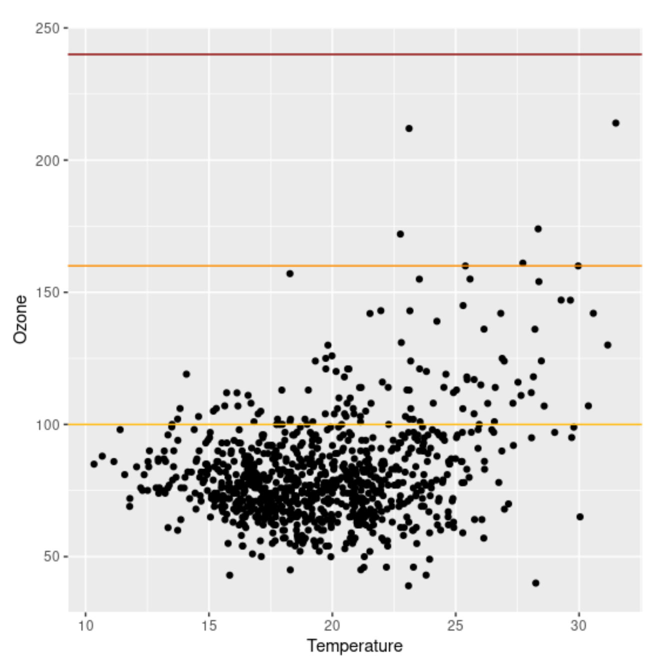
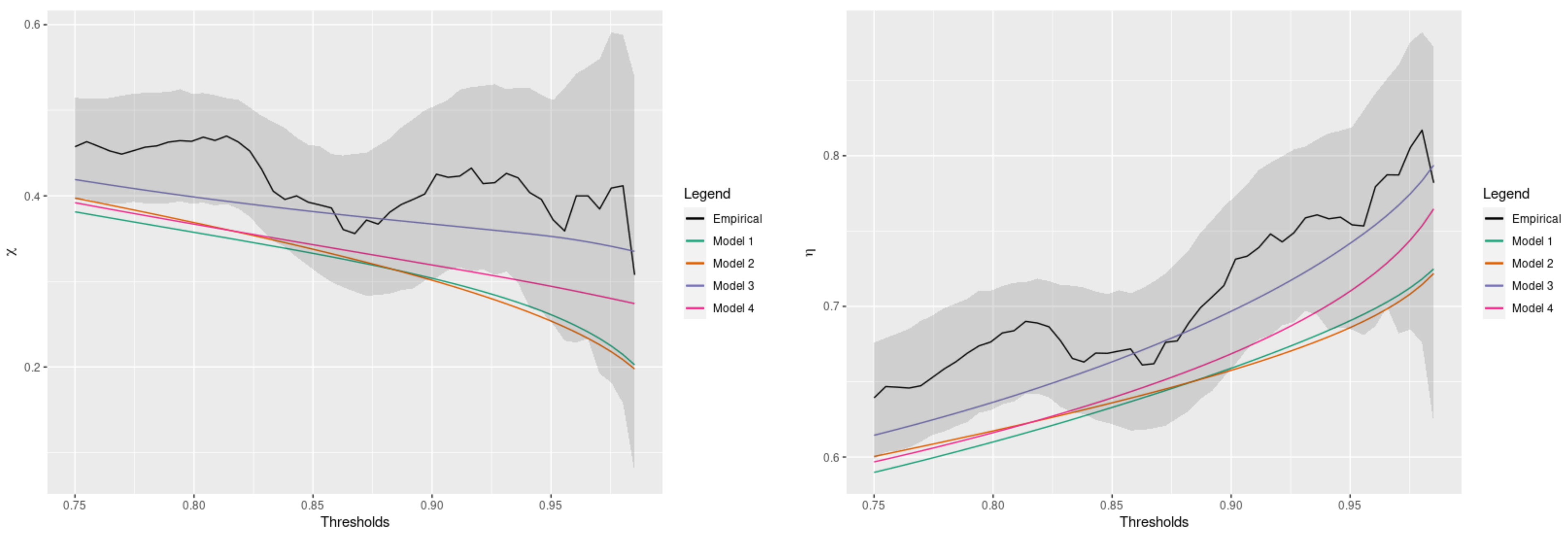


Figure 1: Scatterplot of Daily Maxima of Temperature and Daily Maxima of O_3.

	c_t	c_b
Model 1	Gaussian	Clayton
Model 2	Student t	Frank
Model 3	Gumbel	Joe
Model 4	Coles-Tawn-Dirichlet [3]	Gaussian

Tail Diagnostics



Other Diagnostics

	AIC	Kendall's τ	$P[T \leq 20, O_3 \geq 100]$	$P[T \geq 25, O_3 \geq 160]$
Empirical	-	0.0972	0.0359	0.0034
(**95%** CI)		(0.0507 , 0.1411)	(0.0269 , 0.0493)	(0.0011 , 0.0067)
Model 1	-99.841	0.0921	0.0514	0.0032
Model 2	-113.349	0.0740	0.0492	0.0040
Model 3	-114.713	0.1473	0.0413	0.0060
Model 4	-124.383	0.0749	0.0498	0.0034

Comments

- $\chi > 0$ and $\eta \neq 1$, so Ozone and Temperature are **asymptotic independent** but with positive association ($\tau > 0$).
- **Model 3** seems to capture best the behaviour in the tails; however overall all models seem to behave relatively well
- Stationarity is assumed throughout - relaxing this is an interesting extension to this work

References

[1] Ledford, A. W., and Tawn, J. A. (1996). Statistics for Near Independence in Multivriate Extreme Values. *Biometrika, 83*, 169-187.

[2] Gouldsbrough, L., Hossaini, R., Eastoe, E. and Young, P. J. (2022). A Temperature Dependent Extreme Value Analysis of UK Surface Ozone, 1980-2019. *Atmospheric Environment, 273*.

[3] Coles, S. G. and Tawn, J. A. (1991). Modelling Extreme Multivariate Events. *Journal of the Royal Statistical Society, Series B, 53*, 377-392.

 l.andre@lancaster.ac.uk