# Modelling and inference for the body and tail regions of multivariate data

Lídia Maria Branco Correia Martins André,

B.Sc., M.Sc, M.Res


Submitted for the degree of Doctor of Philosophy

at Lancaster University.

January 2025

STOR-i

excellence with impact

# Abstract

When an accurate representation of multivariate data is required across both the body (described by non-extreme observations) and the tail (defined by the extreme observations) regions, it is crucial to have a model that is able to characterise the joint behaviour across both regions. In this thesis, we focus on developing dependence models that represent the entire distribution without the need to explicitly define each region.

We propose two dependence models that fit the body and tail regions. For the first model, we construct the copula from a mixture of two copulas that are defined on the whole support of the data, and blended through a dynamic increasing weighting function. In this way, we give more weight to a copula tailored to the body for lower values, and more weight to a copula tailored to the tail for larger values. This ensures that there is a smooth transition between the two regions. For the second model, we construct a copula model based on a standard mixture of Gaussian distributions. As opposed to the first model, we avoid choosing a priori which copula families to include in the model, and are only required to determine the number of mixture components in the model. Moreover, we show that it scales relatively well to dimensions beyond the bivariate case. For both models, we derive (sub-)asymptotic dependence properties for specific model configurations, and show that they are flexible in capturing a broad range of extremal dependence structures through simulation studies.

Motivated by the computational resources required to evaluate the likelihood func-

tion of the proposed models, we also explore likelihood-free approaches that use neural networks to perform inference. In particular, we assess the performance of neural Bayes estimators in estimating the model parameters, both for one of the models introduced for the joint body and tail, and further complex extremal dependence models. We also propose using neural networks as classifiers for model selection. In this way, we provide a toolbox for simple fitting and model selection of complex extremal dependence models.

Methods to estimate extremal probabilities of complex environmental phenomena are presented; these result from participation in a challenge at the 2023 EVA conference. We propose using generalised additive models as well as a conditional extremes approach to estimate such quantities.

# Acknowledgements

First and foremost, I would like to say a huge thank you to my academic supervisor Jenny Wadsworth. Thank you for all your patience, kindness and support, I couldn't have asked for a better, more knowledgeable, mentor to guide me throughout the PhD. Thank you for not giving up on me and for believing in me, especially when I don't (which is often the case). I feel very lucky to have had the opportunity to work with you. Thank you!

I would also like to say thank you to Jon Tawn. Thank you for all your help and constant support and guidance throughout these four years. Your knowledge about everything stats related is insane and I am always in awe, it was a privilege getting to know you! I am also very grateful to Raphaël Huser (KAUST) for all his help with the work of Chapter 5 of this thesis. I would also like to acknowledge Adrian O'Hagan (UCD) for his contribution in the first year of my PhD, and for welcoming me in Dublin when I visited.

I was fortunate enough to complete my PhD at the EPSRC-funded STOR-i Centre for Doctoral Training and I would like to thank the STOR-i students and staff members for making my experience better. Moving to a different country is never easy, and doing so whilst in the middle of a world pandemic is even scarier. I kept thinking that I would never fit in, that no one would understand me, and that I would end up feeling lonely for the duration of my PhD. I'm glad to say now, four years later, that none of this happened, and I have to thank a few of people for that. I will start by acknowledging

my cohort: it's hard to do the MRes all online, but you made it somewhat easier thanks to the few games nights we had, the online coffee chats, or even the random text message sent to the group chat. A special thanks goes to the extremes research group without whom my PhD experience would have not been the same: thank you for all your support, I will especially miss our coffee chats on Tuesday mornings! I would like to particularly thank my closest friends from Lancaster: Maddie, Rebecca, Eleanor, Carla, Lydia, Katie and Owen. From book club, baking every Saturday, swimming to all the meals and conversations, I appreciate every single one of you. Thank you for your friendship, kindness, and your constant patience and support, I don't think I could have made it without you.

Last but certainly not least I would like to thank my friends and family, and I would like to do so in Portuguese. Costuma-se dizer que só as amizades mais fortes sobrevivem quando alguém muda de país, e por isso considero-me sortuda por vos continuar a ter na minha vida. Inês – quem diria que de MAEG saíria uma amizade tão boa. Obrigada pelo teu constante apoio nestes últimos dez anos, por todas as viagens e concertos, e por todas as chamadas que serviram para desabafar e falar das minhas irritações semanais. Juntamente contigo, Mariana, obrigada por me fazerem sentir melhor sempre que estava mais em baixo, por todas as noites de jogos, e conversas pela noite adentro, sempre que fui a casa. Sou muito grata pela vossa amizade! Sofia – obrigada por me aturares há 25 anos, consegues acreditar?! Um quarto de século! Não podia ter pedido para crescer ao lado de melhor pessoa, és quem me conhece melhor e, mesmo com todos os meus defeitos, continuas aqui. Muito obrigada por todo o teu apoio e amizade, mal posso esperar para ver o que os próximos 25 anos nos irão trazer! Inês – a única que consegue compreender o meu amor por Bologna! A tua paixão pelo ginásio é tão contagiante, que até a mim me influenciaste a (pelo menos) tentar. Obrigada por estares aqui e genuinamente te interessares por mim, e validares todos as minhas preocupações, não sei o que seria de mim sem ti.

Mãe e Pai – nada disto teria sido possível sem vocês. Obrigada por me incentivarem e inspirarem a prosseguir com a minha educação, e mais importante, obrigada por terem feito acontecer. Tenho imensa sorte em vos ter como os meus pais e por ter o privilégio de poder regressar a casa sempre que preciso! Obrigada por todo o vosso apoio quer emocional, quer matemático (e de LaTeX). Obrigada também por tomarem conta do Tootsy sempre que estou fora. Gostaria também de agradecer ao meu irmão Mário (e aos meus gatos) pelo seu apoio e presença constante.

To all of you, my sincere thank you! Muito obrigada!

# Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Chapter 3 has been published as André, L. M., Wadsworth, J. L., and O'Hagan, A. (2024). Joint modelling of the body and tail of bivariate data. Computational Statistics and Data Analysis, 189:107841. https://doi.org/10.1016/j.csda.2023.107841.

Chapter 4 is a draft paper to be submitted for publication as André, L. M., and Tawn, J. A. (2025). Gaussian mixture copulas for flexible dependence modelling in the body and tails of joint distributions.

Chapter 5 is a draft paper to be submitted for publication as André, L. M., Wadsworth, J. L., and Huser, R. (2025). Neural Bayes estimation for complex bivariate extremal dependence models.

Chapter 6 is a result of the Lancaster and Maynooth Universities contribution for a competition as part of the 2023 Extreme Value Analysis conference at the University of Bocconi, Italy. This has been published as André, L. M., Campbell, R., D'Arcy, E., Farrell, A., Healy, D., Kakampakou, L., Murphy, C., Murphy-Barltrop, C. J. R., and Speers, M. (2024). Extreme value methods for estimating rare events in Utopia. Extremes. https://doi.org/10.1007/s10687-024-00498-w. My primary contributions are in Sections 6.4 - 6.5.

The word count for this thesis is approximately 49 448.

Lídia Maria Branco Correia Martins André

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

Extreme events impact most people at some point in their lives. Consider for example the recent floods in Valencia in Spain, the wildfires in Pedrógrão Grande in Portugal in 2017, or even the 2008 financial crisis. Such events have undeniably led to significant humanitarian disasters and financial losses, from property destruction to the loss of life. Understanding such phenomena is crucial to mitigate their effects and to prevent their recurrence. Furthermore, when these rare observations are triggered by other events, whether extreme or not, joint modelling is essential to fully comprehend such processes. Extreme value theory offers powerful statistical tools to model these impactful events – individually or jointly – and help in their prevention.

The typical approach to modelling extreme observations involves defining an extremal region, which usually requires the choice of a threshold value, or vector, above which observations are considered extreme. In a multivariate setting, however, defining such a region is more complex than in the univariate case, as there are various ways of classifying observations as extreme. In addition, the choice of threshold is often arbitrary and may sometimes be too simplistic. These limitations can be concerning as they

might lead to an inaccurate representation of the extremal structure, or introduce extra sensitivity to the results, with small changes in the value of the threshold(s) potentially leading to different outcomes.

These issues can be overcome by considering models that are flexible in bridging between the non-extremal (i.e., 'body') and extremal (i.e., 'tail') regions. Additionally, in some situations, having an accurate fit of the body can be as important as representing the tail correctly. For instance, understanding the effect of one pollutant on another can be as essential as analysing their individual effect, especially when harmful levels of one subset of pollutants occur within the body of the data. Similarly, in financial applications, the most severe losses typically happen in the tail but may be dependent on liabilities that are not extreme. Since modelling the extreme observations requires asymptotically justified models, and empirically estimating the non-extremal region might not be sufficient, having an accurate representation of both regions is not straightforward. Several models that can fit the body and tail regions simultaneously have been proposed in the univariate framework, but relatively scarce work has been done when moving to a multivariate setting. The primary aim of this thesis is to develop dependence models that jointly fit these two regions, while avoiding strong assumptions such as choosing a threshold vector, or the need to define an extremal region.

Owing to their flexibility, we adopt a copula-based framework throughout this thesis (see e.g, Sklar, 1959). To do so, we construct copula models based on different mixture model densities. This process requires inversion of distribution functions, and numerical integration of density functions, both marginal and joint. Such operations result in likelihood functions that are computationally expensive to evaluate, leading to dimensionality restrictions, and potential barriers in the use of these models in practice. The secondary aim of this thesis is to therefore explore alternative methods to perform inference for complex models, enabling their wider application.

It is often the case that generating data from the model is straightforward and com-

putationally efficient, even when evaluating its likelihood is burdensome. Likelihood-free approaches (also known as simulation-based methods) exploit this advantage to estimate the model parameters. In this thesis, we focus on approaches that leverage neural networks. In particular, we explore the utility of neural Bayes estimators (Sainsbury-Dale et al., 2024a) for inferring the parameters of one of the proposed mixture models, as well as of copula models from the bivariate extremes literature, which can interpolate between two regimes of extremal dependence (asymptotic independence and dependence, defined in Section 2.2.4 of Chapter 2).

## 1.2 Overview of thesis

As stated, our primary aim is to develop dependence models that can accurately represent the body and tail regions jointly. While there is a rich body of literature on univariate models which achieve this, there remains significant potential for multivariate models. Given the computational intensity to fit such models, we then focus on aiding their inference process by leveraging neural network-based likelihood-free approaches. This thesis is organised as follows:

Chapter 2 gives an overview of the key modelling strategies used in extreme value theory. We begin by briefly introducing the standard approaches to univariate extremes, followed by an extensive review of univariate mixture models that consider both the body and tail regions. We then introduce the key concepts when moving to the multivariate extremes framework, including the basic concepts of copula theory. Similarly to the univariate case, we provide a review of multivariate mixture models which concern the modelling of the full distribution. We finish by exploring likelihood-free approaches, especially those that leverage neural networks, to perform inference when the evaluation of the likelihood is computationally burdensome.

In Chapter 3, we propose a dependence model that is able to represent the body

and tail regions of bivariate data accurately. The proposed model blends two copulas with different characteristics over the whole range of the data support. In particular, one copula is tailored to the body and the other to the tail, with a dynamic weighting function employed to smoothly transition from one region to the other. The tail dependence properties of the model are investigated numerically, and derived analytically for two particular cases. Through simulation studies we show that the model is identifiable and is sufficiently flexible to capture a wide variety of structures. Finally, we apply the model to study the dependence between ozone concentration and temperature at two sites in the UK, showing that the model is capable of capturing complex dependence structures.

In Chapter 4, we introduce a different copula model, which is based on a mixture of Gaussian distributions. This copula also avoids the need to define an extremal region, and is scalable to dimensions beyond the bivariate case. We derive sub-asymptotic dependence measures for a simplified model specification, and through simulations, we show that the model is identifiable up to dimension $d = 5$, and that it is able to capture a range of extremal dependence structures. Finally, we apply the proposed model to a 5-dimensional seasonal air pollution data set, previously analysed in the multivariate extremes literature. Through pairwise, trivariate and 5-dimensional analyses, we show the flexibility of the Gaussian mixture copula in capturing different joint distributional behaviours and its ability to identify potential graphical structure features, both of which can vary across the body and tail regions.

In Chapter 5, we provide an amortised likelihood-free model selection and inference toolkit which leverages neural networks, whereby the best model is selected for a given data set and its parameters subsequently estimated using neural point estimation. To do so, we start by exploring the properties of neural Bayes estimation (Sainsbury-Dale et al., 2024a) for parameter inference for several flexible bivariate extremal dependence models, with a view to aiding their routine implementation. We focus specifically on the

model proposed in Chapter 3 and on models that are able to interpolate between the two regimes of extremal dependence at an interior point of the parameter space. Owing to the absence of likelihood evaluation in the inference procedure, classical information criteria such as the Bayesian information criterion (BIC) cannot be used to select the most appropriate model. Instead, we propose using neural networks as classifiers for model selection, and examine their performance for model selection comparing with BIC, when such criterion is available. We apply our classifiers and estimators to analyse the pairwise extremal behaviour of changes in horizontal geomagnetic field fluctuations at three different locations. We show that we are able to obtain fast and reliable estimates of the extremal dependence structure for each pair.

Chapter 6 was written following entry of the Lancopula Utopiversity team to tackle the data challenge of the 2023 Extreme Value Analysis conference. The aim of the challenge was to estimate extremal probabilities (or quantiles) of complex environmental phenomena, and it was split into 4 challenges: C1 - C4. Challenges C1 and C2 concern univariate estimation, whereas challenges C3 and C4 comprise of multivariate problems. For C1 and C2, we propose a flexible modelling approach, which relies on generalised additive models, to estimate extreme quantiles of a non-stationary time series. For challenge C3, we propose an extension of the modelling approach of Wadsworth and Tawn (2013) to estimate joint probabilities with non-stationary extremal dependence. Finally, for challenge C4, we identify sub-groups of the data by employing a dimension reduction technique based on exploratory analysis of the pairwise extremal dependence across 50 locations. By assuming that these sub-groups are independent of each other we estimate joint probabilities by applying the conditional extremes approach of Heffernan and Tawn (2004) within each cluster.

In Chapter 7, we summarise the contributions of this thesis as well as discuss some possible avenues for further work.

# Chapter 2

# Literature review

## 2.1 Univariate extreme value theory

In many applications, where we are concerned about the large (extreme) values, we are interested in methodology for modelling the tails of a distribution. In a univariate framework, there are a range of methods available to achieve this goal; these are reviewed in Coles (2001) for instance. We introduce the two most common approaches: the generalised extreme value (GEV) distribution in Section 2.1.1 and the generalised Pareto (GP) distribution in Section 2.1.2. Finally, in Section 2.1.3 we give an overview of available methods (herein referred to as 'extreme value mixture models' or EVMMs) able to model non-extreme as well as extreme values simultaneously; these models are important when the aim is to represent these two regions correctly.

### 2.1.1 Generalised extreme value distribution

Let $X_1, \ldots, X_n$ be a series of independent and identically distributed (i.i.d.) random variables with common distribution function $F$. The upper tail behaviour of $F$ can be characterised by considering the maximum of this series, $M_n = \max\{X_1, \ldots, X_n\}$. Similarly, the lower tail can be modelled by noting that $\min\{X_1, \ldots, X_n\} = -\max\{-X_1, \ldots,$

$-X_n\}$. The distribution of $M_n$ is then given by

$$\begin{aligned}\Pr(M_n \leq z) &= \Pr(X_1 \leq z, \ldots, X_n \leq z) \\ &= \Pr(X_1 \leq z) \ldots \Pr(X_n \leq z) = F^n(z).\end{aligned}$$

However, for any $z < x^F$, where $x^F$ is the upper end-point of $F$, the distribution of $M_n$ is degenerate to a point mass on $x^F$, since $F^n(z) \to 0$ as $n \to \infty$. This degeneracy problem can be overcome by considering the extremal types theorem of Leadbetter et al. (1983). This states that if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \quad \text{as } n \to \infty,$$

where $G$ is a non-degenerate distribution function, then $G$ belongs to the family of generalised extreme value (GEV) distributions. In this situation, the distribution function $F$ of each variable $X_i$ $(i = 1, \ldots, n)$ is said to lie in the domain of attraction of $G$. The distribution function of the GEV distribution is defined as follows

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, \tag{2.1.1}$$

where $x_+ = \max\{x, 0\}$, and $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$ are the location, scale and shape parameters, respectively. The shape parameter $\xi$ determines which GEV family $G$ belongs to: when $\xi > 0$, the GEV corresponds to a Fréchet distribution, whereas if $\xi < 0$, then $G$ is a Weibull distribution. In the limit case $\xi \to 0$, the GEV distribution becomes a Gumbel distribution.

Although the normalising constants $\{a_n > 0\}$ and $\{b_n\}$ are unknown in practice, assuming that equation (2.1.1) holds for large $n$ yields that

$$\Pr(M_n \leq z) \approx G\left(\frac{z - b_n}{a_n}\right) = G^*(z),$$

where $G^*$ is a member of the GEV family with different location and scale parameters $\mu^*$ and $\sigma^*$. In this way, the normalising constants are absorbed into $\mu^*$ and $\sigma^*$, allowing us to use the GEV distribution to model block maxima data. Such data are derived by first being split into $m$ blocks of sequences of observations of size $n$, and then by taking the maximum for each block, obtaining in this way the sequence $M_{n,1}, \ldots, M_{n,m}$. Choosing the size of the blocks constitutes, however, a trade-off between bias and variance as a value of $n$ too small can lead to biased results, whilst choosing a large $n$, will mean fewer blocks $m$, causing higher variability. This approach is usually known as the block maxima approach and is illustrated in the left plot of Figure 2.1.1.

### 2.1.2 Generalised Pareto distribution

A major drawback of just considering the maximum within a block is that it can lead to some extreme observations being disregarded during the inference procedure. This may be the case when several large values occur in the same block for example. A common alternative is to treat the observations of a sequence that exceed some high threshold $u$ as extreme events.

Consider the series of i.i.d. random variables $X_1, \ldots, X_n$ with common distribution function $F$ with upper end-point $x^F$, and let $X$ be an arbitrary element of this sequence. Pickands (1975) states that, if $F$ is in the domain of attraction of the GEV, then there exists a continuous function $g(u) > 0$ such that, as $u \to x^F$

$$\Pr\left(\frac{X-u}{g(u)} \leq z \;\middle|\; X > u\right) \to H(z),$$

for all $z > 0$. In this case, $H$ is the cumulative distribution function of a generalised

Pareto (GP) distribution, and takes the form

$$H(z) = \begin{cases} 1 - \left(1 + \xi \dfrac{z}{\varphi}\right)_+^{-1/\xi}, & \text{if } \xi \neq 0, \\ 1 - \exp\left\{-\dfrac{z}{\varphi}\right\}, & \text{if } \xi \to 0, \end{cases} \tag{2.1.2}$$

where $x_+ = \max\{x, 0\}$, $\varphi > 0$ and $\xi \in \mathbb{R}$ are the scale and shape parameters, respectively. In addition, parameter $\xi$ is the same as the GEV shape parameter from equation (2.1.1). As with the normalising constants in the GEV framework, the function $g$ is unknown in practice. However, assuming that (2.1.2) holds for some large threshold $u$, the threshold exceedances can be modelled as

$$\Pr(X - u \leq z \mid X > u) \approx H\left(\frac{z}{g(u)}\right) = H^*(z),$$

where $H^*$ is a GP distribution with different scale and shape parameters, allowing us to use the GP distribution to model the exceedances above $u$.

Similarly to the GEV case, choosing the threshold $u$ represents a bias-variance trade-off: a value too low for the threshold may result in bias, whereas a high value for $u$ may lead to high variability in the parameter estimates. Overall, the threshold $u$ should be small enough that there is sufficient data to model, but large enough to ensure a valid asymptotic approximation of the results. There are several methods available to select the threshold; see Coles (2001), Scarrott and MacDonald (2012), Wadsworth (2016), Northrop et al. (2017) and Murphy et al. (2024) for instance. This approach is usually known as the peaks over threshold approach and is illustrated in the right plot of Figure 2.1.1.

### Extended generalised Pareto distribution

Papastathopoulos and Tawn (2013) propose an extension of the GP distribution given in equation (2.1.2) that reduces the sensitivity to the threshold choice by allowing a

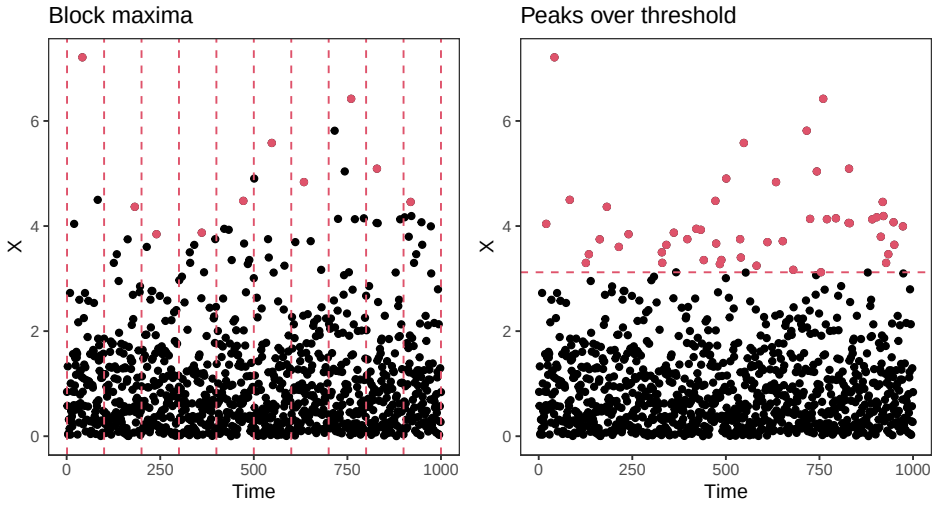



Figure 2.1.1: Block maxima approach (left) and Peaks over threshold approach (right) in simulated data. The red points indicate the data used to fit the GEV and GP distributions, respectively. The vertical lines in the left plot indicate the block length used, and the horizontal line in the right plot indicates the threshold used.

lower threshold to be selected. Thus, they construct parametric models for $X - u \mid X > u$ which are more flexible than the GP distribution, and can be fitted at lower thresholds while ensuring the tails are asymptotically equivalent to the GP distribution. The probability density functions (pdfs) of the three proposed extended GP (EGP) distributions are defined below as

$$\text{EGPD1:} \quad f(z) = \begin{cases} \dfrac{|\xi|}{\text{Be}(\kappa, |\xi|^{-1})} \left[1 - \left(1 + \xi \dfrac{z}{\sigma_u}\right)_+^{-\frac{|\xi|}{\xi}}\right]^{\kappa-1} h(z), & \xi \neq 0, \\ \dfrac{x^{\kappa-1}}{\Gamma(\kappa)} h(z), & \xi \to 0, \end{cases} \tag{2.1.3}$$

$$\text{EGPD2:} \quad f(z) = \begin{cases} \dfrac{1}{\Gamma(\kappa)} \left[\dfrac{1}{\xi} \log\left(1 + \xi \dfrac{z}{\sigma_u}\right)_+\right]^{\kappa-1} h(z), & \xi \neq 0, \\ \dfrac{x^{\kappa-1}}{\Gamma(\kappa)} h(z), & \xi \to 0, \end{cases} \tag{2.1.4}$$

$$\text{EGPD3:} \quad f(z) = \begin{cases} \kappa \left[1 - \left(1 + \xi \dfrac{z}{\sigma_u}\right)_+^{-\frac{1}{\xi}}\right]^{\kappa-1} h(z), & \xi \neq 0, \\ \kappa \left(1 - \exp\left\{-\dfrac{z}{\sigma_u}\right\}\right)^{\kappa-1} h(z), & \xi \to 0, \end{cases} \tag{2.1.5}$$

where Be is the Beta function, $\Gamma$ is the Gamma function and $h$ is the density function

of the GP distribution given in equation (2.1.2). Through the parameter $\kappa > 0$, more flexibility is added to the body of the distribution whilst the behaviour of the tails remains unchanged. The three EGP distributions reduce to the GP distribution when $\kappa = 1$. More recently, Gamet and Jalbert (2022) propose three complementary models to the EGP distributions whose density at the threshold is positive and finite, which might not always be the case with the EGP distributions proposed by Papastathopoulos and Tawn (2013). Further details can be found in Papastathopoulos and Tawn (2013) and Gamet and Jalbert (2022).

### 2.1.3 Extreme value mixture models

As mentioned in Section 2.1.2, the choice of threshold $u$ not only represents a bias-variance trade-off, but can also be a subjective choice. Furthermore, once the choice is made, the inherent uncertainty in subsequent inferences is often ignored. A way of overcoming these issues is by considering models that choose a distribution for the non-extreme observations (known as the 'bulk') along with an appropriate fit for the tail region of a data set; by implicitly or explicitly estimating $u$, these models aim at accounting for the uncertainty in the choice of threshold. A review of several of these approaches, herein referred to as 'extreme value mixture models' (EVMMs), is given in Scarrott and MacDonald (2012).

In general, it is standard practice for EVMMs to model the extreme region using a GP distribution, whilst the modelling of the bulk is implemented in a parametric, semi-parametric or non-parametric way. Furthermore, care is needed to guarantee that neither the bulk nor tail have a big influence on one another. In spite of that, the two regions cannot be fully disjoint as information is shared between them. Depending on the construction of each model, defining a threshold $u$ above which the GP distribution is fitted to the data is still necessary; however, this threshold is often treated as a parameter to be estimated.

There exist several EVMMs available in the literature. For instance, Frigessi et al. (2002) propose fitting two distributions (one of which being the GP distribution) to the whole data, avoiding the need for estimating a threshold $u$. Through a dynamic weighting function $p$ that lies in $(0, 1]$, more weight is given to the bulk distribution at low ranges of the data, and more weight is given to the GP distribution in the upper tail. The density of the proposed model is defined as below

$$f(x) = \frac{[1 - p(x)]d(x) + p(x)h(x)}{K}, \quad x \geq 0, \tag{2.1.6}$$

where $d$ is the density of a distribution with light tail, $h$ is the density of the GP distribution, $p$ is increasing in $x$, and

$$K = \int_0^\infty [1 - p(x)]d(x) + p(x)h(x)\mathrm{d}x$$

is a normalising constant. In general, a smooth transition between the bulk and tail models is achieved with model (2.1.6); however, this does not always happen as choosing the unit step function as the weighting function $p$ leads to discontinuity between the two regions. Therefore, care is needed when choosing $p$ if a smooth transition is the goal.

In turn, de Mendes and Lopes (2004) propose a data driven approach where both tails are modelled using a GP distribution, and the bulk is modelled using a left and right truncated Gaussian. To do so, the authors first standardise the data using $Y_i := (X_i - \text{median}(\boldsymbol{X}))/ \mid X_i - \text{median}(\boldsymbol{X}) \mid$ for $i = 1, \ldots, n$. The proposed density is then given as follows

$$f(y) = p_l h_l(y - u_l) + (1 - p_l - p_u)d(y - u_u) + p_u h_u(y),$$

where $p_l$ and $p_u$ represent the proportions of the data in the lower and upper tails, $h_l$

and $h_u$ are densities of a GP distribution fitted to the lower and upper tails, respectively, with $h_l(y) = -h_u(y)$, $d$ is the density of a standard Gaussian distribution truncated at $u_l$ and $u_u$, and $u_l < 0$ and $u_u > 0$ are the thresholds for the GP distribution fit to the lower and upper tails, respectively. More details can be found in de Mendes and Lopes (2004).

The model proposed by Behrens et al. (2004) assumes a distribution $D$ below threshold $u$, that needs to be estimated, and a GP distribution above. The density of the model is given as

$$f(x) = \begin{cases} d(x), & \text{if } x < u, \\ \varphi_u h(x-u), & \text{if } x \geq u, \end{cases} \tag{2.1.7}$$

where $h$ is the density of a GP distribution as defined in equation (2.1.2) and $\varphi_u = 1 - D(u)$ is the probability of exceedances. This model exhibits a discontinuity at threshold $u$ and performs poorly when there is a smooth transition at $u$. Furthermore, a poor fit in the bulk region will affect the location of the threshold, which in turn will impact the GP fit to the tail; see Behrens et al. (2004) for more details.

Tancredi et al. (2006) propose modelling the bulk distribution as a mixture of uniform distributions. In addition, the authors consider two different thresholds, the usual threshold $u$ used to fit the GP distribution, which is estimated through the inference procedure, and a threshold $u_0 < u$, which is known to be too low, and acts as the starting point for the modelling procedure. More specifically, the density of the proposed model is defined as follows

$$f(x) = \begin{cases} (1-\varphi_u)\displaystyle\sum_{j=1}^{k} \omega_i \mathbb{1}_{[a_j, a_{j+1})}(x), & \text{if } u_0 < x < u, \\ \varphi_u h(x-u), & \text{if } x \geq u, \end{cases}$$

where $\sum_{j=1}^{k} w_j(a_{j+1} - a_j) = 1$, and $a_j < a_{j+1}$ with $a_1 = u_0$ and $a_k = u$, $w_j \in [a_j, a_{j+1})$, $j = 1, \ldots, k$, are the unknown parameters of the model. As before, $h$ is the density of

a GP distribution and $\varphi_u$ is the probability of exceedances.

The discontinuity at the threshold present in model (2.1.7) is avoided in the model proposed by Carreau and Bengio (2009) by forcing continuity up to the first derivative of the density function. Additionally, the distribution fitted below threshold $u$ is taken as the Gaussian distribution. The probability density function of the proposed model is given as

$$f(x) = \begin{cases} \gamma^{-1} d(x), & \text{if } x < u, \\ \gamma^{-1} h(x-u), & \text{if } x \geq u, \end{cases}$$

where $\gamma$ is a re-weighting parameter that ensures that function $f$ integrates to one, $d$ is the pdf of a Gaussian distribution and $h$ is the pdf of the GP distribution. Due to the continuity constraints enforced, the threshold $u$ and the GP scale parameter $\sigma$ are computed implicitly as a function of the remaining model parameters. More details can be found in Carreau and Bengio (2009).

The model introduced in Cabras and Castellanos (2010) builds upon the model of Behrens et al. (2004). However, a distribution $D$ is not assumed for the bulk; instead, the non-extreme observations are first binned into equally spaced counts, and then modelled by a Poisson generalised linear model with a smoother polynomial assumed for the mean parameter. In this way, a more flexible fit to the bulk is achieved. The model proposed by do Nascimento et al. (2012) is built similarly to the model in equation (2.1.7), where the bulk is now modelled using a weighted mixture of Gamma densities. In particular, the authors take $D$ and $d$ to be the distribution and density functions of a mixture of $k$ Gamma distributions given by

$$d(x) = \sum_{j=1}^{k} \omega_j \pi_j(x), \tag{2.1.8}$$

where $\omega_j$ are the mixture weights, and $\pi_j$ are the probability density functions of a

Gamma distribution, $j = 1, \ldots, k$.

A kernel density estimator is used to model the bulk in the model proposed by MacDonald et al. (2011), which is taken as a mean zero Gaussian probability density function. The proposed density is given by

$$f(x) = \begin{cases} (1 - \varphi_u)\dfrac{d(x)}{D(u)}, & \text{if } x \leq u, \\ \varphi_u h(x - u), & \text{if } x > u, \end{cases}$$

where $h$ is the density of a GP distribution, $\varphi_u$ represents the probability of being above threshold $u$ as before, and $d$ and $D$ are non-parametric density and distribution functions. In particular, $d$ is approximated by $\hat{d}$, given as

$$\hat{d}(x) = \frac{1}{n}\sum_{i=1}^{n} K_\lambda(x - x_i),$$

where $K_\lambda$ is a zero mean Gaussian density function with standard deviation $\lambda$, and $n$ is the sample size.

An extension to the EGP distributions given in equations (2.1.3), (2.1.4) and (2.1.5) was proposed by Naveau et al. (2016) in which the authors aim at modelling the low, moderate and large values of a data set jointly. For that, Naveau et al. (2016) construct a model that resembles a GP distribution in the lower and upper tails while avoiding the threshold choice, since this threshold brings discontinuities between low and moderate observations, and between moderate and extreme observations. Let $U$ be a uniform random variable defined on $[0, 1]$ and let $D$ represent a continuous cumulative distribution function (cdf) on $[0, 1]$. Naveau et al. (2016) construct a family of random variables defined by $X := H^{-1}\left(D^{-1}(U)\right)$, whose density is given as

$$f(x) = d[H(x)]h(x),$$

for all $x > 0$. In this equation, $H$ is the GP distribution with scale parameter $\sigma$, $h$ is its density, and $d$ is the probability density function corresponding to distribution $D$. Additionally, distribution $D$ satisfies some constraints such that the upper tail is equivalent to a Pareto tail, and it resembles a GP distribution with negative shape and finite upper-end point. Lastly, the lower and upper tails are bridged together with the bulk via function $D$; see Naveau et al. (2016) for further details.

In the approach proposed by Huang et al. (2019), the full density is modelled using a cubic spline in the histogram of the data. Furthermore, while not imposing a parametric form to fit the bulk of the data, the proposed model still guarantees that the upper tail behaves like a GP distribution. More specifically, the authors consider $X$ to be a heavy-tailed continuous random variable with distribution function $F$, and are interested in modelling the logarithm of the random variable, i.e. $\log(X)$, whose distribution function is $D$. Thus, Huang et al. (2019) propose modelling the density of $X$ as

$$f(x) = \frac{1}{x} \exp\{d(\log(x))\}, \quad x > 0,$$

where $d$ is assumed to belong to the family of natural cubic splines. In this way, the tail is in compliance with a GP distribution and the bulk is modelled in a flexible manner. See Huang et al. (2019) for more details.

In turn, Tencaliec et al. (2020) present an extension to the model of Naveau et al. (2016). Instead of assuming a parametric form for $D$, the authors propose approximating this function using Bernstein polynomials. The density of the model is then as follows

$$f(x) = \hat{d}_{m,n}[H(x)]h(x),$$

where $H$ and $h$ are the distribution and density functions of a GP distribution, respec-

tively, and $\hat{d}_{m,n}$ is given as

$$\hat{d}_{m,n}(t) = \sum_{k=1}^{m} \omega_{k,m}\beta_{k,m-k+1}(t), \quad t \in [0,1],$$

with $m > 0$ denoting the degree of the Bernstein estimator, $n$ denoting the sample size of the data set, $\omega_{k,m} = D_n(k/m) - D_n((k-1)/m)$, and $\beta_{a,b}(t)$ is the probability density function of a Beta distribution with parameters $(k, m-k+1)$, $k = 1, \ldots, m$. In addition, the authors show that the cdf $\hat{D}_{m,n}$ associated with $\hat{d}_{m,n}$ satisfies the constraints imposed in Naveau et al. (2016). More details can be found in Tencaliec et al. (2020).

Similarly to the model proposed by Naveau et al. (2016), a family of models that rely on composition of functions is introduced by Stein (2021). Let $(\phi_i, \tau_i, \kappa_i)$ denote the location, shape and scale parameters for the lower $(i = 1)$ and upper $(i = 2)$ tails, and let $\psi = \log(1 + \exp\{x\})$. The density of the proposed models is defined as follows

$$f(x) = t_\nu(D(x))D'(x),$$

where $t_\nu$ is the density of a $t$ distribution with $\nu$ degrees of freedom, and

$$D(x) = \left[1 + \kappa_2\psi\left(\frac{x-\phi_2}{\tau_2}\right)\right]^{1/\kappa_2} - \left[1 + \kappa_1\psi\left(\frac{x-\phi_1}{\tau_1}\right)\right]^{1/\kappa_1}.$$

This family of distributions has support in $(a, b) \subset \mathbb{R}$, with $a$ and $b$ depending on the values of $\kappa_i$: if $\kappa_2 < 0$, then $b = \phi_2 + \tau_2\psi^{-1}(-1/\kappa_2)$, whilst $b = \infty$ if $\kappa_2 \geq 0$, and, similarly, if $\kappa_1 < 0$, then $a = \phi_1 - \tau_1\psi^{-1}(-1/\kappa_1)$, whereas $a = -\infty$ if $\kappa_1 \geq 0$. Furthermore, the proposed models satisfy some constraints; in particular, for any value of $\nu > 0$, the distribution function $F$ behaves like a GP distribution in the lower and upper tails. This work is extended to the non-stationary case by Krock et al. (2022) by allowing the scale and location parameters, $\tau_i$ and $\phi_i$ $(i = 1, 2)$ respectively, to vary with season and any long-term trends present in the data. More details can be found

in Stein (2021) and Krock et al. (2022).

Finally, Castro-Camilo et al. (2019) and Yadav et al. (2021) propose modelling the bulk and tail regions jointly through a hierarchical model. Similarly to the work of Opitz et al. (2018), Castro-Camilo et al. (2019) propose fitting the data in three stages. In the first step, a Gamma distribution is used to model the non-extreme region; this stage will aid the estimation of the threshold above which the GP distribution is fitted as well. The second step concerns obtaining the probability of excesses above the estimated threshold; this is achieved using a Bernoulli distribution. The last stage then corrects the tail by fitting a GP distribution to the exceedances data. Whilst Opitz et al. (2018) assume a generalised additive modelling framework, Castro-Camilo et al. (2019) assume a latent Gaussian random field that is not only able to describe the trends in the data, but also its underlying dependence structure; see Castro-Camilo et al. (2019) for more details.

In turn Yadav et al. (2021) extend the characterisation of the GP distribution as an exponential mixture (see, e.g., Bopp and Shaby, 2017) to account for more general distribution families. Let $\Lambda \geq 0$ denote a latent random variable with rate parameter $\lambda \geq 0$ such that $X \mid \Lambda \sim F_{X|\Lambda}(\cdot)$. The authors propose a Gamma-Gamma hierarchical model in which both random variables $X \mid \Lambda$ and $\Lambda$ follow a Gamma distribution with parameters $(\lambda, \beta_1)$ and $(\alpha, \beta_2)$, respectively, for $\alpha, \beta_1, \beta_2 > 0$. The GP distribution case then arises when $\beta_1 = 1$. Since this model is suitable for data with heavy tails, Yadav et al. (2021) also propose an extension for when the GP shape parameter $\xi$ is close to 0. In this case, $X^{1/k} \mid \Lambda$ follows a Gamma distribution with parameters $(\lambda, \beta_1)$, whilst the latent variable $\Lambda$ now follows a generalised inverse Gaussian distribution with parameters $(\alpha/2, b, \beta_2)$. Lastly, when $b = 0$, $k = 1$ and $\beta_2 > 0$, this construction generalises to the Gamma-Gamma hierarchical model; see Yadav et al. (2021) for more details.

## 2.2 Multivariate extreme value theory

In a multivariate setting, defining extreme events is not as straightforward as in the univariate framework. For instance, due to the lack of ordering between two vectors, there is no clear way of determining which is largest. Barnett (1976) suggests different ways to define an extreme event, with the best approach depending on the context of the application. One of the most common approaches is to consider componentwise maxima, where the maximum of each variable over a block is taken as an extreme value for that variable; this can be seen as an extension to the multivariate setting of the block maxima approach introduced in Section 2.1.1. Alternative definitions include constructing a convex hull around the data, in which the points are deemed extreme if lie on this region or beyond; considering as extreme events the observations which contain the block maximum of at least one variable; or defining a one-dimensional structure variable that allows the problem to be reduced to a univariate framework as considered by Coles and Tawn (1994).

We introduce copulas, which are models able to capture the overall dependence separately from the marginal distributions, in Section 2.2.1. We then define in more detail two approaches to define extreme events: componentwise maxima in Section 2.2.2 and structure variables in Section 2.2.3. Modelling strategies for asymptotic dependence and asymptotic independence are presented in Sections 2.2.4 and 2.2.5, respectively. In Sections 2.2.6 and 2.2.7 two approaches suitable for both regimes of extremal dependence are introduced. Finally, existing approaches to extend EVVMs to a multivariate setting are given in Section 2.2.8.

### 2.2.1 Copula theory

When moving to the multivariate framework, not only is the marginal modelling important, but also capturing the dependence between the variables, since the behaviour

of one variable can have an influence on another variable. A way of capturing this dependence is by means of copulas, which allow the modelling of the dependence to be independent of the marginal specification. This is not the situation when the dependence is measured through the Pearson's correlation coefficient, for example, since the association here relies on the choice of margins. Furthermore, instead of giving an overall idea of the dependence structure, with copulas it is possible to know where in the support of the data the association between the variables is the weakest/strongest. A detailed description of copula modelling can be found in Joe (1997), Nelsen (2006) or Joe (2014).

Consider the vector of random variables $\boldsymbol{X} = (X_1, \ldots, X_d)$ with distribution function $F$ and let $X_i \sim F_{X_i}$ for $i = 1, \ldots, d$ and $d \geq 2$. According to Sklar's theorem (Sklar, 1959), the joint distribution of $\boldsymbol{X}$ can be written as the composition of the marginal distributions $F_{X_i}$ and a copula $C : [0,1]^d \to [0,1]$. In particular, we have

$$F(x_1, \ldots, x_d) = C\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right). \tag{2.2.1}$$

If the margins are continuous, then Sklar's theorem states that copula $C$ is unique. Furthermore, when it exists, the copula density $c\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right)$ can be obtained by taking the $d^{\text{th}}$ order derivative with respect to the variables $F_{X_1}(x_1), \ldots, F_{X_d}(x_d)$.

Given that $U_i = F_{X_i}(x_i) \sim \text{Unif}(0,1)$, by the probability integral transform (PIT), it can be seen from equation (2.2.1) that the copula $C$ is a multivariate distribution with standard uniform margins. Additionally, the dependence structure can be captured by a copula on any standardised margins, since due to the PIT, we have $F^{-1}_{X_i}(U_i) \sim F_{X_i}$, where $F^{-1}_{X_i}$ is the inverse cdf of margin $X_i$, $i = 1, \ldots, d$. This result allows for an arbitrary choice of margins, depending on the context of each problem. In particular, some margins might highlight some features of the extreme values that others are not able to. Figure 2.2.1 highlights this difference in simulated data from a bivariate logistic distribution with dependence parameter $\alpha = 1/2$ in standard uniform, Fréchet

and exponential margins, respectively, from left to right.

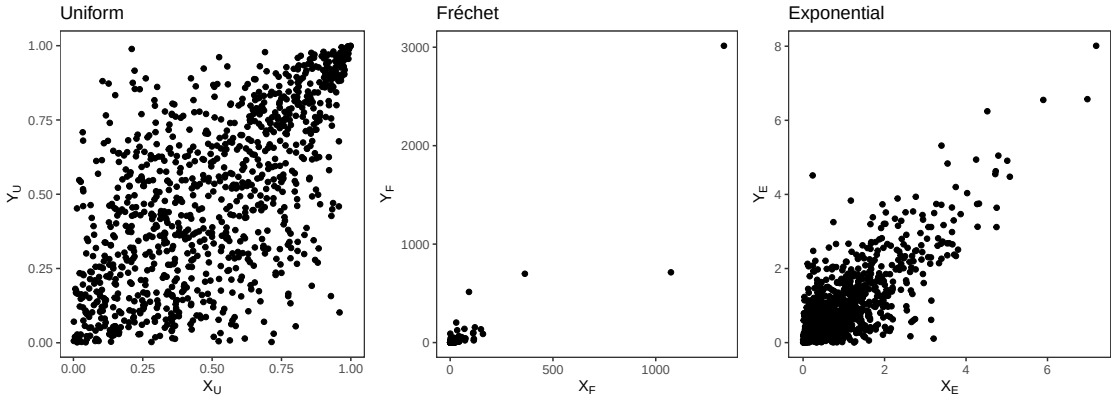



Figure 2.2.1: Example of draws from a copula $C$ in standard uniform (left), standard Fréchet (middle) and standard exponential (right) margins.

### 2.2.2 Componentwise maxima

As mentioned in the introduction of this section, one way of ordering multivariate data is considering componentwise maxima. Let $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n$ be $n$ independent $d$-dimensional random vectors, i.e., $\boldsymbol{X}^j = (X_1^j, \ldots, X_d^j)$ for $j = 1, \ldots, n$, with common distribution $F$. The vector of componentwise maxima $\boldsymbol{M}_n$ is defined as the vector derived by taking the maximum over the $n$ repetitions of each $i = 1, \ldots, d$ variables. That is, $\boldsymbol{M}_n = (M_{1,n}, \ldots, M_{d,n})$, with $M_{i,n} = \max_{1 \leq j \leq n}(X_i^j)$ for $i = 1, \ldots, d$. Similarly to the block maxima approach introduced in Section 2.1.1, if there exists sequences $\{\boldsymbol{a}_n > 0\}$ and $\{\boldsymbol{b}_n\}$ with $\boldsymbol{a}_n = (a_{1,n}, \ldots, a_{d,n})$ and $\boldsymbol{b}_n = (b_{1,n}, \ldots, b_{d,n})$, such that, as $n \to \infty$, the distribution function of the normalised componentwise maxima

$$\Pr\left(\frac{M_{1,n} - b_{1,n}}{a_{1,n}} \leq z_1, \ldots, \frac{M_{d,n} - b_{d,n}}{a_{d,n}} \leq z_d\right) \to G(z_1, \ldots, z_d), \tag{2.2.2}$$

where $G$ is a distribution function, that is non-degenerate in each margin, then $G$ is known as the multivariate extreme value distribution. The margins of $G$ are GEV distributions (as defined in equation (2.1.1)). However, unlike the univariate setting,

there is not a single parametric form for $G$. The componentwise maxima approach is illustrated in the left panel of Figure 2.2.2; the red point shows that componentwise maximum does not always corresponds to an actual observation.

Although not necessary, the usual practice is to set common margins for $X_i$, $i = 1, \ldots, d$. In particular, if we consider standard Fréchet margins, i.e., $F_{X_i}(x_i) = \exp\{-1/x_i\}$ for $x_i > 0$ and $i = 1, \ldots, d$, and set $\boldsymbol{a}_n = \boldsymbol{n}$ and $\boldsymbol{b}_n = \boldsymbol{0}$, then the distribution function $G$ of $\boldsymbol{X}$ takes the form

$$G(x_1, \ldots, x_d) = \exp\{-V(x_1, \ldots, x_d)\}, \tag{2.2.3}$$

where $V$ is known as the exponent measure and is a homogeneous function of order $-1$, that is $V(a\boldsymbol{x}) = a^{-1}\boldsymbol{x}$ with $a > 0$. Furthermore, the exponent measure $V$ can be expressed as

$$V(x_1, \ldots, x_d) = d \int_{\mathcal{S}_{d-1}} \max\left\{\frac{w_1}{x_1}, \ldots, \frac{w_d}{x_d}\right\} \mathrm{d}H(\boldsymbol{w}), \tag{2.2.4}$$

where $\mathcal{S}_{d-1} = \{\boldsymbol{w} \in [0,1]^d : \sum_{i=1}^d w_i = 1\}$ denotes the $d$-dimensional unit simplex, and $H$ is a distribution function on $\mathcal{S}_{d-1}$ known as the spectral measure. For $i = 1, \ldots, d$, this measure satisfies a moment constraint, specifically

$$\int_{\mathcal{S}_{d-1}} w_i \mathrm{d}H(\boldsymbol{w}) = \frac{1}{d}.$$

The spectral measure $H$ gives information about the extremal dependence structure. Consider the bivariate framework; in the case of independence between $X_1$ and $X_2$, $H(\{0\}) = H(\{1\}) = 1/2$, and $V(x_1, x_2) = x_1^{-1} + x_2^{-1}$. On the other hand, $H(\{1/2\}) = 1$ if $X_1$ and $X_2$ are perfectly dependent; in this case, $V(x_1, x_2) = \max\{x_1^{-1}, x_2^{-1}\}$.

Several parametric forms for $G$ have been proposed in the literature; for instance the logistic distribution proposed by Gumbel (1960) has $V(\boldsymbol{x}) = \left(\sum_{i=1}^d x_i^{-1/\alpha}\right)^\alpha$ for $\alpha \in (0, 1]$. This distribution may also be known as the Gumbel copula with parameter

$\delta = 1/\alpha \geq 1$. The cases of independence and perfect dependence occur at the boundary of the parameter set; specifically, the variables are independent if $\alpha = 1$ ($\delta = 1$), and complete dependent when $\alpha \to 0$ ($\delta \to \infty$). Alternative forms for the exponent measure lead to other known distributions such as the asymmetric logistic (Tawn, 1988), the negative logistic (Galambos, 1975), the Hüsler-Reiss (Hüsler and Reiss, 1989), and the Dirichlet, which may also be known as Coles-Tawn or Coles-Tawn-Dirichlet (Coles and Tawn, 1991) distributions.

### 2.2.3 Regular variation

Consider again that vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ has common standard Fréchet margins. The tail of $\boldsymbol{X}$ can alternatively be represented in terms of pseudo-polar coordinates $(R, \boldsymbol{W})$, defined as follows

$$R = \sum_{i=1}^{d} X_i \quad \text{and} \quad \boldsymbol{W} = \frac{\boldsymbol{X}}{\sum_{i=1}^{d} X_i},$$

where $R > 0$ and $\boldsymbol{W} \in \mathcal{S}_{d-1} = \{\boldsymbol{w} \in [0,1]^d : \sum_{i=1}^{d} w_i = 1\}$ are known as the radial and angular components of $\boldsymbol{X}$. In a similar way to how the componentwise maxima can be seen as an extension of the block maxima approach, defining the tail in this way can be viewed as an extension of the peaks over threshold approach introduced in Section 2.1.2.

The vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ is said to be multivariate regularly varying if

$$\lim_{t \to \infty} \Pr(\boldsymbol{W} \in B, R > tr \mid R > t) = r^{-1} H(B), \tag{2.2.5}$$

where $r \geq 1$, $B$ is a measurable subset of $\mathcal{S}_{d-1}$, for which $H(\partial B) = 0$ with $\partial B$ denoting the boundary of $B$, and $H$ is the spectral measure defined in equation (2.2.4). Under limit (2.2.5), and for large values of the radial component, variables $R$ and $\boldsymbol{W}$

are approximately independent. Similarly to the componentwise maxima approach, the extremal dependence of $\boldsymbol{X}$ in a multivariate regular variation framework is characterised through the spectral measure $H$. Additionally, the distribution of $\boldsymbol{X}$ satisfies limit (2.2.2) if $\boldsymbol{X}$ is multivariate regularly varying. The middle and right panels of Figure 2.2.2 illustrate the regular variation assumption in the original scale and in pseudo-polar coordinates, respectively.

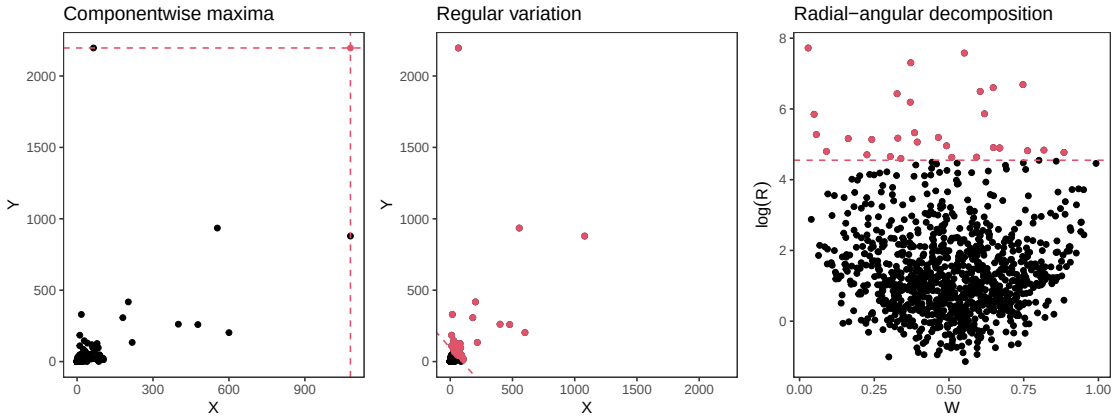



Figure 2.2.2: Componentwise maxima (left), regular variation (middle) and radial-angular decomposition (right) in simulated data. The red values represent the extreme values and the dashed lines indicate the marginal maxima and the radial threshold $u$, respectively for the left and middle and right plots. The right plot shows the independence of $R$ and $\boldsymbol{W}$ given the large threshold value $u$ of $R$.

### 2.2.4 Modelling asymptotic dependence

An important consideration when modelling multivariate extremes is accurately characterising the extremal dependence structure of the data. In particular, the interest lies in determining if the largest values of a data set occur together, or not. Consider a bivariate random vector $(X_1, X_2)$ with $X_i \sim F_{X_i}$ for $i = 1, 2$. A way of quantifying the extremal dependence structure of $(X_1, X_2)$ is through the measure $\chi$ (Joe, 1997, Coles et al., 1999). This coefficient is defined via the limit $\chi := \lim_{u \to 1} \chi(u) \in [0, 1]$, where it

exists, with

$$\chi(u) = \Pr[F_{X_2}(X_2) > u \mid F_{X_1}(X_1) > u] = \frac{\Pr[F_{X_1}(X_1) > u, F_{X_2}(X_2) > u]}{1-u}, \quad (2.2.6)$$

where $u \in (0,1)$. If $\chi > 0$, the variables $X_1$ and $X_2$ are said to be asymptotically dependent (AD), whereas when $\chi = 0$ they are asymptotically independent (AI). Furthermore, equation (2.2.6) can be rewritten by means of copula $C$ as

$$\chi(u) = \frac{1 - 2u + C(u,u)}{1-u},$$

where $C$ is defined as in equation (2.2.1). In the specific case of bivariate extreme value distributions, the coefficient $\chi$ can be derived as $\chi = 2 - V(1,1)$, where $V$ is the exponent measure introduced in equation (2.2.3).

In a multivariate setting, care is needed when defining these two regimes of extremal dependence. Consider now a $d$-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ with $X_i \sim F_{X_i}$ for $i \in D = \{1, \ldots, d\}$; Wadsworth and Tawn (2013) define the $d$-dimensional joint tail dependence through the limit

$$\chi(D) := \lim_{u \to 1} \frac{\Pr[F_{X_i}(X_i) > u : \forall i \in D]}{1-u}, \quad u \in (0,1),$$

where the limit exists. When $\chi(D) > 0$, then all components of $\boldsymbol{X}$ can be large simultaneously, and the variables $\boldsymbol{X}$ are asymptotically dependent. In the case of $\chi(D) = 0$, the variables cannot all take the largest values together; however, there might exist lower dimensional subvectors $\boldsymbol{X}_C = \{X_i : i \in C\}$, where $C \subset D$, that exhibit asymptotic dependence with $\chi(C) > 0$ (Simpson et al., 2020).

### 2.2.5 Modelling asymptotically independent data

Several of the available models in the (multivariate) literature are only suitable for one extremal dependence regime. Recall that in the bivariate case the spectral measure $H$ places all mass on $\{0\}$ and $\{1\}$ when the maxima are independent; this is also the case when $(X_1, X_2)$ are asymptotically independent, resulting in $H$ not being able to distinguish between exact independence and asymptotic independence. Therefore, the componentwise maxima and regular variation approaches introduced in Sections 2.2.2 and 2.2.3, respectively, are only designed for modelling asymptotic dependence. Furthermore, since misspecifying the extremal structure leads to inaccurate representations of the extremal region and, therefore, incorrect extrapolations, it is important to categorise data as belonging to the correct dependence class.

The first approach to modelling asymptotically independent data was introduced by Ledford and Tawn (1996). Consider a bivariate random vector $(X_1, X_2)$ with $X_i \sim F_{X_i}$, $i = 1, 2$, as before. Given a function $\mathcal{L}$ that is slowly-varying at zero[1], the joint tail is assumed to behave as

$$\Pr[F_{X_2}(X_2) > u \mid F_{X_1}(X_1) > u] = \mathcal{L}(1-u)(1-u)^{1/\eta - 1}, \tag{2.2.7}$$

as $u \to 1$, and $\eta \in (0, 1]$. The extremal dependence structure can then be quantified through $\eta$, which is often denoted as the residual tail dependence coefficient. In particular, if $\eta = 1$ and $\mathcal{L}(1-u) \not\to 0$ as $u \to 1$, then the variables $X_1$ and $X_2$ are asymptotically dependent, and asymptotically independent otherwise. Moreover, the coefficient $\eta$ provides information about the strength of asymptotic independence of a given vector; if $\eta \in (0, 1/2)$, then the variables are negatively associated in the extremes, whilst if $\eta \in (1/2, 1)$ they are positively associated in the extremes. Exact independence is obtained if $\eta = 1/2$ and $\mathcal{L}(1-u) = 1$, and near independence achieved

[1]For any $c > 0$, $\mathcal{L}(cx)/\mathcal{L}(x) \to 1$ as $x \to 0$.

if $\eta = 1/2$ and $\mathcal{L}(1-u) \neq 1$.

Similarly to measure $\chi(u)$ given in equation (2.2.6), it is possible to quantify the dependence at sub-asymptotic levels. Given a particular value $u \in (0,1)$, $\eta(u)$ can be obtained as

$$\eta(u) = \frac{\log(1-u)}{\log \Pr[F_{X_1}(X_1) > u, F_{X_2}(X_2) > u]},$$

with $\eta = \lim_{u\to 1} \eta(u)$.

An extension of equation (2.2.7) is also possible to a $d$-dimensional setting. Consider again a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ where $X_i \sim F_{X_i}$ for $i \in D$; Eastoe and Tawn (2012) define the $d$-dimensional joint tail behaviour through

$$\Pr[F_{X_i}(X_i) > u : \forall i \in D] = \mathcal{L}(1-u)(1-u)^{-1/\eta(D)-1},$$

as $u \to 1$, and $\eta(D) \in (0,1]$. The variables $\boldsymbol{X}$ exhibit asymptotic dependence if $\eta(D) = 1$, and $\lim_{u\to 1} \mathcal{L}(1-u) > 0$. However, if $\eta(D) = 1$ and $\lim_{u\to 1} \mathcal{L}(1-u) = 0$, or if $\eta(D) < 1$, then the variables cannot all be extreme together, but subvectors $\boldsymbol{X}_C = \{X_i : i \in C\}$, where $C \subset D$, of lower dimension than $d$ can still be asymptotically dependent as mentioned in Section 2.2.4 (Simpson et al., 2020).

These coefficients are only informative in regions where all variables are extreme, which may not always be the case. For situations where the interest may be in regions where just one variable is extreme, Wadsworth and Tawn (2013) propose an extension of equation (2.2.7). Given standard exponentially distributed variables $X_i$, that is $F_{X_i}(x_i) = 1 - \exp\{-x_i\}$ for $i \in D$, the joint tail behaviour of $\boldsymbol{X} = (X_1, \ldots, X_d)$ is captured through function $\lambda(\boldsymbol{w})$ via the assumption

$$\Pr(X_i > w_i v : \forall i \in D) = \mathcal{L}(\exp\{v\}; \boldsymbol{w}) \exp\{-\lambda(\boldsymbol{w})v\},$$

as $v \to \infty$, where $\boldsymbol{w} \in \mathcal{S}_{d-1} = \{\boldsymbol{w} \in [0,1]^d : \sum_{i=1}^d w_i = 1\}$, the function $\mathcal{L}(\cdot; \boldsymbol{w})$

is slowly-varying at infinity for each $\boldsymbol{w}$, and $\lambda(\boldsymbol{w}) \geq \max\{\boldsymbol{w}\}$ is known as the angular dependence function (ADF). When the variables $\boldsymbol{X}$ are asymptotically dependent, then $\lambda(\boldsymbol{w}) = \max\{\boldsymbol{w}\}$ for all $\boldsymbol{w} \in \mathcal{S}_{d-1}$. Moreover, the residual tail dependence coefficient $\eta(D)$ can be obtained through the ADF by setting $\eta(D) = [d\lambda(\mathbf{1}/d)]^{-1}$ where $\mathbf{1}/d = (1/d, \ldots, 1/d) \in \mathbb{R}^d$.

### 2.2.6 Conditional extreme value model

An alternative method, suitable for studying the regions where only a subset of variables is extreme, is the conditional approach of Heffernan and Tawn (2004). In this approach the interest lies in studying the joint behaviour of random variables given that one of them is large. Consider the random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ on standard Laplace margins, that is $X_i \sim F_{X_i}$ where

$$F_{X_i}(x) = \begin{cases} \frac{1}{2}\exp\{x\}, & \text{if } x \leq 0, \\ 1 - \frac{1}{2}\exp\{-x\}, & \text{if } x \geq 0, \end{cases}$$

for $i = 1, \ldots, d$. Additionally, let us condition on variable $X_i$ being larger than some high threshold $u_i$, i.e. $X_i > u_i$. Assuming that there exist parameter vectors $\boldsymbol{\alpha}_{|i} \in [-1, 1]^{d-1}$ and $\boldsymbol{\beta}_{|i} \in (-\infty, 1]^{d-1}$, Heffernan and Tawn (2004) and Keef et al. (2013a) show that, for a wide variety of underlying dependence structures and all fixed $\boldsymbol{z}_{|i}$, we have

$$\Pr(\boldsymbol{Z}_{|i} \leq \boldsymbol{z}_{|i}, X_i - u_i > x \mid X_i > u_i) \to G_{|i}(\boldsymbol{z}_{|i})\exp\{-x\}, \quad x > 0, \tag{2.2.8}$$

as $u_i \to \infty$, for some non-degenerate function $G_{|i}$ and standardised residuals

$$\boldsymbol{Z}_{|i} = \frac{\boldsymbol{X}_{-i} - \boldsymbol{\alpha}_{|i} X_i}{X_i^{\boldsymbol{\beta}_{|i}}},$$

where $\boldsymbol{X}_{-i}$ denotes the random vector $\boldsymbol{X}$ excluding its $i^{\text{th}}$ component, for $i = 1, \ldots, d$. It follows from equation (2.2.8) that variables $\boldsymbol{Z}_{|i}$ and $X_i - u_i$ are independent for $u_i \to \infty$ and $X_i > u_i$, and variable $X_i - u_i$ is standard exponentially distributed. Furthermore, the conditional extremes approach is able to capture both extremal dependence regimes; we have asymptotic dependence when $\boldsymbol{\alpha}_{|i} = \mathbf{1}$ and $\boldsymbol{\beta}_{|i} = \mathbf{0}$, and the variables cannot all be large together otherwise. Finally, negative dependence between the variables is captured for $\boldsymbol{\alpha}_{|i} < \mathbf{0}$ (Keef et al., 2013a).

### 2.2.7 Random scale constructions

The majority of statistical models available for modelling multivariate extremes usually suit only one regime of extremal dependence, or both with asymptotic dependence as a boundary case. Depending on the specification of each variable in the model, bivariate dependence models using a random scale representation may be able to capture both asymptotic dependence and independence, with the transition between the two dependence regimes achieved smoothly at interior points of the parameter space. Examples of such models include those proposed by Wadsworth et al. (2017), Huser and Wadsworth (2019) and Engelke et al. (2019). A detailed overview of the dependence properties of models constructed using a random scale representation is given in Engelke et al. (2019).

Let $R > 0$ follow a non-degenerate distribution and be independent of a random vector $(W_1, W_2) \subseteq \mathbb{R}^2$. A random scale construction is defined as the bivariate random vector $(X_1, X_2)$ constructed as

$$(X_1, X_2) = R(W_1, W_2).$$

For the model proposed by Wadsworth et al. (2017), the variable $R$ follows a GP distribution with scale parameter 1 and shape parameter $\xi \in \mathbb{R}$. In addition, the random

vector $(W_1, W_2)$ is constructed as follows

$$(W_1, W_2) = \frac{(W, 1-W)}{\|(W, 1-W)\|_m} \in \mathcal{S}^m = \{\boldsymbol{w} \in \mathbb{R}_+^2 : \|\boldsymbol{w}\|_m = 1\},$$

where $W$ is a random variable with cdf $F_W$. Furthermore, the norm $\|\cdot\|_m$ and $F_W$ are modelling choices. With $\|\cdot\|_m$ taken as the $L_\infty$ norm, that is $\|(W, 1-W)\|_\infty = \max(W, 1-W)$, this model is able to interpolate smoothly between asymptotic independence and asymptotic dependence through the GP shape parameter $\xi$. In particular, for $\xi > 0$, the random vector $(X_1, X_2)$ exhibits asymptotic dependence with

$$\chi = \mathrm{E}\left(\min\left\{\frac{W_1^{1/\xi}}{\mathrm{E}\left(W_1^{1/\xi}\right)}, \frac{W_2^{1/\xi}}{\mathrm{E}\left(W_2^{1/\xi}\right)}\right\}\right) > 0, \text{ and } \eta = 1.$$

When $\xi \leq 0$, then the vector $(X_1, X_2)$ is asymptotically independent with $\chi = 0$. Moreover, if $\xi = 0$, then $\eta = \|(1,1)\|_\infty^{-1}$, and when $\xi < 0$, $\eta = (1-\xi)^{-1}$.

In turn, Huser and Wadsworth (2019) assume that all variables are Pareto distributed with different shape parameters. More specifically, for $\delta \in (0, 1)$, the cdf of variable $R$ is given as $F_R(r) = 1 - r^{-1/\delta}$, whilst we have $F_{W_i}(w_i) = 1 - w_i^{-1/(1-\delta)}$ for variables $W_i$, $i = 1, 2$. Furthermore, $(W_1, W_2)$ follows an an asymptotically independent copula with residual tail dependence coefficient $\eta_W < 1$. Asymptotic dependence is present when $\delta > 1/2$ with

$$\chi = \mathrm{E}\left(\min\left\{\frac{W_1^{1/\delta}}{\mathrm{E}\left(W_1^{1/\delta}\right)}, \frac{W_2^{1/\delta}}{\mathrm{E}\left(W_2^{1/\delta}\right)}\right\}\right) > 0, \text{ and } \eta = 1.$$

Asymptotic independence is present when $\delta \leq 1/2$, i.e., $\chi = 0$, while $\eta$ can take the

following values

$$\eta = \begin{cases} 1, & \text{if } \delta = 1/2, \\ \delta/(1-\delta), & \text{if } \eta_W/(1+\eta_W) < \delta < 1/2, \\ \eta_W, & \text{if } \delta \leq \eta_W/(1+\eta_W). \end{cases}$$

Different bivariate copulas can be assumed for the random vector $(W_1, W_2)$, with the Gaussian copula being a natural choice used in Huser and Wadsworth (2019).

Two different models were proposed by Engelke et al. (2019); in both, the variable $W$ is assumed to follow a Beta distribution with shape parameter $\alpha > 0$, i.e. $W \sim \text{Beta}(\alpha, \alpha)$. For the first proposed model, the variable $R$ is assumed to follow a Weibull distribution with shape parameter $\beta > 0$ and scale parameter 1, that is $F_R(r) = 1 - \exp\{-r^\beta\}$. The random vector $(W_1, W_2)$ is constructed in a similar way to the approach of Wadsworth et al. (2017); more specifically, we have

$$(W_1, W_2) = \frac{(W, 1-W)}{\nu(W, 1-W)},$$

where $\nu(W, 1-W) = \theta \max(W, 1-W) + (1-\theta)\min(W, 1-W)$ with $\theta \geq 1/2$. The extremal dependence of this model is controlled by parameter $\theta$; we have asymptotic independence with $\chi = 0$ and $\eta = \theta^\beta$ when $\theta \leq 1$, and asymptotic dependence with $\chi = 2(\theta-1)/(2\theta-1)$ and $\eta = 1$ when $\theta > 1$.

For the second model proposed by Engelke et al. (2019), the variable $R$ follows again a GP distribution with scale parameter 1 and shape parameter $\xi \in \mathbb{R}$. In addition, the variables $W_1$ and $W_2$ are now independent of each other and follow a Beta distribution with shape parameter $\alpha > 0$. Like the model of Wadsworth et al. (2017), the GP shape parameter $\xi$ determines the extremal dependence class. If $\xi > 0$, then variables $X_1$ and $X_2$ are asymptotically dependent with

$$\chi = \frac{\text{E}\left(\min\{W_1, W_2\}^{1/\xi}\right)}{\text{E}\left(W_1^{1/\xi}\right)} \quad \text{and} \quad \eta = 1.$$

When $\xi \leq 0$, variables $X_1$ and $X_2$ exhibit asymptotic independence, where $\chi = 0$. Moreover, $\eta = 1$ if $\xi = 0$, and $\eta = (1 - \xi\alpha)/(1 - 2\xi\alpha)$ if $\xi < 0$.

In all of the above random scale models, the inference procedure is done by exploiting the copula $C$ of the models in order to uniquely capture the dependence structure independently of the marginal distribution of each variable involved in the models. Furthermore, the inference procedure is performed via censored likelihood-based methods; given that the interest lies in capturing the tail behaviour of $(X_1, X_2)$, non-extreme contributions need to be excluded from the inference procedure to prevent bias.

### 2.2.8 Approaches to model the body and tail jointly

Similarly to the univariate setting, in the multivariate extremes literature it is usually the case that a multivariate threshold is needed to define an extremal region in which an asymptotically-motivated distribution (or copula) is assumed to hold. A way to avoid such a choice is to consider models that are able to jointly represent the body and tail regions. Additionally, such models are required when modelling the whole region accurately is of interest.

Several methods to create models with more flexible dependence structures are available in the copula literature, especially in financial applications. One such method is known as transformation or distortions of copulas, in which known copulas are transformed into a new copula by means of a bijection on the unit interval; such copulas provide a more accurate representation of the dependence structure that might be necessary in certain scenarios (see e.g., Durrleman et al., 2000, Morillas, 2005, Klement et al., 2005, Durante et al., 2010). In particular, given a bijection function $\gamma : [0, 1] \to [0, 1]$, a bivariate copula $C : [0, 1]^2 \to [0, 1]$ can be transformed to a new copula $C_\gamma : [0, 1]^2 \to [0, 1]$ as $C_\gamma(x, y) = \gamma^{-1}\left(C(\gamma(x), \gamma(y))\right)$. While these methods focus more on constructing new copula families, and not on accurately representing both the body and tail regions of a data set, Durrleman et al. (2000) and Durante et al. (2010) show that imposing

specific conditions on the bijection function $\gamma$ changes the dependence structure of the new copula $C_\gamma$ when compared to the original copula $C$. More specifically, in Durrleman et al. (2000), $C_\gamma$ preserves the same tail behaviour as $C$, while the overall dependence captured with Kendall's concordance coefficient (Kendall, 1938) of $C_\gamma$ is different than that of $C$. In contrast, with the bijection $\gamma$ defined by Durante et al. (2010), the new copula $C_\gamma$ exhibits different extremal behaviour than the original copula $C$.

Similarly to the transformation of copulas approaches, methods based on piecewise constructions (see for instance Hummel, 2009) or convex combinations (see e.g., Bacigál et al., 2010, Shamiri et al., 2011) focus on creating new copula families, which might be more flexible in certain applications than the standard copula families. By nesting copulas in each other so that box copulas are constructed, Hummel (2009) is able to control and, therefore modify, the extremal behaviour of a data set. On the other hand, Bacigál et al. (2010) and Shamiri et al. (2011) propose convex combinations of known copulas, especially belonging to the Archimedean family; the former considers additive generators of binary copulas, whilst Shamiri et al. (2011) constructs a Clayton-Gumbel copula, based on a standard mixture of these two copulas. In this way, Shamiri et al. (2011) are able to represent any asymmetry that might be present in the data and capture strong dependence in both tails.

An alternative way of defining new copulas that are able to capture more complicated dependence structures is by considering patchwork copulas (Pfeifer and Ragulina, 2021); these include copulas based on ordinal sums (Alsina et al., 2006), gluing copulas (Mesiar et al., 2008, Siburg and Stoimenov, 2008) and copulas based on rectangular representations (Durante et al., 2009). In Durante et al. (2013) a generalised method to construct such copulas is proposed; specifically, given a copula $C$, the probability mass distribution of its patchwork copula only differs from that of copula $C$ within a $d$-dimensional box $B \subseteq [0,1]^d$. This modification aims to overcome the difficulty of representing the tail region when considering the full data set by allowing the patchwork

copula to have different tail dependencies properties than copula $C$.

In turn, Pfeifer et al. (2016) propose generalised partition-of-unity copulas, which were later formalised for the discrete and continuous cases by Pfeifer et al. (2017) and Pfeifer et al. (2019), respectively. More specifically, by considering infinite, as opposed to finite, partitions-of-unity, the proposed copulas are able to capture tail dependence and asymmetry in the data. This is achieved by approximating the density of the model by an infinite mixture of functions, and the modelling is done in a data driven procedure.

In the approach of Hu and O'Hagan (2021), several different copula families are fit to the full data set and, by averaging over these families, the authors aim to obtain a more robust estimate of the tail dependence of a data set; this is achieved by using Bayesian model averaging (BMA). Where the quantity of interest is the extremal dependence measure $\chi$, and $C_i$ $(i = 1, \ldots, K)$ are $K$ fitted copulas to the data set $\mathcal{D}$, the posterior distribution of $\chi$ given the data $\mathcal{D}$ is given by

$$\Pr(\chi \mid \mathcal{D}) = \sum_{i=1}^{K} \Pr(\chi \mid C_i, \mathcal{D}) \Pr(C_i \mid \mathcal{D}).$$

Assuming that all copulas are equally likely a priori, the weight for copula $C_i$, given by $W_{C_i}$, is approximated using the Bayesian Information Criterion (BIC) as follows

$$W_{C_i} = \Pr(C_i \mid \mathcal{D}) \approx \frac{\exp\{-\text{BIC}_i/2\}}{\sum_{j=1}^{K} \exp\{-\text{BIC}_j/2\}},$$

for $i = 1, \ldots, K$. By using BMA, the authors aim at improving the estimation of the tail behaviour; the use of BIC in the weight calculations, however, assigns the focus on the body and not on the tail of the data.

The modelling of the full distribution when the focus is on the extreme events is considered in the spatial literature by Gräler (2014), Krupskii et al. (2018) and Zhang et al. (2022b), for instance. In particular, Gräler (2014) considers convex combinations

of bivariate copulas when constructing a spatial vine copula. In this way, different dependence properties between each location are captured, allowing for a better extrapolation of the tail. In turn Krupskii et al. (2018) and Zhang et al. (2022b) propose modelling the whole data set, not specifically aiming at an accurate representation of both regions, but as a way of avoiding the computational burden inherent with censored likelihoods for extremes. In spite of that, both methods show reasonable flexibility in capturing the body and tail of the data set.

Methods which aim specifically at representing the body and tail regions of the data set in an accurate manner are still scarce in the literature. For instance, Vrac et al. (2007) propose a bivariate extension of the univariate model of Frigessi et al. (2002) to jointly model precipitation intensities. Let $X_1$ and $X_2$ be two positive and heavy tailed random variables, and consider the radial and angle variables $R = X_1 + X_2$ and $W = X_2/R$, respectively; a bivariate event is considered extreme when variable $R$ is large. Since an observed angle has support in the unit interval, i.e. $w \in [0, 1]$, it is assumed that $W \sim \text{Beta}(\beta_1, \beta_2)$ with $\beta_1, \beta_2 > 0$. Vrac et al. (2007) propose modelling the bulk of the distribution with bivariate Gamma random vectors whilst the tail region is modelled in the pseudo-polar coordinate system $(R, W)$. The authors bypass the need for selecting a threshold when transitioning from the body region to the tail by considering a weighting function, $p$, that varies only with $R$. The proposed model has density given as follows

$$f(x_1, x_2; \boldsymbol{\theta}) = k_{\boldsymbol{\theta}} \left[ (1 - p(r; \mu) g(x_1, x_2; \boldsymbol{\gamma}) + p(r; \mu) h(r; \varphi, \xi) b(w; \beta_1, \beta_2) \right],$$

where $k_{\boldsymbol{\theta}}$ is a normalising constant, $\mu$ is the location parameter of the weighting function, $h$ is the density function of the GP distribution given in equation (2.1.2) with scale and shape parameters $\varphi$ and $\xi$, respectively, $g$ is the density function of a bivariate Gamma distribution with vector of parameters $\boldsymbol{\gamma}$, $b$ is the density function of a Beta distribution with shape parameters $\beta_1$ and $\beta_2$, and $\boldsymbol{\theta}$ is the vector of model parameters.

In turn, Aulbach et al. (2012a) and Hu et al. (2024) propose two different extensions to the multivariate setting of the model of Behrens et al. (2004). The former define a new copula by joining two $d$-dimensional copulas; in particular, they fit one copula to the data and substitute its upper tail region with a different copula. While in theory any copula can be used for the upper tail region, they propose using a copula of a multivariate GP distribution in their model, meaning that the proposed copula is only suitable to capture asymptotic dependence. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ be two independent random vectors, and $C_1$ and $C_2$ be two copulas. Further, assume that $\boldsymbol{X} \sim C_1$ and $\boldsymbol{Y} \sim C_2$, where $C_1$ and $C_2$ are two copulas with shifted support in $[-1, 0]^d$, and let $\boldsymbol{t} = (t_1, \ldots, t_d)$ denote a vector of appropriate threshold values. Aulbach et al. (2012a) construct a random vector $\boldsymbol{Q}$, whose $i^{th}$ element is given by

$$Q_i := Y_i \mathbb{1}_{Y_i \leq t_i} - t_i X_i \mathbb{1}_{Y_i > t_i},$$

for $i = 1, \ldots, d$. Then, the random vector $\boldsymbol{Q}$ follows a copula with support on $[-1, 0]^d$. Moreover, its copula coincides with $C_1$ on the region $(t_1, 0] \times \ldots \times (t_d, 0]$ and with $C_2$ on $[-1, t_1] \times \ldots \times [-1, t_d]$. In a later work of Aulbach et al. (2012b), an exact representation of this copula is provided.

On the other hand, Hu et al. (2024) propose a mixture model where the body is modelled with a parametric distribution in the max-domain of attraction (MDA) of a multivariate generalised extreme value (mGEV) distribution as given in equation (2.2.2), and the tail is described by a multivariate generalised Pareto (mGP) distribution. Furthermore, they define an extremal region where at least one of the components is larger than a threshold vector $\boldsymbol{t}$. Let $F_b$ denote the distribution function fitted to the bulk region; Hu et al. (2024) assume that $F_b$ is a multivariate Gaussian distribution in their work, although any distribution in the MDA of the mGEV is permitted. In a similar manner to the univariate model of Behrens et al. (2004), the density function of the

proposed model is given as follows

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \begin{cases} f_b(\boldsymbol{x};\boldsymbol{\theta}_b), & \text{if } \boldsymbol{x} \leq \boldsymbol{t}, \\ [1 - F_b(\boldsymbol{t};\boldsymbol{\theta}_b]h(\boldsymbol{x}-\boldsymbol{t};\boldsymbol{\theta}_t), & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_b, \boldsymbol{\theta}_t, \boldsymbol{t})$ is the vector of model parameters, $f_b$ is the density function of the distribution assumed for the bulk region, whose parameters are $\boldsymbol{\theta}_b$, and $h$ is the density function of the mGP distribution with parameters $\boldsymbol{\theta}_t$; see Hu et al. (2024) for the specific characterisation of density $h$. In order to avoid choosing an appropriate vector of threshold values, these are treated as parameters in the model, as opposed to the approach of Aulbach et al. (2012a). However, given that the upper tail region is described by a mGP distribution, this method is only suitable for asymptotically dependent data.

Finally, Leonelli and Gamerman (2020) propose a semi-parametric approach suitable for both asymptotic dependence and independence; they do so by using a copula-based framework, allowing for the separation of the marginal modelling from the dependence structure. More specifically, the authors model each variable using the EVMM proposed by do Nascimento et al. (2012), and given in equation (2.1.8), while the dependence structure is modelled by a mixture of copulas, which are assumed to belong to the elliptical family. Although the model is suitable for both regimes of extremal dependence, the model is only able to capture asymptotic independence if all the copula terms in the mixture are suitable for AI data.

## 2.3 Neural likelihood-free inference

Several models available in the multivariate extremes literature, such as those from Section 2.2.7, have likelihood functions that rely heavily on numerical integration, inversion of functions, or a combination of the two; this results in computationally

expensive likelihood-based inference, which may constitute a potential barrier to the routine use of such models in practice. However, even when the likelihood function is intractable, it is usually possible to simulate from the model. In such cases, simulation-based techniques, which are often likelihood-free, are an appealing alternative avenue to perform inference. The most commonly used simulation-based methods are mentioned in Section 2.3.1, and we introduce simulation-based techniques which leverage neural networks in Section 2.3.2.

### 2.3.1 Simulation-based approaches

As stated by Cranmer et al. (2020), in simulation-based methods, the statistical model is often defined through the simulator function; such function takes the vector of model parameters as inputs, and outputs data observations generated from the model. There are several simulation-based procedures available in the literature, with density estimation (Diggle and Gratton, 1984) and approximate Bayesian computation (ABC; see e.g. Lintusaari et al., 2017 or Sisson et al. (2018) for a review) being among the most commonly used methods.

In density estimation methods, histograms or kernel density estimates of simulated data are used to approximate the log-likelihood function. Furthermore, although the simulation and estimation of the density steps are computationally expensive, once the likelihood function is approximated, new observations can be evaluated efficiently. This leads to an amortised inference procedure as the computationally expensive steps need not be repeated for new data. Owing to this, density estimation methods are well suited for handling replicated data (Cranmer et al., 2020).

In turn, ABC methods are seen as rejection sampling algorithms, whereby observed and simulated data are compared based on some distance measure, which often involves some summary statistics. More specifically, the model parameters are drawn from a prior distribution, and subsequently accepted if the simulated data are sufficiently close

to the observed data, and rejected otherwise. Moreover, the acceptance probability is determined for a tolerance level $\varepsilon > 0$; when $\varepsilon \to 0$, then the inference procedure becomes exact. Choosing a suitable prior distribution, defining the similarity between the samples, and choosing a suitable tolerance value constitute some drawbacks of using ABC to perform inference. For instance, while uninformative priors are in general more applicable, large distances between the data can be produced by most of parameter values; this leads to the need for more simulated data sets, which comes at a computational cost (Grazian and Fan, 2020). Regarding the tolerance level, smaller values of $\varepsilon$ results in a higher number of simulations that are infeasible. On the other hand, an increase in sample efficiency obtained by a large $\varepsilon$ comes at the expense of inference quality. Contrary to density estimation procedures, ABC methods are not amortised; in fact, the majority of the algorithm steps have to be undertaken for new observations, which makes ABC more suitable for cases where the number of i.i.d. observations is small (Cranmer et al., 2020).

Both density estimation and ABC methods suffer from the curse of dimensionality, performing poorly for high dimensional data. Alternatively, the synthetic likelihood method proposed by Wood (2010) is computationally more efficient in higher dimensional settings. More specifically, the synthetic likelihood method constructs an approximate likelihood function by assuming that the summary statistics follow a multivariate Gaussian distribution, whose mean and variances are approximated by averages over a set of simulated summary statistics (Grazian and Fan, 2020). Whilst the approximated likelihood becomes more accurate when the number of simulated data sets increases, simulating new data sets is particularly expensive. Furthermore, the quality of inference depends on how close the Gaussian assumption is to the truth. The synthetic likelihood method is usually easier to tune than ABC methods, but the Gaussian assumption may lead to sub-optimal inferences in some situations. More recently, and still assuming that the summary statistics follow a multivariate Gaussian distribution,

Price et al. (2018) propose using the synthetic likelihood method to target the posterior distribution of the data.

Alternative simulation-based methods include indirect inference or pseudo-marginal Markov chain Monte Carlo (MCMC). Indirect inference methods involve using an auxiliary parametric model, which is analytically or computationally more tractable than the likelihood function. Then, by deriving some summary statistics from this auxiliary model, the relationship between the parameters of the auxiliary model and those of the likelihood function is analysed, and point estimates for the model parameters can be obtained (Gourieroux et al., 1993). In pseudo-marginal MCMC, the likelihood function is approximated by unbiased estimates, often obtained through importance sampling. Since the expected value of these estimates correspond to the true likelihood function, the pseudo-marginal MCMC algorithm is able to correctly sample from it; see for instance Beaumont (2003) or Andrieu and Roberts (2009) for more details.

### 2.3.2 Neural simulation-based methods

In recent years, simulation-based approaches which leverage neural networks have been emerging in the literature; Zammit-Mangion et al. (2025) review several of these methods. Briefly, a neural network works as a function approximator, mapping the inputs to the outputs in a non-linear way. An application of neural simulation-based methods is to use neural networks to obtain parameter point estimates. For instance, Gerber and Nychka (2021) use neural networks to estimate the covariance parameters of a spatial Gaussian process; more specifically, the network learns by taking synthetically generated fields and their corresponding (known) covariance parameters as inputs. In the approach of Lenzi et al. (2023), deep learning techniques are used to perform inference for max-stable processes given that simulating from such models is often fast and tractable. In this, data generated from the max-stable model serves as the input to a convolutional neural network (CNN), which maps the data to the parameter space and

outputs the parameter estimates.

In turn, Majumder et al. (2024) construct a process mixture model based on a convex combination of a max-stable and a Gaussian processes, whose likelihood function is intractable. The authors use feed-forward neural networks to approximate the distribution at one spatial location given a neighbouring set, and develop synthetic likelihood functions to approximate the likelihood function. An extension of the process mixture model which aims to account for non-stationarity is proposed later by Majumder and Reich (2023).

Alternatively, Sainsbury-Dale et al. (2024a) propose using neural Bayes estimators (NBEs) to estimate the vector of model parameters for a general parametric model. Similarly to the approach of Lenzi et al. (2023), the neural network takes as inputs data, $\boldsymbol{X}$, generated by the model, which are then mapped to a point summary of the posterior distribution, returning an estimate of the vector of model parameters, $\hat{\boldsymbol{\theta}}$. More specifically, permutation-invariant neural networks are used to approximate Bayes estimators, which minimise a weighted average of the Bayes risk, $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$.

Given $n$ i.i.d. replicates of the data $\boldsymbol{X}_1, \dots \boldsymbol{X}_n \in \mathcal{S}$ and parameter space $\Theta$, a deep neural network is used to approximate point estimators $\hat{\boldsymbol{\theta}} : \mathcal{S}^n \to \Theta$. Such an approximation is achieved by finding the best parameters of the neural network (often known as the 'weights' and 'biases') that minimise a loss function, $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$, which penalises the distance between the observed data and the realisations of the output. In particular, $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \mathrm{E}\left[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})\right]$. Additionally, the proposed approach is able to handle replicated data with different sample sizes.

This methodology was further developed by Richards et al. (2024) to account for censored observations as inputs; when the interest lies in the extremal behaviour, methodologies able to handle censored data are crucial since we do not wish non-extreme observations to affect the extremal dependence estimation. The proposed approach by Richards et al. (2024) further accounts for cases where the censoring level is not

assumed fixed; this is achieved by considering the censoring level as an extra input to the neural network. More recently, Sainsbury-Dale et al. (2024b) propose using graph neural networks to handle irregular spatial data, which are collected over an arbitrary set of spatial locations.

Instead of considering a point estimation approach, Walchessen et al. (2024) design a binary classification task where a CNN learns the likelihood function of spatial processes from which it is possible to simulate. In particular, simulations from the model are used to create two classes, each consisting of pairs of model parameters and their corresponding field realisations. The classifier is then trained to discriminate between these two classes. Additionally, the classifier allows to produce not only the learned likelihood surfaces, but also estimates of the model parameters and confidence regions.

Finally, Lenzi and Rue (2023) avoid the bias towards the parameter region of the training data, often present in neural point approaches, by proposing an automatic iterative approach; by dynamically updating the distribution of the parameters, the training data are iteratively modified until the region of the parameters corresponding to the actual data is achieved. Similarly to the approach of Sainsbury-Dale et al. (2024a), the proposed method is also able to handle replicated data with different sample sizes. Finally, while the training step of a neural network is computationally expensive, once completed, inference on the model parameters is typically achieved in fraction of seconds; in particular, the trained network can be applied every time there is new data (Richards et al., 2024). This results in an amortised estimation process over time, which is particularly useful in online inference, for example.

The use of neural networks as a modelling technique has also been growing in the extremes literature. For instance, in a univariate framework, Cannon (2010), Cannon (2011), Vasiliades et al. (2015), Bennett et al. (2015) and Shrestha et al. (2017) use neural networks to estimate the parameters of a GEV distribution when accounting for non-stationarity in the data. More specifically, a conditional density network is

used to specify the GEV parameters as a function of covariates, allowing the model to capture a wide range of non-stationary relationships. Ceresetti et al. (2012) use neural networks to facilitate the estimation of return levels (i.e., the value we expect to be exceeded, on average, once every $1/p$ observations, where $p$ is a small probability), whereas Carreau and Vrac (2011) propose a new class of conditional mixture models, in which the EVMM of Carreau and Bengio (2009) is included (given in Section 2.1.3), that builds on neural networks. In the multivariate framework, this area of research is mainly predominant in a spatial setting. For instance, Ahmed et al. (2022) and Wixson and Cooley (2024) leverage CNNs as classification tools. Contrary to the approach of Walchessen et al. (2024), these aim at distinguish between the two regimes of extremal dependence defined in Section 2.2.4. More recently, Ahmed et al. (2024) propose using such neural networks for model selection of appropriate extremal dependence structures in a spatial setting. Murphy-Barltrop et al. (2024) use neural networks to learn about the extremal dependence structure of higher-dimensional data. More specifically, the authors consider a geometric approach, where the joint behaviour can be inferred from a star-shaped limit set; this limit set is estimated using a multi-layer perceptron neural network. In a regression context, Cisneros et al. (2024) propose using graph CNNs and a GP distribution to model the full distribution of wildfire spread. On the other hand, Pasche and Engelke (2024) propose an extreme quantile regression network which estimates the scale and shape parameters of a non-stationary GP distribution, that hence depend on covariates, through a neural network, whilst Richards and Huser (2022) create partially-interpretable neural networks to perform extreme quantile regression.

# Chapter 3

# Joint modelling of the body and tail of bivariate data

## 3.1 Introduction

### 3.1.1 Motivation

When dealing with environmental phenomena such as high temperatures, wind speeds or air pollution, or with financial applications such as insurance losses, interest often lies in modelling the extreme observations, which are typically scarce. For such cases, a model with focus on the tail of the distribution is required as common statistical models that may be used to fit the entire data set lead to poor estimates of the extremes. To overcome this issue, models based on extreme value theory (EVT) can be applied; these aim to quantify the behaviour of a process at extremely large (or small) values of a series. Typically, the generalised extreme value (GEV) distribution is fitted to block maxima, often annual maxima, or the generalised Pareto distribution (GPD) is fitted to data exceeding a high threshold. The former can be seen as a wasteful approach if there are more data on extremes available, while the latter usually requires a subjective choice of threshold, which inevitably leads to uncertainty, with different choices leading

to different results; see Coles (2001).

However, in some cases, interest not only lies in modelling the extreme observations accurately but also fitting the non-extremes well, meaning a flexible model over the whole support of the distribution is required. For instance, the concentration of pollutants in the air may be so high that harmful levels are actually in the body of the data set. Thus, from a public health perspective, we care not only about the probability of exceeding extreme, and potentially more dangerous, pollutant levels but also about the probability of exceeding harmful yet locally moderate levels. Fitting a model to both the bulk (i.e., the non-extreme observations) and tail (i.e., the extreme observations) of a data set has been dealt with in the univariate framework but little work has been done in extending to a multivariate setting. In this work, we outline an approach that offers dependence models for the bulk and tail, while ensuring a smooth transition between the two.

### 3.1.2 Background

In the univariate setting, several models have been proposed to join one distribution for the bulk to a GPD for the tail. Scarrott and MacDonald (2012) review several of these approaches, hereafter referred to as extreme value mixture models, or EVMMs. These models aim to account for the uncertainty in the choice of threshold, by implicitly or explicitly estimating it. With EVMMs, care is needed so that the bulk and tail are not excessively influenced by each other, though they cannot be fully disjoint since they share information. Parametric EVMMs entail fitting a specified distribution to the bulk and a GPD to the tail, while semi-parametric models fit a GPD to the tail with a more flexible model in the bulk. Behrens et al. (2004) propose a parametric model, which exhibits discontinuity at the threshold; Carreau and Bengio (2009) avoid this by forcing continuity up to the first derivative of the density function. On the other hand, Frigessi et al. (2002) fit two distributions to the whole data, giving more weight to the bulk at

low ranges in the support and to the GPD in the upper tail by means of a dynamic weighting function $p(x;\theta) \in (0,1]$. The density of their model is defined as

$$h(x;\theta,\boldsymbol{\beta},\boldsymbol{\alpha}) = \frac{[1-p(x;\theta)]f(x;\boldsymbol{\beta}) + p(x;\theta)g(x;\boldsymbol{\alpha})}{K(\theta,\boldsymbol{\beta},\boldsymbol{\alpha})},$$

where $g(x;\boldsymbol{\alpha})$ is the density of the GPD with vector of parameters $\boldsymbol{\alpha}$, $f(x;\boldsymbol{\beta})$ is a density with a lighter tail and vector of parameters $\boldsymbol{\beta}$, $K(\theta,\boldsymbol{\beta},\boldsymbol{\alpha})$ is a normalising constant and $p(x;\theta)$ is increasing in $x$ for all $\theta$. Because $p(x;\theta)$ depends on $x$, it favours the GPD in the upper tail whilst the lower tail is controlled by $f(x;\boldsymbol{\beta})$. However, careful choice of the weighting function is needed since some functions, such as the unit step function, may lead to a discontinuity in the transition between the two distributions; see Frigessi et al. (2002) for details. More recently, methods introduced by Naveau et al. (2016) and Stein (2021) aim to model the lower and upper tails of the data with GPDs, while ensuring a smooth transition between the regions. The former achieve this by constructing a model relying on compositions of functions, where one is a cumulative distribution function (CDF) of a GPD, and the other is a CDF that satisfies certain constraints to ensure both tails follow a generalised Pareto-type distribution. The model proposed by Stein (2021) also assumes a composition of functions, where one is a monotone-increasing function that controls both the lower and upper tails, and the other is a Student t CDF. Finally, Krock et al. (2022) extend the latter approach to incorporate non-stationarity. The methods proposed by Frigessi et al. (2002), Naveau et al. (2016), Stein (2021) and Krock et al. (2022) avoid the choice of threshold.

In a semi-parametric framework, Cabras and Castellanos (2010) approximate the bulk distribution by an equi-spaced binning of the data followed by a Poisson log-link generalised linear model fit to the counts with a polynomial smoother for the mean parameter. do Nascimento et al. (2012) define the bulk distribution as a weighted mixture of gamma densities, extending the method proposed by Behrens et al. (2004), while Huang et al. (2019) estimate the log-density by first transforming the data and

then applying a cubic spline to the histogram. Tencaliec et al. (2020) propose a method based on the extension of the GPD proposed by Naveau et al. (2016). Finally, Tancredi et al. (2006) and MacDonald et al. (2011) propose non-parametric fits to the data. In the former, the bulk model is fitted via a mixture of uniform distributions whereas in the latter a kernel density estimator is used instead.

When we move to the multivariate setting, there is an extra difficulty; not only is it important to model the margins of the data correctly, but the dependence between the variables is also of interest since the behaviour of one variable can influence the behaviour and value of another. It is common practice to measure this relationship using correlation coefficients, such as Pearson's linear correlation or Kendall's concordance (Kendall, 1938). However, these only give information about the association between variables as a whole. An alternative is to use copulas, which fully capture the dependence between two or more variables. According to Sklar's Theorem (Sklar, 1959), the multivariate distribution function, $F$, of the random vector $(X_1, \ldots, X_d)$ can be written as the composition of a copula, $C$, and the marginal distributions of each $X_i$, $F_{X_i}(X_i)$, $i = 1, \ldots, d$, $d \geq 2$, as follows

$$F(x_1, \ldots, x_d) = C\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right).$$

If the variables are continuous, then the copula $C$ is unique. One advantage of copulas is that they are able to describe the dependence structure of two or more variables in a way that does not depend on the margins. Where it exists, the copula density $c\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right)$ can be obtained by taking the $d^{\text{th}}$ order derivative with respect to the variables $F_{X_1}(x_1), \ldots, F_{X_d}(x_d)$.

There is a large literature on dependence modelling for extremes, which usually involves defining a multivariate threshold above which an asymptotically-motivated copula is assumed to hold. However, models specifically aimed at capturing the behaviour of extremes as well as the body of the data, while permitting a likelihood-based ap-

proach to inference, are scarce in the literature. Both defining and performing inference on such models can be challenging compared to univariate models.

Methods for constructing more flexible copula families have been increasing in recent years, especially in financial applications. For instance, Durrleman et al. (2000), Morillas (2005), Klement et al. (2005) and Durante et al. (2010) propose transforming known copulas, especially from the Archimedean family, by means of bijections on $[0, 1]$. In particular, the methods proposed by Durrleman et al. (2000) and Durante et al. (2010) allow for a more accurate fit of the dependence structure. Given a bijection $\gamma : [0, 1] \to [0, 1]$, the copula $C$ is transformed into a new copula $C_\gamma$ in the following way $C_\gamma(x, y) = \gamma^{-1}(C(\gamma(x), \gamma(y)))$. Moreover, depending on specific conditions imposed on $\gamma$, the dependence structure of $C_\gamma$ contrasts with that of $C$ in different ways. Specifically, in the method proposed by Durrleman et al. (2000), changes in the overall dependence measures of $C$, such as Kendall's $\tau$, are possible while $C$ and $C_\gamma$ share the same extremal behaviour. On the other hand, Durante et al. (2010) study how the dependence in the extremes changes from $C$ to $C_{\gamma^{-1}}$, while the fit in the body remains the same between the two.

Other possibilities for building new copula families rely on piecewise constructions or convex combinations. For the former, by constructing box copulas (i.e., copulas nested in each other), Hummel (2009) is able to control and modify the dependence in the tail. For the latter, Bacigál et al. (2010) propose new construction techniques through additive generators of binary Archimedean copulas, whereas Shamiri et al. (2011) construct a Clayton-Gumbel copula, where, by means of a standard mixture model, two individual copulas are joined into one. This model allows for asymmetry in the data while being able to capture strong dependence in both tails.

Methods based on transformation of copulas or convex combinations allow for different, more flexible, dependence structures beyond the usual copulas. However, their main focus lies in providing a way of constructing new copula families, rather than

offering an accurate representation of the bulk and tail regions simultaneously.

Alternatively, patchwork copulas can offer a way to capture dependence structures that are not well suited to standard copulas. These allow for different copula models to be fitted to several regions of $[0,1]^2$ based on their characteristics; see for example Pfeifer and Ragulina (2021). Particular cases of patchwork copulas include those based on ordinal sums (Alsina et al., 2006); gluing copulas, where two or more copulas are scaled back to boxes in a region of the unit square and glued together along some hyperplane (Mesiar et al., 2008; Siburg and Stoimenov, 2008); and copulas based on rectangular constructions, where it is possible to have a copula in the body and another in the upper tail by defining two rectangles (disjoint up to their boundaries) over the diagonal, for example; see Durante et al. (2009) for more details. A generalised method to construct patchwork copulas that include the above mentioned cases is given in Durante et al. (2013). Given a copula $C$, a patchwork copula derived from it features the same probability mass distribution as $C$, excluding a $d$-dimensional box ($\subseteq [0,1]$) in which the probability mass is distributed differently. These models can be used to modify the extremal behaviour of a copula in two or more corners of $[0,1]^d$, and allow strong positive tail dependence to be induced if the application requires it. In this way, patchwork copulas aim to overcome the issue of misrepresentation of the extremes, when considering the whole data set. However, the transition between the non-extreme and the extreme regions is not smooth and therefore may be unsuitable in many real applications.

Aulbach et al. (2012a,b) suggest an extension to the multivariate setting of the model proposed by Behrens et al. (2004). They define a novel copula model by joining two $d$-dimensional ($d \geq 2$) copulas, one for the upper tail and the other for the body, in a manner that produces a new copula. Specifically, the authors assume two independent random vectors, each of which follow an arbitrary copula, that is $\boldsymbol{V} = (V_1, \ldots, V_d) \sim C_1$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_d) \sim C_2$. It is also required that the copulas are defined in $[-1,0]^d$,

which is not a problem since, if $\boldsymbol{U}$ follows a copula $C : [0,1]^d \to [0,1]$, then $\widetilde{\boldsymbol{U}} = \boldsymbol{U} - 1$ follows a copula $\widetilde{C}$ with shifted support i.e., $\widetilde{C} : [-1,0]^d \to [0,1]$. Then, by an appropriate choice of threshold vector $\boldsymbol{t} = (t_1, \ldots, t_d)$, they construct a random vector $\boldsymbol{Q}$, whose $i^{\text{th}}$ element is given by

$$Q_i \coloneqq Y_i \mathbb{1}_{Y_i \leq t_i} - t_i V_i \mathbb{1}_{Y_i > t_i}, \qquad i = 1, \ldots, d. \tag{3.1.1}$$

The authors prove that $\boldsymbol{Q}$ also follows a copula with support on $[-1,0]^d$, which coincides with $C_1$ on the region $(t_1, 0] \times \ldots \times (t_d, 0]$ and with $C_2$ on the region $[-1, t_1] \times \ldots \times [-1, t_d]$. An exact representation of the method is presented in Aulbach et al. (2012b). However, the model not only requires a choice of cut-off values $t_i$, $i = 1, \ldots, d$, to define the regions to fit each copula but, as with patchwork copulas, the transition between the two copulas may not be smooth. Figure 3.1.1 displays an example of a data set simulated according to equation (3.1.1); the discontinuity at the threshold is evident. Moreover, this method does not offer a convenient formulation of the likelihood, which results in difficulties for inference.

More recently, Pfeifer et al. (2017) and Pfeifer et al. (2019) propose infinite discrete and continuous partition-of unity copulas, respectively; these are flexible in higher dimensions and can be applied when there is asymmetry in the data. Similar to patchwork copulas, these copulas allow for implementing positive dependence in the tails; the density of the proposed model is approximated by an infinite mixture of functions, and careful choice of these functions can modify the tail behaviour if required.

A different type of approach was taken by Hu and O'Hagan (2021), who consider averaging different copula families that have been fitted to the whole distribution, in order to obtain a more robust estimate of the tail dependence of the data set. However, the use of BIC in the calculation of the weights assigned to each copula places the focus on the body and not on the tail of the data.

In a spatial context, Gräler (2014) proposes capturing the dependence of skewed

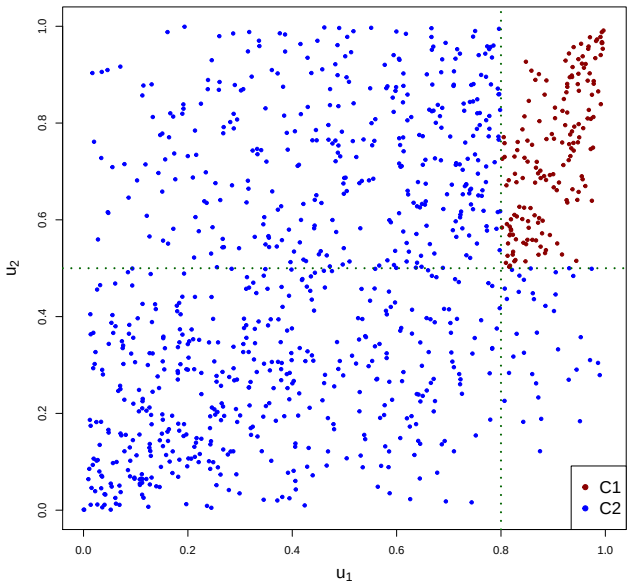



Figure 3.1.1: Example of $\boldsymbol{Q}$ simulated according to equation (3.1.1) with a Gumbel copula with parameter $\alpha = 2$ selected for the upper tail copula $C_1$ and Gaussian copula with parameter $\rho = 0.6$ selected for the body copula $C_2$ of the model proposed by Aulbach et al. (2012a). For illustration purposes, the vector of thresholds was chosen to be $\boldsymbol{t} = (0.8, 0.5)$.

spatial random fields (that display extreme events) by considering convex combinations of bivariate copulas in the construction of a spatial copula. In this way, between each location, a different dependence model is obtained. More recently, Krupskii et al. (2018) and Zhang et al. (2022b) each propose models fitted to both the body and tail of a distribution. The former outlines a copula model based on the assumption that there exists a common factor which affects the joint dependence of all the observations of the underlying process, and which is able to model both tail dependence and asymmetry. Numerical integration over this factor variable leads to a likelihood that can be fitted to all data. The latter propose using the generalised hyperbolic copula, which is flexible due to having a relatively large number of parameters. For both of these models, the authors show that there is reasonable flexibility for capturing both body and tail, yet a primary motivation for fitting to all data is the desire to avoid the computational difficulty involved in using censored likelihoods for extremes.

### 3.1.3 Extremal dependence properties

When the focus lies on extreme values, studying the extremal dependence between the variables is of interest. Two variables are said to be asymptotically dependent (AD) if joint extremes occur at a similar frequency to marginal extremes, or asymptotically independent (AI) otherwise. This dependence can be quantified through the measure $\chi = \lim_{r \to 1} \chi(r) \in [0, 1]$, where the limit exists, with

$$\chi(r) = P[F_Y(Y) > r \mid F_X(X) > r] = \frac{1 - 2r + C(r, r)}{1 - r}, \qquad r \in (0, 1), \tag{3.1.2}$$

where $C$ is the copula of $(X, Y)$; see Joe (1997) or Coles et al. (1999). The random variables $X$ and $Y$ are asymptotically independent if $\chi = 0$, whereas if $\chi > 0$ they are asymptotically dependent.

A complementary measure to $\chi$ is the residual tail dependence coefficient $\eta \in (0, 1]$ proposed by Ledford and Tawn (1996). For a function $\mathcal{L}$ that is slowly-varying at zero, they assume that the joint tail can be written as

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

The variables are asymptotically dependent if $\eta = 1$ and $\mathcal{L}(1 - r) \not\to 0$ as $r \to 1$, and asymptotically independent otherwise. Additionally, if $\eta \in (0, 1/2)$ , the variables show negative extremal association; positive extremal association if $\eta \in (1/2, 1]$ and they exhibit near extremal independence if $\eta = 1/2$.

Similarly to $\chi(r)$, for a particular value of $r \in (0, 1)$, $\eta(r)$ can be obtained as

$$\eta(r) = \frac{\log\left(P[F_X(X) > r]\right)}{\log\left(P[F_X(X) > r, F_Y(Y) > r]\right)}, \tag{3.1.3}$$

with $\eta = \lim_{r \to 1} \eta(r)$.

This paper is organised as follows: in Section 3.2 we present our proposed model

and its properties. Inference for the model is studied in Section 3.3, complemented by a simulation study to demonstrate performance in correctly specified and misspecified scenarios. We then apply our methodology to ozone and temperature data in the UK in Section 3.4 and conclude with a discussion in Section 3.5.

## 3.2 Weighted copula model

### 3.2.1 Model definition

Our interest lies in accurately modelling both the bulk and the tail of the whole distribution. From existing literature in the dependence context, Hummel (2009), Aulbach et al. (2012a,b), Durante et al. (2013) and Pfeifer et al. (2017, 2019) are concerned with representing both regions correctly. However, our model differs from these approaches in that we aim for a smooth transition between the two regions and allow for likelihood-based inference. To do so, we propose a mixture model where we fit two copulas to the whole range of the support and blend them by means of a dynamic weighting function $\pi$; in this way, data can be allowed to favour the "best" copula for each region, avoiding the subjective choice of thresholds often present in EVT applications. This approach can be seen as an extension to the multivariate framework of the model proposed by Frigessi et al. (2002) mentioned in Section 3.1.2.

Although our ideas could theoretically be applied in higher dimensions, we restrict ourselves to the bivariate setting for computational simplicity. Let $c_t$ and $c_b$ be copula densities representing the tail and the body, with vectors of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. For $(u^*, v^*) \in [0,1]^2$, we define a new density $c^*$ by

$$c^*(u^*, v^*; \boldsymbol{\gamma}) = \frac{\pi(u^*, v^*; \theta) c_t(u^*, v^*; \boldsymbol{\alpha}) + [1 - \pi(u^*, v^*; \theta)] c_b(u^*, v^*; \boldsymbol{\beta})}{K(\boldsymbol{\gamma})}, \tag{3.2.1}$$

where $\boldsymbol{\gamma} = (\theta, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is the vector of model parameters and

$$K(\boldsymbol{\gamma}) = \int_0^1 \int_0^1 [\pi(u^*, v^*; \theta) c_t(u^*, v^*; \boldsymbol{\alpha}) + (1 - \pi(u^*, v^*; \theta)) c_b(u^*, v^*; \boldsymbol{\beta})] \, \mathrm{d}u^* \, \mathrm{d}v^*$$

is a normalising constant. The weighting function $\pi$ depends on the data, and is specified such that, for small values of $u^*$ and $v^*$, more weight is given to $c_b$ and, for larger values, more weight is given to $c_t$. Thus, for a fixed value of the parameter $\theta$, the function $\pi : (0,1)^2 \to (0,1)$ should be increasing in $u^*$ and $v^*$. We note that having a dynamic weighting function is a modelling choice, but without this equation (3.2.1) simply represents a standard mixture model. Moreover, $\pi$ is not required to be monotonic and can be defined based on the application, which might make more sense outside of the extreme value context.

A direct consequence of $\pi(u^*, v^*; \theta)$ depending on the data is that the margins of the density $c^*$ are non-uniform; this leads to complications for inference. That is, we cannot fit $c^*$ directly to the data as it is not a copula density. We overcome these issues by fitting the copula of the density in equation (3.2.1), which requires numerical integration to calculate. The first stage is to obtain the true margins of $(U^*, V^*) \sim c^*$ as

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

and similarly for $F_{V^*}$, and then the corresponding inverse functions, $F_{U^*}^{-1}$ and $F_{V^*}^{-1}$ so that we can transform the margins to Uniform$(0,1)$ via the probability integral transform. The resulting copula is thus represented as

$$c(u, v; \boldsymbol{\gamma}) = \frac{c^* \left( F_{U^*}^{-1}(u), F_{V^*}^{-1}(v); \boldsymbol{\gamma} \right)}{f_{U^*} \left( F_{U^*}^{-1}(u) \right) f_{V^*} \left( F_{V^*}^{-1}(v) \right)}, \tag{3.2.2}$$

where $f_{U^*}$ and $f_{V^*}$ are the marginal probability density functions of $c^*$ and $\boldsymbol{\gamma} = (\theta, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is the vector of model parameters, common to the density in equation (3.2.1). Note

that each of $f_{U^*}$, $f_{V^*}$, $F_{U^*}$ and $F_{V^*}$ depends on $\boldsymbol{\gamma}$, but this is suppressed in the notation for readability.

### 3.2.2 Simulation

It is important to be able to sample from the proposed model so that it can be validated. To do so, we first note that we can rewrite the density (3.2.1) as a standard mixture of two densities

$$c^*(u^*, v^*; \boldsymbol{\gamma}) = \frac{K_t}{K} f_t(u^*, v^*; \theta, \boldsymbol{\alpha}) + \left(1 - \frac{K_t}{K}\right) f_b(u^*, v^*; \theta, \boldsymbol{\beta}),$$

where $K = K(\boldsymbol{\gamma})$ and

$$\begin{aligned} f_t(u^*, v^*; \theta, \boldsymbol{\alpha}) &= \frac{\pi(u^*, v^*; \theta) c_t(u^*, v^*; \boldsymbol{\alpha})}{K_t}, \\ f_b(u^*, v^*; \theta, \boldsymbol{\beta}) &= \frac{[1 - \pi(u^*, v^*; \theta)] c_b(u^*, v^*; \boldsymbol{\beta})}{K_b}, \\ K_t &= \int_0^1 \int_0^1 \pi(u^*, v^*; \theta) c_t(u^*, v^*; \boldsymbol{\alpha}) \mathrm{d}u^* \, \mathrm{d}v^*, \\ K_b &= \int_0^1 \int_0^1 [1 - \pi(u^*, v^*; \theta)] c_b(u^*, v^*; \boldsymbol{\beta}) \mathrm{d}u^* \, \mathrm{d}v^*. \end{aligned}$$

Note that $K = K_t + K_b$. Thus, to simulate from $c^*(u^*, v^*; \boldsymbol{\gamma})$ we need to be able to sample from the two densities $f_t(u^*, v^*; \theta, \boldsymbol{\alpha})$ and $f_b(u^*, v^*; \theta, \boldsymbol{\beta})$, which are non-standard as they depend on the weighting function $\pi(u^*, v^*; \theta)$ as well as the copula densities. However, as we can sample from the densities $c_t(u^*, v^*; \boldsymbol{\alpha})$ and $c_b(u^*, v^*; \boldsymbol{\beta})$, we can use a rejection sampling scheme to simulate from the required densities $f_t$ and $f_b$.

Note that, since the weighting function $\pi(u^*, v^*; \theta)$ is in $(0, 1)$, it is the case that

$$\sup_{(u^*, v^*) \in (0,1)^2} \frac{f_t(u^*, v^*; \boldsymbol{\alpha})}{c_t(u^*, v^*; \boldsymbol{\alpha})} = \sup_{(u^*, v^*) \in (0,1)^2} \frac{\pi(u^*, v^*; \theta) c_t(u^*, v^*; \boldsymbol{\alpha})}{K_t c_t(u^*, v^*; \boldsymbol{\alpha})} = \frac{\pi(u^*, v^*; \theta)}{K_t} \leq \frac{1}{K_t}.$$

Similarly, the ratio $f_b/c_b$ is bounded by $1/K_b$. The rejection algorithm for sampling

from $c^*$ via $f_t$ and $f_b$ is then as follows:

1. Simulate $n$ draws from $c_t(u^*, v^*; \boldsymbol{\alpha})$ and keep each with probability

$$\frac{f_t(u^*, v^*; \theta, \boldsymbol{\alpha})}{(1/K_t)c_t(u^*, v^*; \boldsymbol{\alpha})} = \frac{K_t \pi(u^*, v^*; \theta) c_t(u^*, v^*; \boldsymbol{\alpha})}{K_t c_t(u^*, v^*; \boldsymbol{\alpha})} = \pi(u^*, v^*; \theta).$$

   The expected number of returned draws from $f_t$ is $nK_t$.

2. Simulate $n$ draws from $c_b(u^*, v^*; \boldsymbol{\beta})$ and keep each with probability

$$\frac{f_b(u^*, v^*; \theta, \boldsymbol{\beta})}{(1/K_b)c_b(u^*, v^*; \boldsymbol{\beta})} = \frac{K_b[1 - \pi(u^*, v^*; \theta)] c_b(u^*, v^*; \boldsymbol{\beta})}{K_b c_b(u^*, v^*; \boldsymbol{\beta})} = 1 - \pi(u^*, v^*; \theta).$$

   The expected number of returned draws from $f_b$ is $nK_b$.

The total expected number of draws from both distributions together is $n(K_t+K_b) = nK$; these are in proportions $K_t/K$ and $K_b/K = 1-K_t/K$, and consequently we have a random sample from density $c^*$. To get a fixed sample size $n'$, we simply take sufficiently large $n$ and keep $n'$ draws at random.

Figure 3.2.1 illustrates two examples of random samples from our weighted copula model with different weighting functions. In each case we take a Gumbel copula with $\alpha = 2$ as $c_t$ and a Gaussian copula with $\rho = 0.6$ as $c_b$, which are the same components as the example in Figure 3.1.1. See A.1 for a directory of copula models and their parameterisations. Contrary to the Aulbach et al. (2012a) approach, we see that there is no cut-off between the two regions, with a smooth transition from data points mainly derived from $c_b$ in the bottom left to those mainly derived from $c_t$ in the top right. The influence of the choice of weighting function is also visible; for the same value of $\theta$, a preference for $c_t$ over $c_b$ is shown in the right plot.

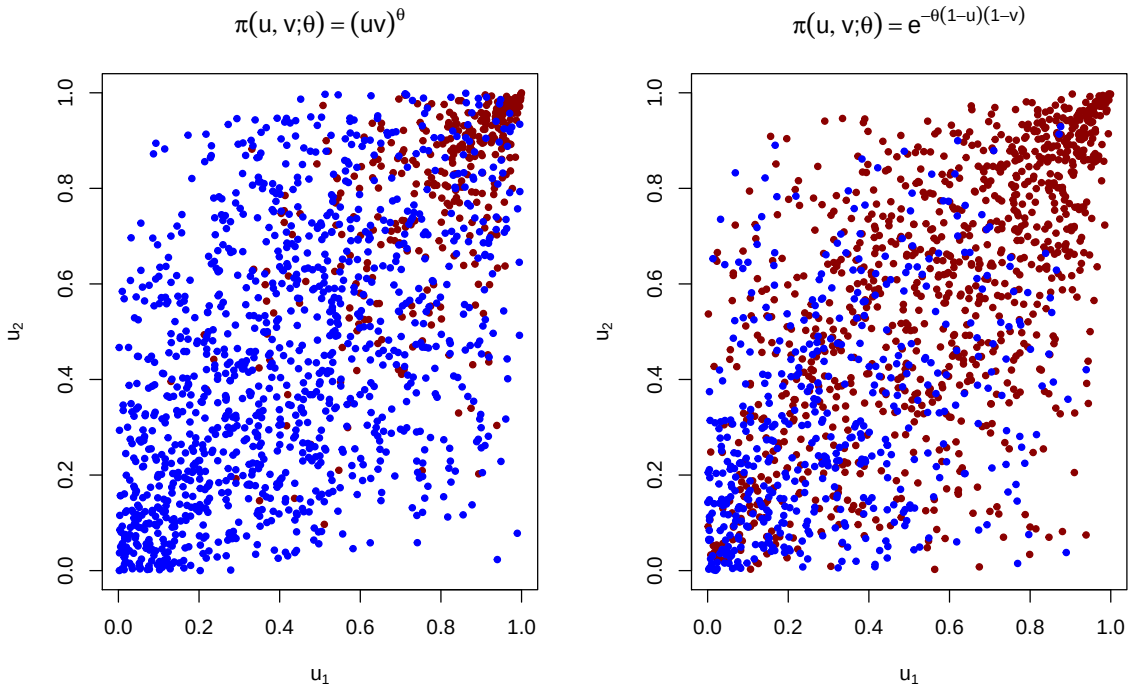



Figure 3.2.1: Example of data points from two weighted copula models simulated according to the sampling procedure detailed in Section 3.2.2. In both cases, a Gumbel copula with parameter $\alpha = 2$ is taken as $c_t$ and a Gaussian copula with parameter $\rho = 0.6$ as $c_b$. Two weighting functions are used with $\theta = 1.5$ in both: $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$ (left) and $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$ (right). Points in blue originate from $c_b$ and points in red originate from $c_t$.

### 3.2.3 Extremal dependence properties

We are interested in understanding the extremal dependence properties of the proposed model and, to do so, we compute the dependence measures $\chi$ and $\eta$ mentioned in Section 3.1.3. However, since they are defined in terms of the joint survival function of $(F_X(X), F_Y(Y))$, which we do not have, and the integral of the density in equation (3.2.1) is intractable, $\chi$ and $\eta$ are mainly obtained numerically. We have, however, derived these measures for one particular case with two different weighting functions; these are presented in the Supplementary Material. For a set of bivariate copulas, Heffernan (2000) and Joe (2014) study these dependence measures; a selection of which are summarised in Table 3.2.1.

Table 3.2.1: $\chi$ and $\eta$ for a selection of copulas; $\rho$ is the parameter of the Gaussian copula, and $\alpha$ the parameter of the Gumbel and Hüsler-Reiss copulas.

| Copula | $\chi$ | $\eta$ |
|---|---|---|
| Gaussian | 0 | $(1+\rho)/2$ |
| Frank | 0 | $1/2$ |
| Gumbel | $2-2^{1/\alpha}$ | 1 |
| Hüsler-Reiss | $2-2\Phi(1/\alpha)$ | 1 |

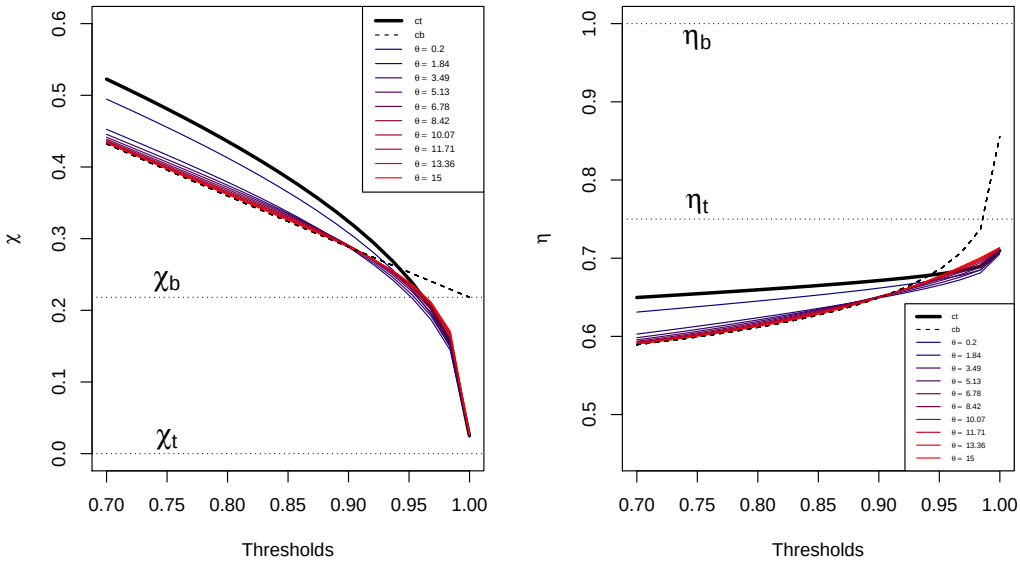
We consider mixtures of these four copulas to study the dependence properties of our model. In addition, we study the influence of the weighting function $\pi(u^*, v^*; \theta)$ and its parameter $\theta$. Thus, we consider two functions, $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$ and $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$, each with $\theta \in [0.2, 15]$. The dependence measures $\chi(r)$ and $\eta(r)$ were computed for 10 different threshold values $r$ ranging from 0.7 to 0.9998779, which is $1-(2\times \text{Machine Epsilon})^{0.25}$ in R, according to equations (3.1.2) and (3.1.3). For small $\theta$, the weighting functions are closer to 1 at lower levels $u^*$ and $v^*$, meaning that the tail copula dominates over a larger region, and vice versa for large $\theta$. In general, we expect that, in the limit $r \to 1$ and with a weighting function that goes to 1 with $u^*$ and $v^*$, the dependence properties of our model are dominated by those from the copula tailored to the tail, with similarities to the body copula for large $\theta$ and smaller $r$. Table 3.2.2 shows the theoretical values for $\chi$ and $\eta$ for each of the copulas used in the four weighted copula models, and Figure 3.2.2 shows the outcomes of our numerical investigations for Case 3. The remaining results are shown in the Supplementary Material. For use in Table 3.2.2 and beyond, we let $\eta_t$ and $\chi_t$ represent $\eta$ and $\chi$ for the tail copula, and similarly $\eta_b$ and $\chi_b$ for the body copula.

Table 3.2.2: Theoretical values for $\chi$ and $\eta$ for each of the copulas considered in the weighted copula models studied based on Table 3.2.1. AD denotes "asymptotically dependent"; AI denotes "asymptotically independent".

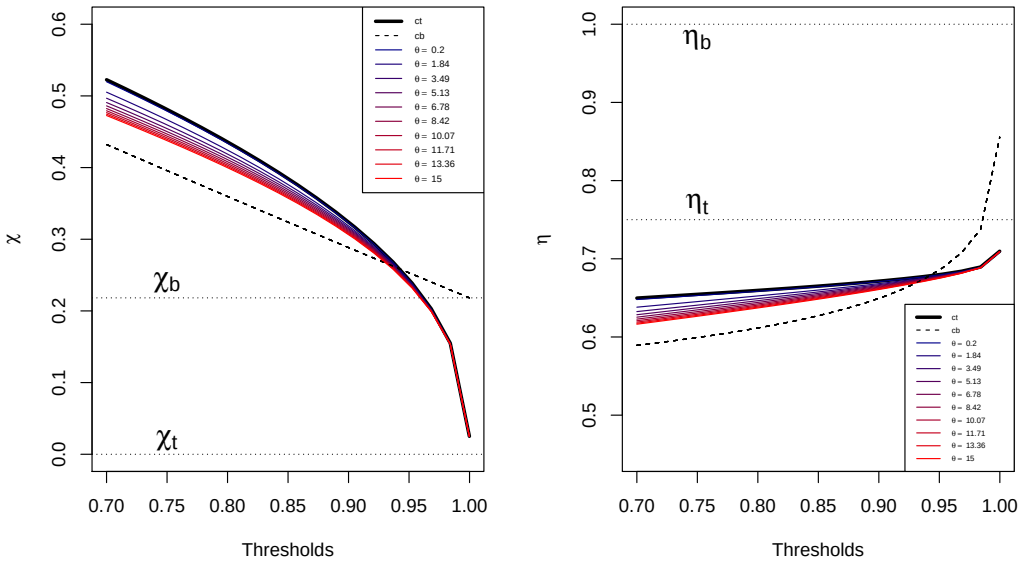
| Case | Body Copula $c_b$ | | Tail Copula $c_t$ | | $\chi_t$ | $\chi_b$ | $\eta_t$ | $\eta_b$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Frank (AI) | $\alpha = 2$ | Gaussian (AI) | $\rho = 0.6$ | 0 | 0 | 0.8 | 0.5 |
| 2 | Frank (AI) | $\alpha = 1$ | Gumbel (AD) | $\alpha = 3$ | 0.74 | 0 | 1 | 0.5 |
| 3 | Gumbel (AD) | $\alpha = 1.2$ | Gaussian (AI) | $\rho = 0.5$ | 0.22 | 0 | 1 | 0.75 |
| 4 | Gumbel (AD) | $\alpha = 2$ | Hüsler-Reiss (AD) | $\alpha = 2$ | 0.62 | 0.59 | 1 | 1 |

We can see from Figure 3.2.2 that, in the limit $r \to 1$, $\chi(r)$ and $\eta(r)$ of the weighted copula model tend towards $\chi_t$ and $\eta_t$ for both weighting functions. However, the results in the Supplementary Material suggest that this does not hold true for each of the combinations we consider. Depending on the weighting function, our investigations suggest that $c_b$ has an influence on the extremal dependence properties of the model in some cases. In particular, if $c_t$ is an asymptotically dependent copula and the weighting function is $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$, we observe that the limiting value of $\chi$ for the weighted copula model is dominated by $\chi_t$ with an influence from $\chi_b$. For an asymptotically independent tail copula and/or the weighting function $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$, our investigations suggest that the limiting extremal dependence properties of the model are those from $c_t$. Moreover, the influence of the parameter $\theta$ differs since $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$ grows more slowly than $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$ as $u^*$, $v^* \to 1$. When $\theta$ is larger, $\chi(r)$ and $\eta(r)$ are closer to $\chi_b(r)$ and $\eta_b(r)$, particularly for smaller $r$, where $\chi_b(r)$ and $\eta_b(r)$ are the sub-asymptotic extremal dependence measures $\chi(r)$ and $\eta(r)$ for $c_b$.

We note that this investigation suggests that there are some interesting subtleties in the tail dependence of models constructed in this way, and does not provide general conclusions. As shown theoretically for some of the considered cases, the weighted copula model has some intriguing features, such as the influence that the body copula might have when the tail component is asymptotically dependent for a given weighting function, which are worth investigating further. However, for specific cases, similar numerical or theoretical investigations can be carried out for any copulas and weighting functions of interest.




(a) $\chi(r)$ and $\eta(r)$ with weighting function $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$.




(b) $\chi(r)$ and $\eta(r)$ with weighting function $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$.

Figure 3.2.2: $\chi(r)$ and $\eta(r)$ for different thresholds $r \in [0.7, 1)$ for the proposed model with both $\pi(u^*, v^*; \theta)$ when $c_b$ is Gumbel (AD) and $c_t$ is Gaussian (AI). The coloured lines represent the 10 different models depending on different values of $\theta$; the thick black lines represent the single copula models - Gumbel (dashed) and Gaussian (solid). The theoretical values for the Gumbel and Gaussian copulas based on Table 3.2.2 are represented by the horizontal dashed lines.

## 3.3 Inference

### 3.3.1 Parameter estimation

In order to estimate $\boldsymbol{\gamma}$, we maximise the log-likelihood function of model (3.2.2),

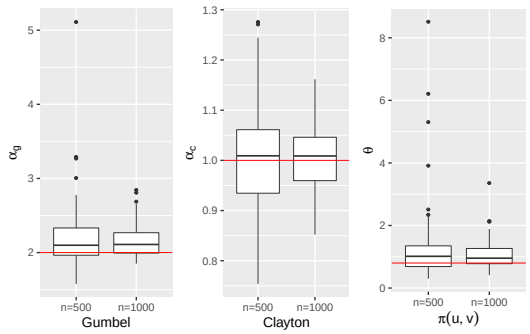
$$\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \log c(u_i, v_i; \boldsymbol{\gamma}), \qquad u_i, v_i \in [0,1]^2,\ i = 1, \ldots, n, \tag{3.3.1}$$

assuming $n$ independent observations from the copula. Because $F_{U^*}^{-1}$ and $F_{V^*}^{-1}$ are computationally expensive to obtain by a root finding algorithm, these are approximated using a smooth spline, following Zhang et al. (2022a). We found the spline approximation produces results with a similar degree of precision to the root finding algorithm, while reducing the computational time considerably.

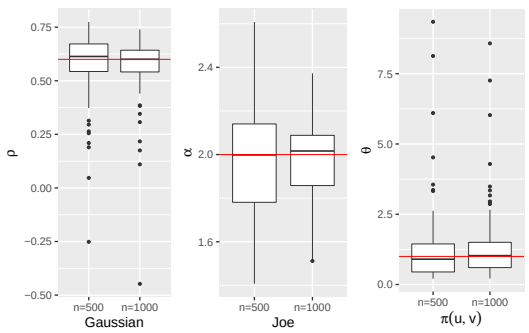
We conduct a simulation study to verify that inference on the proposed model produces reasonable estimates for the vector of model parameters $\boldsymbol{\gamma}$, and their inherent uncertainty. To do so, we consider two examples with different sample sizes: 500 and 1000 data points. Data are sampled from density (3.2.1) via the sampling procedure outlined in Section 3.2.2.

For the first case, we take $c_b$ to be the Clayton copula density with $\alpha = 1$, and $c_t$ to be the Gumbel copula density with $\alpha = 2$. For the second example, $c_b$ is taken as the Joe copula density with $\alpha = 2$ and $c_t$ is the Gaussian copula density with $\rho = 0.6$. The parameter of the weighting function is set to be $\theta = 0.8$ in the first example and $\theta = 1$ in the second case. Each data set is simulated 100 times.

Figure 3.3.1 displays the results of the simulation study. For each parameter, the left boxplot shows the spread of estimates when $n = 500$, and the right boxplot displays this for $n = 1000$. We observe that estimation seems generally unbiased and uncertainty reduces when the sample size increases.



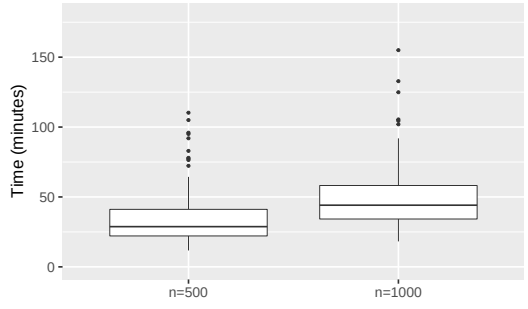

(a) Parameter estimates of $c_t$ (left plot), $c_b$ (middle plot), and the weighting parameter $\theta$ (right plot) for $n = 500$ and $n = 1000$. The true values for the parameters are shown in red.



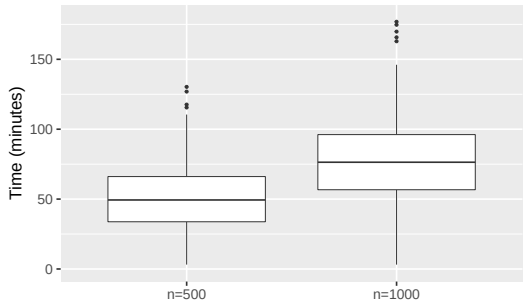

(b) Parameter estimates of $c_t$ (left plot), $c_b$ (middle plot), and the weighting parameter $\theta$ (right plot) for $n = 500$ and $n = 1000$. The true values for the parameters are shown in red.

Figure 3.3.1: Estimation variability obtained by simulating each case 100 times.

Because the copula density (3.2.2) relies on numerical integration to obtain $F$, $f$ and $F^{-1}$, it is important to assess the computational effort required to perform inference. Figure 3.3.2 displays the time taken to optimise the likelihoods on an internal computing node running CentOS Linux, with an Intel CPU running at 500GB of RAM. We can see that, for each of the models, the time taken increases with the sample size, which is to be expected. It also varies with the chosen copulas; for example, to evaluate the likelihood with $n = 500$ data points, the first model took around 30 minutes while the second took around 50 minutes.




(a) Case when $c_t$ is Gumbel and $c_b$ is Clayton for $n = 500$ (left) and $n = 1000$ (right).



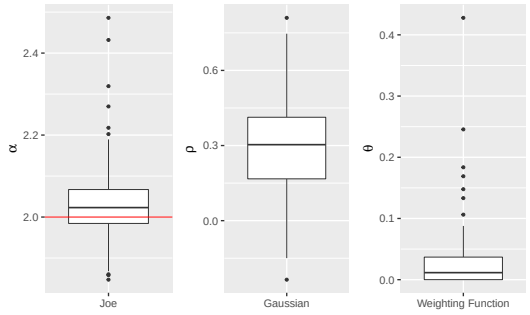

(b) Case when $c_t$ is Gaussian and $c_b$ is Joe for $n = 500$ (left) and $n = 1000$ (right).

Figure 3.3.2: Time (minutes) taken to optimise the log-likelihood (3.3.1) for each simulation.

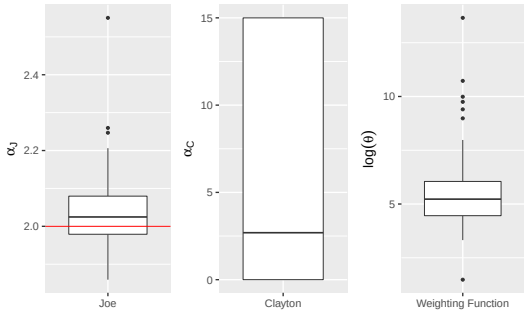
### 3.3.2 Model misspecification

In addition to checking if inference on the model produces reasonable estimates for $\boldsymbol{\gamma}$, we study the ability of the model to capture a misspecified dependence structure. We consider two situations: the case where the underlying data set comes from a single copula and we fit our model with this copula as one of the components; and the case where the fitted model does not contain the true copula. In the first case, we investigate whether the estimate of the parameter of the weighting function $\theta$ agrees with the true data. Since $\pi(u^*, v^*; \theta)$ is increasing in $(0, 1)$, we expect $\hat{\theta}$ to be large (small) when the true copula is tailored to the body (tail) of the distribution. In the second case, we investigate whether our model still produces reliable estimates of various dependence summaries even though the true dependence structure cannot be captured.

For the first case, we generate 1000 data points from a Joe copula with $\alpha = 2$ and fit two weighted copula models: one with the true copula as $c_t$ and a Gaussian copula as $c_b$, and the other with the true copula as $c_b$ and a Clayton copula as $c_t$. As before, 100 simulations for each case were performed and the results are shown in the boxplots in Figure 3.3.3.



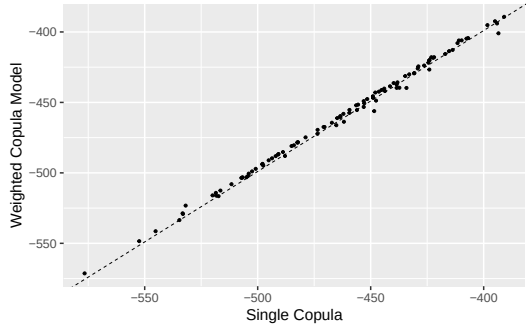

(a) Parameter estimates when $c_t$ is taken as the true copula (left) and $c_b$ is taken as the Gaussian copula (middle). The true value for the parameter is shown in red. Estimates of $\theta$ are shown in the right boxplot.



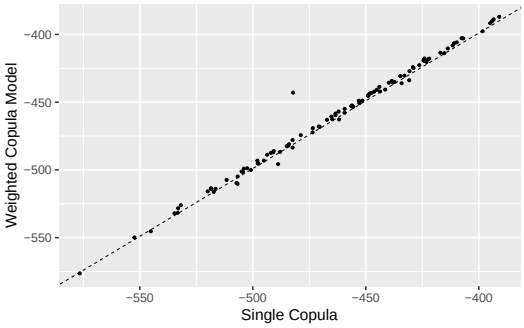

(b) Parameter estimates when $c_b$ is taken as the true copula (left) and $c_t$ is taken as the Clayton copula (middle). The true value for the parameter is shown in red. Estimates of $\log(\theta)$ are shown in the right boxplot.

Figure 3.3.3: Estimation variability obtained by simulating each case 100 times.

We observe that, when the Joe copula is taken as $c_t$, the estimates for $\theta$ are all less than 1, and when it is taken as $c_b$, these are considerably larger (here we use the logarithm of $\theta$ for ease of visualisation). Looking at the estimates for the parameter of the true copula, although they show some bias, they are fairly close to the true values, represented by the red lines. Finally, the estimates for the parameters of the misspecified copula show larger variability, which is to be expected as most of the weight is on the true copula. Figure 3.3.4 shows a comparison between the AIC of the true and weighted copula models, respectively. In the majority of cases (89% for the first and 92% for the second), the true model outperforms the weighted copula model in terms of AIC, as expected.




(a) Case with the true copula as $c_t$ and a Gaussian copula as $c_b$.




(b) Case with true copula as $c_b$ and a Clayton copula as $c_t$.

Figure 3.3.4: Comparison between the AIC of the true model and the fitted model.

For our second experiment, to evaluate the outcome of not being able to capture the true dependence structure, we simulate 1000 data points from a Gaussian copula with $\rho = 0.65$ and from a Galambos copula with $\alpha = 2$. For both cases, we generate 50 repetitions of the data set and fit a variety of weighted copula models, selecting the best model based on the average AIC values. In order to assess if the selected weighted copula model is flexible enough to capture the dependence of the true data sets, we compute three measures of dependence: Kendall's $\tau$, and $\chi(r)$ and $\eta(r)$ from equations (3.1.2) and (3.1.3), respectively, at several thresholds $r \in (0, 1)$. We show how the model performs by comparing with the theoretical values of the underlying models; the results

are shown in Figures 3.3.5 and 3.3.6.

Figure 3.3.5 displays the results for the weighted copula model where $c_t$ is inverted Gumbel, $c_b$ is Student t, and the true underlying structure is Gaussian. The results for the second model where the true underlying structure is Galambos and the selected weighted copula model is Coles-Tawn as $c_t$ and Frank as $c_b$ are shown in Figure 3.3.6. In both cases, we observe that the misspecified models capture the three dependence measures fairly well.

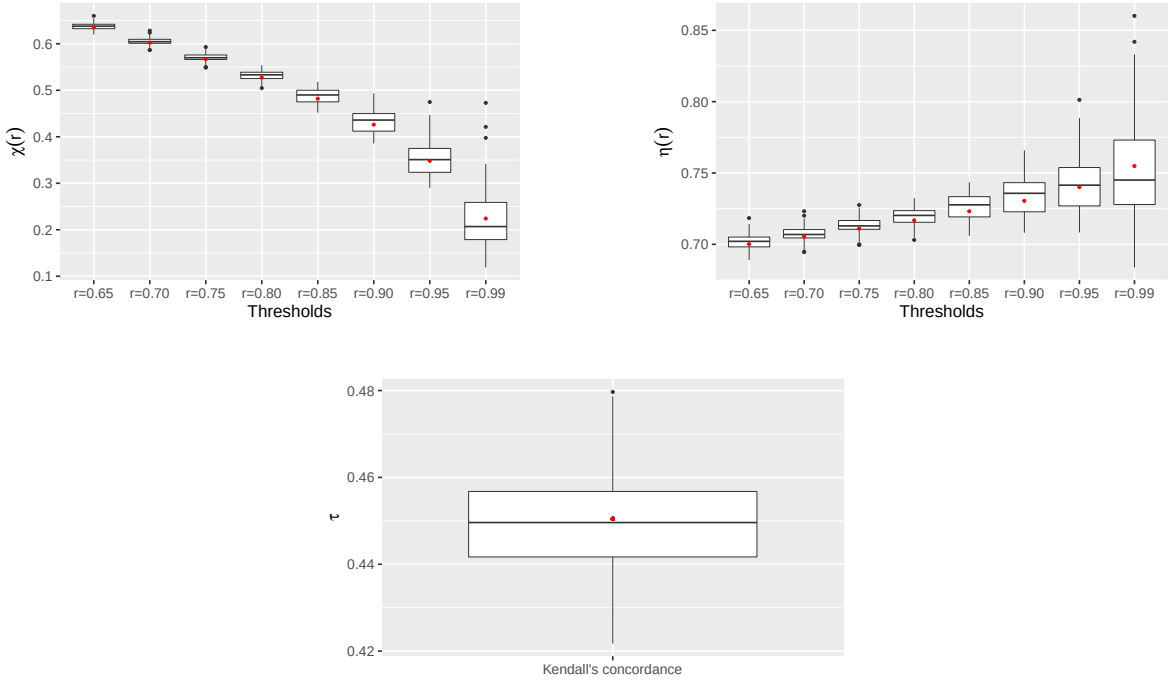



Figure 3.3.5: Model and theoretical (in red) $\chi(r)$ (top left) and $\eta(r)$ (top right) at levels $r \in \{0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.99\}$, and Kendall's $\tau$ (bottom) for the selected model when the true model is Gaussian with $\rho = 0.65$.

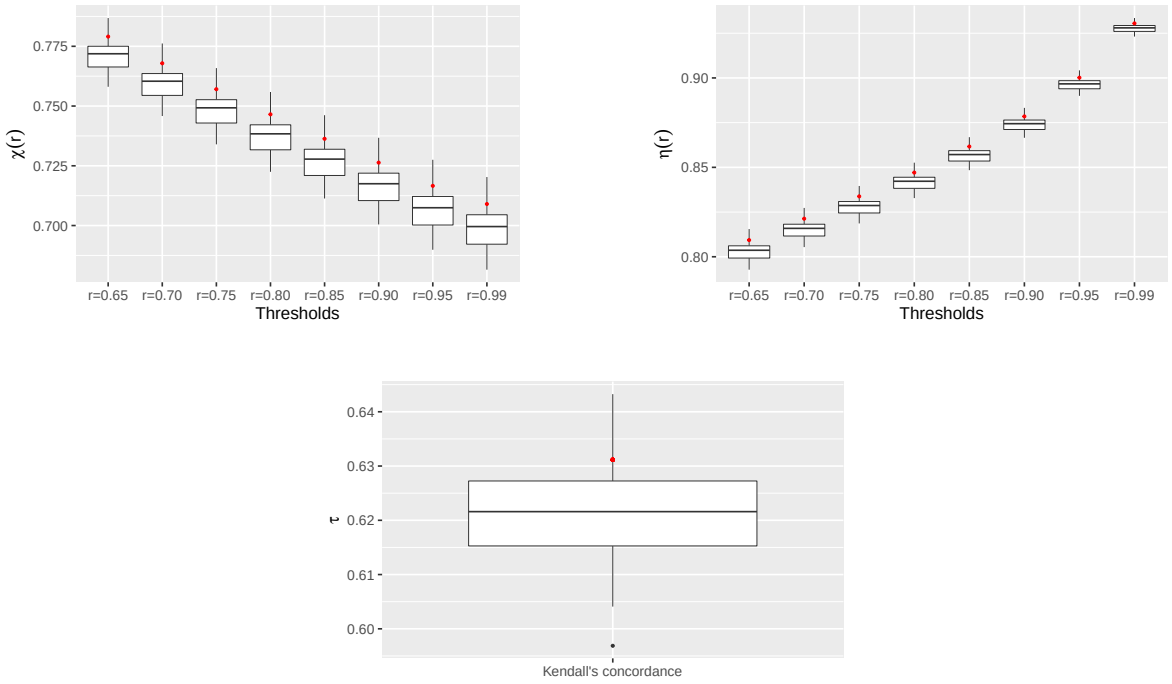



Figure 3.3.6: Model and theoretical (in red) $\chi(r)$ (top left) and $\eta(r)$ (top right) at levels $r \in \{0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.99\}$, and Kendall's $\tau$ (bottom) for the selected model when the true model is Galambos with $\alpha = 2$.

## 3.4 Case study: ozone and temperature data

### 3.4.1 Data and background

The relationship between ozone concentration and temperature has been analysed previously in the literature. For instance, Finch and Palmer (2020) show that there is an increase of exceeding regulated thresholds for ozone when the temperature is high. More recently, Gouldsbrough et al. (2022) study how extreme levels of ozone concentration are influenced by temperature in the UK by applying a temperature-dependent univariate extreme value model. They show that, with the increase in temperatures, the probability of exceeding a moderate regulated threshold of ozone concentration has increased over the last decade; this leads to this event no longer being considered extreme. The analysis of Gouldsbrough et al. (2022) only considers the univariate dis-

tribution of ozone extremes conditional upon the value of temperature. Since both temperature and ozone concentration are measurements of random variables, we can apply our weighted copula model to learn about the relationship between these variables at all levels. Specifically, we study the dependence between temperature and ozone concentration at two UK sites: Blackpool (urban background) and Weybourne (rural background). Table 3.4.1 shows the regulated threshold indexes for the levels of air pollution for Ozone in the UK.

Table 3.4.1: Daily Air Quality Index (DAQI) for ozone ($O_3$) concentrations in the UK.

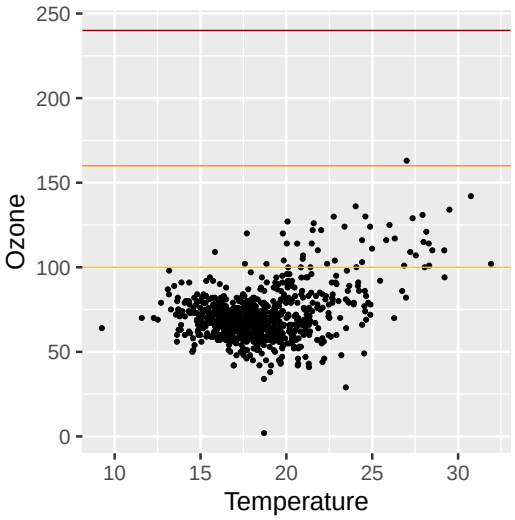
| Levels | Low | Moderate | High | Very High |
|---|---|---|---|---|
| $O_3\,(\mu g/m^3)$ | $[0, 100]$ | $[101, 160]$ | $[161, 240]$ | $> 240$ |

We took the daily maxima from 8-hour running means ozone concentration available on the UK's Automatic Urban and Rural Network (AURN) (https://uk-air.defra.gov.uk) and obtain the corresponding daily maximum temperature data from the Centre for Environmental Analysis (CEDA) archive (https://archive.ceda.ac.uk). Since higher temperatures are expected during summer, and in order to overcome the non-stationarity often present in temperature data, we restrict our analysis to the summer months (June-August). Based on the available data, we consider the years from 2011 to 2019 for Blackpool and from 2010 to 2019 for Weybourne; this results in 827 and 892 observations, respectively. Figure 3.4.1a shows the scatterplot of the daily maxima of temperature and the daily maxima of ozone for the summers of 2011 to 2019 in Blackpool and the respective regulated UK thresholds, while Figure 3.4.1b shows the relationship between the variables when transformed to uniform margins using a semi-parametric approach with a GPD fit to the tail of both distributions. That is, we

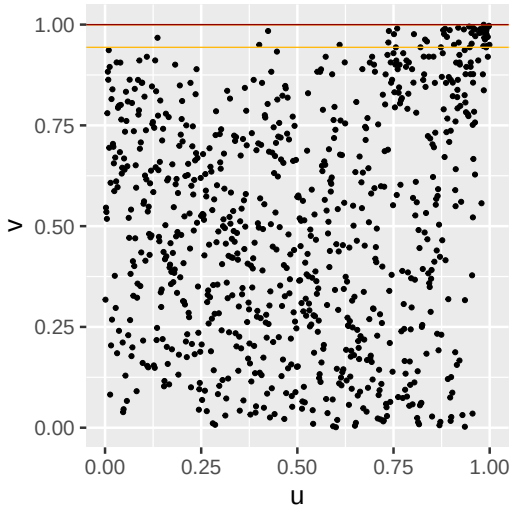
estimate the CDF of each marginal distribution via

$$F(x) = \begin{cases} \widetilde{F}(x), & x \leq r, \\ 1 - \phi_r \left[1 + \frac{\xi(x-r)}{\sigma}\right]_+^{-1/\xi}, & x > r, \end{cases}$$

where $\widetilde{F}(x)$ is the empirical distribution function, $\phi_r$ is the probability of exceeding a selected high threshold $r$, and $\xi$ and $\sigma$ are the GPD shape and scale parameters, respectively. The corresponding analysis for Weybourne is presented in the Supplementary Material; the results show similar conclusions to the analysis for Blackpool.




(a) Daily maxima of temperature and ozone. The moderate, high and very high DAQI are represented by the yellow, orange and red lines, respectively.




(b) Daily maxima of temperature (u) and ozone (v) on uniform margins. The corresponding moderate, high and very high DAQI are represented by the yellow, orange and red lines, respectively.

Figure 3.4.1: Summer data from 2011 to 2019 for Blackpool, UK.

### 3.4.2 Model fitting

We start by fitting a single copula model to the whole data set for comparison with the weighted copula model. Looking at Figure 3.4.1b, the variables seem to exhibit

positive correlation when they are both extreme, but negative dependence otherwise. We anticipate that the weighted copula model may be flexible enough to capture this, whereas a single copula is likely to be too rigid. Table 3.4.2 shows the MLEs obtained by fitting a range of copulas and the corresponding AIC values. From the copulas considered, the only ones capable of capturing negative dependence are the Gaussian and Frank, when their parameters are negative, and the Student t (which also exhibits lower and upper tail dependence). However, all parameter estimates are positive. In terms of AIC, the best fit is the Joe, followed by the Galambos, Hüsler-Reiss, Gumbel and Coles-Tawn copulas; these are all known to be asymptotically dependent copulas, which appears to agree with the dependence in the upper tail shown in Figure 3.4.1b. As a further diagnostic, we compute the dependence measure $\eta(r)$ from equation (3.1.3) for $r \in (0, 1)$ empirically, as well as for the five best models in terms of AIC, and for the Gaussian and Frank copulas; this is shown in Figure 3.4.2. The confidence intervals in Figure 3.4.2 were obtained via block bootstrapping the data with a block length of 14 days, to reflect temporal dependence in the extremes. It is evident that none of the copulas fit the model well in the whole support based on this measure. However, the Joe copula (in orange) appears to give the best fit in the tail, consistent with its AIC value being lowest.

Table 3.4.2: MLEs for ten copulas and their AIC values. Lower AIC values are preferred.

| Copula | Parameter | AIC |
|---|---|---|
| Clayton | $1.22 \times 10^{-8}$ | 2.0 |
| Frank | 0.92 | -15.8 |
| Gumbel | 1.20 | -97.4 |
| Inverted Gumbel | 1.04 | 0.1 |
| Galambos | 0.46 | -99.0 |
| Gaussian | 0.19 | -28.6 |
| Joe | 1.41 | -143.6 |
| Student t | 0.16 4.52 | -52.8 |
| Hüsler-Reiss | 0.82 | -99.1 |
| Coles-Tawn | 0.24 0.22 | -95.9 |

We next fit the weighted copula model to the whole data set taking the weighting

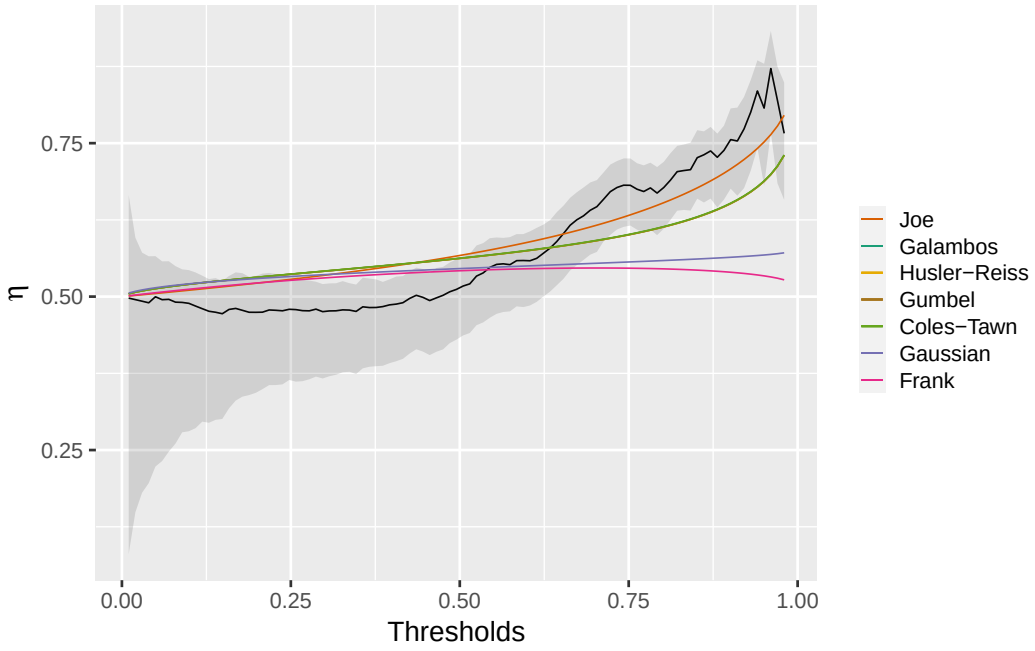



Figure 3.4.2: Empirical $\eta(r)$ (in black) and $\eta(r)$ for seven copulas (in colour) for $r \in (0, 1)$. The 95% confidence bands were obtained by block bootstrapping. Note that the $\eta(r)$ for the Galambos, the Hüsler-Reiss, the Gumbel and the Coles-Tawn copulas overlap.

function $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$. We consider several copulas with different extremal dependence characteristics to fit both $c_b$ and $c_t$; Table 3.4.3 shows the MLEs obtained by optimising the log-likelihood (3.3.1) and their AIC values for some of the models considered. According to AIC, there is a preference for models with the Gaussian and Frank as candidates for $c_b$ and AD copulas, such as the Galambos, Hüsler-Reiss, Joe and Coles-Tawn copulas, as $c_t$. In contrast to the single copula fits, the parameter estimates for the Gaussian and the Frank copulas are negative, which mirror the negative association visible in the body of Figure 3.4.1b.

We next consider a different weighting function, $\pi(u^*, v^*; \theta) = \exp\{-\theta(1 - u^*)(1 - v^*)\}$, in the five models with the lowest AICs. The MLEs and the AIC values are shown in Table 3.4.4. In terms of AIC, these models are all better fits to the data, while the negative correlation is still captured by $c_b$, and is now stronger. Because these models represent a better fit based on AIC, we focus on them for the rest of the analysis.

Table 3.4.3: MLEs for different weighted copula models and their AIC values when the weighting function used is $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$. Lower AIC values are preferred.

| Model | $c_t$ | $c_b$ | $\hat{\boldsymbol{\alpha}}$ | $\hat{\beta}$ | $\hat{\theta}$ | AIC |
|---|---|---|---|---|---|---|
| Model 1 | Hüsler-Reiss | Gaussian | 1.24 | -0.40 | 0.35 | -176.1 |
| Model 2 | Galambos | Gaussian | 0.79 | -0.41 | 0.34 | -172.1 |
| Model 3 | Coles-Tawn | Gaussian | 0.35 2.86 | -0.33 | 0.43 | -158.4 |
| Model 4 | Coles-Tawn | Frank | 0.33 4.80 | -2.52 | 0.37 | -163.2 |
| Model 5 | Joe | Frank | 1.61 | -4.11 | 0.18 | -184.9 |
| Model 6 | Clayton | Gaussian | 12.10 | -0.20 | 2.10 | -129.9 |
| Model 7 | Inverted Gumbel | Gaussian | 2.65 | -0.29 | 0.90 | -153.4 |
| Model 8 | Hüsler-Reiss | Joe | 1.28 | 1.30 | 3.18 | -145.6 |
| Model 9 | Student t | Galambos | 0.72 4.98 | 0.28 | 2.59 | -125.0 |
| Model 10 | Gaussian | Clayton | 0.81 | $3.38 \times 10^{-4}$ | 2.80 | -132.6 |
| Model 11 | Gumbel | Joe | 1.52 | 1.18 | 0.91 | -145.1 |

Table 3.4.4: MLEs for five weighted copula models and their AIC values when the weighting function used is $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$. Lower AIC values are preferred.

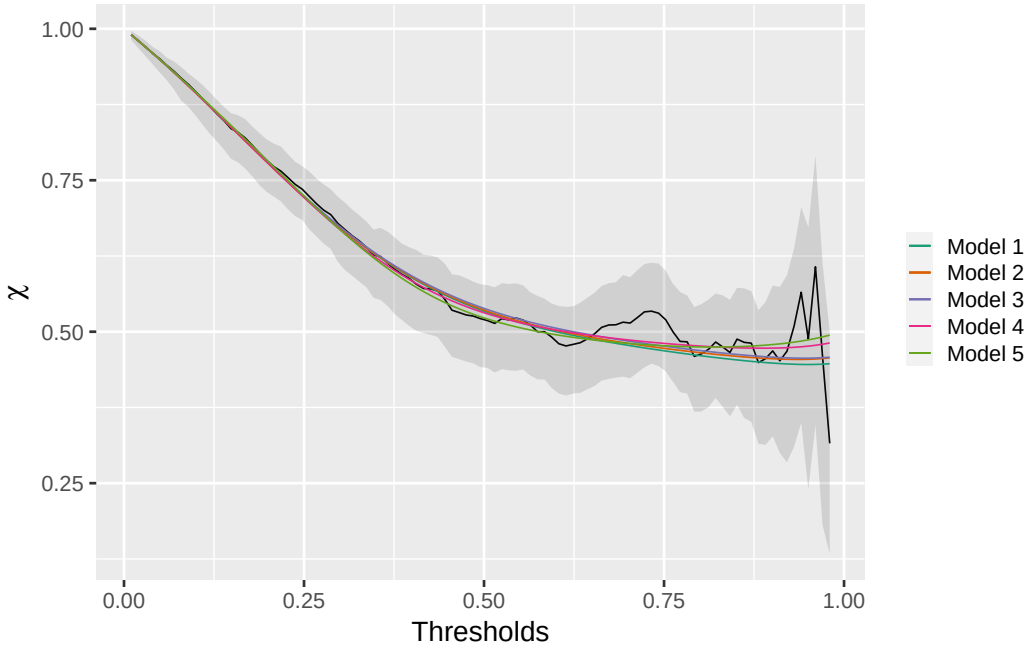
| Model | $c_t$ | $c_b$ | $\hat{\boldsymbol{\alpha}}$ | $\hat{\beta}$ | $\hat{\theta}$ | AIC |
|---|---|---|---|---|---|---|
| Model 1 | Hüsler-Reiss | Gaussian | 1.33 | -0.74 | 3.32 | -240.1 |
| Model 2 | Galambos | Gaussian | 0.90 | -0.72 | 3.55 | -237.2 |
| Model 3 | Coles-Tawn | Gaussian | 0.85 0.79 | -0.74 | 3.25 | -234.8 |
| Model 4 | Coles-Tawn | Frank | 0.869 1.02 | -4.51 | 4.33 | -235.7 |
| Model 5 | Joe | Frank | 1.72 | -6.49 | 2.45 | -232.9 |

### 3.4.3 Diagnostics

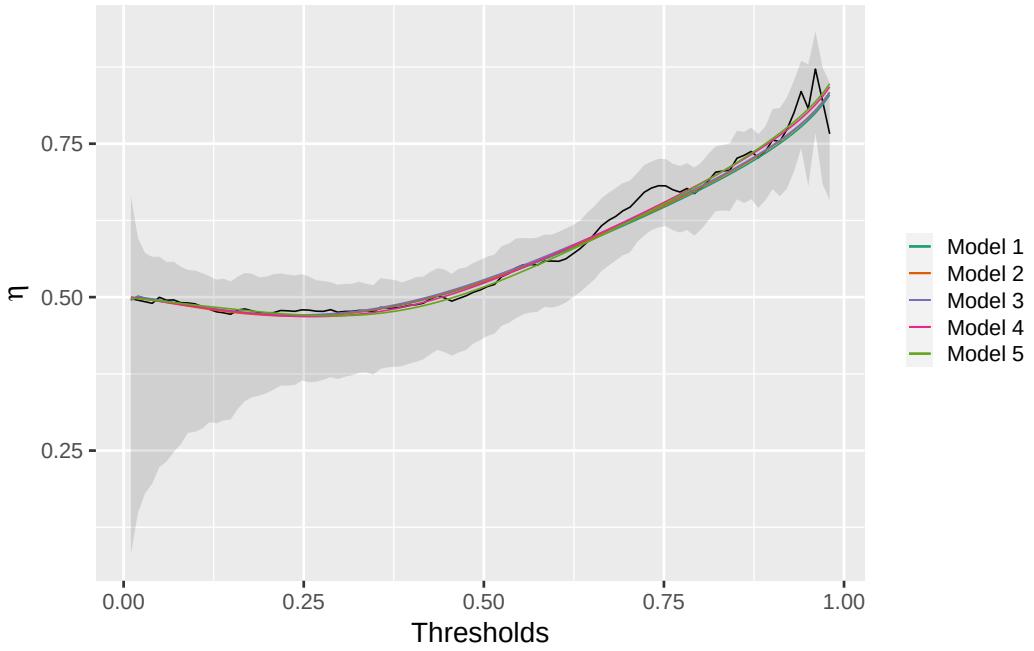
To check the adequacy of the model fits, we compare a variety of empirical dependence measures to their model-based counterparts. These include Kendall's $\tau$, the dependence measures $\chi(r)$ and $\eta(r)$ for $r \in (0, 1)$, and some probabilities of interest. Specifically, we look at the probability of ozone concentrations exceeding the so-called moderate threshold (i.e., 100 $\mu g/m^3$) when the temperature is high or low, and the probability of $O_3$ exceeding this and the higher threshold of 160 $\mu g/m^3$, knowing that the temperature is in a specific range.

Figure 3.4.3 displays $\chi(r)$ and $\eta(r)$ for $r \in (0, 1)$. A clear improvement from the single copula models shown in Figure 3.4.2 can be seen as now all five models offer a reasonable fit throughout the whole support. In addition, model 5 (in light green)

seems to provide slightly better $\chi(r)$ and $\eta(r)$ estimates at median values of $r$ and in the tail.




(a) Empirical $\chi(r)$ (in black) and $\chi(r)$ for the five models (in colour) for $r \in (0, 1)$. The 95% confidence bands were obtained by block bootstrapping.




(b) Empirical $\eta(r)$ (in black) and $\eta(r)$ for the five models (in colour) for $r \in (0, 1)$. The 95% confidence bands were obtained by block bootstrapping.

Figure 3.4.3: Dependence measures $\chi(r)$ and $\eta(r)$.

The average temperature in summer in Blackpool is between 17°C and 20°C and the observed 90th, 95th and 99th percentiles of the temperature are approximately 22°C, 24°C and 28°C, respectively. Thus, we focus on probabilities based on these values of temperature; these are presented with Kendall's $\tau$ in Table 3.4.5. We can see that the five models give very similar probabilities and they are all inside the 95% confidence interval of the empirical values, except for $P[T \leq 16, O_3 \geq 100]$ and $P[O_3 \geq 160 \mid 28 \leq T \leq 29]$. The empirical probability and its 95% confidence interval of the latter are explained by the low number of observations present in the data set. When there are no observations in a certain region then this will be true of each bootstrap sample as well. Gouldsbrough et al. (2022) obtained the mean probability of exceeding the high threshold 160 $\mu g/m^3$ at the 99th percentile of temperature for urban and rural backgrounds across the UK. These were 0.0002 ([0, 0.0004]) for an urban background and 0.006 ([0.003, 0.009]) for a rural background. We obtained higher probabilities of exceeding this threshold given that the temperature is close to the observed 99th percentile (we refer readers to the Supplementary Material for the results for Weybourne). This might be due to having only considered two sites within the UK, and potentially some of the characteristics of the relationship between temperature and ozone being better captured with the weighted copula model than with the univariate conditional model.

An advantage of this modelling approach in comparison to the conditional univariate modelling of Gouldsbrough et al. (2022) is that we are able to extrapolate and consider probabilities of ozone exceeding certain thresholds at temperature values that have not been observed in the data set. In this way, we can consider probabilities such as $P[O_3 \geq 160 \mid 33 \leq T \leq 35]$, which we estimate to be 0.6944 for Model 1, for example.

Table 3.4.5: Diagnostics for the best five models based on their AIC values. The 95% confidence intervals for the empirical values were obtained by block bootstrapping.

| Model | Kendall's $\tau$ | $P[T \leq 16, O_3 \geq 100]$ | $P[T \geq 22, O_3 \geq 100]$ |
|---|---|---|---|
| Empirical | 0.0821 | 0.0012 | 0.0363 |
| (95% CI) | (0.0173 , 0.1867) | (0.0000 , 0.0011) | (0.0170 , 0.0601) |
| Model 1 | 0.0690 | 0.0036 | 0.0332 |
| Model 2 | 0.0663 | 0.0040 | 0.0336 |
| Model 3 | 0.0770 | 0.0039 | 0.0338 |
| Model 4 | 0.0779 | 0.0035 | 0.0348 |
| Model 5 | 0.0718 | 0.0036 | 0.0353 |

| Model | $P[T \geq 24, O_3 \geq 100]$ | $P[O_3 \geq 100 \mid 22 \leq T \leq 23]$ | $P[O_3 \geq 160 \mid 28 \leq T \leq 29]$ |
|---|---|---|---|
| Empirical | 0.0302 | 0.1330 | 0.0000 |
| (95% CI) | (0.0147 , 0.0544) | (0.0227 , 0.1944) | (0.0000 , 0.0000) |
| Model 1 | 0.0246 | 0.1441 | 0.0070 |
| Model 2 | 0.0250 | 0.1412 | 0.0062 |
| Model 3 | 0.0251 | 0.1429 | 0.0061 |
| Model 4 | 0.0262 | 0.1392 | 0.0055 |
| Model 5 | 0.0267 | 0.1366 | 0.0050 |

## 3.5 Conclusions and discussion

In this paper, we introduced a dependence model that is able to capture both the body and tail of a bivariate data set. This is important when we aim to obtain an accurate representation of the data in both regions. The model has the advantage of not requiring a choice of thresholds above which we fit the copula tailored to the extreme observations. Moreover, it offers a smooth transition between the two copulas. Through simulation studies, we have shown that the model behaves as expected when only a single dependence structure is present, and that it is sufficiently flexible to capture misspecified dependence structures. We applied the weighted copula model to study the relationship between temperature and concentrations of air pollution in the UK and showed that this model performs substantially better than fitting a single copula model to the data. In fact, in this particular application, we were able to capture the negative dependence exhibited by the bulk and the positive association present in the

upper tail, which was not possible through fitting a single copula.

A drawback of the weighted copula model is that it is computationally expensive due to the need for numerical integration and inversion. As shown in the simulation studies in Sections 3.3.1 and 3.3.2, for a sample size of 1000, optimising the log-likelihood takes more than one hour to compute, although the run time also varies depending on the chosen copulas. Whilst in principle the weighted copula model could be extended to higher dimensions, doing so would exacerbate the computational issues.

For the temperature and ozone data, we have $\chi(r) > 0$ and $\eta(r) < 1$, for the largest values of $r$, which does not allow us to draw conclusions about the extremal dependence. This is a common situation in practice but results in complications if we wish to extrapolate for larger values than the ones observed. Incorporating a more flexible copula as the tail component of the proposed model is a possibility to overcome this issue. Such a copula could be the one proposed by Huser and Wadsworth (2019), which is able to capture both dependence classes with the transition between them occurring at an interior point of the parameter space. However, because it is computationally expensive on its own, when applied as the tail component in our model, the computational time required was not feasible.

It would be an advantage to have a copula model that could accommodate changes in the dependence structure due to covariates over the whole support of the distribution. Until now, we have been assuming stationarity, which is rarely the case in real world situations. Non-stationary multivariate extreme value methods naturally focus on capturing trends present in the extreme observations. However, data may be extreme in only one variable and thus studying the trends present in the body of the data is of importance as well. Incorporating covariates in the proposed model would also be an interesting avenue for future work.

Finally, some theoretical aspects of the weighted copula model remain open for further work. For instance, it would be interesting to investigate bounds on differences

between $c_b$ and/or $c_t$ with the copula $c$ of $c^*$, or whether we could identify the family of the resulting copulas in specific cases such as when both $c_t$ or $c_b$ are from the same family. Further theoretical exploration of extremal dependence properties of the weighted copula model would also be valuable as only particular cases were considered.

# Chapter 4

# Gaussian mixture copulas for flexible dependence modelling in the body and tails of joint distributions

## 4.1 Introduction

### 4.1.1 Motivation

Preventing impactful events such as high temperatures, floods, market crashes is crucial, requiring the need for models able to characterise well such large, and rare, observations. When these rare observations are triggered by another event, the joint modelling of such phenomena is needed to understand the effect of one occurrence on another. Since the interest is on the joint behaviour in the tail, it is typical in the literature to define an extreme region on which inference is based; this often requires selecting a threshold above which the observations are deemed extreme. Such a choice, however, is often arbitrary and might lead to inaccurate representations of the extremal dependence

structure. This can be overcome if models that show flexibility in jointly modelling the body and tail regions of the data are considered instead. This approach is particularly useful in contexts such as air pollutant concentrations, where studying the effect of one pollutant on another is as crucial as analysing them individually. Since harmful levels might occur not only in the tails but also in the regions where only a subset of air pollutants are extreme, modelling the body is as important as capturing the tail behaviour accurately.

Modelling the entire data set with common statistical distributions might sometimes lead to a poor fit of the data. For instance, in a univariate framework, distributions such as the Gaussian are unfit to capture the tail behaviour well even if they are a reasonable fit to the body of the distribution. Therefore, when the interest is in the extreme observations of a data set, a model justified for the tail is required. In the univariate framework, the common practice is to fit asymptotically justified models: either a generalised extreme value (GEV) distribution, if the underlying data are block maxima, or a generalised Pareto (GP) distribution, if the underlying data are exceedances over a threshold (Coles, 2001). Likewise, asymptotically justified models are required to fit the tail region of a multivariate data set, with such models requiring marginal and dependence structures supported by asymptotic arguments. While the margins are either GEV or GP distributed, extra considerations are needed for the dependence structure of a multivariate data set. Additionally, such asymptotically justified models are unlikely to be an appropriate fit for non-extreme observations. Since often these observations provide no useful information about the extremes, the body of the distribution has been modelled empirically (Coles and Tawn, 1991). However, relying solely on thresholds to determine which observations are useful for inferring the extremal behaviour of a data set can be sometimes too simplistic. Specifically, the choice of threshold can be not only subjective, but also introduce a high degree of sensitivity to the results, with small changes in its value potentially leading to different outcomes. Furthermore, defining an

extremal region in a multivariate setting is not as straightforward as with the univariate framework, as there are various ways to represent it, potentially requiring a number of thresholds. Lastly, in some cases, having an accurate fit of the body can be as important as representing the tail correctly. Thus, empirically estimating the non-extremal region might not be feasible and more suitable models might be required in some situations.

### 4.1.2 Univariate extreme value mixture models

In univariate extremes, the body and upper tail regions of a random variable $X$ are defined by the events $\{X < u\}$ and $\{X \geq u\}$, respectively, for a suitably defined threshold $u$ (Coles, 2001). While the common practice is to fit an asymptotically justifiable model above $u$ and model the observations below empirically, various univariate models that fit parametric distributions to both of these regions jointly have been proposed; Scarrott and MacDonald (2012) review several of these models, henceforth referred to as extreme value mixture models (EVMMs). Typically, these involve fitting a GP distribution to the upper tail, while a different, more suitable, model is chosen for the bulk region. Moreover, by implicitly or explicitly treating the threshold $u$ as a parameter of the model, the majority of these models aim to account for the uncertainty around this threshold choice in the inference procedure.

A substantial literature of EVMM has been developed, with the following covering the core examples. Frigessi et al. (2002) proposes a dynamically weighted mixture model where a GP and a lighter tailed (such as the Weibull) distributions are fit to the full support of the data. By means of a non-decreasing weighting function that depends on the data, the GP distribution is tailored to the tail while the lighter tailed distribution will be predominant in the bulk region. Contrarily, the models introduced by Behrens et al. (2004), Carreau and Bengio (2009), Tancredi et al. (2006) and MacDonald et al. (2011) fit the bulk region in a parametric, semi-parametric or non-parametric way, respectively, whilst a GP distribution is used to model the exceedances above a large

threshold; in these, the threshold $u$ is usually estimated within the modelling framework. de Mendes and Lopes (2004), Naveau et al. (2016), Tencaliec et al. (2020), Stein (2021) and Krock et al. (2022) propose modelling both the lower and upper tails with a GP distribution, and bridge them through either a different distribution for the bulk or by means of a composition of functions. Finally, modelling the bulk and tail regions jointly in a hierarchical way has also been proposed in the literature; for instance, Bottolo et al. (2003) assume that the exceedances of different clusters are generated by a Poisson process, whereby each parameter is modelled by a hierarchical mixture prior. In this way, the presence of heterogeneity is accounted for in the modelling. More recently, Castro-Camilo et al. (2019) construct a latent Gaussian model based on a spliced Gamma-GP distribution to describe the body and tail regions, respectively, while Yadav et al. (2021) propose a Gamma-Gamma hierarchical model for which the GP distribution is a special case.

### 4.1.3 Dependence modelling

When moving to a multivariate framework, the dependence between variables presents an additional challenge to those set out in Section 4.1.2. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a $d$-dimensional random vector with $d \geq 2$ and $X_i$ be the $i^{th}$ marginal variables for $i \in D = \{1, \ldots, d\}$. Sklar's theorem (Sklar, 1959) states that, if $\boldsymbol{X}$ has joint distribution function $F_{\boldsymbol{X}}$, marginal distribution functions $F_{X_i}$ $(i \in D)$, and is a jointly continuous variable, then there exists a unique copula $C_{\boldsymbol{X}} : [0,1]^d \to [0,1]$ such that, for all $(u_1, \ldots, u_d) \in [0,1]^d$,

$$C_{\boldsymbol{X}}(u_1, \ldots, u_d) = F_{\boldsymbol{X}}\left(F_{X_1}^{-1}(u_1), \ldots, F_{X_d}^{-1}(u_d)\right).$$

The dependence structure of $\boldsymbol{X}$ is then fully captured through the copula $C_{\boldsymbol{X}}$ independently of the margins. Additionally, where it exists, the copula density $c_{\boldsymbol{X}}(u_1, \ldots, u_d)$

can be obtained by taking the $d^{th}$ order mixed derivative of $C_{\boldsymbol{X}}$ with respect to the variables $u_i$, $i \in D$. A review of a range of parametric copulas is provided by Joe (1997), and methods for construction of flexible copula parametric families, which are tailored to capture different dependence structures throughout the support $[0,1]^d$ with $d \geq 2$, are reviewed by André et al. (2024).

In multivariate extremes, capturing the extremal behaviour of a data set is key to correctly analysing and drawing appropriate conclusions about the data set in hand. In particular, we are interested in models that are flexible enough to accommodate the two regimes of extremal dependence: asymptotic dependence (AD), where the variables are likely to occur together at an extreme level, or asymptotic independence (AI), otherwise. The joint tail behaviour of a random vector $\boldsymbol{X}$ can be quantified through the measure $\chi_D$ (Joe, 1997; Coles et al., 1999), which is defined, where it exists, via the limit $\chi_D = \lim_{r \to 1} \chi_D(r) \in [0,1]$ with

$$\chi_D(r) = \Pr[F_{X_i}(X_i) > r : \forall i \in D \setminus \{1\} \mid F_{X_1}(X_1) > r] = \frac{\Pr[F_{X_i}(X_i) > r : \forall i \in D]}{1-r}, \tag{4.1.1}$$

for $r \in (0,1)$. If $\chi_D > 0$, the variables in $\boldsymbol{X}$ are said to be AD, whilst in the case where $\chi_D = 0$, the variables cannot take all their largest values together. Moreover, larger values of $\chi_D$ indicate stronger AD levels. In the bivariate case, when $\chi_D = 0$, it can be said that $X_1$ and $X_2$ are AI; however, care is needed in higher dimensions as a lower dimensional subvector $\boldsymbol{X}_C = \{X_i : i \in C\}$, where $C \subset D$, could still exhibit AD and have $\chi_C > 0$ even though $\chi_D = 0$ (Simpson et al., 2020).

A complementary measure to $\chi_D$ was introduced by Ledford and Tawn (1996) in the bivariate case, and presented in $d$-dimensions by Eastoe and Tawn (2012). Given a function $\mathcal{L}_D$ that is slowly-varying at infinity, the joint tail of $\boldsymbol{X}$ can be characterised as

$$\Pr[F_{X_i}(X_i) > r : \forall i \in D \setminus \{1\} \mid F_{X_1}(X_1) > r] \sim \mathcal{L}_D((1-r)^{-1})(1-r)^{1/\eta_D - 1}, \tag{4.1.2}$$

as $r \to 1$, and $\eta_D \in (0, 1]$. The extremal dependence structure is quantified through the measure $\eta_D$; when $\eta_D = 1$ and $\mathcal{L}_D(x) \not\to 0$ as $x \to \infty$, then the variables in $\boldsymbol{X}$ are AD, and if $\eta_D < 1$, then they cannot all be extreme together. When $d = 2$, the vector $(X_1, X_2)$ is AI in the latter case. Furthermore, the coefficient $\eta_D$ provides insight about the strength of AI of a given random vector. In particular, if $\eta_D = 1/d$ and $\mathcal{L}_D(x) = 1$ ($\mathcal{L}_D(x) \neq 1$) then independence (near independence) is achieved, whereas when $\eta_D > 1/d$ ($\eta_D < 1/d$), there is evidence of positive (negative) dependence in the extremes. Similar to $\chi_D$, the coefficient $\eta_D$ is taken as the limit $\eta_D = \lim_{r\to 1} \eta_D(r)$, where it exists, with

$$\eta_D(r) = \frac{\log(1-r)}{\log\left(\Pr[F_{X_i}(X_i) > r : \forall i \in D]\right)}, \quad r \in (0, 1). \tag{4.1.3}$$

The two measures $(\chi_D, \eta_D)$ of extremal dependence, however, are only informative when studying the joint tail behaviour, i.e., when all the variables are extreme together. An extension of expression (4.1.2) was proposed by Wadsworth and Tawn (2013) for when the interest lies instead in regions where variables are not required to be equally extreme over different margins. Specifically, the relative level of extremity across $\boldsymbol{X}$ is represented by $\boldsymbol{w} = (w_1, \ldots, w_d) \in \mathcal{S}_{d-1} := \{\boldsymbol{w} \in [0, 1]^d : \sum_{i=1}^d w_i = 1\}$. In this approach, the joint tail behaviour of $\boldsymbol{X}$ is captured through the function $\lambda_D(\boldsymbol{w})$ via the assumption that, for any $\boldsymbol{w} \in \mathcal{S}_{d-1}$ with $w_t > 0$, for some selected $t \in D$, we then have

$$\Pr[F_{X_i}(X_i) > 1 - (1-r)^{w_i/w_t} : \forall i \in D] \sim \mathcal{L}_{\boldsymbol{w}}[(1-r)^{-1/w_t}](1-r)^{\lambda_D(\boldsymbol{w})/w_t} \tag{4.1.4}$$

as $r \to 1$, where function $\mathcal{L}_{\boldsymbol{w}}[(1-r)^{-1/w_t}]$ is slowly-varying at infinity (implying that $\mathcal{L}_{\boldsymbol{w}}(x)$ is slowly-varying at infinity for all $\boldsymbol{w} \in \mathcal{S}_{d-1}$ with $w_t > 0$). The function $\lambda_D(\boldsymbol{w})$ satisfies a number of properties, including $\lambda_D(\boldsymbol{w}) \geq \max\{\boldsymbol{w}\}$ for all $\boldsymbol{w} \in \mathcal{S}_{d-1}$. In the boundary case then $\lambda_D(\boldsymbol{w}) = \max\{\boldsymbol{w}\}$, and the variables in the random vector $\boldsymbol{X}$ exhibit AD. In addition, complete independence is achieved when $\lambda_D(\boldsymbol{w}) = 1$

for all $\boldsymbol{w} \in \mathcal{S}_{d-1}$. Furthermore, the coefficient $\eta_D$ can be obtained from the function $\lambda_D(\boldsymbol{w})$ by setting $\boldsymbol{w} = \mathbf{1}_d/d$ where $\mathbf{1}_d = (1, \ldots, 1)$ is a $d$-dimensional vector as then $\eta_D = [d\lambda_D(\mathbf{1}_d/d)]^{-1}$. In a similar way to $\chi_D$ and $\eta_D$, it can be shown by rearrangement of expression (4.1.4) that $\lambda_D(\boldsymbol{w})$ can be taken as the limit $\lambda_D(\boldsymbol{w}) = \lim_{r\to 1} \lambda_D(\boldsymbol{w}, r)$, where it exists, with

$$\lambda_D(\boldsymbol{w}, r) = w_t \frac{\log(\Pr[F_{X_i}(X_i) > 1 - (1-r)^{w_i/w_t} : \forall i \in D])}{\log(1-r)}, \quad r \in (0,1), \qquad (4.1.5)$$

for any $\boldsymbol{w} \in \mathcal{S}_{d-1}$ with $w_t > 0$ for $t \in D$.

### 4.1.4 Multivariate extreme mixture models

In the univariate framework, the tail region is often defined by the observations exceeding a high threshold value $u$; however, when moving to a multivariate setting, there are several ways of defining such a region. For instance, by assuming a threshold vector $\boldsymbol{u} = (u_1, \ldots, u_d) \in \mathbb{R}^d$, the extremal region can be defined by the observations that jointly exceed $\boldsymbol{u}$ (Ledford and Tawn, 1996), or by the components of $\boldsymbol{X} \in \mathbb{R}^d$ that exceed $\boldsymbol{u}$ in at least one component (Heffernan and Tawn, 2004).

Recent work has been carried out to try and capture the body and tail regions accurately, either to avoid the choice of a threshold vector $\boldsymbol{u}$, or due to the need for modelling the whole distribution, or both. For instance, in the context of precipitation modelling Vrac et al. (2007) propose a bivariate extension of the model of Frigessi et al. (2002) with a bivariate Gamma model for the bulk region. For the upper tail region, they transform $(X_1, X_2)$ into pseudo-coordinates $(R, W)$ by setting $R = X_1 + X_2$ and $W = X_1/(X_1 + X_2)$, and model $R$ with a GP distribution and $W$ with a Beta distribution. This joint modelling approach results in previously unspecified margins. Similarly to Frigessi et al. (2002), a dynamic weighting function is used to bridge both regions, with this function only depending on the variable $R$, while avoiding the choice of

threshold vector. Aulbach et al. (2012a,b) extend the model of Behrens et al. (2004) to a multivariate setting. More specifically, given uniformly distributed variables $\boldsymbol{U} \in [0,1]^d$ and a pre-defined $d$-dimensional threshold vector $\boldsymbol{v} := (v_1, \ldots, v_d) \in [0,1]^d$, one copula is fit for the bulk region defined by the observations $\{\boldsymbol{U} \not\leq \boldsymbol{v}\}$, i.e., where not all components exceed their respective threshold, and the copula of a multivariate extreme value (EV) distribution, is used for the upper tail region $\{\boldsymbol{U} > \boldsymbol{v}\}$. Similarly to Aulbach et al. (2012a,b), the approach of Hu et al. (2024) builds upon the model of Behrens et al. (2004); however, the authors do not work on a copula-based framework. Instead, Hu et al. (2024) model the bulk, defined by the events that are jointly below a quantile vector, with a multivariate GEV distribution whilst a multivariate GP distribution is fitted to the upper tail region, defined by the events where at least one variable exceeds a quantile level. Moreover, the threshold vector is treated as model parameters which need to be estimated. Leonelli and Gamerman (2020) propose modelling the marginal variables using EVMMs while the dependence structure is captured through a mixture of elliptical copulas, which are fit to the full range of the data so that there is no need to define an extremal region. Finally, André et al. (2024) extend the model of Frigessi et al. (2002) to a multivariate setting, whereby two copulas are fitted to the full support of the data and are then blended through a non-decreasing dynamic weighting function. Similarly to Leonelli and Gamerman (2020), they do not require the definition of an extremal region. While the model of Vrac et al. (2007) is limited to the bivariate case, the remaining models can be extended to higher dimensions ($d \geq 2$); however, due to computational constraints, their practical application is restricted to the bivariate setting.

The above listed models do not necessarily cover both AD and AI even in the bivariate case. Specifically, the models of Vrac et al. (2007), Aulbach et al. (2012a,b) and Hu et al. (2024) are only suitable to modelling AD data, whereas those proposed by Leonelli and Gamerman (2020) and André et al. (2024) capture both regimes, although

the former can only capture AI if all the copulas in the mixture exhibit AI. This restriction of the Leonelli and Gamerman (2020) approach makes the analysis highly sensitive as it is necessary to determine which extremal dependence regime is appropriate to fit the data prior to fitting their model. Additionally, Leonelli and Gamerman (2020) and André et al. (2024) require choosing a priori which copula families to include in the mixture model, which then requires multiple model fits using a range of copula combinations.

We propose a copula model constructed from a mixture of multivariate Gaussian distributions which overcomes the limitations of these existing approaches. It accommodates both AI and AD while avoiding the selection of a threshold vector $\boldsymbol{u}$ and, subsequently, the need for defining an extremal region. The model scales relatively well to dimensions $d > 2$, e.g., contrary to the approaches of Leonelli and Gamerman (2020) and André et al. (2024). In addition, we only need to specify the number of Gaussian mixture components to incorporate in the model, for which we develop diagnostics tools to guide this choice. Therefore, while avoiding the choice of copulas or distributions to take, this copula model is suitable to model the two regimes of extremal dependence, and is fast to evaluate, even in a 5-dimensional setting.

This paper is organised as follows: in Section 4.2 we define our proposed model, and introduce its properties, in terms of $(\chi_D, \eta_D)$, and its inference and diagnostic procedures. Section 4.3 presents simulation studies performed to assess the performance of the model. We apply our methodology on the 5-dimensional seasonal air pollution data set analysed by Heffernan and Tawn (2004) in Section 4.4 and conclude in Section 4.5.

## 4.2 Methodology

### 4.2.1 Model definition and inference for copula

Let us consider a $d$-dimensional random vector $\boldsymbol{Y} := (Y_1, \ldots, Y_d) \in \mathbb{R}^d$. We propose a dependence model for the copula of $\boldsymbol{Y}$ based on a mixture of multivariate Gaussian distributions. To do so, we first transform the margins of $\boldsymbol{Y}$ into uniform margins through $\boldsymbol{U} = T(\boldsymbol{Y})$, where $T : \mathbb{R}^d \to [0,1]^d$ is applied componentwise, and then fit a copula to the random vector $\boldsymbol{U}$. Working in a copula-based framework instead of considering the original scale is not novel; see for instance Wadsworth et al. (2017), Huser and Wadsworth (2019), Engelke et al. (2019) or André et al. (2024).

Assume now that we have a mixture of $k \geq 1$ components, where the $j^{th}$ component is a $d$-dimensional random variable $\boldsymbol{Z}_j := (Z_j^1, \ldots, Z_j^d)$ where $j \in K = \{1, \ldots, k\}$. Variables from different mixture components, i.e., $Z_j^i$ and $Z_{j'}^{i'}$ for any $j \neq j' \in K$, are taken to be independent for all $i$, $i' \in D$. Moreover, we assume that $\boldsymbol{Z}_j$ follows a multivariate Gaussian distribution, i.e., $\boldsymbol{Z}_j \sim \text{MVN}(\boldsymbol{\mu}_j, \Sigma_j)$, with mean vector $\boldsymbol{\mu}_j = (\mu_j^1, \ldots, \mu_j^d)'$ and variance-covariance matrix

$$\Sigma_j = \begin{pmatrix} \sigma_{1j}^2 & \rho_j^{1,2}\sigma_{1j}\sigma_{2j} & \ldots & \rho_j^{1,d}\sigma_{1j}\sigma_{dj} \\ \rho_j^{1,2}\sigma_{1j}\sigma_{2j} & \sigma_{2j}^2 & \ldots & \rho_j^{2,d}\sigma_{2j}\sigma_{dj} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_j^{1,d}\sigma_{1j}\sigma_{dj} & \rho_j^{2,d}\sigma_{2j}\sigma_{dj} & \ldots & \sigma_{dj}^2 \end{pmatrix},$$

where $\rho_j^{m,n} \in [-1, 1]$ for $j \in K$ and $m \neq n \in D$ is the correlation between the $m^{th}$ and $n^{th}$ variables of the $\boldsymbol{Z}_j^{th}$ mixture component, and $\sigma_{ij} > 0$ for all $j \in K$ and $i \in D$. Consequently, we have that $Z_j^i \sim N(\mu_j^i, \sigma_{ij}^2)$ for $i \in D$ and $j \in K$. As with any mixture model some conditions need to be imposed to ensure identifiability of the parameters of the mixture terms. Identifiability of our model is further complicated by the sole use of the copula structure, leading to other parameters of our model for $\boldsymbol{Y}$ not being

identifiable on the copula scale. For these two reasons, respectively, we impose that $\mu_{j-1}^1 < \mu_j^1$ for $j = 2, \ldots, k$, and $(\boldsymbol{\mu}_1, \sigma_{11}^2) = (\mathbf{0}, 1)$. The latter condition ensures some referencing point when moving to the copula framework, as copulas are invariant to additive and scale transformations of $\boldsymbol{Y}$.

The random vector $\boldsymbol{Y}$, whose distribution function is $F_{\boldsymbol{Y}}$ and $Y_i \sim F_{Y_i}$ $(i \in D)$, is then defined as below

$$\boldsymbol{Y} = \begin{cases} \boldsymbol{Z}_1 = \left(Z_1^1, Z_1^2, \ldots, Z_1^d\right), & \text{with probability } p_1, \\ \boldsymbol{Z}_2 = \left(Z_2^1, Z_2^2, \ldots, Z_2^d\right), & \text{with probability } p_2, \\ \qquad \vdots & \qquad \vdots \\ \boldsymbol{Z}_k = \left(Z_k^1, Z_k^2, \ldots, Z_k^d\right), & \text{with probability } p_k, \end{cases}$$

where $p_1 \in (0, 1]$, $0 \leq p_j \leq 1$ for $j = 2, \ldots k$, and $\sum_{j=1}^k p_j = 1$. Furthermore, the survivor marginal distribution for random variable $Y_i$ $(i \in D)$ can be defined as

$$\overline{F}_{Y_i}(y) = \sum_{j=1}^{k} p_j \overline{\Phi}\left(\frac{y - \mu_j^i}{\sigma_{ij}}\right), \quad \text{for } y \in \mathbb{R}, \tag{4.2.1}$$

where $\overline{\Phi}$ denotes the standard Gaussian survivor distribution function and $k \geq 1$. Similarly, the joint survivor distribution function of $\boldsymbol{Y}$ is given as

$$\overline{F}_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{j=1}^{k} p_j \Pr[Z_j^1 > y_1, \ldots, Z_j^d > y_d], \quad \text{for } \boldsymbol{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d. \tag{4.2.2}$$

In order to be able to work on a copula-based framework, we first need to transform the margins into uniform margins; this can be achieved via the probability integral transform. More specifically, if $T(\boldsymbol{Y}) = (F_{Y_1}(Y_1), \ldots, F_{Y_d}(Y_d))'$ with $Y_i \sim F_{Y_i}$ for $i \in D$, then $\boldsymbol{U} = T(\boldsymbol{Y})$ is a random vector with standard uniform margins. Therefore, we can

fit the copula density of $\boldsymbol{Y}$ as follows

$$c_{\boldsymbol{Y}}(\boldsymbol{u};\boldsymbol{\theta}) = \frac{f_{\boldsymbol{Y}}\left(F_{Y_1}^{-1}(u_1),\ldots,F_{Y_d}^{-1}(u_d);\boldsymbol{\theta}\right)}{\prod_{i=1}^{d} f_{Y_i}\left(F_{Y_i}^{-1}(u_i)\right)}, \quad \boldsymbol{u}\in[0,1]^d, \tag{4.2.3}$$

where $f_{Y_i}$ and $f_{\boldsymbol{Y}}$ are the density functions of $Y_i$ and $\boldsymbol{Y}$, respectively, $F_{Y_i}^{-1}$ is the inverse cumulative distribution function (cdf) for $i \in D$, and $\boldsymbol{\theta} = (\boldsymbol{p}, (\boldsymbol{\mu}_j, \boldsymbol{\sigma}_{\Sigma_j}, \boldsymbol{\rho}_{\Sigma_j}) : j \in K)$ is the vector of model parameters, where $\boldsymbol{p} = (p_1,\ldots,p_{k-1})$, $\boldsymbol{\mu}_j = (\mu_1^j,\ldots,\mu_d^j)$, $\boldsymbol{\sigma}_{\Sigma_j} = (\sigma_{1j},\ldots,\sigma_{dj})$ and $\boldsymbol{\rho}_{\Sigma_j} = (\rho_j^{1,2},\ldots,\rho_j^{d-1,d})$ from the variance-covariance matrix $\Sigma_j$. This results in $k\,(1 + d(d-3)/2) - d - 2$ model parameters that need to be estimated, which increases with $k$, but more predominantly with dimension $d$.

When the margins of a random variable $\boldsymbol{X} \in \mathbb{R}^d$ are unknown, which is often the case, then $\boldsymbol{U} \in [0,1]^d$, with uniform $[0,1]$ margins, is obtained through componentwise rank transform of the data $\boldsymbol{X}$. For $n$ independent and identically distribution (i.i.d.) observations from $\boldsymbol{X}$, which are assumed to have copula family $c_{\boldsymbol{Y}}(\boldsymbol{u};\boldsymbol{\theta})$, inference on model (4.2.3) is performed by maximum likelihood estimation (MLE) of the vector of parameters $\boldsymbol{\theta}$ with the log-likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} c_{\boldsymbol{Y}}(\boldsymbol{u}_t;\boldsymbol{\theta}), \tag{4.2.4}$$

where $\boldsymbol{u}_t \in [0,1]^d$, $t = 1,\ldots,n$, $n \geq 2$ are the transformed sample of $n$ $\boldsymbol{X}$ variables to uniform margins.

The identifiability constraints on the parameters are imposed within the log-likelihood function; more specifically, every time the optimisation algorithm evaluates a parameter value that fails to satisfy the constraints, a value of $\ell(\boldsymbol{\theta})$ of $-\infty$ is returned. In the case of the mixing probabilities, the estimated $p_k$ ($k \geq 2$) is obtained implicitly as $\boldsymbol{p} \in \mathcal{S}_{d-1}$. In a higher-dimensional setting ($d \geq 2$), the optimisation of the log-likelihood (4.2.4) is initially performed in a lower-dimensional setting to ensure (faster)

convergence to a global maximum; specifically, all pairwise parameter estimates are obtained, and then these are used as initial values for the parameters in the higher-dimensional optimisation. When these initial values do not meet higher dimensional constraints, for example leading to non semi-positive definite $\Sigma_j$, small perturbations on $\boldsymbol{\rho}_{\Sigma_j}$ $(j \in K)$ are added.

To aid the inference procedure, information about the proposed graphical structure of $\boldsymbol{Z}_j$ $(j \in K)$ components can be exploited, enabling both expert judgement coupled with a reduction in the dimensionality of the parameters space of our copula model. For instance, if it is considered appropriate to model variables $Z_j^m$ and $Z_j^n$ for $m \neq n \in D$ for some $j \in K$ as conditionally independent given the remaining $d-2$ variables from the mixture component $\boldsymbol{Z}_j$, this information can be embedded in the likelihood function. Alternatively, fewer parameters need to be estimated if we have pairwise exchangeability, that is, for each mixture component $\boldsymbol{Z}_j$, all means and variances are assumed to be identical, i.e., $\mu_j^1 = \ldots = \mu_j^d$ and $\sigma_{1j}^2 = \ldots = \sigma_{dj}^2$ for each $j \in K$; under this assumption, only a single mean and a single variance are inferred for each one of the $k$ mixture components, rather than $2d$ parameters per component. We note, however, that we are still required to estimate $d(d-1)/2$ correlation parameters.

### 4.2.2 Extremal dependence properties

The sub-asymptotic measure $\chi_D(r)$, given by expression (4.1.1), can be found numerically for any vector of model parameters $\boldsymbol{\theta}$. Although it is known that for a Gaussian copula, $\chi_D(r) \to 0$ and $\eta_D(r) \to (\mathbf{1}_d' \Sigma_{\boldsymbol{\rho}}^{-1} \mathbf{1}_d)^{-1}$, as $r \to 1$, where $\Sigma_{\boldsymbol{\rho}} \in \mathbb{R}^{d \times d}$ is the underlying Gaussian correlation matrix with $d \geq 2$ (Joe, 2014), we will show that $\chi_D(r)$ can be arbitrarily close to 1 for any $r \in (0,1)$ with the Gaussian mixture copula, and thus being able to capture key features of the data at sub-asymptotic levels, even for AD variables. In particular, assume that $\boldsymbol{\mu}_j = \mathbf{0}_d$, where $\mathbf{0}_d = (0, \ldots, 0)$ is a $d$-dimensional vector of zeros, for $j = 1, \ldots, k-1$ and $\boldsymbol{\mu}_k = \mu \mathbf{1}_d$ with $\mu > 0$, and $\boldsymbol{\sigma}_{\Sigma_j} = \mathbf{1}_d$ for

all $j \in K$. Additionally, consider $\rho_j^{m,n} = \rho$ for all $m \neq n \in D$ and $j \in K$, and let $p_j = (k-p)/(k(k-1))$ for $j = 1, \ldots, k-1$ and $p_k = p/k$ for $0 \leq p \leq 1$. This structure allows us to represent our model with $k-1$ similar mixture terms. Furthermore, less weight is assigned to the $k^{th}$ mixture component given that $0 \leq p_k \leq 1/k$. While these are rather simplistic assumptions, the same arguments hold if we take $\boldsymbol{\mu}_j = j\varepsilon$ for small $\varepsilon > 0$ (satisfying in this way the ordering condition) for example, therefore ensuring an identifiable model.

Given this structure, the survivor marginal distribution of $Y_i$ ($i \in D$) from expression (4.2.1) simplifies to

$$\overline{F}_{Y_i}(y) = \left(1 - \frac{p}{k}\right) \overline{\Phi}(y) + \frac{p}{k}\overline{\Phi}(y-\mu), \quad y \in \mathbb{R}. \tag{4.2.5}$$

Similarly, we can define the joint survivor distribution of $\boldsymbol{Y}$, given in expression (4.2.2), as

$$\overline{F}_{\boldsymbol{Y}}(y\mathbf{1}_d) = \left(1 - \frac{p}{k}\right) \overline{\Phi}_d(y\mathbf{1}_d; \Sigma_\rho) + \frac{p}{k}\overline{\Phi}_d((y-\mu)\mathbf{1}_d; \Sigma_\rho),$$

where $\overline{\Phi}_d$ is the standard multivariate Gaussian survivor distribution function with correlation matrix $\Sigma_\rho$, i.e., with all off-diagonal entries $\rho$, and $d \geq 2$.

Consider now a large enough $\mu > 0$. As $y \to \infty$, we have that

$$\overline{F}_{Y_i}(y) \sim \frac{p}{k}\overline{\Phi}(y-\mu) \quad \text{and} \quad \overline{F}_{\boldsymbol{Y}}(y\mathbf{1}_d) \sim \frac{p}{k}\overline{\Phi}_d((y-\mu)\mathbf{1}_d; \Sigma_\rho).$$

Recall that $\chi_D(r)$, defined in expression (4.1.1), is in terms of standard uniform variables. So that we are able to determine $\chi_D(r)$ under the imposed conditions, we need to express it in terms of the variables $Y_i$ for all $i \in D$. Owing to the assumptions made on the parameters, we have common margins; hence we let $y = F_{Y_i}^{-1}(r)$ for $i \in D$.

Expression (4.1.1) can then be rewritten as

$$\chi_D(r) = \frac{\Pr[Y_i > F_{Y_i}^{-1}(r) : \forall i \in D]}{\Pr[Y_1 > F_{Y_1}^{-1}(r)]}, \quad \text{for } r \in (0,1).$$

With the imposed conditions on the mean, variance and correlation parameters, it follows that the sub-asymptotic extremal dependence measure $\chi_D(r)$, can be explicitly written as

$$\chi_D(r) \sim \frac{\overline{\Phi}_d\left((F_{Y_i}^{-1}(r) - \mu)\mathbf{1}_d; \Sigma_\rho\right)}{\overline{\Phi}\left(F_{Y_i}^{-1}(r) - \mu\right)}, \quad \text{as } r \to 1. \tag{4.2.6}$$

Now consider letting $\mu \to \infty$ and $p \to 0$ as $r \to 1$, with $\overline{\Phi}(\mu)/p \to 0$ also as $r \to 1$. This ensures that the marginal tail of the distribution of $Y_i$ ($i \in D$) is dominated by the $k^{th}$ mixture component. Specifically, we have that $1 - r = \overline{F}_{Y_i}(y)$, so from expression (4.2.5), when $\mu = y$,

$$1 - r = \frac{p}{k}\overline{\Phi}(0) + \left(1 - \frac{p}{k}\right)\overline{\Phi}(\mu) = \frac{p}{2k} + \left(1 - \frac{p}{k}\right)\overline{\Phi}(\mu),$$

since $\overline{\Phi}(0) = 1/2$. As $p \to 0$ and $\mu \to \infty$, with $\overline{\Phi}(\mu)/p \to 0$ as $r \to 1$, we then have that

$$1 - r = \frac{p}{2k} + \mathcal{O}(\overline{\Phi}(\mu)) = \frac{p}{2k}\left(1 + \mathcal{O}\left(\frac{\overline{\Phi}(\mu)}{p}\right)\right) = \frac{p}{2k}\left(1 + o(1)\right).$$

Thus, $1 - r \sim p/(2k)$ as $r \to 1$, and $F_{Y_i}^{-1}(r) \sim \mu$ as $r \to 1$. It then follows from expression (4.2.6) that

$$\chi_D(r) \sim \frac{\overline{\Phi}_d\left(\mathbf{0}_d; \Sigma_\rho\right)}{\overline{\Phi}(0)} = 2\overline{\Phi}_d\left(\mathbf{0}_d; \Sigma_\rho\right).$$

Thus, we have that

$$\chi_D(r) \to 2\overline{\Phi}_d\left(\mathbf{0}_d; \Sigma_\rho\right), \quad \text{as } r \to 1.$$

Given that $\overline{\Phi}_d(\mathbf{0}_d; \Sigma_0) = (1/2)^d$ and $\overline{\Phi}_d(\mathbf{0}_d; \Sigma_1) = 1/2$, by suitable changes in $\rho$, the measure $\chi_D(r)$ can exceed any arbitrary level up to 1. This is possible when the mode

$\mu$ is sufficiently larger than the others, and the $k^{th}$ mixture probability $p_k$ approaches 0. Relaxing some or all of these constraints will only allow for more general and richer joint behaviour. We note that similar results can be derived for measures $\eta_D(r)$ and $\lambda_D(\boldsymbol{w}, r)$ from expressions (4.1.3) and (4.1.5), respectively.

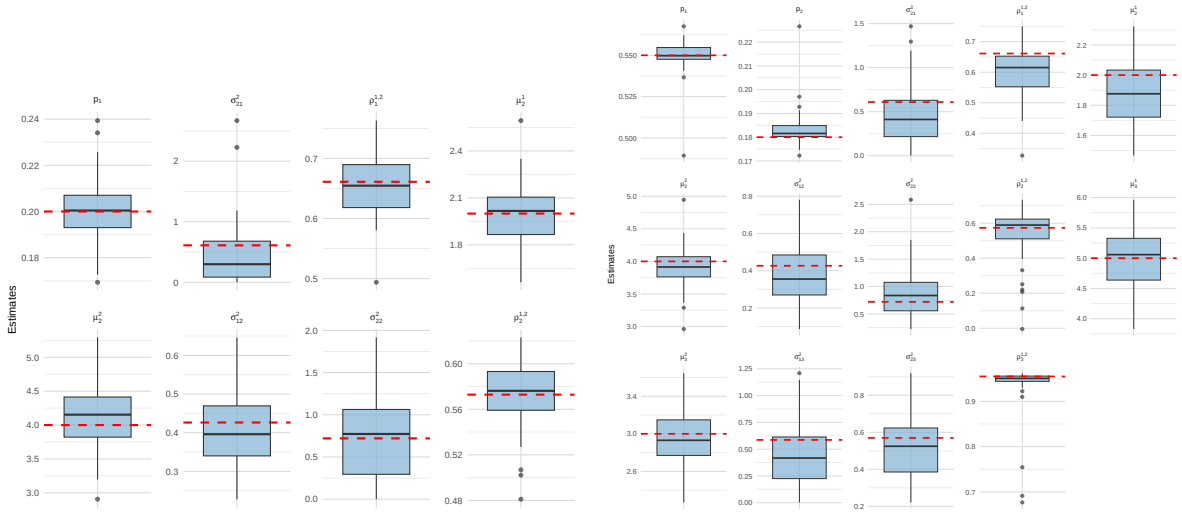
## 4.3 Simulation Studies

### 4.3.1 Model inference

We showcase the identifiability and inference performance of the Gaussian mixture copula by performing a simulation study, illustrating the performance of the sampling distribution of the MLE of $\boldsymbol{\theta}$ over i.i.d. replicated samples. To do so, we consider three Gaussian copula model specifications with $(d, k) = (2, 2)$ (Case I), $(d, k) = (2, 3)$ (Case II) and $(d, k) = (5, 2)$ (Case III) with parameters denoted by $\boldsymbol{\theta}_{\text{I}}$, $\boldsymbol{\theta}_{\text{II}}$ and $\boldsymbol{\theta}_{\text{III}}$ parameters, respectively. In all cases, i.i.d. realisations from model (4.2.3) are generated with a sample size of 1000, and each sample is simulated 50 times. Examples of Cases I-III with pairwise exchangeability, given in Figure B.2.3 of the Supplementary Material, indicate that identifiability is not an issue when a simplified model specification is assumed. Specifically, most estimates are concentrated around the true values for all cases.

For Case I, we set $p_1 = 0.20$, $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = (2, 4)$, $\boldsymbol{\sigma}_{\Sigma_1} = (1.00, 0.61)$, $\boldsymbol{\sigma}_{\Sigma_2} = (0.43, 0.72)$, $\rho_{\Sigma_1} = 0.66$ and $\rho_{\Sigma_2} = 0.57$. In Case II, when an extra mixture component is added, we retain the models for the $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ mixture components, and for the extra mixture component we take $(p_1, p_2) = (0.55, 0.18)$, $\boldsymbol{\mu}_3 = (5, 3)$, $\sigma_{\Sigma_3} = (0.59, 0.57)$ and $\rho_{\Sigma_3} = 0.96$. Figure 4.3.1 displays the results of the simulation study for Cases I and II in the left and right panels, respectively, and the results for Case III are shown in Figure B.2.1 of the Supplementary Material. From the findings of Cases I-III, there is indication that model identifiability is not a concern. Additionally, it can

be seen that the MLE estimates seem to be concentrated around the true values for most parameters, particularly when the model includes fewer parameters. This is to be expected since less parameters often leads to smaller variability in the estimation and parameter dependencies. When moving to a higher dimensional setting, estimation becomes computationally more expensive. Furthermore, as shown in Figure B.2.1 of the Supplementary Material, a few of the estimates appear to deviate further from the true values, particularly those associated with the $\boldsymbol{Z}_2$ mixture component, which seem to show higher variability. Given the high number of parameters to estimate, and that the numerical maximiser converged, without any convergence concerns, for all the 35 parameters in the model, this is not considered an issue with the model or its parameterisation.




(a) (b)

Figure 4.3.1: Boxplots of estimates of the Gaussian mixture copula model based on 50 replicated data sets: (a) Case I and (b) Case II. The true parameter values are indicated by the red lines.

To assess the computational effort required to evaluate the log-likelihood function given in expression (4.2.4), especially when moving to a higher dimensional setting, we record the times taken to optimise the log-likelihood function across the three cases; these are shown in Figure B.2.2 of the Supplementary Material. Additionally, the log-

likelihood function is evaluated using an internal computing node running Linux on an Intel Ice Lake CPU with 200GB RAM memory. As should be expected, the time to optimise one log-likelihood increases with both $d$ and $k$. While the optimisation time increases, on average, in 1.6 minutes when one extra mixture component is added when $d = 2$, in the case of a higher dimension such as $d = 5$, this computational time increases in 6.9 hours, on average.

### 4.3.2 Model fit and diagnostics

#### Overview

The fit of the proposed model (4.2.3) is assessed in a range of data sets exhibiting different dependence structures; we show several cases here and refer the reader to Section B.2.2 of the Supplementary Material for additional cases. For the model selection procedure, we use the Akaike information criterion (AIC; Akaike, 1974). We compare estimates of the extremal dependence measures $\chi_D(r)$ and $\eta_D(r)$ from expressions (4.1.1) and (4.1.3), respectively, obtained through the model fit with their empirical counterparts and their true values. The marginal and joint empirical probabilities needed to estimate the numerator and denominator of the empirical measures in expressions (4.1.1) and (4.1.2) are computed by considering the proportion of points lying in the regions $(u_1, 1)$ and $(u_1, 1) \times \ldots \times (u_d, 1)$, respectively. Furthermore, pointwise 95% confidence intervals are obtained for both measures by computing the empirical $\chi_D(r)$ and $\eta_D(r)$ for $B$ bootstrap samples of the data, with $B$ set to the sample size used in each application.

When we are interested in regions where at least one variable is extreme, we compare probabilities obtained with the model fit and their empirical values in regions of the form $(u_1, 1) \times (0, u_2) \times \ldots \times (0, u_d)$ when considering $U_1$ being extreme (and hence $u_1$ is close to 1), for example. Although this relates with function $\lambda_D(\boldsymbol{w}, r)$ given in expression (4.1.5), we instead obtain such probabilities by considering extremal regions $A_{\boldsymbol{w}}$,

for $\boldsymbol{w} \in \mathcal{S}_{d-1}$, with this region defined by standard exponentially distributed variables. More specifically, in a bivariate setting we have that

$$A_w = \left\{ X_1^E > \max\left\{ \frac{w}{1-w}, 1 \right\} u^E, X_2^E > \max\left\{ \frac{1-w}{w}, 1 \right\} u^E \right\}$$

where $(X_1^E, X_2^E)$ is a 2-dimensional random vector with standard exponential random variables $X_i^E$ for $i = 1, 2$, and $u^E$ is a threshold level for $\max\{X_1^E, X_2^E\}$; see Appendix B.1 for more details. When moving to a $d$-dimensional setting, we consider the probability $\Pr((X_1^E, \ldots, X_d^E) \in A_{\boldsymbol{w}} \mid \max_{i \in D}\{X_i^E\} > u^E)$.

We consider a range of copula families that exhibit different dependence structures to assess the performance of the Gaussian mixture copula. More specifically, for the case where the underlying data are AI, we consider an inverted extreme value copula with logistic dependence structure (Ledford and Tawn, 1997), since this copula is known to have $\chi_D = 0$. Following the same reasoning, we consider an extreme value copula with logistic dependence structure (Gumbel, 1960) to assess the fit given by our model when the data are AD, since this copula has $\chi_D > 0$. To show the performance of the Gaussian mixture copula with non-exchangeable underlying data (i.e., data showing asymmetries), an extreme value copula with an asymmetric logistic dependence structure (Tawn, 1988) is used. Finally, we assess the fit of our copula model with more complex type data by considering a particular specification of the weighted copula model (henceforth referred to as WCM) proposed by André et al. (2024). In all cases, the non-exchangeable Gaussian mixture copula model is used. However, in the AI and AD cases, the performance of the model may improve if the exchangeable model is used instead.

### Asymptotically independent data

The performance of the Gaussian mixture copula is first assessed on bivariate data, $d = 2$, generated from a bivariate inverted extreme value copula with logistic depen-

dence structure, parameter $\alpha_{IL} = 0.6$ and $n = 5000$. This copula has $\chi_D = 0$ and so exhibits AI. We consider Gaussian mixture copulas with $k = 1, 2$ and 3 mixture components. Not surprisingly given the AI nature of the underlying data, all the three specifications provide good fits even though the fitted model does not contain the true copula class as a special case. The decrease in AIC with $k > 1$ in relation to when $k = 1$ is $-171.77$ for $k = 2$ and $-222.06$ for $k = 3$, which indicates the best fit over $k = 1-3$ is given by the copula with $k = 3$ components. The dependence measure $\chi_D(r)$ computed from the three model fits for $r \in (0, 1)$ is shown in the top left panel of Figure 4.3.2, where a comparison with the true $\chi_2(r)$ is given. In addition, we present the results for $\eta_2(r)$ zoomed in for $r \in [0.99, 1)$ in the top right panel. There are differences, though small, between the three fits with $k = 1$ slightly over-estimating the empirical and true $\chi_2(r)$ for higher values of $r$. Given that the three models seem to capture the joint behaviour for all $r$, it can be argued that it is sufficient to consider the simplest model configuration. However, the closeness of fit for $\chi_2(r)$ may not be representative of other joint distribution characteristics, given the clear differences in AIC values. The plot for $\eta_2(r)$ across all $r$, given in Figure B.2.4 of the Supplementary Material, shows similar findings. We also consider a smaller sample size ($n = 1000$) with $k = 1, 2, 3$, where the results shown in Figure B.2.5 of the Supplementary Material indicate a very good fit for all $k = 1, 2, 3$ mixtures.

We also study the $d = 5$ case with $n = 1000$ and a dependence parameter of $\alpha_{IL} = 0.3$. In addition, only $k = 1$ and 2 mixture components are consider. With a decrease in AIC of $-1221.46$ for $k = 2$ in relation to $k = 1$, the model with $k = 2$ is the preferred one to fit the data. This is also visible in the bottom left panel of Figure 4.3.2 with the $k = 2$ model capturing the joint tail behaviour well for all levels $r \in (0, 1)$, whilst with $k = 1$ the model under-estimates the empirical and true $\chi_5(r)$ measures for levels $r < 0.75$. We can see from the plot for $\eta_D(r)$ for $r \in [0.99, 1)$ on the bottom right, however, that the $k = 1$ model is closer to the true $\eta_5(r)$ as $r \to 1$. The results for $\eta_5(r)$

across all $r$ is given in Figure B.2.4 of the Supplementary Material. In both studies, the $\eta_D(r)$ plots given in the right panel of Figure 4.3.2 show that for values $r$ very close to 1, the empirical estimates fail to characterise the joint behaviour, whereas the Gaussian mixture copulas with $k = 1 - 3$ components are all able to extrapolate far into the tail. This sudden drop of the empirical estimates and their pointwise confidence intervals for $r > 0.966$ and $r > 0.99$ for $d = 2$ and $d = 5$, respectively, is due to the lack of observations that are jointly bigger than $r$, resulting in $\eta_D(r)$ not being defined.

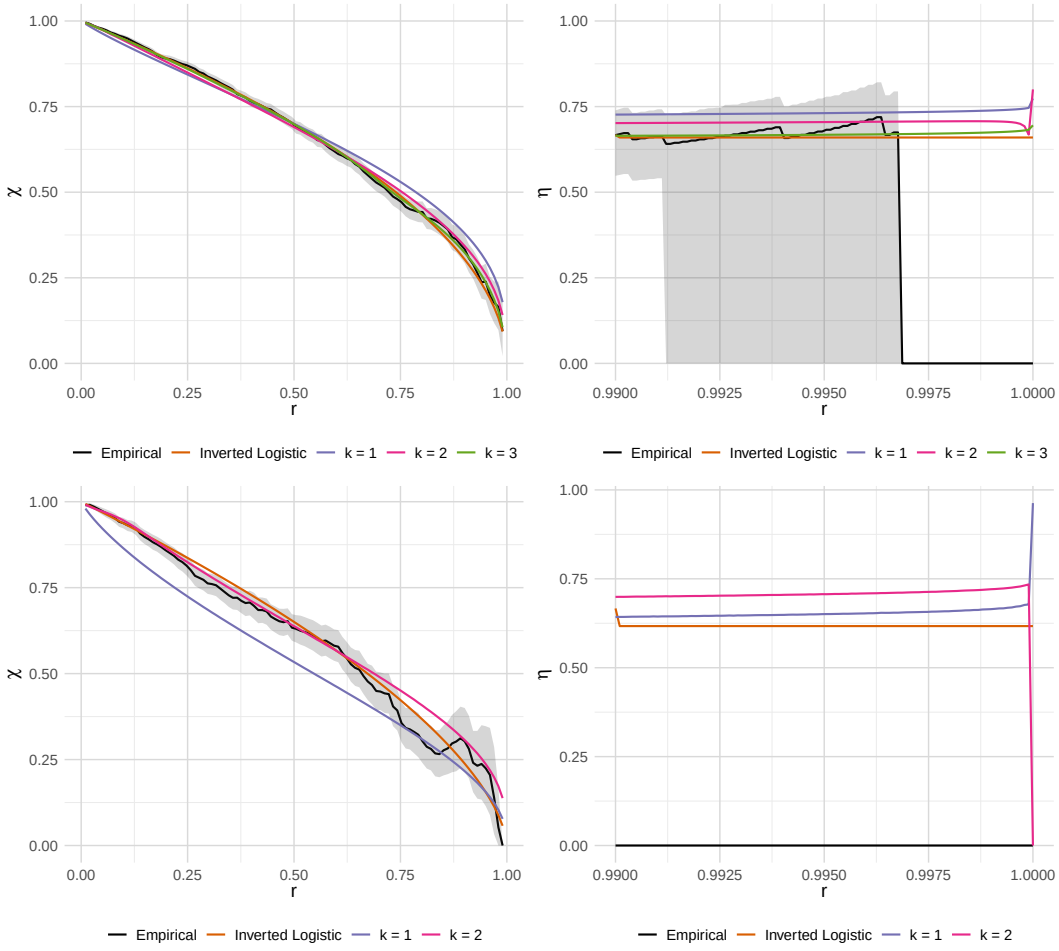



Figure 4.3.2: Estimates of $\chi_D(r)$ for $r \in (0.1)$ with true (in orange) and empirical (in black) values also shown. The corresponding results for $\eta_D(r)$ are zoomed in for $r \in [0.99, 1)$ on the right. The pointwise 95% confidence intervals for the empirical $\chi_D(r)$ are obtained through bootstrap. When $d = 2$ (top), models with $k = 1, 2$ and 3 mixture components are considered, whereas when $d = 5$ (bottom) models with only $k = 1$ and 2 mixture components are studied.

**Asymptotically dependent data**

We then consider an AD copula, specifically the extreme value copula with logistic dependence structure. When $d = 2$, data are generated with dependence parameter $\alpha_L = 0.6$, and sample size $n = 5000$. As mentioned previously, this model exhibits AD with $\chi_2 = 2 - 2^{\alpha_L}$. As before, we compare the fits with $k = 1, 2$ and 3 mixture components. The decrease in AIC with $k > 1$ relative to when $k = 1$ is $-219.28$ for $k = 2$ and $-226.18$ for $k = 3$, indicating that the copula with $k = 3$ mixture components is the one that best fits the data. This is further supported when comparing measure $\chi_2(r)$ for $r \in (0, 1)$ obtained with these three model fits. The results are shown in the top panel of Figure 4.3.3, where in the right $\chi_2(r)$ is zoomed in for $r \in [0.99, 1)$. We can see that in the case where $k = 1$, the fitted model is AI, thus clearly under-estimates $\chi_2(r)$ for $r > 0.6$, with the bias increasing as $r \to 1$. On the other hand, the Gaussian mixture copula with $k = 2$ and 3 are able to account for the behaviour of the joint tail at levels $r$ very close to 1 with values of $\chi_2(r)$ close to the true value over this region. This approximate finding of AD is consistent with the underlying data which is known to exhibit AD. The corresponding plot for $\eta_2(r)$ is given in the left panel of Figure B.2.6 of the Supplementary Material, showing similar findings. A similar study with a smaller sample size ($n = 1000$) is presented in Figure B.2.6 of the Supplementary Material. It can be seen that now the model with $k = 2$ components is not able to capture the extremal behaviour, i.e., at levels of $r$ close to 1. Higher sample sizes may improve the flexibility of this model specification and its ability to capture $\chi_2(r)$ at levels of $r$ very close to 1, as shown by the case when $n = 5000$.

Consider now a $d = 5$ dimensional setting with $n = 1000$ instead. Due to the larger number of parameters we study only $k = 1$ and 2 mixture components. When considering $k = 2$ components, a mixing probability estimate of $\hat{p}_1 = 0.98$ is obtained. Despite $\hat{p}_1$ being so close to one, the extra component adds more flexibility to the modelling of the data, resulting in a decrease in AIC of $-148.69$ in relation to $k = 1$.

This is also visible in the bottom right panel of Figure 4.3.3, where, as before, $\chi_5(r)$ is zoomed in for $r \in [0.99, 1)$. Although the results suggest that for this setting we probably need $k > 2$, and a larger $n$, to get a better estimate of $\chi_5(r)$ from the Gaussian mixture copula, the model with $k = 2$ components is able to capture the joint tail behaviour well for levels $r \in (0.9, 0.99]$, despite it under-estimating $\chi_5(r)$ for lower values of $r$. The results for $\eta_D(r)$ are given in the right panel of Figure B.2.7 of the Supplementary Material, and show similar findings. Similarly to before, the empirical estimates and their pointwise confidence intervals are 0 for $r > 0.998$ and $r > 0.996$ for $d = 2$ and $d = 5$, respectively, since there are no observations that jointly exceed such values. Thus, the empirical $\chi_D(r)$ fails to characterise the joint behaviour beyond these values of $r$. As shown by the right panel, this is not the case for the Gaussian mixture copula, particularly for $d = 2$, as the estimates of $\chi_2(r)$ for the $k = 2$ and $k = 3$ models lie close to the truth for $r$ very close to 1.

**Non-exchangeable data**

To show the performance of the Gaussian mixture copula with non-exchangeable data, we generate $n = 5000$ samples from a bivariate extreme value copula with asymmetric logistic dependence structure with dependence parameter $\alpha_A = 0.2$ and asymmetry parameters $t_1 = 0.2$ and $t_2 = 0.8$ This copula has $\chi_2 = t_1 + t_2 - \left(t_1^{1/\alpha_A} + t_2^{1/\alpha_A}\right)^{\alpha_A}$. As with the previous cases, we consider the Gaussian mixture copula with $k = 1 - 3$. From the results shown in Figure 4.3.4, we see that the $k = 1$ model is not able to capture the extremal behaviour of the data, while the $k = 2$ and $k = 3$ models provide a good fit overall for $\chi_2(r)$ up to $r$ very close to 1. This is in agreement with the AIC values, where a decrease of $-978.76$ for $k = 2$ and of $-1020.02$ for $k = 3$ relatively to the $k = 1$ model is observed, so there is not much difference in AIC for $k = 2$ or $k = 3$. In a similar argument to the model for the AI study, it is sufficient to consider a simpler model with $k = 2$ in this particular case. The results for $\eta_2(r)$, given in Figure B.2.8 of

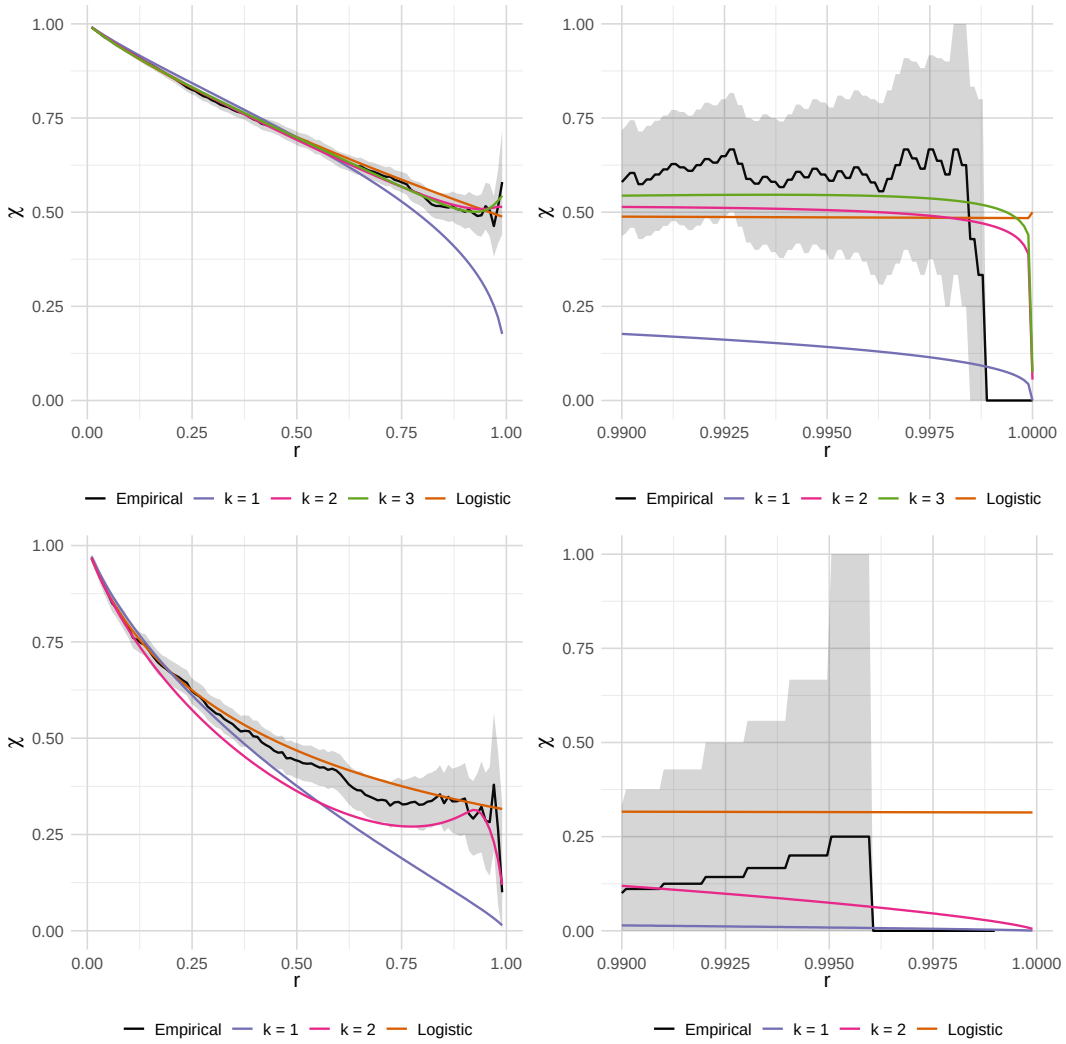



Figure 4.3.3: Estimates of $\chi_D(r)$ for $r \in (0.1)$ with true (in orange) and empirical (in black) values also shown. These are zoomed in for $r \in [0.99, 1)$ on the right. The pointwise 95% confidence intervals for the empirical $\chi_D(r)$ are obtained through bootstrap. When $d = 2$ (top), models with $k = 1, 2$ and 3 mixture components are considered, whereas when $d = 5$ (bottom) models with only $k = 1$ and 2 mixture components are studied.

the Supplementary Material, show similar conclusions.

Further, we assess the performance of the Gaussian mixture copula along different rays $w \in \mathcal{S}_1$ and compute the probability $\Pr(A_w \mid \max_{i=1,2}\{X_i^E\} > u^E)$. The results for the 0.75 and 0.90 quantiles $u^E = \{1.4, 2.3\}$ of $\max_{i=1,2}\{X_i^E\}$, respectively, are shown in Figure 4.3.5 in the left and right panels, respectively. Similarly to measures $\chi_2(r)$ and $\eta_2(r)$, the $k = 2$ and $k = 3$ models capture the extremal behaviour at all $w$ considered

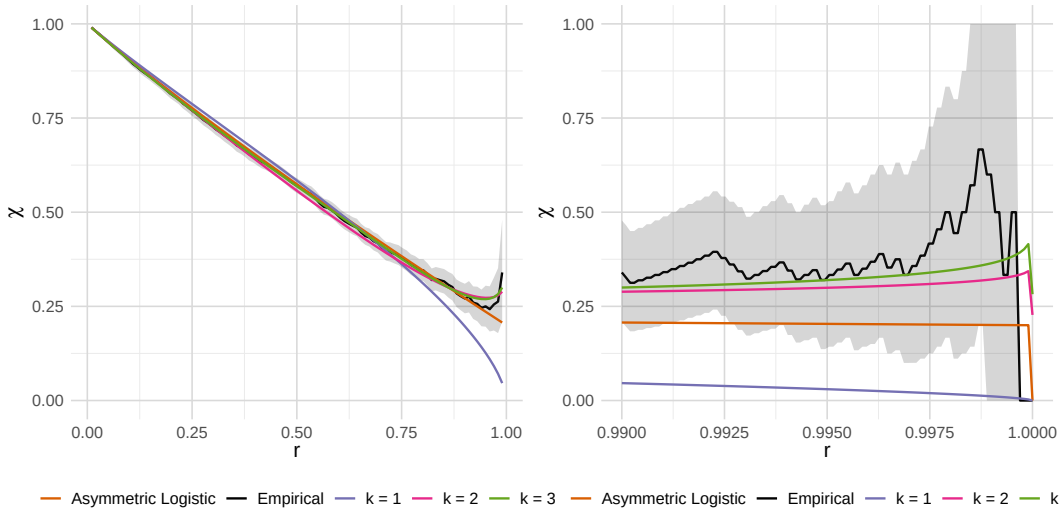



Figure 4.3.4: Estimates of $\chi_2(r)$ for $r \in (0.1)$ with true (in orange) and empirical (in black) values also shown. These are zoomed in for $r \in [0.99, 1)$ on the right. The pointwise 95% confidence intervals for the empirical $\chi_2(r)$ are obtained through bootstrap.

for either $u^E$. In particular, they lie within the pointwise 95% confidence intervals for the empirical probabilities. On the other hand, the Gaussian mixture copula with $k = 1$ under-estimates the joint behaviour for $w \leq 0.5$, and over-estimates otherwise, lying outside of the pointwise 95% confidence intervals for the most $w$. This is particularly pronounced for higher $u^E$, as shown by the right panel.

### Weighted copula model

Finally, we assess the fit of the Gaussian mixture copula in more complex type data. To do so, we consider data generated from a configuration of the WCM. In particular, we take the copula tailored to the tail, $c_t$, to be a bivariate extreme value copula with logistic dependence structure with dependence parameter $\alpha_L = 0.3$, and the copula tailored to the body, $c_b$, to be a Frank copula (Frank, 1979) with parameter $\alpha_F = 2$. Furthermore, we use the dynamic weighting function $\pi(\boldsymbol{v}; \theta) = (v_1 v_2)^\theta$, $\boldsymbol{v} = (v_1, v_2) \in [0, 1]^2$, with $\theta = 1.5$, and $n = 5000$. Similarly to the previous cases, the decrease in AIC with $k > 1$ relative to when $k = 1$ is $-974.27$ for $k = 2$ and $-1017.67$ for $k = 3$, indicating

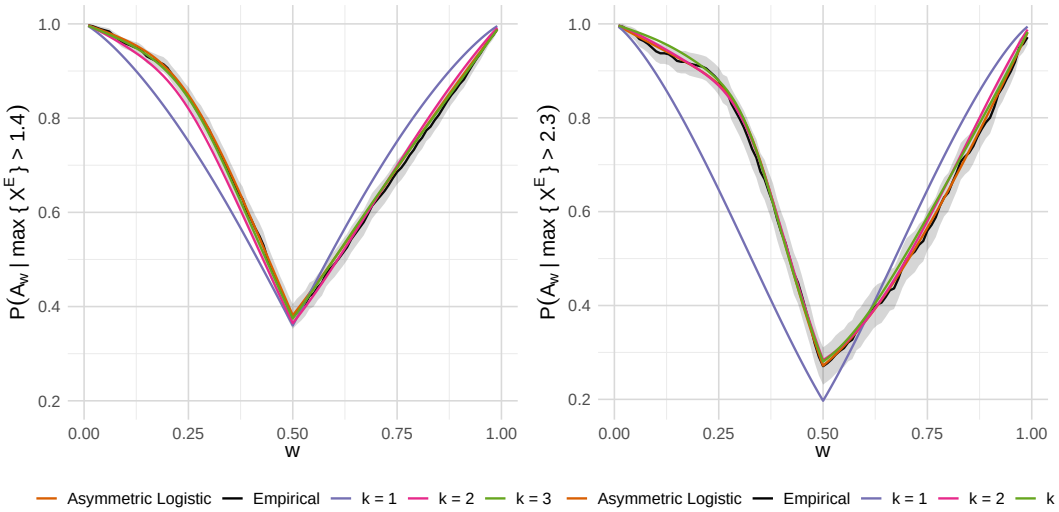



Figure 4.3.5: Comparison between the estimates of probabilities $\Pr(A_w \mid \max_{i=1,2}\{X_i^E\} > u^E)$ for two large values $u^E = \{1.4, 2.3\}$ with true (in orange) and empirical (in black) values also shown. The pointwise 95% confidence intervals for the empirical probabilities are obtained through bootstrap.

that the $k = 3$ model provides the best fit to the data. Likewise to the previous case, the difference in AIC between the $k = 2$ and $k = 3$ models is very small, meaning that the Gaussian mixture copula with $k = 2$ may be sufficient to fit the underlying data. This is also in agreement with the results for $\chi_2(r)$ shown in Figure 4.3.6, and for $\eta_2(r)$ given in Figure B.2.9 of the Supplementary Material. While the $k = 1$ model clearly under-estimates $\chi_2(r)$ from $r > 0.5$, the Gaussian mixture copulas with $k = 2$ and $k = 3$ lie closely to the true $\chi_2(r)$ for the full distribution. Moreover, the results for $\chi_2(r)$ shown in the right panel of Figure B.2.9 indicate that the empirical estimates, and their pointwise confidence intervals, become uninformative, and therefore unreliable, for $r > 0.9975$. This is not the case for the $k = 2$ and $k = 3$ models, for which the estimates for $\chi_2(r)$ remain stable and close to the truth.

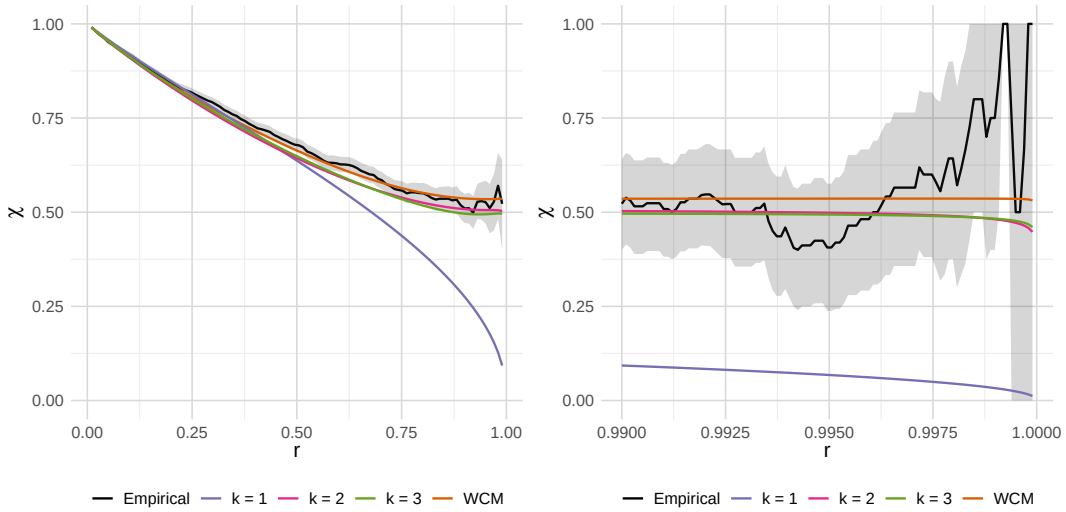



Figure 4.3.6: Estimates of $\chi_2(r)$ for $r \in (0.1)$ with true (in orange) and empirical (in black) values also shown. These are zoomed in for $r \in [0.99, 1)$ on the right. The pointwise 95% confidence intervals for the empirical $\chi_2(r)$ are obtained through bootstrap.

## 4.4 Case study: air pollution data

### 4.4.1 Data description and previous analysis

We apply the Gaussian mixture copula to the 5-dimensional seasonal air pollution data set analysed by Heffernan and Tawn (2004), which consider the joint behaviour of random variables conditionally on one of them being large. Contrary to the Gaussian mixture copula, the conditional approach requires the definition of an extremal region of the form $\{X_2, \ldots, X_d\} \mid \{X_1 > u\}$ for some large marginal threshold $u$, for example. In their study, Heffernan and Tawn (2004) take $u$ to be the 0.9 marginal quantile. However, as stated by Liu and Tawn (2014), this conditional approach is not self-consistent as considering different conditioning variables, i.e., given $\{X_j > u\}$ not given $\{X_i > u\}$ for $j = 2, \ldots, d$, may lead to different conclusions in the joint region $\{X_i > u, X_j > u\}$ $(i \in D)$, which is not the case for the Gaussian mixture copula model.

The data set includes daily maxima of the hourly means of ground level measurements of ozone ($O_3$), nitrogen dioxide ($NO_2$), nitrogen oxide ($NO$), sulphur dioxide

($SO_2$) and particulate matter ($PM_{10}$) recorded at Leeds, UK, from 1994 to 1998. In order to remove the temporal non-stationarity, Heffernan and Tawn (2004) divide the data set into two seasons, winter from the months of November to February, and summer from the months of April to July. In their analysis, the pairs $(NO_2, NO)$, $(NO, PM_{10})$ and $(NO_2, PM_{10})$ were judged to exhibit AD in the winter season, with the remaining pairs (in both seasons) indicating the presence of AI. In our analysis, we denote the variables after rank transformation to uniform $(0, 1)$ variables as $O_3^*$, $NO_2^*$, $NO^*$, $SO_2^*$ and $PM_{10}^*$.

### 4.4.2 Pairwise analysis

We apply our Gaussian mixture copula with $k = 1$ and 2 mixture components to the three pairs that Heffernan and Tawn (2004) identified as being potentially AD to determine whether we obtain similar results. The change observed in the AIC values shown in Table 4.4.1 (denoted by $\text{AIC}_{k_1-k_2}$) suggest that $k = 2$ is the most suitable model for all pairs except $(NO, NO_2)$, for which there is a small increase in AIC for $k = 2$ when compared to the $k = 1$ model. These results are in agreement with the model-based $\chi_2(r)$ obtained for the three pairs for $r \in (0, 1)$, as shown in Figure 4.4.1. In particular, the estimated $\chi_2(r)$ given by the mixture model with $k = 2$ closely aligns with the behaviour of the empirical measure across all $r \in (0.1)$. On the other hand, it is clear that the $k = 1$ model is not able to capture the asymptotic behaviour of pairs $(NO_2, PM_{10})$ and $(NO, PM_{10})$, as it under-estimates the empirical $\chi_2(r)$ for $r > 0.6$ by approaching 0 quicker. However, it appears sufficient for pair $(NO_2, NO)$. Although the estimated mixing probabilities are far from 0 or 1, the AIC and $\chi_2(r)$ results for pair $(NO_2, NO)$ indicate that adding an extra component is not necessary, as little to no difference is notable in the considered diagnostics. Furthermore, for pairs $(NO, NO_2)$ and $(NO_2, PM_{10})$, the empirical $\chi_2(r)$ is clearly positive, which is also mirrored by the sub-asymptotic model-based $\chi_2(r)$ obtained by the $k = 1$ and $k = 2$

models for pair $(NO, NO_2)$, and by the $k = 2$ for pair $(NO_2, PM_{10})$. These results agree with the findings of Heffernan and Tawn (2004). Lastly, the results for $\eta_2(r)$, given in Figure B.3.1 of the Supplementary Material, lead similar conclusions. For pair $(NO, PM_{10})$, the estimated $\chi_2(r)$ from the $k = 2$ model approaches 0 as $r \to 1$, suggesting AI. In this case, measure $\eta_2(r)$ provides more insight. More specifically, $\eta_2(r) \to 0.75$ as $r \to 1$, meaning that the extremes of $(NO, PM_{10})$ exhibit positive dependence.

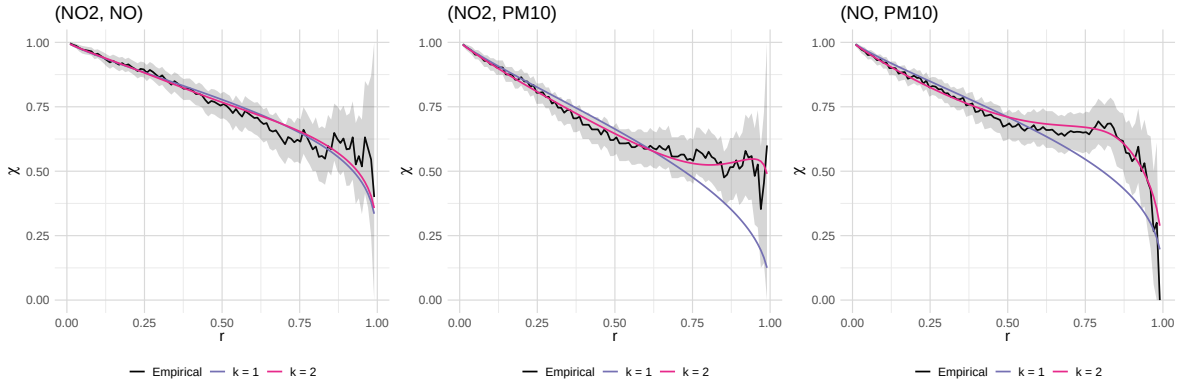



Figure 4.4.1: Estimates of $\chi_2(r)$ for $r \in (0.1)$ with empirical (in black) values also shown for pairs $(NO_2, NO)$ (left), $(NO_2, PM_{10})$ (middle) and $(NO, PM_{10})$ (right). The pointwise 95% confidence intervals for the empirical $\chi_2(r)$ are obtained through bootstrap.

Table 4.4.1: Change in AIC values obtained for the Gaussian mixture copula for $k = 2$ relative to when $k = 1$ for pairs $(NO_2, NO)$, $(NO_2, PM_{10})$ and $(NO, PM_{10})$. The estimated mixing probabilities $(\hat{p}_1, \hat{p}_2)$ are reported for the $k = 2$ model. All the values are rounded to 2 decimal places.

| Pair | $\text{AIC}_{k_1-k_2}$ | $(\hat{p}_1, \hat{p}_2)$ |
|---|---|---|
| $(NO_2, NO)$ | 4.01 | (0.37, 0.63) |
| $(NO_2, PM_{10})$ | −34.40 | (0.91, 0.09) |
| $(NO, PM_{10})$ | −50.87 | (0.78, 0.22) |

### 4.4.3 Trivariate analysis

Before analysing the full data set, we apply the Gaussian mixture copula with $k = 1$ and 2 mixture components to the triple consisting of the pollutants studied in the

bivariate setting, i.e., $(NO_2, NO, PM_{10})$, in the winter season to assess if the triple provides evidence for AD. Even if each of these pairs were AD, the triple being AD does not necessarily follow. However, if one pair (e.g., $(NO, PM_{10})$) were AI, then the triple must also be AI. The decrease in AIC for $k = 2$ relative to when $k = 1$ is of $-61.22$, meaning that the $k = 2$ provides the best fit to the triple according to this criterion. In addition, the mixing probabilities obtained for the $k = 2$ model are $(\hat{p}_1, \hat{p}_2) = (0.73, 0.27)$, indicating that an extra Gaussian component allows for a more flexible fit. This can also be seen with the $\chi_3(r)$ estimates given in Figure 4.4.2. Whilst the true $\chi_3(r)$ is unknown, when comparing the model-based estimates with the empirical values, the $k = 2$ model is able to capture the joint behaviour for all $r \in (0, 1)$. The same is not true with $k = 1$, as it appears to over-estimate the empirical $\chi_3(r)$ for smaller $r$ and clearly under-estimate $\chi_3(r)$ for $r > 0.75$. Given that pair $(NO, PM_{10})$ exhibits AI according to the pairwise analysis, it is not a surprise that both $k = 1$ and $k = 2$ indicate that $NO_2$, $NO$ and $PM_{10}$ cannot all be extreme at the same time, which is consistent with our findings from the three pairwise analysis. The results for $\eta_3(r)$ are shown in Figure B.3.2 of the Supplementary Material, for which the same conclusions can be drawn. Similarly to the pair $(NO, PM_{10})$, the extremes of the triple $(NO_2, NO, PM_{10})$ are jointly positively dependent as $\eta_3(r) \to 0.62$ as $r \to 1$ for both $k = 1 - 2$ models.

We further assess the performance of the Gaussian mixture copula by considering the behaviour of the remaining variables when conditioning on one variable being large. More specifically, we are interested in probabilities where at least one variable is extreme, e.g., of the form $\Pr(NO^* > v, PM_{10}^* > v \mid NO_2^* > u)$ with $v \in (0, 1)$ and some large $u$. Considering such probabilities are key to learn about the risk of one pollutant, in this case $NO_2$, exceeding a large level, as well as its impact on other pollutants, whether they too exceed or not a high level. Similarly to the measure $\chi_3(r)$, we compare the probabilities for both model fits with their empirical counterpart for $u = \{0.75, 0.90\}$;

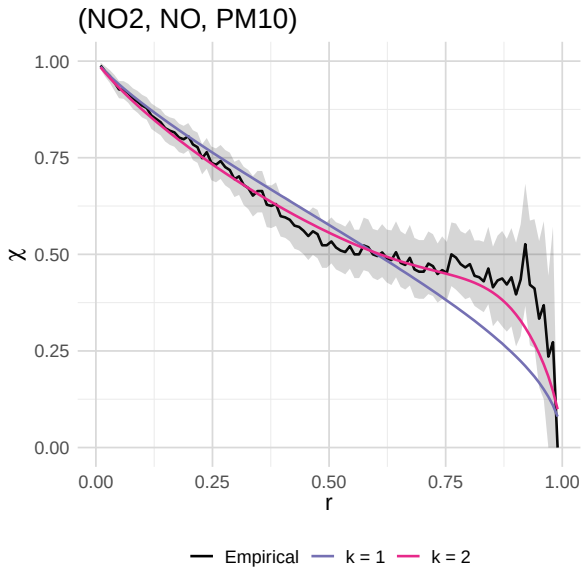



Figure 4.4.2: Estimates of $\chi_3(r)$ for $r \in (0.1)$ with empirical (in black) values also shown for the triple $(NO_2, NO, PM_{10})$. The pointwise 95% confidence intervals for the empirical $\chi_2(r)$ are obtained through bootstrap.

the results are shown in Figure 4.4.3. There is a clear the difference between the $k = 1$ and $k = 2$ models, with an improvement shown by $k = 2$ when $u = 0.9$. In particular, the probabilities across all $v \in (0, 1)$ lie within the empirical pointwise 95% confidence intervals for both $u$. The same is not true for the $k = 1$ model when $u = 0.90$, suggesting that the $k = 1$ model may perform poorly when at least one variable exceeds a very high level, such as 0.90. We note that for $v \leq 0.25$ and $u = 0.9$, the lower and upper bounds of the confidence intervals for the empirical probability coincide and are equal to 1.

Further conclusions about the dependence between the variables can be drawn by exploring the graphical structure of the fit provided by each model. To do so, we analyse the precision matrices estimated from the $k = 1 - 2$ models, denoted by $\Sigma^{-1}_{\boldsymbol{\rho}\,(k=1)}$ and $\Sigma^{-1}_{\boldsymbol{\rho},j\,(k=2)}$ for $j = 1, 2$, respectively; their off-diagonal values are given in Table 4.4.2. From $\Sigma^{-1}_{\boldsymbol{\rho}\,(k=1)}$, estimated with the $k = 1$ model, the entry for $(NO_2, PM_{10})$ is close to 0, which might suggest that $PM_{10}$ is conditionally independent to $NO_2$ given $NO$. From the fitted model with $k = 2$ components, we have $\hat{\boldsymbol{\mu}}_2 = (0.83, 0.90, 2.73)$, meaning that

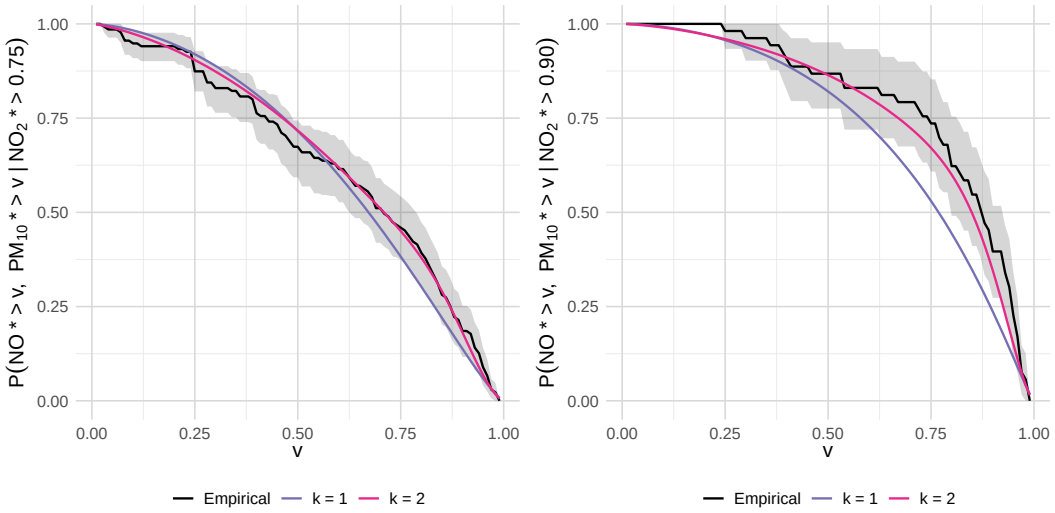



Figure 4.4.3: Comparison between model-based probabilities $\Pr(NO^* > v, PM_{10}^* > v \mid NO_2^* > u)$ for two large values $u = \{0.75, 0.90\}$ given by the Gaussian mixture copula with $k = 1$ (in purple) and $k = 2$ (in pink) components. The empirical probability is given in black, and its pointwise 95% confidence intervals are obtained through bootstrap.

the second mixture component is further in the tail region, as all $\mu_2^i > 0$ for $i = 1, 2, 3$. In addition, the entry for $(NO_2, PM_{10})$ of $\Sigma^{-1}_{\boldsymbol{\rho},1\,(k=2)}$ remains close to 0, suggesting that $PM_{10}$ might be conditionally independent of $NO_2$ given $NO$ in the body of the data. However, in $\Sigma^{-1}_{\boldsymbol{\rho},2\,(k=2)}$, the entry for $(NO_2, PM_{10})$ is no longer close 0, whereas the entry for $(NO, PM_{10})$ is. This might indicate that variable $PM_{10}$ is potentially conditionally independent to $NO$ given $NO_2$ in the extremes. This interpretation would closely agree with the pairwise analysis given that $(NO, PM_{10})$ are potentially AI. This conclusion would benefit from input from atmospheric scientists as it would be reassuring to know if there was a physical basis for the transition of conditional independence from the body to the tails of the joint distribution.

Table 4.4.2: Off diagonal values of the estimated precision matrices $\Sigma^{-1}_{\boldsymbol{\rho}\,(k=1)}$ and $\Sigma^{-1}_{\boldsymbol{\rho},j\,(k=2)}$ for $j = 1, 2$ for triplet $(NO_2, NO, PM_{10})$. All values are rounded to 2 decimal places.

| Model | | $\Sigma^{-1}_{(NO_2,NO)}$ | $\Sigma^{-1}_{(NO_2,PM_{10})}$ | $\Sigma^{-1}_{(NO,PM_{10})}$ |
|---|---|---|---|---|
| $k = 1$ | | $-1.80$ | $\mathbf{-0.09}$ | $-0.92$ |
| $k = 2$ | $(j = 1)$ | $-1.61$ | $\mathbf{-0.07}$ | $-0.43$ |
| | $(j = 2)$ | $-1.68$ | $-0.73$ | $\mathbf{0.03}$ |

### 4.4.4 Higher dimensional analysis

A similar analysis is performed for the full data set ($d = 5$), where, contrary to the pairwise and trivariate analysis, the summer season is also presented. In this case, when considering all the pollutants jointly, the Gaussian mixture copula with only $k = 1$ is the preferred one. In particular, for the summer season, a mixing probability $\hat{p}_1$ of exactly one is obtained when considering $k = 2$ components, meaning that adding an extra component only adds complexity to the model. This is visible in Table 4.4.3 and Figure 4.4.4 with the changes in AIC values and model-based $\chi_5(r)$ obtained. For the summer season, the $k = 2$ model reduces to the $k = 1$ model according to the estimated mixing probabilities, with the larger number of parameters reflected on the change in AIC. Exploring the graphical structure of the underlying data could help reducing the dimensionality in such cases. In particular, potentially conditional independence between variables could be taken into account during the analysis.

Table 4.4.3: Change in AIC values obtained for the Gaussian mixture copula for $k = 2$ relative to when $k = 1$ for $(O_3, NO_2, NO, SO_2, PM_{10})$ for the winter and summer seasons. The mixing probabilities $(\hat{p}_1, \hat{p}_2)$ are reported for the $k = 2$ model. All the values are rounded to 3 decimal places.

| Season | $\text{AIC}_{k_1-k_2}$ | $(\hat{p}_1, \hat{p}_2)$ |
|---|---|---|
| Winter | 25.519 | (0.997, 0.003) |
| Summer | 40.773 | (1.000, 0.000) |

Similarly to the trivariate case, we report the off-diagonal values of the estimated precision matrices from the $k = 1 - 2$ models for the winter and summer seasons in Table 4.4.4. Given that the estimated mixing probability for the summer was $\hat{p}_1 = 1$, the results for the $k = 2$ model are not presented. For the winter, the entry for $(NO, SO_2)$ of $\Sigma^{-1}_{\boldsymbol{\rho}(k=1)}$ is close to 0, which might suggest that $NO$ and $SO_2$ are conditionally independent given $O_3$, $NO_2$ and $PM_{10}$. The same entry remains close to 0 for the first mixture component $j = 1$ from the $k = 2$ model, which would still indicate that these variables are conditionally independent given the remaining pollutants. In addition,

the entry $(NO, PM_{10})$ of $\Sigma^{-1}_{\boldsymbol{\rho},2\,(k=2)}$ is near 0, which suggests that, given the remaining variables, $NO$ and $PM_{10}$ are potentially conditionally independent further in the tail. However, given that $\hat{\boldsymbol{\mu}}_2 = (2.36, 3.18, -0.20, 0.16, 3.96)$ and thus not all $\mu_2^i > 0$ for $i \in D$, it is not clear from $\hat{\boldsymbol{\mu}}_2$ alone if the second mixture component is further in the joint tail. Simulation from the mixture copula, with different mixture components being identified, shows that the second mixture component is allowing for asymmetries in each pair. Similarly to the trivariate results, this interpretation would closely agree with the findings of the pairwise analysis for this pair. For summer, the entries for $(O_3, SO_2)$ and $(NO, SO_2)$ are close to 0. In this case, the results suggest that $O_3$ and $SO_2$ may be conditionally independent given $NO_2$, $NO$ and $PM_{10}$, and the same for variables $NO$ and $SO_2$. Finally, the results from Table 4.4.4 suggest that $(NO, SO_2)$ might be conditionally independent given the remaining variables across both seasons with all models considered. The same is not true for pair $(O_3, SO_2)$; in particular, the results indicate that these variables might potentially be conditionally independent given $NO_2$, $NO$ and $PM_{10}$ in summer but not in winter. As before, it would be beneficial to have atmospheric scientific expertise to help to better understand why this change between seasons is occurring.

Table 4.4.4: Off diagonal values of the estimated precision matrices $\Sigma^{-1}_{\boldsymbol{\rho}\,(k=1)}$ and $\Sigma^{-1}_{\boldsymbol{\rho},j\,(k=2)}$ for $j = 1, 2$ for $(O_3, NO_2, NO, SO_2, PM_{10})$. All values are rounded to 2 decimal places.

| | Model | $\Sigma^{-1}_{(O_3,NO_2)}$ | $\Sigma^{-1}_{(O_3,NO)}$ | $\Sigma^{-1}_{(O_3,SO_2)}$ | $\Sigma^{-1}_{(O_3,PM_{10})}$ | $\Sigma^{-1}_{(NO_2,NO)}$ |
|---|---|---|---|---|---|---|
| W | $k=1$ | $-0.75$ | 0.88 | 0.68 | 0.14 | $-1.86$ |
| | $k=2$ $(j=1)$ | $-0.81$ | 0.91 | 0.66 | 0.38 | $-2.03$ |
| | $k=2$ $(j=2)$ | $-0.31$ | 0.33 | 0.11 | 0.60 | 0.25 |
| S | $k=1$ | $-0.54$ | 0.80 | **0.00** | $-0.24$ | $-1.45$ |
| | Model | $\Sigma^{-1}_{(NO_2,SO_2)}$ | $\Sigma^{-1}_{(NO_2,PM_{10})}$ | $\Sigma^{-1}_{(NO,SO_2)}$ | $\Sigma^{-1}_{(NO,PM_{10})}$ | $\Sigma^{-1}_{(SO_2,PM_{10})}$ |
| W | $k=1$ | $-0.38$ | $-0.38$ | **0.08** | $-0.58$ | $-0.44$ |
| | $k=2$ $(j=1)$ | $-0.18$ | $-0.20$ | **0.02** | $-0.58$ | $-0.37$ |
| | $k=2$ $(j=2)$ | $-0.20$ | 0.20 | 0.90 | **0.04** | 0.42 |
| S | $k=1$ | $-0.48$ | $-0.40$ | **0.03** | $-0.28$ | $-0.54$ |

From Figure 4.4.4, it is clear that $\chi_5(r) \to 0$ as $r \to 1$ when considering the joint behaviour of all pollutants for both seasons indicating that all the pollutants cannot be large together; given that some pollutants are AI between pairs, this is not surprising. The model-based $\chi_5(r)$ obtained with both $k = 1$ and $k = 2$ lie within pointwise 95% confidence intervals for the empirical estimate of $\chi_5(r)$, especially in the winter season. Moreover, both model $\chi_5(r)$ estimates are close to the empirical values, indicating that either model is a good fit to the data. Although not as pronounced as in the winter season, similar conclusions can be drawn for the summer season. The corresponding results for $\eta_5(r)$ are presented in Figure B.3.3. Whilst for the summer season, the model estimates of $\eta_5(r)$ approach 0.35 as $r \to 1$, for the winter season $\eta_5(r) \to 0.2$ with the $k = 1$ model, and $\eta_5(r) \to 0.15$ with the $k = 2$ model. These results indicate that, in the summer season, the extremes of $(O_3, NO_2, NO, SO_2, PM_{10})$ are positively dependent, but in the winter season, they either nearly independent according to the $k = 1$ model, or negatively dependent based on the mixture model with $k = 2$ components. We note that there are no points that are jointly bigger than $r > 0.75$, which results in $\eta_5(r)$ not being defined (recall expression (4.1.3)). Thus, a drop in the empirical $\eta_5(r)$ and corresponding pointwise confidence intervals values is observed.

Similarly to the $d = 3$ case, we assess the performance of the Gaussian mixture copula by considering the behaviour of the remaining variables when conditioning on one variable being large. More specifically, we condition on $O_3^*$ being larger than $u = \{0.75, 0.90\}$, and compare the model-based probabilities with their empirical counterpart; these are shown in Figure 4.4.5. Analogous to the trivariate case, such probabilities inform us about the joint behaviour of the remaining pollutants when, in this case, $O_3$ exceeds large levels. Figure 4.4.5 shows that, for each season, the fitted models seem to capture the conditioning behaviour for all levels $v \in (0, 1)$, especially when $u = 0.75$. However, for $u = 0.9$, and particularly for the summer season, there is evidence that the model fit can be improved as the probabilities estimated by the model

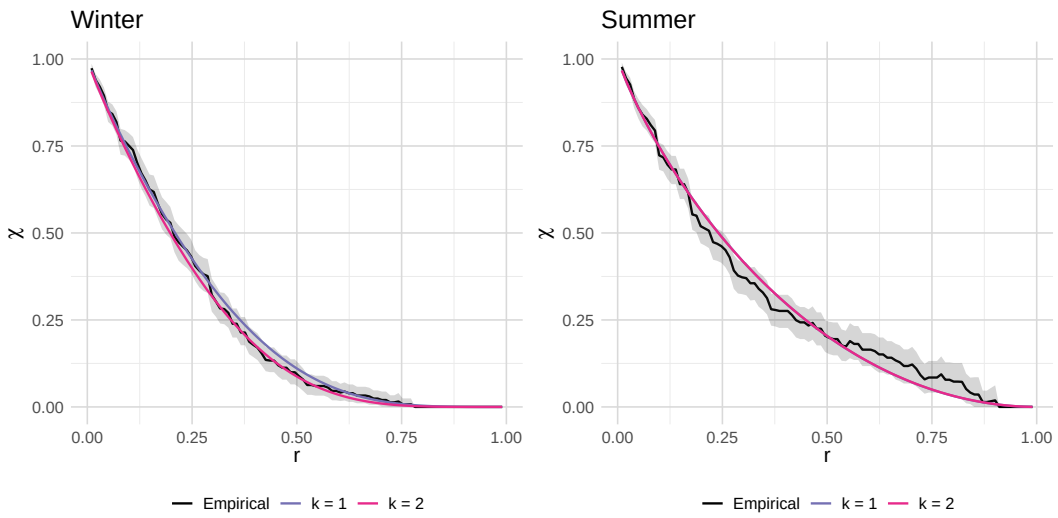



Figure 4.4.4: Estimates of $\chi_5(r)$ for $r \in (0.1)$ with empirical (in black) values also shown for $(O_3, NO_2, NO, SO_2, PM_{10})$ in the winter season (left) and the summer season (right). The pointwise 95% confidence intervals for the empirical $\chi_5(r)$ are obtained through bootstrap. Note that $\chi_5(r)$ for $k = 1$ and $k = 2$ overlap in the right panel.

lie outside the pointwise 95% confidence intervals.

## 4.5 Conclusions and discussion

We proposed a copula model based on a mixture of multivariate Gaussian distributions to represent the body and tail regions of multivariate data. This copula model avoids the need to specify a threshold vector which defines an extremal region, and is able to represent a broad range of complex extremal dependence structures. While the model exhibits asymptotic independence in the limit, theory and the simulation studies performed showed that the Gaussian mixture copula is able to capture asymptotic dependence at quantiles on uniform margins $r$ approaching 1, particularly for models with $k = 3$ mixture components, or $k = 2$ with larger sample sizes. Additionally, we showed that the Gaussian mixture copula is flexible enough to fit more complex data structures, including non-exchangeable data; in particular, the model represents the joint tail for levels $r$ very close to 1, and captures the sub-asymptotic extremal

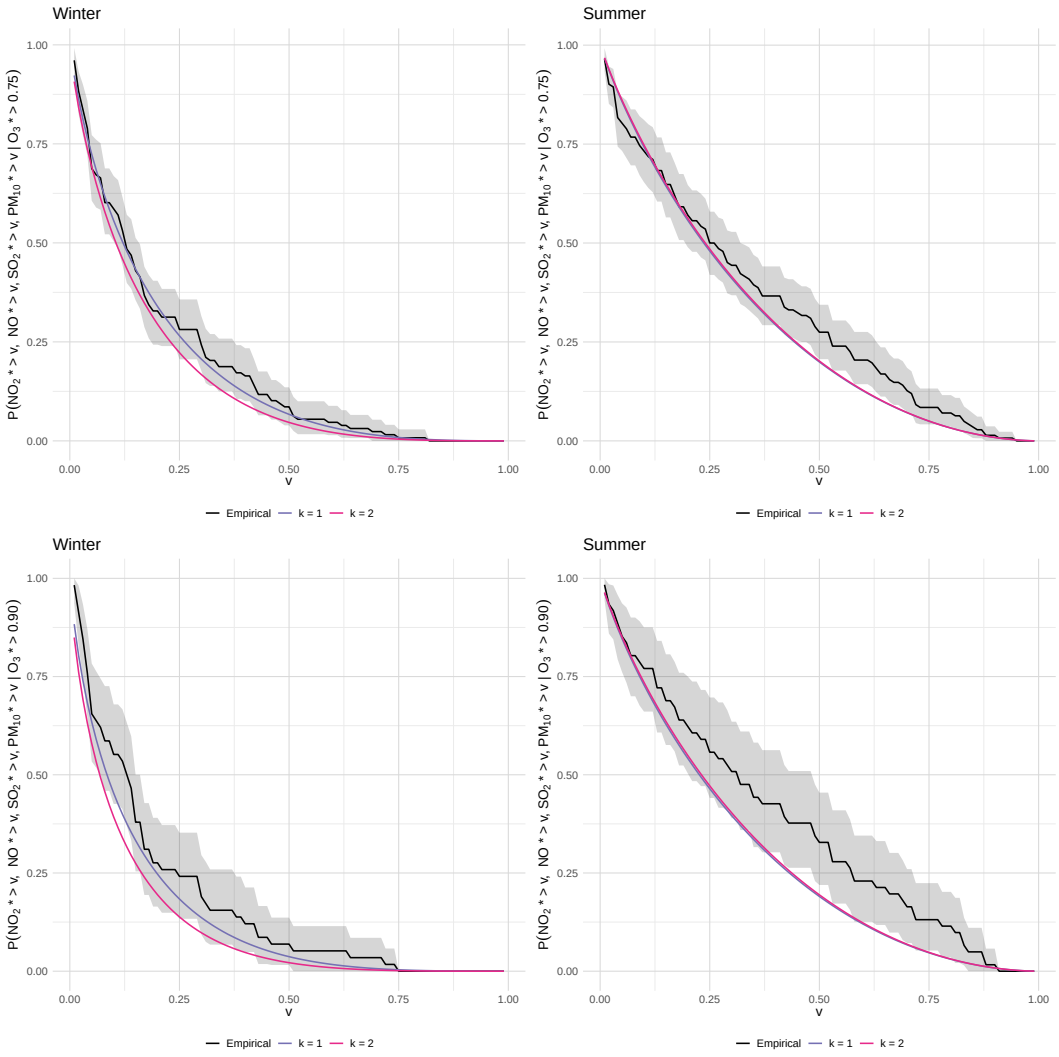



Figure 4.4.5: Comparison between model-based probabilities $\Pr(NO_2^* > v, NO^* > v, SO_2^* > v, PM_{10}^* > v \mid O_3^* > u)$ for $v \in (0,1)$, for two large values $u = \{0.75, 0.90\}$ given by the Gaussian mixture copula with $k = 1$ (in purple) and $k = 2$ (in pink) components. The empirical probability is given in black, and its pointwise 95% confidence intervals are obtained through bootstrap. Note that for the summer season, the $k = 1$ and $k = 2$ model probabilities overlap.

behaviour along different rays accurately.

We showcased the performance of the Gaussian mixture model by applying it to the 5-dimensional seasonal air pollution data set analysed by Heffernan and Tawn (2004). We started by performing a bivariate analysis on the pairs of pollutants identified by Heffernan and Tawn (2004) as exhibiting asymptotic dependence. When applying the proposed copula model, we obtained similar findings for sub-asymptotic levels;

more specifically, we showed that using a model with $k = 2$ mixture components, the joint behaviour could be effectively characterised for values of $r$ very close to 1. Before studying the full data set, we extended the analysis to the triple of pollutants used in the bivariate case. In higher dimensions, it becomes evident that the fitted Gaussian mixture copula exhibits asymptotic independence, which is consistent with the empirical evidence based on non-parametric estimates. Nevertheless, it provides a more accurate representation of the joint behaviour with $k = 2$ components when compared to $k = 1$. This conclusion was further supported by examining the conditional behaviour of the variables at various levels, given one variable being large. For each analysis, we constructed the copula model based on the dimension of each data set. Alternatively, it would be interesting to evaluate the performance of the copula model in fitting the pairs and triple by marginalising the 5-dimensional copula over the variables not included in the joint vector of interest. Finally, as shown by the pairwise study and noted by Simpson et al. (2020), there are variables indexed by $C \subset D$ that exhibit $\chi_C(r) > 0$ even though $\chi_D(r) = 0$ when $r = 0.99$.

Although the Gaussian mixture copula scales relatively well to higher dimensions, the evaluation of its log-likelihood becomes increasingly computationally expensive when $d \geq 2$. For instance, when moving from a bivariate to a 5-dimensional setting, our simulation studies showed that the computational time increased in 6.9 hours on average for a model with $k = 2$ mixture components. This is heavily due to the high number of correlation parameters in the model, but also due to the need for inversion of functions when constructing the copula model. These issues lead to complications in the inference procedure, particularly when we wish to consider adding an extra mixture component, or moving to an even higher dimensional setting. Since simulation from the model is straightforward and efficient, the computational burden of the inference procedure can be mitigated by employing simulation-based methods. Such methods, often referred to as likelihood-free approaches, do not rely on the knowledge of a likeli-

hood function. Examples include approximate Bayesian computation (ABC; e.g., Sisson et al., 2018) or neural network-based techniques (e.g., Zammit-Mangion et al., 2025). Alternatively, the number of parameters in the model could be reduced by exploring data reduction methods for the covariance structure, such as those used in the Gaussian mixture models considered by McNicholas and Murphy (2008).

The literature in mixture modelling is vast. While we have focused on Gaussian mixture models, considering Dirichlet mixture models (see, e.g., Ferguson, 1974, Escobar and West, 1995, De Iorio et al., 2009, Inácio de Carvalho et al. 2017, Quintana et al., 2022) instead would allow, for example, to incorporate non-stationarity, by extending the modelling framework to a regression context. Within a regression framework, alternative approaches include mixture of experts models (Gormley and Frühwirth-Schnatter, 2019), whereby the parameters of the mixture model vary with covariates, or the heavy-tailed normalised generalised Gamma-mixture models proposed by Ramírez et al. (2024).

# Chapter 5

# Neural Bayes estimation for complex bivariate extremal dependence models

## 5.1 Introduction

Recent developments in multivariate extreme value modelling have produced new classes of models that allow for interpolation between the two key tail dependence regimes of asymptotic dependence and asymptotic independence. These models simplify the approach to bivariate extremal modelling, by eliminating the need to pre-determine a dependence regime using unreliable empirical diagnostics. However, their likelihoods often rely on numerical integration and inversion of functions, as well as censoring of non-extreme values, which makes likelihood evaluation burdensome. In other situations, the likelihood function might not be available at all. However, despite the likelihood function being intractable or unavailable, it is often possible to simulate data from the model; this allows for the use of simulation-based likelihood-free algorithms to estimate model parameters.

One simulation-based approach is the pseudo-marginal Markov chain Monte Carlo (MCMC) sampler proposed first by Beaumont (2003) and later formalised by Andrieu and Roberts (2009). When the target distribution (e.g., a likelihood function) is intractable, pseudo-marginal MCMC is able to approximate the target function using an unbiased estimator obtained through importance sampling. The expected value of such an estimator corresponds to the true target distribution, enabling the algorithm to correctly sample from it. Another commonly used likelihood-free procedure is approximate Bayesian computation (ABC; see, e.g., Lintusaari et al., 2017 and Sisson et al., 2018). Specifically, this can be seen as a rejection sampling algorithm, where the model parameters are generated from a prior distribution and subsequently accepted or not based on the distance between the simulated sample and the original sample, often evaluated based on informative summary statistics. Choosing a suitable prior distribution, and defining how similar the samples are, constitute major drawbacks of using ABC to perform inference. A poor choice of prior distribution might lead to misleading posterior estimates, particularly in situations where the selected summary statistics are not very informative. Conversely, a prior distribution which is too informative can result in a posterior that is skewed or biased towards the prior distribution, even if it leads to less variable estimates. Alternatively, Wood (2010) propose the 'synthetic likelihood' method, which constructs an approximate likelihood function by assuming that user-defined summary statistics follow a multivariate normal distribution. This approach is usually easier to tune than ABC and computationally more efficient, especially with higher dimensional data sets (Price et al., 2018), but its underlying Gaussian assumption makes it inflexible in some cases, which may lead to sub-optimal inferences.

More recently, there has been a growing interest in likelihood-free estimation methods which use neural networks; see Zammit-Mangion et al. (2025) for an in-depth review. The extremes literature has also started to be impacted by this new inference paradigm, mostly in the spatial setting; see for instance Lenzi et al. (2023), Majumder

and Reich (2023), Majumder et al. (2024), Richards et al. (2024), Sainsbury-Dale et al. (2024a,b) and Walchessen et al. (2024). An approach to likelihood-free inference using neural networks is to obtain a point estimate of the vector of model parameters through a neural Bayes estimator (Sainsbury-Dale et al., 2024a). It can be argued that training the neural network to build such an estimator is computationally expensive; however, this step only needs to be done once, with estimates subsequently obtained in milliseconds with new data, using a single graphics processing unit (GPU). As mentioned in Zammit-Mangion et al. (2025), this means that neural Bayes estimators are amortised, which allows for their repeated use at almost no extra computational cost (see, e.g. Richards et al., 2024 for a compelling data illustration). This makes neural Bayes estimation an appealing, and much faster, avenue to performing inference compared to state-of-the-art likelihood-based methods. Moreover, unlike other likelihood-free methods, such as ABC, this approach automatically learns the relevant summary statistics for the inference problem at hand. Given the computational complexity of the models of interest, this is the approach we take in this paper. Whilst the methodology developed for neural Bayes estimation has been mostly applied in the spatial and temporal contexts, we are interested in exploring its applicability in a simple bivariate setting, while allowing for censored data inputs. In order to achieve this, an appropriate neural network architecture will need to be designed, along with suitable prior choices. While the number of parameters to estimate may be similar to typical spatial models, the inference procedure may be more challenging in a bivariate setting compared to a spatial context, since there are fewer distinctive features (e.g., location, distance) in the data that the neural Bayes estimator can learn from.

When neural Bayes estimators are adopted for inference, typical model selection techniques, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), are often not available as they require knowledge of the likelihood function. Therefore, having a likelihood-free way of selecting the best model, for a given

set of candidate models fitted to a certain data set, is desirable. Radev et al. (2023) show how the marginal likelihood (and hence, Bayes factors useful for model selection) may be approximated using a neural network-based approach. Similarly, neural methods targeting the likelihood-to-evidence ratio (see, e.g., Cranmer et al., 2016, Hermans et al., 2020) could be used for model selection using likelihood ratios. However, these methodologies rely on full posterior and/or full likelihood neural approximations, which are harder to train than neural Bayes estimators. Ahmed et al. (2024) propose using neural networks for model selection of the extremal dependence structure, but only in the spatial context. Furthermore, it is based on a relatively simple neural network, which — similarly to ABC — relies on user-defined summary statistics (used therein as input to the neural network). Such a neural classifier thus loses discriminatory power if these summary statistics are not sufficient. Alternatively, if the set of candidate models is finite and exhaustive, model selection can naturally be seen as a classification problem. This is the approach we propose, for which we consider a neural network architecture that is analogous to that used for parameter estimation. More specifically, a neural classifier is designed to learn the distinguishing features of each model. Once trained, this classifier is able to estimate the probability of a data set arising from a certain model.

In this paper, we aim to provide a toolbox for simple fitting and comparison of complex bivariate extremal dependence models, which avoids the subjective, and often awkward, selection of summary statistics. We start by exploring the utility of neural Bayes estimation in this specific setting; this is done through assessment of the estimation accuracy of the model parameters and key dependence measures. We then examine the success of the neural model selection classifier; when available, we compare its performance to a likelihood-based information criterion. The end goal is to make the entire statistical pipeline amortised. First, the best model for a given data set is selected through the neural classifier, and then estimates of the model parameters are

obtained through a neural Bayes estimator.

This paper is organised as follows: in Section 5.2, we introduce the methodology used for parameter estimation using neural networks for both uncensored and censored data, and describe our model selection procedure based on a classification task. Section 5.3 presents an overview of bivariate extreme value modelling and introduces the models of interest for which likelihood-based inference is burdensome. Simulation studies assessing our proposed inference and model selection frameworks are discussed in Section 5.4. We then apply the proposed toolbox to study the pairwise extremal behaviour of the changes in horizontal geomagnetic field fluctuations between three locations in Section 5.5, followed by a conclusion in Section 5.6.

## 5.2 Inference methodology

In this section, we review background on neural point estimation using the methodology developed by Sainsbury-Dale et al. (2024a) and Richards et al. (2024) and describe our neural approach to model selection. In Section 5.2.1, we introduce neural Bayes estimators; the basic approach outlined here is suitable when the sample size is fixed and known, and data are fully observed (i.e., uncensored). In Sections 5.2.2 and 5.2.3, we explain how to adapt the estimation procedure to account for variable sample size and censored data, respectively. In Section 5.2.4, we present our neural classifier for model selection and describe how to construct it. Finally, in Section 5.2.5, we give implementation details.

### 5.2.1 Neural Bayes estimators

Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \in \mathcal{S} \subseteq \mathbb{R}^d$ be $n$ independent and identically distributed random vectors with density $f(\boldsymbol{z}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^p$ is the vector of parameters, and let their collection be represented by $\boldsymbol{Z} = (\boldsymbol{Z}_1', \ldots, \boldsymbol{Z}_n')'$. A point estimator $\hat{\boldsymbol{\theta}}(\cdot)$ maps data $\boldsymbol{Z}$ to parameter

estimates from the parameter space $\Theta$, i.e., $\hat{\boldsymbol{\theta}} : \mathcal{S}^n \to \Theta$. Given a non-negative loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\cdot))$, a Bayes estimator minimises a weighted average of the risk at $\boldsymbol{\theta}$, $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\cdot))$, which may be expressed as

$$r_\Omega(\hat{\boldsymbol{\theta}}(\cdot)) = \int_\Theta R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\cdot)) \mathrm{d}\Omega(\boldsymbol{\theta}) = \int_\Theta \int_{\mathcal{S}^n} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\boldsymbol{z})) f(\boldsymbol{z} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{z} \mathrm{d}\Omega(\boldsymbol{\theta}), \tag{5.2.1}$$

where $\Omega(\cdot)$ is a prior measure for $\boldsymbol{\theta}$. Equation (5.2.1) is known as the Bayes risk. Under suitable regularity conditions and the squared error loss, these estimators are consistent and asymptotically efficient; see for instance Lehmann and Casella (1998, Ch. 5 and 6).

In practice, however, Bayes estimators are rarely available in closed form, and the Bayes risk in equation (5.2.1) is difficult to evaluate. This can be overcome by approximating these estimators using a neural network, since these are universal function approximators (Hornik et al., 1989, Sainsbury-Dale et al., 2024a). In this situation, a neural point estimator $\hat{\boldsymbol{\theta}}(\cdot; \boldsymbol{\gamma})$ is constructed as a neural network (with parameters $\boldsymbol{\gamma}$) that returns a point estimate from data input $\boldsymbol{Z}$. Bayes estimators may thus be approximated with $\hat{\boldsymbol{\theta}}(\cdot; \boldsymbol{\gamma}^*)$ where $\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} r_\Omega(\hat{\boldsymbol{\theta}}(\cdot; \boldsymbol{\gamma}))$. Since equation (5.2.1) can rarely be evaluated, it is usually approximated using Monte Carlo techniques as follows

$$r_\Omega(\hat{\boldsymbol{\theta}}(\cdot; \boldsymbol{\gamma})) \approx \frac{1}{KJ} \sum_{\boldsymbol{\theta} \in \Upsilon} \sum_{\boldsymbol{Z} \in \mathcal{Z}_{\boldsymbol{\theta}}} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\boldsymbol{Z}; \boldsymbol{\gamma})), \tag{5.2.2}$$

where $\Upsilon$ is a set of $K$ samples $\boldsymbol{\theta} \sim \Omega$ from the prior and $\mathcal{Z}_{\boldsymbol{\theta}}$ is a set of $J$ samples $\boldsymbol{Z} \sim f(\boldsymbol{z}; \boldsymbol{\theta})$ for each given $\boldsymbol{\theta}$. A neural point estimator $\hat{\boldsymbol{\theta}}(\cdot; \boldsymbol{\gamma}^*)$ is called a neural Bayes estimator (NBE) as it minimises a Monte Carlo approximation of the Bayes risk; see Sainsbury-Dale et al. (2024a) for more details.

The discrepancy between the neural Bayes estimator and the true Bayes estimator will depend on a few factors, one of which is the neural network architecture (Sainsbury-Dale et al., 2024a). Through a judicious choice of architecture, NBEs can be enforced to satisfy the fundamental property that $\hat{\boldsymbol{\theta}}(\boldsymbol{Z}; \boldsymbol{\gamma}) = \hat{\boldsymbol{\theta}}(\boldsymbol{Z}^*; \boldsymbol{\gamma})$, for any permutation

$\boldsymbol{Z}^*$ of the independent replicates in $\boldsymbol{Z}$; this can be achieved by exploiting a neural network architecture known as DeepSets (Zaheer et al., 2017). Let $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^q$ and $\boldsymbol{\phi} : \mathbb{R}^q \to \mathbb{R}^p$ be two multi-layer neural networks parametrised by $\boldsymbol{\gamma}_{\boldsymbol{\psi}}$ and $\boldsymbol{\gamma}_{\boldsymbol{\phi}}$, respectively, and $\boldsymbol{a} : (\mathbb{R}^q)^n \to \mathbb{R}^q$ be a permutation-invariant set function where each element $a_s(\cdot)$ returns the elementwise average over its input set for $s = 1, \ldots, q$. The NBE is then represented as

$$\hat{\boldsymbol{\theta}}(\boldsymbol{Z}; \boldsymbol{\gamma}) = \boldsymbol{\phi}\left(\boldsymbol{T}(\boldsymbol{Z}; \boldsymbol{\gamma}_{\boldsymbol{\psi}}); \boldsymbol{\gamma}_{\boldsymbol{\phi}}\right) \quad \text{with} \quad \boldsymbol{T}(\boldsymbol{Z}; \boldsymbol{\gamma}_{\boldsymbol{\psi}}) = \boldsymbol{a}\left(\{\boldsymbol{\psi}(\boldsymbol{Z}_i; \boldsymbol{\gamma}_{\boldsymbol{\psi}}) : i = 1, \ldots, n\}\right), \tag{5.2.3}$$

where $\boldsymbol{T}$ denotes a vector of learnt summary statistics, and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_{\boldsymbol{\phi}}, \boldsymbol{\gamma}'_{\boldsymbol{\psi}})'$ are the parameters of the neural networks $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$. In the multivariate unstructured setting, a dense neural network (DNN) may be used for both $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$; see Sainsbury-Dale et al. (2024a) for more details. A schematic of the DeepSets architecture is shown in Section C.1 of the Supplementary Material.

### 5.2.2 Variable sample size

When training a neural Bayes estimator on a data set with a fixed number, $n$, of replicates, this estimator will generally not be Bayes for a data set with a different sample size $\tilde{n} \neq n$. Therefore, in order to ensure that the trained NBE approximately minimises the Bayes risk $r_\Omega(\hat{\boldsymbol{\theta}}(\cdot; \boldsymbol{\gamma}))$, Sainsbury-Dale et al. (2024a) propose two approaches: either obtaining a piecewise neural Bayes estimator by pre-training the estimator for specific fixed sample sizes (Goodfellow et al., 2016), or treating the sample size as a random variable $N$; we adopt the latter in our work.

Let us assume that the sample size $N$ follows a discrete uniform distribution, that is $N \sim \text{Unif}(\{n_1, n_1 + 1, \ldots, n_2\})$ where $\Pr(N = n) = 1/(n_2 - n_1 + 1)$ for $n \in (n_1, n_2)$ and $n_1, n_2 \in \mathbb{N}$. Further, the sample size $N$ is assumed independent of the model parameters $\boldsymbol{\theta}$. This extra random variable modifies the Bayes risk function (5.2.1), which can now

be approximated as follows

$$r_\Omega(\hat{\boldsymbol{\theta}}(\cdot;\boldsymbol{\gamma})) \approx \frac{1}{KJ}\sum_{\boldsymbol{\theta}\in\Upsilon}\sum_{n\in\mathcal{N}}\sum_{\boldsymbol{Z}\in\mathcal{Z}_{\boldsymbol{\theta},n}} \Pr(N=n)L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\boldsymbol{Z};\boldsymbol{\gamma})),$$

where $\mathcal{N}$ is the set of sample sizes drawn from the Unif$(\{n_1, n_1+1,\dots,n_2\})$, and $\mathcal{Z}_{\boldsymbol{\theta},n}$ is a set of $J$ data sets of size $n$ drawn from the model for the sampled parameters $\boldsymbol{\theta}$. Thus, during training, it is now necessary to simulate the sample size along with model parameters from the prior and replicated data from the model; however, the general method remains the same.

### 5.2.3 Censored data

The methodology proposed by Sainsbury-Dale et al. (2024a) is not able to handle censored data as input of the neural network; however, this is essential in multivariate models aimed at capturing the extremal dependence structure. In these models, low observations are often censored to prevent these non-extreme values from affecting the estimation of this tail dependence. Thus, Richards et al. (2024) propose an adaptation of the neural Bayes estimators in order to include this type of data.

Consider the random vector $\boldsymbol{Z}_i = (Z_{i1},\dots,Z_{id})'$, $i = 1,\dots,n$, and let $F_j^{-1}$ be the inverse cumulative distribution function (cdf) of variable $Z_{i,j}$, $j = 1,\dots,d$. There are various censoring schemes that can be adopted. One possibility, used by Richards et al. (2024), is to censor the observations that fall below a high marginal quantile $\tau$, i.e., if $Z_{i,j} < F_j^{-1}(\tau;\boldsymbol{\theta})$, then $Z_{i,j}$ is treated as censored. Instead, in this paper, we censor the observations only if all the components are below the marginal quantile, i.e., if $\max_{j=1,\dots,d} Z_{i,j} < F_j^{-1}(\tau;\boldsymbol{\theta})$, then the entire vector $\boldsymbol{Z}_i$ is treated as fully censored, otherwise, if at least one component of $\boldsymbol{Z}_i$ has a value above its marginal quantile, the entire vector is treated as uncensored. In order for the neural Bayes estimator to account for censored-type data, Richards et al. (2024) propose standardising data

$\boldsymbol{Z}_i$ $(i = 1, \ldots, n)$ to a common margin and setting the censored observations to some constant $c \in \mathbb{R}$, yielding $\boldsymbol{Z}_i^* \in \mathbb{R}^d$ $(i = 1, \ldots, n)$. To improve performance of the NBE, this constant $c$ should be set to a value outside of the support of the data. In order for information on the censored observations to be passed to the NBE, Richards et al. (2024) propose creating a one-hot encoded vector $\boldsymbol{I}_i$ that identifies which indices of $\boldsymbol{Z}_i^*$ are censored (with value 1), and which are not (with value 0). Then, the neural Bayes estimator is trained using an augmented data set $\boldsymbol{A}$ containing the data $\boldsymbol{Z}^* = ((\boldsymbol{Z}_1^*)', \ldots, (\boldsymbol{Z}_n^*)')'$ and the indicator vector $\boldsymbol{I} = (\boldsymbol{I}_1', \ldots \boldsymbol{I}_n')'$, that is $\boldsymbol{A} = ((\boldsymbol{Z}^*)', \boldsymbol{I}')'$. Passing $\boldsymbol{A}$ as the input to the neural network in place of $\boldsymbol{Z}$ in equation (5.2.3) is sufficient to ensure the information about the censoring scheme is given to the NBE. Lastly, to handle the augmented data set $\boldsymbol{A}$, a dense bilinear layer is used as the input layer of the neural network $\boldsymbol{\psi}(\cdot)$; this allows for a full connection between two inputs (here, $\boldsymbol{Z}^*$ and $\boldsymbol{I}$) and the output.

Similarly to the sample size, these NBEs are only (approximately) optimal when applied to data sets where the censoring level is kept the same as the one used for training. When this is not the goal, having an estimator which performs well for any valid $\tau \in (0, 1)$ is desirable. This can be achieved by feeding $\tau$ as an extra input to the outer neural network $\boldsymbol{\phi}(\cdot)$ as

$$\hat{\boldsymbol{\theta}}(\boldsymbol{A}; \boldsymbol{\gamma}, \tau) = \boldsymbol{\phi}\left(\boldsymbol{T}(\boldsymbol{A}; \boldsymbol{\gamma}_{\boldsymbol{\psi}}, \tau); \boldsymbol{\gamma}_{\boldsymbol{\phi}}\right) \quad \text{with} \quad \boldsymbol{T}(\boldsymbol{A}; \boldsymbol{\gamma}_{\boldsymbol{\psi}}, \tau) = \left(\boldsymbol{T}(\boldsymbol{A}; \boldsymbol{\gamma}_{\boldsymbol{\psi}})', \tau\right)',$$

where $\boldsymbol{T}(\boldsymbol{A}; \boldsymbol{\gamma}_{\boldsymbol{\psi}}) = \boldsymbol{a}\left(\{\boldsymbol{\psi}(\boldsymbol{A}_i; \boldsymbol{\gamma}_{\boldsymbol{\psi}}) : i = 1, \ldots, n\}\right)$ as in (5.2.3) with $\boldsymbol{A}_i$ in place of $\boldsymbol{Z}_i$.

In this situation, the censoring level is treated as a random variable T, which requires an additional prior. Since we are not interested in censoring too low, the prior $\mathrm{T} \sim \mathrm{Unif}(\tau_1, 1)$ with $\tau_1 > 0$ seems the most straightforward choice. Thus, each vector of parameters $\boldsymbol{\theta}$ now has a censoring level associated with it, making it possible to have different censoring levels for the $K$ training parameter vectors and corresponding $J$ samples. The Monte Carlo approximation of the Bayes risk in this case takes now

the form

$$r_\Omega(\hat{\boldsymbol{\theta}}(\cdot;\tau,\boldsymbol{\gamma})) \approx \frac{1}{KJ}\sum_{\boldsymbol{\theta}\in\Upsilon}\sum_{n\in\mathcal{N}}\sum_{\tau\in\mathcal{T}}\sum_{\boldsymbol{A}\in\mathcal{A}_{\boldsymbol{\theta},n,\tau}} \Pr(\mathrm{T}=\tau)\Pr(N=n)L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\boldsymbol{A};\boldsymbol{\gamma},\tau)). \quad (5.2.4)$$

In equation (5.2.4), $\mathcal{T}$ is the set of censoring levels drawn from the Unif$(\tau_1, 1)$ distribution with $\Pr(\mathrm{T}=\tau) = 1/(1-\tau_1)$, and $\mathcal{A}_{\boldsymbol{\theta},n,\tau}$ is a set of $J$ data sets of size $n$ drawn from the model for given sampled parameters $\boldsymbol{\theta}$, masked at the censoring level $\tau$ and augmented with the threshold exceedances indicator vector as described above.

### 5.2.4 Model selection

In a likelihood-free setting, we can treat model selection as a multiclass classification problem. Consider $M \geq 2$ candidate models indexed by $\zeta \in \{1,\ldots,M\}$, and let $p_\zeta \in [0,1]$ denote the prior probability of each data set $\boldsymbol{Z}$ being generated by the model with index $\zeta$, where $p_1 + \ldots + p_M = 1$. Let us also assume that, a priori, each data set has equal probability of being assigned to each model (herein referred to as a 'class'). Thus, our prior on the class index $\zeta$ is $\zeta \sim \text{Multinomial}(1/M,\ldots,1/M)$.

Similarly to the parameter estimation procedure, a dense neural network (DNN) is used to train the model selection classifier. Note that, as before, a prior on the vector of model parameters is also required and it should be in agreement with the prior distributions used for training of the NBEs described in Section 5.2.1. More specifically, our neural classifier takes data $\boldsymbol{Z}$ as input, and maps it using a DNN to an estimate $\hat{\boldsymbol{p}} = (\hat{p}_1,\ldots,\hat{p}_M)'$ of $\boldsymbol{p}$, giving the posterior probabilities of $\boldsymbol{Z}$ belonging to each class index $\zeta \in \{1,\ldots,M\}$. In the multiclass classification problem, a natural choice of loss function is the categorical cross-entropy function, which can be thought of as stemming from a multinomial likelihood. This loss function is given by

$$L(\boldsymbol{\zeta}^*, \hat{\boldsymbol{p}}) = -\sum_{m=1}^{M} \zeta_m^* \log(\hat{p}_m),$$

where $\boldsymbol{\zeta}^* = (\zeta_1^*, \ldots, \zeta_M^*)'$ with $\zeta_m^* \in \{0, 1\}$ being the indicator of the true model corresponding to the training data set $\boldsymbol{Z}$, that is $\zeta_m^* = 1$ if $\boldsymbol{Z}$ is generated from model $m$, and $\zeta_m^* = 0$ otherwise. In addition, $\hat{p}_m$ is the estimated probability outcome of each data set $\boldsymbol{Z}$, for $m = 1, \ldots, M$. This loss function measures the dissimilarity between the estimated outcomes and actual classes, adjusting the training of the model by penalising incorrect predictions. Finally, the DeepSets architecture introduced in Section 5.2.1 can also be applied to the model selection procedure, and similar techniques to those used for parameter estimation can be employed to ensure the model selection classifier remains suitable for censored data, variable sample sizes, and/or variable censoring levels.

We note, however, that while in likelihood-based inference the classical model selection criteria, such as the BIC, are linked with the parameter estimation (i.e., the parameter estimates are used to calculate the BIC value), the same does not hold true with the proposed toolbox. Instead, two separate neural networks need to be trained for the model selection and parameter estimation procedures. Once a model is selected by the neural classifier, the parameters of that model are estimated using the techniques from Section 5.2.1.

### 5.2.5 Implementation details

Two key components of a standard dense neural network are its activation function and parameters $\boldsymbol{\gamma}$ (often referred to as weights and biases). The activation function introduces non-linearity into the model, allowing the network to learn and represent the complexities of the underlying data. The parameters determine the strength (weights) of the connection between two neurons (i.e., nodes of the neural network), and shift the input of the activation function (biases). Let $w_i \in \mathbb{R}$ represent the weights of each neuron $i$, $i = 1, \ldots, n$, and $b$ denote the bias; an activation function $\sigma$ transforms the weighted sum of the inputs of each neuron and the bias as

$\sigma\left(\sum_{i=1}^{n} w_i \boldsymbol{z}_i + b\right)$. Unless stated otherwise, the Rectified Linear Unit (ReLU) activation function is used when training the neural Bayes estimator, except in the final layer. Assuming $x = \sum_{i=1}^{n} w_i \boldsymbol{z}_i + b \in \mathbb{R}$, the ReLU function returns the same value if $x > 0$, and 0 otherwise, i.e, $\sigma(x) = \max\{x, 0\}$. For the final layer, different activation functions are used so that constraints on the parameters support are satisfied. In particular, for strictly positive parameters, the softplus activation function, i.e., $\sigma(x) = \log(1 + \exp\{x\})$, is used. Where the parameter is bounded in the interval $[a, b]$, a layer compression is used instead; such layer uses a logistic function $\sigma(x) = a + (b-a)/(1+\exp\{-kx\})$, that restricts input $x$ to be within $[a, b]$. The identity function, $\sigma(x) = x$, is used for parameters whose support is on the real line. For the model selection procedure, the softmax activation function is used in the final layer so that the output $\hat{\boldsymbol{p}}$ is a valid vector of probabilities, i.e. is non-negative and $\sum_{m=1}^{M} \hat{p}_m = 1$. More specifically, given a particular class $m$ and assuming that $\boldsymbol{x} = (x_1, \ldots, x_M)' \in \mathbb{R}^M$ is the output from the penultimate layer, the softmax activation function takes the form $\sigma(\boldsymbol{x})_m = \exp\{x_m\} / \sum_{m^*=1}^{M} \exp\{x_{m^*}\}$.

As mentioned in Section 5.2.1, neural Bayes estimators are trained by minimising the empirical Bayes risk with respect to the parameters $\boldsymbol{\gamma}$; more specifically, the training process involves learning the optimal parameters $\boldsymbol{\gamma}$ that map the data inputs $\boldsymbol{Z}$ to the parameter estimates $\hat{\boldsymbol{\theta}}$. This optimisation is done via back-propagation and the stochastic gradient descent (SGD) algorithm, where parameters are iteratively updated to minimise the objective function. Moreover, during training, two types of data sets are used: the training and validation sets. Both data sets are passed through the network and are refreshed after every epoch; here an epoch is defined as one full cycle through all the input data in the training set during the SGD process. In particular, the training set contains data used to train the model, which are used to update the parameters of the network. On the other hand, the validation set does not contribute to the parameter updates; instead, it is used to assess the abil-

ity of the model to generalise to new data, thus avoiding overfitting, and also used to define an early-stopping criterion for the algorithm. Finally, the performance of the trained estimator to model new data is further assessed with a test set, which is not to be used during training. All the computations are performed using the `NeuralEstimators` (Sainsbury-Dale et al., 2024a) and `Flux` (Innes, 2018) packages in `julia`; see `https://msainsburydale.github.io/NeuralEstimators.jl/dev/` and `https://fluxml.ai/Flux.jl/stable/`, respectively, for the full documentation.

## 5.3 Bivariate models of interest

From now on, we work in the bivariate setting. Specifically, we focus on the modelling of the joint tail behaviour of the random vector $\boldsymbol{Z} = (Z_1, Z_2)'$. We note, however, that the inference methodology can be applied to higher dimensions. We are particularly interested in bivariate models that are suitable for the modelling of both types of extremal dependence structures. More specifically, we focus on flexible models that allow interpolation between asymptotic dependence and independence, as well as on the weighted copula model (WCM) proposed by André et al. (2024). We aim to provide a tool for fast inference for a variety of bivariate models exhibiting complex dependence structures. Evaluation of their likelihood functions relies heavily on numerical integration and inversion of functions; this results in computationally costly likelihood-based inference procedures and may otherwise limit the use of these models in practice. Furthermore, likelihood-based inference for the WCM (which is a mixture model) is currently infeasible when one of its components is taken as one of the models able to interpolate between the two classes of extremal dependence. Section 5.3.1 reviews basics of copula modelling, while a background on extremal dependence measures is given in Section 5.3.2. The models of interest are introduced in Sections 5.3.3 and 5.3.4.

### 5.3.1 Copula modelling

The dependence between variables $Z_1$ and $Z_2$ can be characterised by means of copulas. Let $F_{Z_1}$ and $F_{Z_2}$ be the marginal cumulative distribution functions of variables $Z_1$ and $Z_2$, respectively, i.e., $Z_1 \sim F_{Z_1}$ and $Z_2 \sim F_{Z_2}$, and let $F_{Z_1,Z_2}$ denote their joint distribution function. According to Sklar's theorem (Sklar, 1959), the underlying copula $C : [0,1]^2 \to [0,1]$ of $\boldsymbol{Z} = (Z_1, Z_2)'$ can be obtained as

$$C(u_1, u_2) = f_{Z_1,Z_2}\left(F_{Z_1}^{-1}(u_1), F_{Z_2}^{-1}(u_2)\right), \quad (u_1, u_2)' \in [0,1]^2.$$

When $Z_1$ and $Z_2$ are continuous variables, the copula $C$ is unique and represents the joint distribution function of $\boldsymbol{U} = (U_1, U_2)'$, where $U_1 = F_{Z_1}(Z_1)$ and $U_2 = F_{Z_2}(Z_2)$ are Unif$(0,1)$ random variables. This result is useful since it is sufficient to marginally transform the data to a uniform scale through the probability integral transform and represent their joint behaviour via a copula model $C$. Finally, when it exists, the copula density $c(u_1, u_2)$ can be obtained by taking the second derivative of $C$ with respect to $u_1$ and $u_2$, as

$$c(u_1, u_2) = \frac{F_{Z_1,Z_2}\left(F_{Z_1}^{-1}(u_1), F_{Z_2}^{-1}(u_2)\right)}{f_{Z_1}\left(F_{Z_1}^{-1}(u_1)\right) f_{Z_2}\left(F_{Z_2}^{-1}(u_2)\right)}, \quad (u_1, u_2)' \in [0,1]^2,$$

where $f_{Z_i}$ is the probability density function of variable $Z_i$ for $i = 1, 2$. When $F_{Z_i}$ does not have an explicit form, it often needs to be computed through numerical integration in the original scale; this is also necessary for $f_{Z_i}$. Additionally, inversion techniques are required to compute $F_{Z_i}^{-1}$. All of these can require substantial computational resources.

### 5.3.2 Bivariate extremal dependence measures

When interest lies in the joint extremes of a bivariate random vector, a key element is to correctly identify its extremal dependence behaviour, i.e., whether large values in

different components of this vector are likely to occur simultaneously or not. Misidentifying the extremal dependence structure may indeed lead to inaccurate representations of the extremes, and incorrect extrapolations. Intuitively speaking, asymptotic dependence (AD) is present if the most extreme values of the components of the random vector $(Z_1, Z_2)'$ can occur together, and asymptotic independence (AI) is present otherwise. This extremal behaviour is often quantified through the tail dependence coefficient $\chi \in [0, 1]$ (see, e.g., Joe, 1997) and/or through the residual tail dependence coefficient $\eta \in (0, 1]$ (Ledford and Tawn, 1996). The coefficient $\chi$ can be obtained as $\chi = \lim_{y \to 1} \chi(y)$, when it exists, with

$$\chi(y) = \frac{\Pr(F_{Z_1}(Z_1) > y, F_{Z_2}(Z_2) > y)}{1 - y}, \quad y \in (0, 1). \tag{5.3.1}$$

The vector $(Z_1, Z_2)'$ is asymptotically independent if $\chi = 0$, and asymptotically dependent if $\chi > 0$. Given a function $\mathcal{L}$, that is slowly-varying at zero (i.e, for any $c > 0$, $\mathcal{L}(cx)/\mathcal{L}(x) \to 1$ as $x \to 0$), Ledford and Tawn (1996) assume the joint tail may be expressed as

$$\Pr(F_{Z_2}(Z_2) > y \mid F_{Z_1}(Z_1) > y) = \mathcal{L}(1 - y)(1 - y)^{1/\eta - 1}, \quad \eta \in (0, 1],\ y \to 1. \tag{5.3.2}$$

If $\eta = 1$ and $\mathcal{L}(1 - y) \not\to 0$ as $y \to 1$, then $(Z_1, Z_2)'$ is asymptotically dependent with $\chi = \lim_{y \to 1} \mathcal{L}(1 - y)$, otherwise it is asymptotically independent, with larger values of $\eta \in (0, 1]$ indicating stronger dependence. Similarly to $\chi$, a sub-asymptotic version of $\eta$ can be obtained from equation (5.3.2).

Taken together, $\chi > 0$ provides a summary measure of dependence within the AD class (with $\eta = 1$), while $\eta \leq 1$ provides a summary within the AI class (with $\chi = 0$). Available models in the multivariate extremes literature are often only suitable for one extremal dependence class. That is, aside from boundary points of the parameter space, traditional models either yield $\{\chi > 0, \eta = 1\}$ or $\{\chi = 0, \eta \leq 1\}$, but cannot span across

both classes. However, in recent years, more flexible methods which are able to capture both types of dependence have been proposed; see for instance Wadsworth et al. (2017), Huser and Wadsworth (2019) and Engelke et al. (2019). Each of these can be written as a random scale construction, which we introduce in the following section.

### 5.3.3 Random scale construction models

A range of bivariate dependence models can be constructed using a random scale representation based on 'radial' and 'angular' coordinates. More specifically, a random scale mixture vector $(Z_1, Z_2)'$ is constructed as follows

$$(Z_1, Z_2)' = R(V_1, V_2)', \quad R \perp\!\!\!\perp (V_1, V_2)', \tag{5.3.3}$$

where $R > 0$ is the 'radial' variable and is assumed to follow a non-degenerate distribution, and $(V_1, V_2)' \in \mathcal{V} \subseteq \mathbb{R}^2$ is the vector of 'angular' components. Different models can be obtained by varying the distributions and constructions of $R$ and $(V_1, V_2)'$. Depending on the precise specification, these may be able to interpolate between the two regimes of extremal dependence; see Engelke et al. (2019) for a detailed overview of the dependence properties arising from this construction. Interest lies in exploiting the copula $C$ of these flexible models. Moreover, since these aim at capturing the extremal dependence of the vector $(Z_1, Z_2)'$, non-extreme values are often censored to prevent their influence on the joint tail. We now present four particularly interesting bivariate models with construction (5.3.3), each of which can yield both $\{\chi > 0, \eta = 1\}$ and $\{\chi = 0, \eta \leq 1\}$ (depending on their parameter vectors), with the transition between these two regimes occurring at interior points of the parameter space.

**Model W.** Let $V \sim \text{Beta}(\alpha, \alpha)$, and let $R$ follow a generalised Pareto distribution (GPD) with scale parameter 1 and shape parameter $\xi \in \mathbb{R}$, i.e, $R \sim \text{GPD}(1, \xi)$.

Wadsworth et al. (2017) propose a model with

$$(V_1, V_2)' = \frac{(V, 1-V)'}{\max(V, 1-V)} \in \Sigma = \{\mathbf{v} \in \mathbb{R}_+^2 : \max(v, 1-v) = 1\}.$$

Given the model construction, $C(u_1, u_2)$ is assumed to hold only when $\|(Z_1, Z_2)'\|_\infty$ is large. This model is able to smoothly interpolate between the two classes of asymptotic dependence through the parameter $\xi$. For $\xi > 0$, $(Z_1, Z_2)'$ is asymptotically dependent with

$$\chi_W = \mathrm{E}\left(\min\left\{\frac{V_1^{1/\xi}}{\mathrm{E}\left(V_1^{1/\xi}\right)}, \frac{V_2^{1/\xi}}{\mathrm{E}\left(V_2^{1/\xi}\right)}\right\}\right) > 0 \quad \text{and} \quad \eta_W = 1.$$

If $\xi \leq 0$, then asymptotic independence is present with $\eta_W = (1-\xi)^{-1} \leq 1$ and $\chi_W = 0$. More details can be found in Wadsworth et al. (2017).

**Model HW.** Huser and Wadsworth (2019) detail a model construction providing flexible extremal dependence structures for spatial processes. However, they also propose a bivariate model able to transition between different types of dependence. In this model, both $R$ and the vector $(V_1, V_2)'$ are marginally Pareto distributed with different shape parameters. More specifically, the marginal cdfs of each variable are given as $F_R(r) = 1 - r^{-1/\delta}$, $r \geq 1$, and $F_{V_1}(v_1) = 1 - v_1^{-1/(1-\delta)}$, $v_1 \geq 1$, with $F_{V_2}(v_2)$ defined analogously, for $\delta \in (0, 1)$. Additionally, we assume here that $(V_1, V_2)'$ follows a bivariate Gaussian copula with correlation parameter $\omega \in (-1, 1)$.

This model is able to interpolate between the two classes of extremal dependence through the parameter $\delta$; when $\delta > 1/2$, the tail of $R$ is heavier than $(V_1, V_2)'$ and so $(Z_1, Z_2)'$ is asymptotically dependent with

$$\chi_{HW} = \mathrm{E}\left(\min\left\{\frac{V_1^{1/\delta}}{\mathrm{E}\left(V_1^{1/\delta}\right)}, \frac{V_2^{1/\delta}}{\mathrm{E}\left(V_2^{1/\delta}\right)}\right\}\right) > 0 \quad \text{and} \quad \eta_{HW} = 1.$$

When $\delta \leq 1/2$, $(Z_1, Z_2)'$ is asymptotically independent with

$$\eta_{HW} = \begin{cases} 1, & \text{if } \delta = 1/2, \\ \delta/(1-\delta), & \text{if } \eta_V/(1+\eta_V) < \delta < 1/2, \\ \eta_V, & \text{if } \delta \leq \eta_V/(1+\eta_V), \end{cases}$$

where $\eta_V < 1$ is the residual tail dependence coefficient (5.3.2) of vector $(V_1, V_2)'$. Although $\eta_{HW} = 1$ for $\delta = 1/2$, the variables $Z_1$ and $Z_2$ are asymptotically independent since $\mathcal{L}(1-y) \to 0$ as $y \to 1$ in equation (5.3.2). More details can be found in Huser and Wadsworth (2019).

The final two models were proposed by Engelke et al. (2019). Similarly to the model introduced by Wadsworth et al. (2017), let $V \sim \text{Beta}(\alpha, \alpha)$, $\alpha > 0$.

**Model E1.** For the first model, $R$ follows a Weibull distribution with distribution function $F_R(r) = 1 - \exp\{-r^\beta\}$, $r$, $\beta > 0$, and the angular components are constructed as follows

$$(V_1, V_2)' = \frac{(V, 1-V)'}{\nu(V, 1-V)'},$$

where $\nu(V, 1-V) = \mu \max(V, 1-V) + (1-\mu)\min(V, 1-V)$ and $\mu \geq 1/2$. For this model, the extremal dependence is controlled by $\mu$; when $\mu \leq 1$, $Z_1$ and $Z_2$ are asymptotically independent with $\chi_{E1} = 0$ and $\eta_{E1} = \mu^\beta$. If $\mu > 1$, they are asymptotically dependent with

$$\chi_{E1} = \frac{2(\mu-1)}{2\mu - 1} \quad \text{and} \quad \eta_{E1} = 1.$$

**Model E2.** For the second model, $R \sim \text{GPD}(1, \xi)$ with $\xi \in \mathbb{R}$ as in Wadsworth et al. (2017). The angular components $V_1$ and $V_2$ are now independent of each other and distributed as $V$, that is $V_1, V_2 \sim \text{Beta}(\alpha, \alpha)$. The extremal dependence of this model is

determined by $\xi$. In particular, if $\xi > 0$,

$$\chi_{E2} = \frac{\mathrm{E}\left(\min\{V_1, V_2\}^{1/\xi}\right)}{\mathrm{E}\left(V^{1/\xi}\right)} \quad \text{and} \quad \eta_{E2} = 1,$$

and $(Z_1, Z_2)'$ are asymptotically dependent. When $\xi \leq 0$, the variables are asymptotically independent with

$$\chi_{E2} = 0 \quad \text{and} \quad \eta_{E2} = \begin{cases} 1, & \text{if } \xi = 0, \\ \frac{1-\xi\alpha}{1-2\xi\alpha}, & \text{if } \xi < 0. \end{cases}$$

### 5.3.4 Weighted copula model

André et al. (2024) propose a model that is able to accurately represent both the body and tail regions of a data set, while ensuring a smooth transition between them. Let $c_t$ denote the density of the copula tailored to the tail and $c_b$ denote the density of the copula tailored to the body. For $(x_1, x_2)' \in [0,1]^2$, the density of the proposed model is given by

$$h(x_1, x_2; \boldsymbol{\theta}) = \frac{\pi(x_1, x_2; \kappa) c_t(x_1, x_2; \boldsymbol{\lambda}_t) + [1 - \pi(x_1, x_2; \kappa)] c_b(x_1, x_2; \boldsymbol{\lambda}_b)}{K(\boldsymbol{\theta})}, \tag{5.3.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\lambda}_t', \boldsymbol{\lambda}_b', \kappa)'$ is the vector of model parameters, $K(\boldsymbol{\theta})$ a normalising constant, and $\pi(x_1, x_2; \kappa) : (0,1)^2 \to (0,1)$ is a dynamic weighting function. Note that model (5.3.4) does not have uniform margins in general, which makes likelihood-based inference difficult. The weighting function depends on the data and is specified such that it is increasing in $x_1$ and $x_2$ for a fixed value of $\kappa$; in particular, more weight is given to $c_b$ for small values of $(x_1, x_2)'$, and more weight is given to $c_t$ for large values of $(x_1, x_2)'$. Similarly to the models introduced in Section 5.3.3, the interest is in exploiting the copula $C(u_1, u_2; \boldsymbol{\theta}) = H\left(H_{X_1}^{-1}(u_1), H_{X_2}^{-1}(u_2); \boldsymbol{\theta}\right)$ of density (5.3.4), where $H_{X_1}$ and $H_{X_2}$ are the marginal cdfs of $H$. As shown in André et al. (2024), the model shows

some interesting features regarding extremal dependence, with $c_b$ potentially having an influence on $\chi$ depending on the weighting function chosen. More details can be found in André et al. (2024).

## 5.4 Simulation studies

Several simulation studies are performed to assess the trained neural Bayes estimators in different scenarios. We present selected studies in this section, and give the remaining ones in the Supplementary Material. All general settings and priors are outlined in Section 5.4.1. We then start by studying the performance of the NBE in estimating the model parameters in Section 5.4.2. In Section 5.4.3, the efficacy of the neural classifier for model selection is assessed, in which we focus on the four models from Section 5.3.3. In Section 5.4.4 we investigate the performance of the NBEs trained for model selection and parameter estimation in misspecified scenarios.

### 5.4.1 General settings

In all simulations, the sample size $n$ and, when applicable, the censoring level $\tau$ are assumed random and independent realisations of variables distributed, respectively, as $N \sim \text{Unif}(\{100, 101, \ldots, 1500\})$ and $\text{T} \sim \text{Unif}(0.5, 0.99)$. Each parameter of the four models mentioned in Section 5.3.3 are also assumed independent a priori and uniformly distributed. In particular, for Model W (Wadsworth et al., 2017) we take $\alpha \sim \text{Unif}(0.2, 15)$ for the parameter of the Beta distribution, and $\xi \sim \text{Unif}(-2, 1)$ for the shape parameter of the GPD. For Model HW (Huser and Wadsworth, 2019), we take $\delta \sim \text{Unif}(0, 1)$ and $\omega \in \text{Unif}(-1, 1)$. For Model E1 (Engelke et al., 2019), we take $\alpha \sim \text{Unif}(0.2, 15)$ for the parameter of the Beta distribution, $\beta \sim \text{Unif}(0, 15)$ for the Weibull parameter, and $\mu \sim \text{Unif}(0.5, 4)$. For Model E2 (Engelke et al., 2019), we assume that $\alpha \sim \text{Unif}(0.2, 15)$ and $\xi \sim \text{Unif}(-2, 1)$. For the weighted copula model

from equation (5.3.4), we take the weighting function to be $\pi(x_1, x_2; \kappa) = (x_1 x_2)^{\exp\{\kappa\}}$, with $\kappa \sim \text{Unif}(-3.51, 1.95)$ as the prior for the weighting function parameter; based on previous analysis of this model, $\kappa \in (-3.51, 1.95)$ ensures that there is a good representation of both copula components, $c_b$ and $c_t$, in each drawn sample.

For each step, training and validation sets for the vector of parameters and corresponding data realisations are generated. More specifically, we use $|\Upsilon|_{\text{train}} = K$ and $|\Upsilon|_{\text{val}} = K/5$ with $K = 100\,000$ (recall equation (5.2.2)) for the training and validation parameter sets, respectively. The number of layers assumed for each neural network $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ (recall equation (5.2.3)) and its parameters are determined experimentally, and in order to reduce the computational intensity, we adopt the 'simulation-on-the-fly' technique where the train and validation sets are refreshed every epoch; see Sainsbury-Dale et al. (2024a) for more details. When censoring is applicable, we adopt the scheme where the observations for which the maximum value is less than $\tau$ are censored. All the simulations are performed using a high-end computing cluster with a NVidia V100 32 GB GPU hardware with 192 GB of memory; see `https://lancaster-hec.readthedocs.io/en/latest/` for more details (last accessed on 20/11/2024). For reproducibility and to make our methodology available to a broad readership, all the trained NBEs for parameter estimation and neural classifiers for model selection are available on `https://github.com/lidiamandre/NBE_classifier_depmodels`.

### 5.4.2 Parameter estimation

We take the mean absolute loss function, $L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$, which targets the marginal posterior medians, and set $J = 5$ in equation (5.2.2). This means that we have 5 data set realisations (and censoring level values when applicable) for each parameter vector $\boldsymbol{\theta}^{(k)}$ $(k = 1, \ldots, K)$. Moreover, the training step finishes if the estimated Bayes risk computed based on the validation set has not decreased in 5 consecutive epochs. We first demonstrate the performance of NBEs for uncensored data from the weighted

copula model from Section 5.3.4, and for censored data from one of the random scale mixture models mentioned in Section 5.3.3. The neural network architecture used for parameter estimation is given in Table C.2.1 of Section C.2 of the Supplementary Material. Finally, the assessment of the NBEs is done on a test parameter set with 1000 parameter vectors $\boldsymbol{\theta}$, and corresponding test data realisations, each of size $n = 1000$.

In the first simulation study, we consider the weighed copula model from equation (5.3.4). In particular, we take $c_b$ to be the Gaussian copula density with correlation parameter $\lambda_b \equiv \rho \in (-1, 1)$, for which we take $\rho \sim \text{Unif}(-1, 1)$, and $c_t$ to be the copula density of Model E1 with $\boldsymbol{\lambda}_t = (\alpha, \beta, \mu)'$. This is a configuration for which likelihood-based inference is simply infeasible, owing to the nesting of two complex models requiring numerical integrals and inversion of functions. Figure 5.4.1 shows the results; the true values for each parameter are compared with their estimated values given by the trained NBE. It can be seen that there is a bit of variability, in particular for the parameters $\alpha$ and $\beta$; we note that these are the parameters with the largest prior range. Analysing the most interesting parameters, we can see that the NBE generally estimates the weighting parameter $\kappa$ well, indicating that it is able to distinguish between the body and tail components of the copula. In addition, estimates of $\mu$ are quite accurate; this means that the two regimes of extremal dependence are well captured. In particular, from the asymptotically dependent samples (i.e., $\mu > 1$), 98.28% were estimated to exhibit AD, whereas 86.15% of the data sets showing asymptotic independence (i.e., $\mu \leq 1$) were estimated to be AI. Finally, estimates of $\rho$ exhibit the best results. For negatively correlated data sets (i.e., $\rho < 0$), 96.06% were estimated to be negatively correlated, whereas 92.28% of the positively correlated data sets were correctly identified.

It is important to assess the uncertainty of the NBE; we consider two different approaches to this task. For the first, a non-parametric bootstrap procedure is adopted, where we generate $B = 400$ bootstrap samples from which model parameters are re-

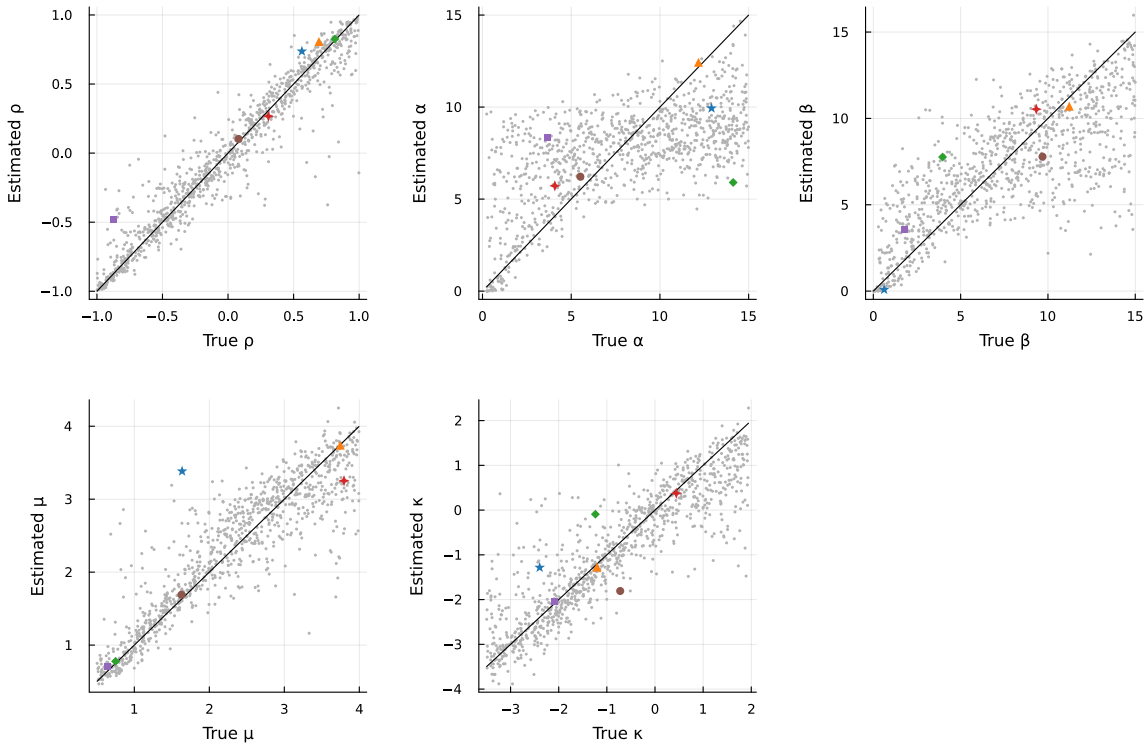



Figure 5.4.1: Assessment of the NBE when $c_b$ is the Gaussian copula with parameter $\rho$, $c_t$ is Model E1 with parameters $\boldsymbol{\lambda}_t = (\alpha, \beta, \mu)'$, and with weighting function $\pi(x_1, x_2; \kappa) = (x_1 x_2)^{\exp\{\kappa\}}$, $x_1, x_2 \in (0, 1)$ for a sample size of $n = 1000$. The points highlighted in different shapes and colours refer to parameter configurations used for further diagnostics (see Figure 5.4.2).

estimated and 95% confidence intervals are obtained. This is done for each parameter configuration and data from the test set. For the second approach, we train an additional estimator, now under the quantile loss function targeting jointly a low and a high posterior quantile; precisely, given a probability level $q \in (0, 1)$, this loss function is defined as $L_q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \sum_{k=1}^{p} (\hat{\theta}_k - \theta_k)(\mathbb{1}_{(\hat{\theta}_k > \theta_k)} - q)$. With this estimator (herein referred to as the neural interval estimator), we are able to approximate marginal 95% central credible intervals by training for $q = \{0.025, 0.975\}$. We evaluate the performance of each approach by computing coverage probabilities, with the results presented in Table 5.4.1. The bootstrap procedure leads to lower than nominal coverage rates, which can be linked to the quality of estimates shown in Figure 5.4.1: the best coverage is obtained for the parameters showing a better correspondence between true and estimated

values, such as $\rho$ or $\mu$. The parameters $\alpha$ and $\beta$ exhibit more bias in their estimation, which is reflected in worse coverage rates. On the other hand, the neural interval estimator is much better calibrated. In the context of NBE, bootstrap-based confidence intervals have been used to account for the uncertainty in the estimation; however, to the best of our knowledge, their coverage rates have not been explored.

Table 5.4.1: Coverage probability and average length of the 95% uncertainty intervals obtained via a non-parametric bootstrap procedure and via the neural interval estimator averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Bootstrap procedure | | Interval estimator | |
|---|---|---|---|---|
| | Coverage | Length | Coverage | Length |
| $\rho$ | 0.80 | 0.27 | 0.97 | 0.57 |
| $\alpha$ | 0.36 | 3.12 | 0.97 | 11.84 |
| $\beta$ | 0.53 | 3.42 | 0.96 | 10.32 |
| $\mu$ | 0.67 | 0.63 | 0.97 | 1.67 |
| $\kappa$ | 0.70 | 1.12 | 0.97 | 2.54 |

We explore whether apparent bias in parameter estimates leads to bias in dependence quantities of interest. To do so, we consider parameter and data sets for which the NBE severely under- or over-estimates at least one parameter and compare the empirical and model-based dependence measure $\chi(y)$ from equation (5.3.1) at several thresholds $y \in [0.01, 0.99]$. For this model configuration, model-based $\chi(y)$ are estimated using a Monte Carlo approximation with 500 000 samples. The parameter configurations considered are highlighted with different shapes and colours in Figure 5.4.1, and the results for $\chi(y)$ are shown in Figure 5.4.2. Despite some parameters being massively under/over-estimated, the dependence structure is still well captured overall, apart from configuration $\boldsymbol{\theta}^{(1)}$ (in blue) for which $\mu$ is over-estimated. Given that this is the parameter that controls the dependence structure of the data, directly determining the value of $\chi_{E1}$, this result is not surprising. Finally, from the $B = 400$ bootstrap samples obtained previously, we compute coverage probabilities of 95% confidence intervals for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$; these are shown in Table 5.4.2. As can be seen, the true value for $\chi(y)$ is within the confidence intervals in more than 79% of the time,

suggesting that this derived feature of the models is well estimated and well calibrated even when the individual parameter estimates obtained by the NBE exhibit bias and display poor coverage properties.

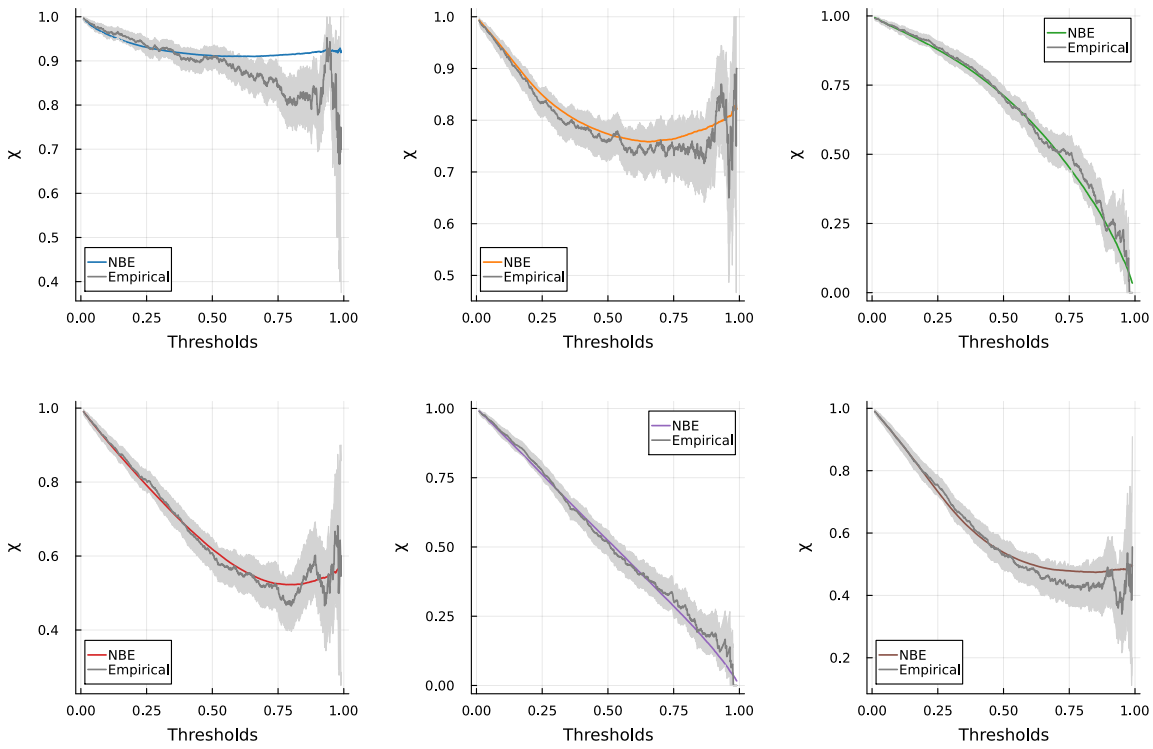



Figure 5.4.2: Empirical $\chi(y)$ (in grey) and model-based $\chi(y)$ for the fitted weighted copula model with parameter configurations: $\hat{\boldsymbol{\theta}}^{(1)}$ (in blue), $\hat{\boldsymbol{\theta}}^{(2)}$ (in orange), $\hat{\boldsymbol{\theta}}^{(50)}$ (in green), $\hat{\boldsymbol{\theta}}^{(78)}$ (in red), $\hat{\boldsymbol{\theta}}^{(100)}$ (in purple) and $\hat{\boldsymbol{\theta}}^{(500)}$ (in brown), for $y \in [0.01, 0.99]$. The 95% confidence bands, representing uncertainty in the empirical estimates, were obtained by boostrapping.

Table 5.4.2: Coverage probability and average length of the 95% confidence intervals for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$ obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.50)$ | 0.86 | 0.09 |
| $\chi(0.80)$ | 0.82 | 0.11 |
| $\chi(0.95)$ | 0.79 | 0.12 |

For the second simulation study, we consider Model W with the priors mentioned in Section 5.4.1; since this model is suitable for censored data, we consider a variable

censoring level $\tau$ drawn from the prior given in Section 5.4.1. Figure 5.4.3 shows the performance of the trained NBE; estimates are quite accurate overall, but with some bias for large $\alpha$. We observe from the right panel of Figure 5.4.3 that the two regimes of extremal dependence are well captured; in particular, from the samples exhibiting asymptotic dependence (i.e., $\xi > 0$), 94.05% were estimated to be AD, while 98.19% of the asymptotically independent samples (i.e., $\xi \leq 0$) were correctly identified as AI data sets.

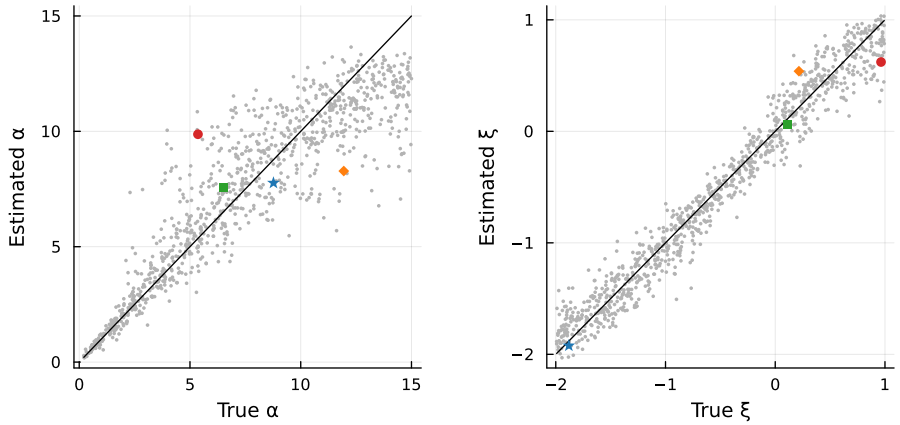



Figure 5.4.3: Assessment of the NBE for Model W with parameters $\boldsymbol{\theta} = (\alpha, \xi)'$ for a sample size of $n = 1000$. The points highlighted in different shapes and colours refer to parameter configurations used for further diagnostics (see Figure 5.4.4).

As with the first study, we assess the uncertainty of the NBE through non-parametric bootstrap and by training a neural interval estimator. Coverage probabilities of the 95% uncertainty intervals and their average length are shown in Table 5.4.3. The coverage probabilities are once again better using the trained neural interval estimator, but at the cost of wider intervals on average. The extremal dependence measure $\chi(y)$ from equation (5.3.1) at several thresholds $y \in [\tau, 0.99]$, where $\tau$ is the censoring level, is computed as a further diagnostic for the four parameter configurations highlighted in Figure 5.4.3; the results are given in Figure 5.4.4. Despite the under/over-estimation by the NBE, e.g., shown by the vector of parameters $\boldsymbol{\theta}^{(980)}$ given in red, the extremal dependence behaviour is well captured. This is further supported by the coverage

probabilities of 95% confidence intervals for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$, which are achieved with new data sets for 1000 parameter configurations, each generated with a fixed censoring level $\tau = 0.8$; the results are shown in Table 5.4.4. As a final diagnostic, we compare the joint behaviour along different rays. Transforming the model variables, $U_1$ and $U_2$, into standard exponentially distributed variables, $X_1^E$ and $X_2^E$, and given ray $w \in [0, 1]$, the joint probability $\Pr(X_1^E > wy, X_2^E > (1-w)y)$ is compared with its empirical counterpart for two different rays $w = \{0.3, 0.8\}$ and $y \in [\tau, 0.99]$. The results for three parameter configurations exhibiting asymptotic independence are given in Figure 5.4.5. It is visible that there is a very good agreement between the estimated and empirical joint probabilities for both rays considered, supporting the efficacy of the trained NBE.

Table 5.4.3: Coverage probability and average length of the 95% uncertainty intervals obtained via a non-parametric bootstrap procedure and via the neural interval estimator averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Bootstrap procedure | | Interval estimator | |
|---|---|---|---|---|
| | Coverage | Length | Coverage | Length |
| $\alpha$ | 0.70 | 3.05 | 0.96 | 6.68 |
| $\xi$ | 0.78 | 0.41 | 0.98 | 0.81 |

Table 5.4.4: Coverage probability and average length of the 95% confidence intervals for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$ obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.80)$ | 0.91 | 0.06 |
| $\chi(0.95)$ | 0.89 | 0.09 |
| $\chi(0.99)$ | 0.88 | 0.09 |

**Comparison with censored maximum likelihood estimation**

Finally, we compare estimates obtained by the NBE and the censored maximum likelihood estimation (CMLE) procedure. We first compute the CMLE for the four parameter configurations highlighted in Figure 5.4.3, and the corresponding estimates for

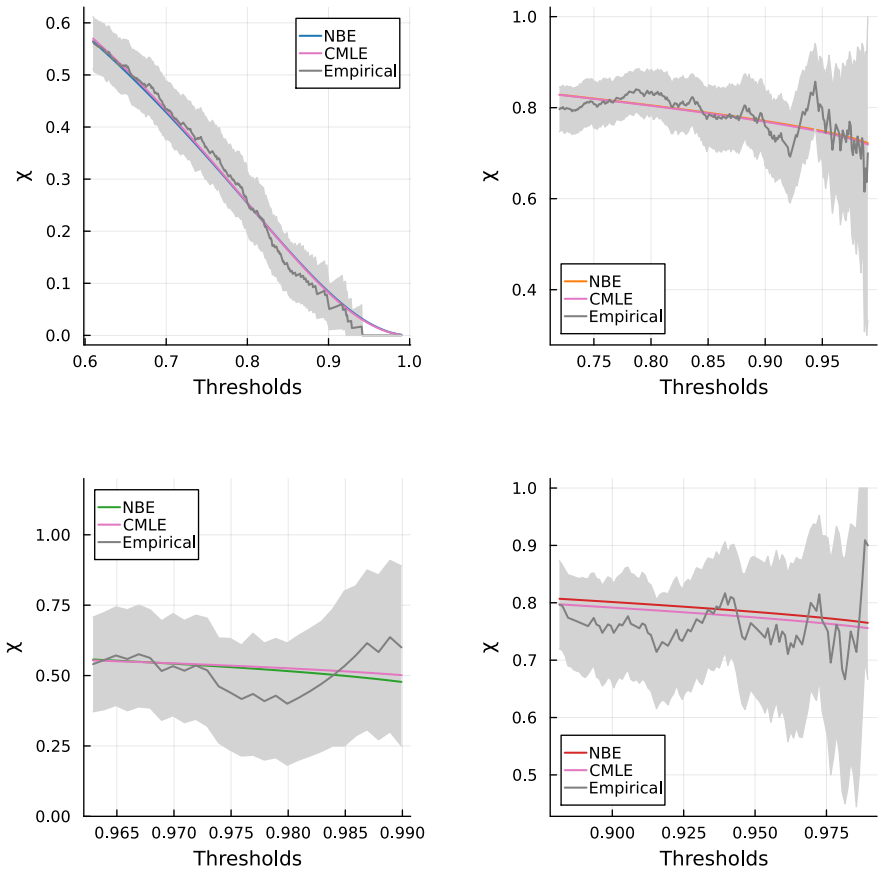



Figure 5.4.4: Empirical $\chi(y)$ (in grey) and model-based $\chi(y)$ estimated via the NBE with parameter configurations: $\hat{\boldsymbol{\theta}}^{(1)}$ (in blue), $\hat{\boldsymbol{\theta}}^{(32)}$ (in orange), $\hat{\boldsymbol{\theta}}^{(403)}$ (in green) and $\hat{\boldsymbol{\theta}}^{(980)}$ (in red) for $y \in [\tau, 0.99]$, where $\tau$ is the corresponding censoring level. A comparison with model $\chi(y)$ estimated via censored maximum likelihood inference is given in pink. The 95% confidence bands, representing uncertainty in empirical estimates, were obtained by boostrapping.

$\chi(y)$ for $y \in [\tau, 0.99]$. The results are shown by the pink lines in Figure 5.4.4. As it can be seen, almost identical results are obtained when using censored maximum likelihood. Then, with the assigned prior distributions, we generate 5 different parameter vectors $\boldsymbol{\theta} = (\alpha, \xi)'$, censoring levels $\tau$ and the corresponding data sets, each of which with a sample size of 1000. Each data set is simulated 100 times. Figure 5.4.6 shows the comparison between the two estimators for two parameter vectors: $\boldsymbol{\theta}_1 = (2.94, 0.11)'$ with censoring level $\tau_1 = 0.79$ and $\boldsymbol{\theta}_2 = (8.87, -1.97)'$ with $\tau_2 = 0.60$ (rounded to 2 decimal places). The remaining three cases are given in Section C.2.2 of the Supplementary

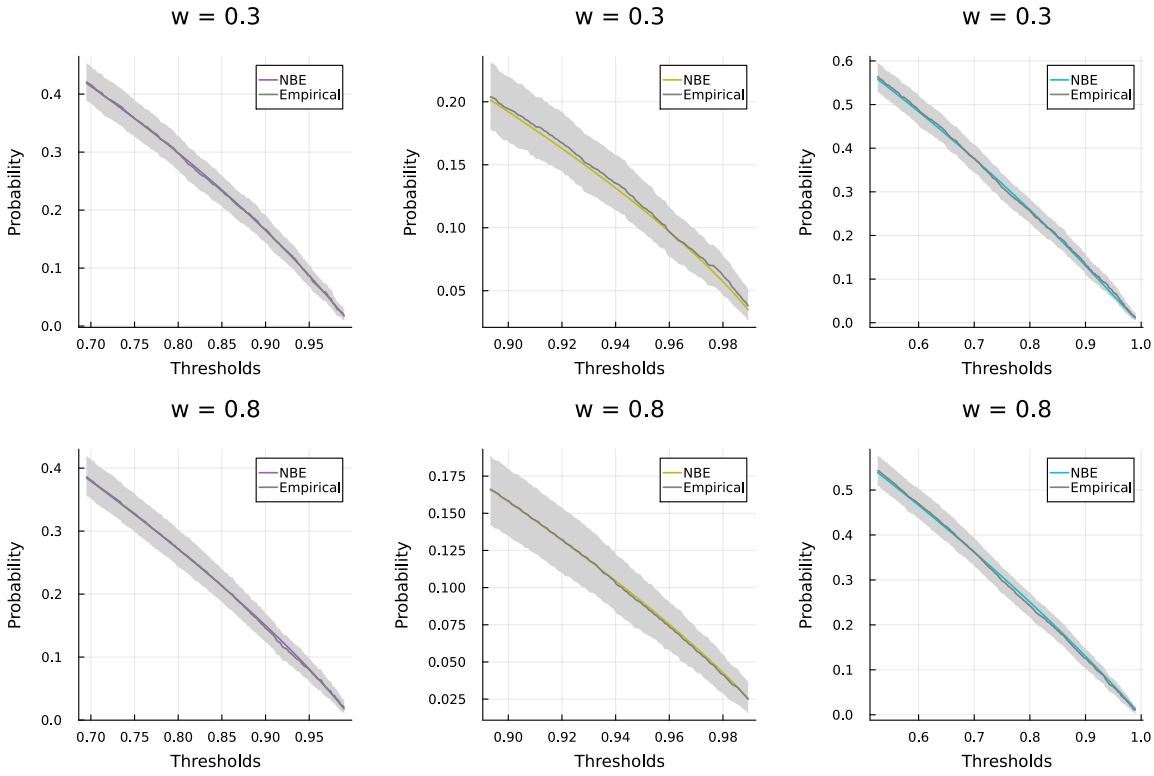
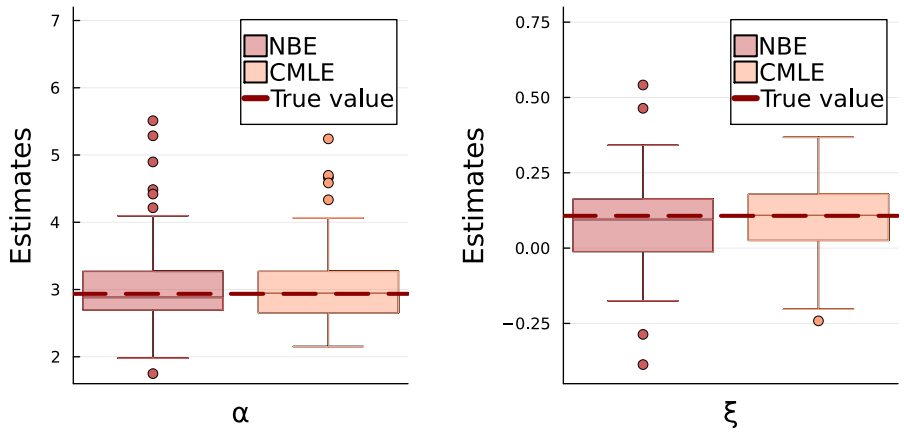


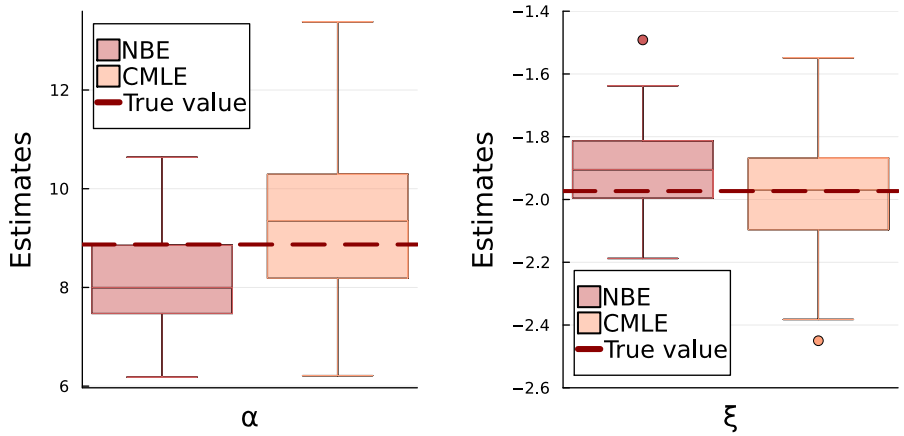

Figure 5.4.5: Empirical (in grey) and model-based estimates of $\Pr(X_1^E > wy, X_2^E > (1-w)y)$ estimated via the NBE with parameter configurations: $\hat{\boldsymbol{\theta}}^{(181)}$ (in purple), $\hat{\boldsymbol{\theta}}^{(272)}$ (in yellow) and $\hat{\boldsymbol{\theta}}^{(983)}$ (in cyan) for $y \in [\tau, 0.99]$, where $\tau$ is the corresponding censoring level. The 95% confidence bands, representing uncertainty in empirical estimates, were obtained by boostrapping.

Material. Estimates given by the NBE are more biased, and generally less variable than the CMLE depending on the case. Despite this small bias, estimates provided by the NBE are still relatively accurate whilst being much faster to obtain. In particular, the NBE took on average 0.676 seconds to evaluate, whereas the censored MLE took 92.611 seconds on average, thus the NBE is about 137 times faster.

To assess the effect of assuming the sample size and/or censoring level to be unknown, we perform a similar study considering fixed censoring level ($\tau = 0.8$) with fixed ($n = 1000$) and variable sample size. The results are presented in Section C.2.2 of the Supplementary Material. Although the overall findings are similar, slightly higher coverage probabilities, and less wider intervals, are obtained in the cases with less unknown quantities.




(a)




(b)

Figure 5.4.6: Comparison between parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\xi})'$ given by CMLE (orange) and by NBE (red) for 100 samples with $n = 1000$. The true parameters are given by the red line. (a) $\boldsymbol{\theta}_1 = (2.94, 0.11)'$ with censoring level $\tau_1 = 0.79$ and (b) $\boldsymbol{\theta}_2 = (8.87, -1.97)'$ with censoring level $\tau_2 = 0.60$.

### 5.4.3 Model selection

Here we investigate the performance of the neural classifier for model selection, outlined in Section 5.2.4. We set $J = 1$ from equation (5.2.2) for both training and validation data sets. For each sampled class index $\zeta^{(k)}$ $(k = 1, \ldots, K)$, data set $\boldsymbol{Z}^{(k)}$ is generated with a random sample size $n$, a random censoring level $\tau$, and a random vector of model parameters $\boldsymbol{\theta}$ using the priors defined in Section 5.4.1. The training finishes if the Bayes risk using the validation set has not decreased in 10 consecutive epochs.

We demonstrate the performance of the model selection classifier for the four models mentioned in Section 5.3.3, and since their focus is in the modelling of the joint tail behaviour, the data are treated as censored. The neural network architecture used for the classification problem is given in Table C.3.1 in Section C.3 from the Supplementary Material. When $M = 2$, all pair combinations, out of the four models, are compared, whereas all models are considered simultaneously when $M = 4$. Finally, in order to assess the neural classifier, we generate a test parameter set with 1000 values for model index $\zeta$, and a test data set of size $n = 1000$. The class which has the highest probability as output is the assigned model for the data set. We also compare the classifier with model selection via the Bayesian information criterion (BIC). This requires optimisation of the CMLE, which is time-consuming. However, we wish to explore how well the neural classifiers perform in comparison to established methods.

Figure 5.4.7 shows the assessment of the classifier when $M = 2$; the left bar plot represents the true counts of data sets generated from each model, whereas the middle and right bar plots represent the percentage of correctly identified models after applying the neural classifier and via BIC, respectively. For this study, the proportions of correctly identified models with the neural classifier are always above 71%, with an average of 87%, whilst through BIC these are above 63% with an average of 86%. Aside from the cases where the choice is between Model W and Model E2 or between Model HW and Model E2, the neural classifier and the likelihood-based BIC do a similar job, with the selection procedure through neural classifier being quicker than using the BIC as no likelihood evaluation is needed. Finally, the uncertainty of the neural classifiers is assessed through a bootstrap procedure. Given $B = 400$ samples, the 95% confidence intervals for the proportions of correctly identified data sets are computed for each model and classifier; these results are reported in the middle bar plots of Figure 5.4.7.

Results for the case when $M = 4$ are displayed in Figure 5.4.8. As before the left bar plot represents the true counts of data sets generated from each model, whilst the middle

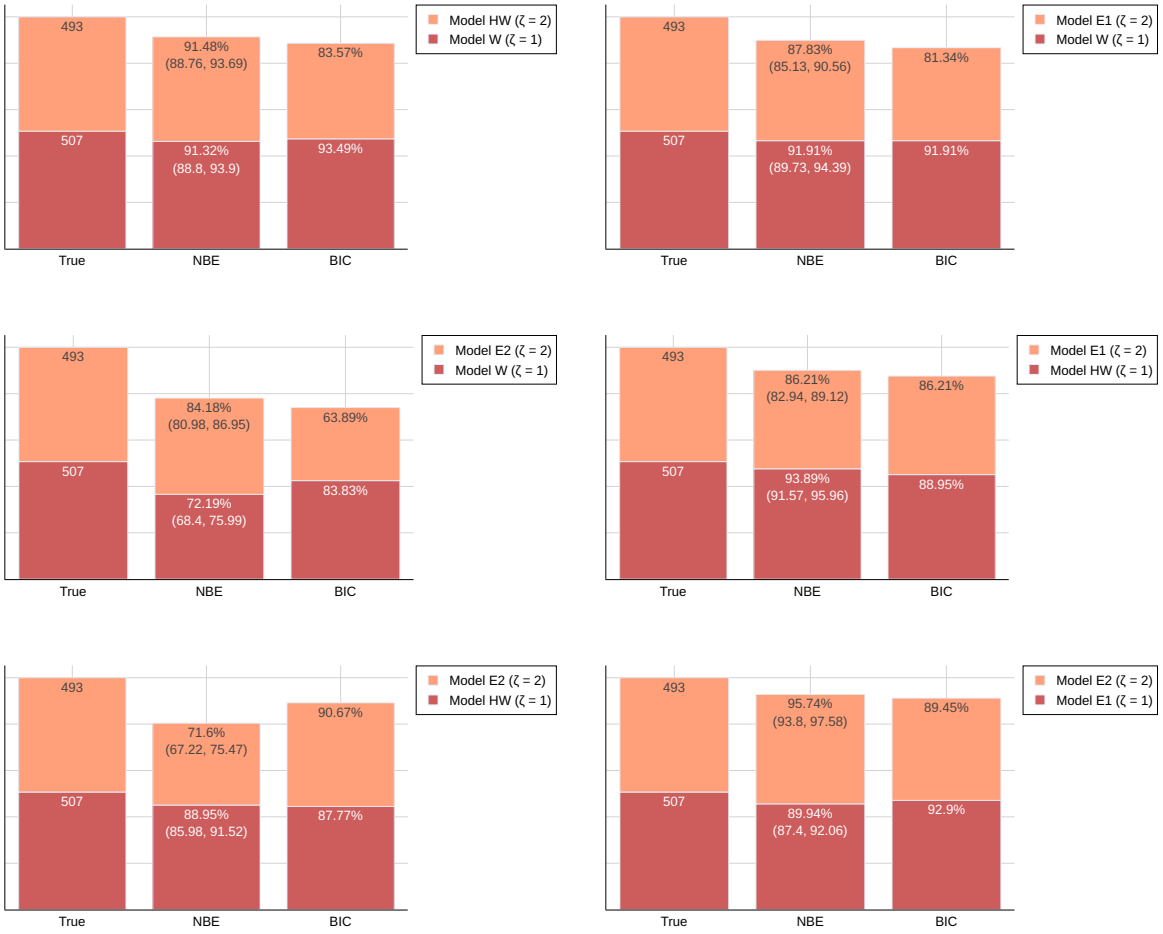



Figure 5.4.7: Proportion (in %) of correctly identified data sets when $M = 2$ through the neural classifier (middle) and through BIC (right) for the six pairs of models considered. The true counts of data sets generated from class index $\zeta = 1$ (red) and from class index $\zeta = 2$ (orange) are given in the left bar plot.

and right bar plots show the percentage of correctly identified models through neural classification and through the BIC, respectively. Similarly to the previous case, the uncertainty of the neural classifier is assessed via a bootstrap procedure with $B = 400$ bootstrap samples; the 95% confidence intervals for the proportions obtained by the neural classifier are shown in the middle bar plot. The resulting proportions achieved through the neural classifier are all above 68% with an average of 78%, whereas via the BIC they are above 61% with and average of 72%. The proportions obtained via the BIC are lower than the ones obtained with the neural classifier apart from Model

W, which indicates that the neural classifier correctly identifies the data sets more frequently than the BIC, and in a quicker way.

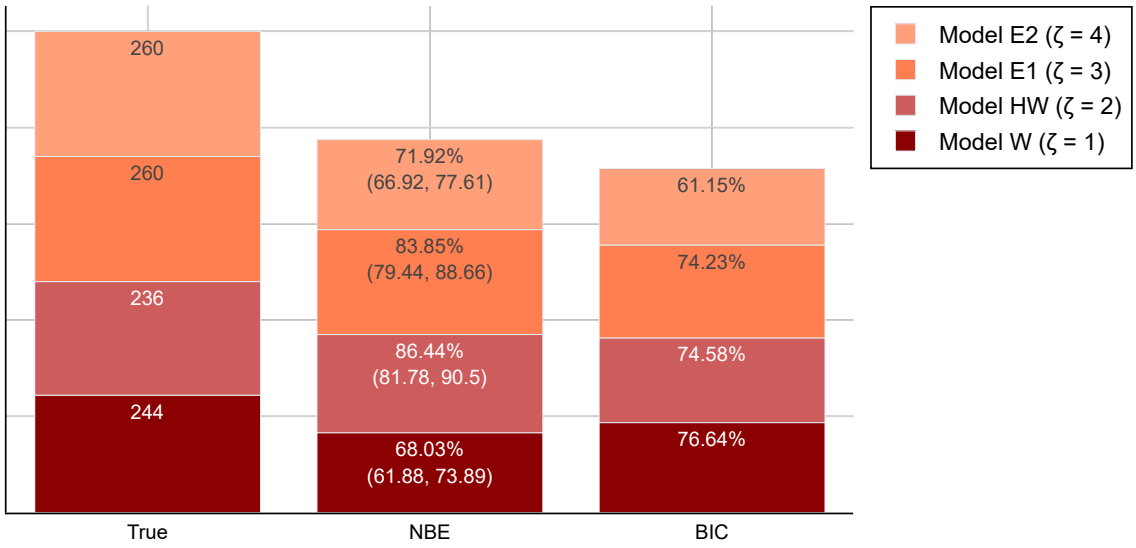



Figure 5.4.8: Proportion (in %) of correctly identified data sets when $M = 4$ through the neural Bayes estimators (middle) and through BIC (right). The true counts of data sets generated from class indices $\zeta = 1$ (red), $\zeta = 2$ (light red), $\zeta = 3$ (orange) and $\zeta = 4$ (light orange) are given in the left bar plot.

### 5.4.4 Misspecified scenarios

We now examine the performance of the model selection and subsequent parameter estimation in a misspecified scenario, where the underlying data do not come from one of the models considered. We consider two different situations and, in both cases, model selection is first done using the neural classifier, followed by estimation of the model parameters using the NBE trained for the selected models. A comparison with classical model selection and inference tools is also provided.

For the first study, we generate 100 samples, each with $n = 1000$, from a Gaussian copula with correlation parameter $\rho = 0.5$, and consider a censoring level of $\tau = 0.75$. The proportion of times each model was selected through the neural classifier and BIC is given on the left of Table 5.4.5, and the proportion of AD and AI samples

identified by the NBE and CMLE is on the right. Model HW is the most suitable according to either selection procedure, with the classifier selecting it 88% of the time and the BIC 69%. This is to be expected given the nature of the underlying data and the model assumptions, i.e., we are assuming that $(V_1, V_2)'$ follows a Gaussian copula. Both inference procedures are able to capture the correct extremal dependence structure. In particular, according to the NBE and CMLE, 97% and 96% of the samples, respectively, exhibit AI, which is in agreement with the underlying data since Gaussian data are known to be AI. According to the NBE, the two samples fitted with Model E1 exhibit AD, with estimated values $\mu = 1.127$ and $\mu = 1.247$. For the CMLE, three of the samples fitted with Model W are AD; for these, the estimates for $\xi$ are close to 0, specifically $\xi = \{0.003, 0.021, 0.043\}$. As a further diagnostic, we compute $\chi(y)$ estimates at three levels $y = \{0.80, 0.95, 0.99\}$ for the selected models, and compare with the true $\chi(y)$ at each level. The results are given in Figure 5.4.9. The estimates obtained with either inference method are concentrated around the true value, indicating that both the NBE and CMLE are able to capture the true extremal dependence structure. In Section C.4 of the Supplementary Material we describe an individual example, where the proposed toolbox for model selection and inference is presented in detail. The results obtained through this individual example agree with the findings of the repeated study.

Table 5.4.5: Proportion of times each model was selected through the neural classifier and through BIC (left), and proportion of AD and AI samples identified by the NBE and CMLE (right). All the values are rounded up to 2 decimal places.

| Model | Neural classifier | BIC |
|---|---|---|
| Model W | 0.02 | 0.30 |
| Model HW | 0.88 | 0.69 |
| Model E1 | 0.02 | 0.00 |
| Model E2 | 0.08 | 0.01 |

| Method | AD | AI |
|---|---|---|
| NBE | 0.02 | 0.98 |
| CMLE | 0.03 | 0.97 |

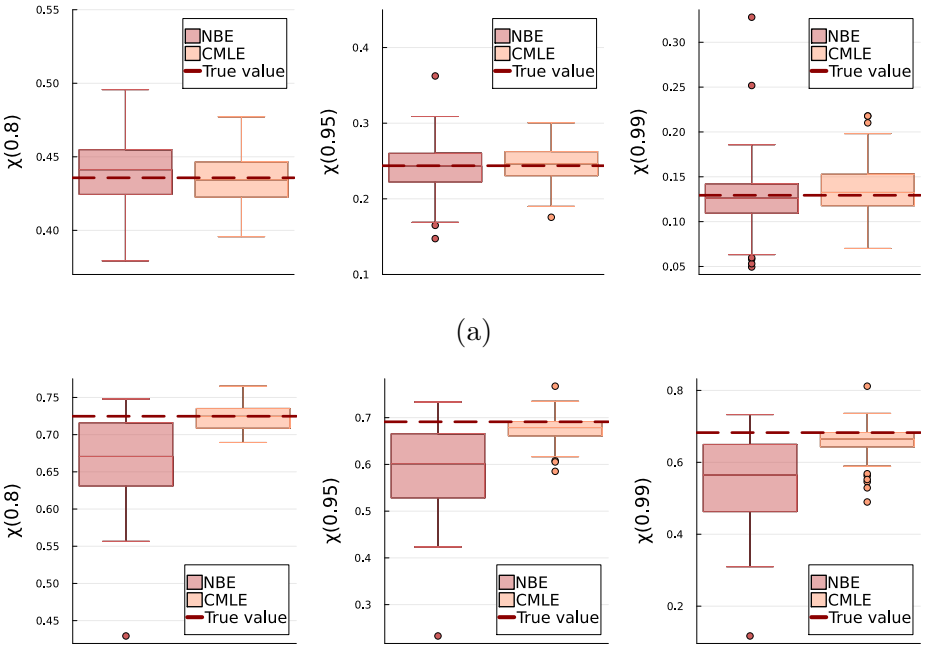
For the second case, we generate 100 samples, each with $n = 1000$, from a logistic distribution (Gumbel, 1960) with dependence parameter $\alpha_L = 0.4$, and consider a censoring level of $\tau = 0.8$. The proportion of times each model was selected and

the proportion of AD and AI samples identified by the NBE and CMLE are given in Table 5.4.6. For this case, the model selection procedures differ, with the neural classifier selecting Model HW 79% of the time, and the BIC selecting Model W as the most suitable in 46% of cases. Additionally, the CMLE correctly identifies all of the samples to be AD, whereas the NBE misclassifies 16 of the samples fitted with Model HW. For these 16 samples, the estimated values for $\delta$ are near 0.5, the boundary point for AI, in 12 samples; specifically, $\delta \in [0.45, 0.5]$. The estimates for $\chi(y)$ at three levels $y = \{0.80, 0.95, 0.99\}$ are shown in Figure 5.4.9. Contrarily to the previous case, the NBE under-estimates the true $\chi(y)$ of the logistic distribution, while the CMLE under-estimates slightly for higher levels. Moreover, the estimates provided by the NBE exhibit higher variability than the CMLE. As with the first specification in the section, an individual example is presented in Section C.4 of the Supplementary Material. Similarly to the repeated study, Model HW is selected as the most suitable to fit the data and seems to under-estimate the true $\chi(y)$ for $y \in [0.8, 0.99]$.

Table 5.4.6: Proportion of times each model was selected through the neural classifier and through BIC (left), and proportion of AD and AI samples identified by the NBE and CMLE (right). All the values are rounded up to 2 decimal places.

| Model | Neural classifier | BIC |
|---|---|---|
| Model W | 0.18 | 0.46 |
| Model HW | 0.79 | 0.24 |
| Model E1 | 0.02 | 0.03 |
| Model E2 | 0.01 | 0.27 |

| Method | AD | AI |
|---|---|---|
| NBE | 0.84 | 0.16 |
| CMLE | 1.00 | 0.00 |




(a)




(b)

Figure 5.4.9: Model-based estimates of $\chi(y)$ given by the NBE (red) and by the CMLE (orange) for levels $y = \{0.80, 0.95, 0.99\}$ and for 100 samples of (a) a Gaussian copula with correlation parameter $\rho = 0.5$ and $\tau = 0.75$, and of (b) a logistic distribution with dependence parameter $\alpha_L = 0.4$ and $\tau = 0.8$. For both cases, the true $\chi(y)$ value is given by the dashed red lines.

## 5.5 Case study: changes in horizontal geomagnetic field fluctuations

### 5.5.1 Data and background

The behaviour of the sun and the consequences of its interaction with the Earth's magnetic field and atmosphere are known as space weather events. Examples include phenomena such as the auroras, often known as Southern and Northern lights, or solar storms. These events can cause large fluctuations in the geomagnetic field, leading to geomagnetically induced currents (GICs), which are electrical currents generated at the surface of the planet by rapid changes in the magnetic field. Furthermore, GICs can cause disruption on power grids, communication systems, railway systems, and other

critical infrastructures. Thus, the modelling of extreme solar activity can help prevent the impacts of GICs.

Following Rogers et al. (2020), we use the rate of change of the horizontal geomagnetic field $\mathrm{d}B_H/\mathrm{d}t$, which is available through the SuperMAG interface (Gjerloev, 2009), as a measure of the magnitude of GICs. In particular, we apply the proposed toolbox to select and infer on the pairwise extremal dependence structure between measurements at three pairs of locations in the northern hemisphere: two in Greenland and one in the east coast of Canada (see Table 5.5.1). We take daily maximum absolute one-minute changes in $\mathrm{d}B_H/\mathrm{d}t$, which results in 7572 complete observations. However, since the estimators are trained for sample sizes between 100 and 1500, we take a subset of the data set with $n = 1500$, retaining every $5^{\text{th}}$ observation in order to reduce the temporal dependence present in the data, and truncate the resulting data set to 1500 observations by removing the last few. Figure 5.5.1 shows the scatterplots of pairwise daily maxima absolute one-minute changes in $\mathrm{d}B_H/\mathrm{d}t$ between the pairs of locations considered. We first transform the data to uniform margins using the semi-parametric approach of Coles and Tawn (1991) with a generalised Pareto distribution fit to the tail of the data. Thus, the cdf of each margin is estimated via

$$F(x) = \begin{cases} \tilde{F}(x), & \text{if } x \leq r, \\ 1 - \phi_r \left[1 + \xi \left(\frac{x-r}{\sigma}\right)\right]_+^{-1/\xi}, & \text{if } x > r, \end{cases}$$

where $\tilde{F}(x)$ is the empirical distribution, $\phi_r$ is the probability of exceeding a selecting high threshold $r$, and $\sigma$ and $\xi$ are the scale and shape parameters of the GPD, respectively.

Table 5.5.1: International Association of Geomagnetism and Aeronomy (IAGA) code, and location of the observatory for the three locations considered.

| IAGA code | Observatory (Country) | Latitude | Longitude |
|---|---|---|---|
| SCO | Scoresby Sund 2 (Greenland) | 70.48 | −21.97 |
| STF | Sdr Stromfjord (Greenland) | 67.02 | −50.72 |
| STJ | St. John's (Canada) | 47.60 | −52.68 |

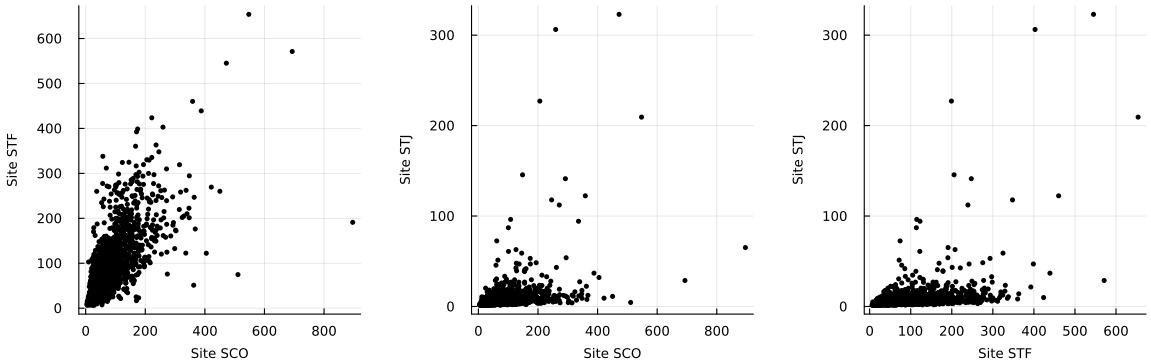



Figure 5.5.1: Daily maxima absolute one-minute changes in $\mathrm{d}B_H/\mathrm{d}t$ measurements between three pairs of locations: (SCO, STF) on the left, (SCO, STJ) in the middle, and (STF, STJ) on the right.

### 5.5.2 Statistical inference

We are interested in modelling the joint extremal behaviour of $\mathrm{d}B_H/\mathrm{d}t$ between each pair of locations, so we focus on the four models mentioned in Section 5.3.3, and censor the non-extreme observations; we do so by taking $\tau = 0.85$ as the censoring threshold. We note, however, that we have considered a range of censoring levels and summarise these results in Section C.5 of the Supplementary Material. We start by applying our neural classifier to select the best model and estimate its parameters through the corresponding trained NBE; the results are shown in Tables 5.5.2, 5.5.3 and 5.5.4 for pairs (SCO, STF), (SCO, STJ) and (STF, STJ), respectively. The fit of the preferred model is then assessed by comparing the estimated dependence measure $\chi(y)$ with its empirical counterparts for levels $y \in [0.85, 0.99]$. As an extra comparison, the estimated measures for the model with the second highest posterior probability are also obtained. Finally, a comparison with the results obtained through censored MLE is also provided; the

results for $\chi(y)$ are shown in Figure 5.5.2, where the confidence bands are obtained via block bootstrap with a block length of 10 to reflect the remaining temporal dependence present.

For pair (SCO, STF), Model HW is clearly the selected model with a probability of 0.992 followed by Model E2 with probability 0.078. As can be seen by the estimates given by the parameters of interest, $\hat{\delta}$ and $\hat{\xi}$, respectively, both models indicate the presence of asymptotic independence in the data. The same is not true for Model W. It can be seen on the left panel of Figure 5.5.2 that models HW and E2 (in blue and orange, respectively) seem to capture well the extremal dependence of data. According to the BIC, however, Model E1 is the most appropriate to fit this pair. Likewise to the NBE estimates, the CMLE estimates for this model indicate the presence of asymptotically independent data, which is in agreement with the estimates obtained by the NBE and CMLE for models HW and E2. Furthermore, the estimated $\chi(y)$ given by the CMLE (in purple) almost overlaps with the obtained $\chi(y)$ with Model E2. When testing a range of censoring thresholds, Model E2 or HW were consistently the preferred choices, both indicating asymptotic independence.

Model E2 is clearly selected as the best model to fit pair (SCO, STJ) with a posterior probability of 0.9. However, as shown by the middle panel of Figure 5.5.2, the estimated $\chi(y)$ measure seems to be under-estimated by this model (in blue) for higher levels $y$. In turn, the estimated measure $\chi(y)$ given by Model HW (in orange) is closer to its empirical counterpart further in the tail, indicating a better fit of the dependence structure in the limit. Although small, the probability for Model HW given by the neural classifier is the second highest, followed closely by Model W. For this pair, all four models indicate the presence of asymptotic independence data through the estimates given by the NBE and CMLE. Contrarily to the neural classifier, BIC selects Model E1 as the best model to fit the data, with a very small difference to model HW and W. The estimated $\chi(y)$, highlighted in purple, is also able to capture the extremal

Table 5.5.2: Model selection procedure obtained through the neural classifier for censoring level $\tau = 0.85$, and parameter estimates given by the trained NBE for pair (SCO, STF). The results through censoring MLE and BIC are given in the bottom table. All the values are rounded to 3 decimal places.

| Model | $\hat{p}$ | $\hat{\boldsymbol{\theta}}_{\text{NBE}}$ | Extremal dependence |
|---|---|---|---|
| W | $1.762 \times 10^{-4}$ | $(\hat{\alpha}, \hat{\xi}) = (2.625, 0.121)$ | AD |
| HW | **0.922** | $(\hat{\delta}, \hat{\omega}) = (0.258, 0.714)$ | AI |
| E1 | $8.539 \times 10^{-11}$ | $(\hat{\alpha}, \hat{\beta}, \hat{\mu}) = (3.771, 0.150, 0.939)$ | AI |
| E2 | 0.078 | $(\hat{\alpha}, \hat{\xi}) = (2.450, -0.179)$ | AI |
| **Model** | **BIC** | $\hat{\boldsymbol{\theta}}_{\text{CMLE}}$ | **Extremal dependence** |
| W | 229.875 | $(\hat{\alpha}, \hat{\xi}) = (2.724, 0.093)$ | AD |
| HW | 218.987 | $(\hat{\delta}, \hat{\omega}) = (0.353, 0.676)$ | AI |
| E1 | **211.927** | $(\hat{\alpha}, \hat{\beta}, \hat{\mu}) = (0.708, 0.339, 0.488)$ | AI |
| E2 | 216.518 | $(\hat{\alpha}, \hat{\xi}) = (2.321, -0.180)$ | AI |

dependence well. Similarly to the first pair, Model HW or E2 were selected as the preferred fit for pair (SCO, STJ) for a range of censoring levels, both models indicating asymptotic independence.

For the final pair (STF, STJ), Model E2 is again selected as the preferred model with a probability of 0.672. Model W closely follows with a posterior probability of 0.308. However, Model E2 has the highest BIC value, indicating that, in the likelihood framework, any of the other models would provide a better fit for this pair. The right panel of Figure 5.5.2 shows that Model E2 (in blue) under-estimates measure $\chi(y)$ for all levels $y$ considered, whilst Model E1 (obtained by CMLE and given in purple) over-estimates slightly the empirical value of $\chi(y)$. On the other hand, Model W (in orange) best captures the extremal dependence structure over most of the range. Despite this, all models indicate asymptotically independent variables; this is also in agreement with the estimates obtained by censored maximum likelihood inference. Interestingly enough, the estimated value of parameter $\mu$ of Model E1 given by either approach is (very close to) 1. When considering different censoring thresholds, the neural classifier consistently chose either Model HW or E2 as the best fit for the pair (STF, STJ), both models indicating asymptotic independence across all censoring levels.

Table 5.5.3: Model selection procedure obtained through the neural classifier for censoring level $\tau = 0.85$, and parameter estimates given by the trained NBE for pair (SCO, STJ). The results through censoring MLE and BIC are given in the bottom table. All the values are rounded to 3 decimal places.

| Model | $\hat{p}$ | $\hat{\boldsymbol{\theta}}_{\text{NBE}}$ | Extremal dependence |
|---|---|---|---|
| W | 0.029 | $(\hat{\alpha}, \hat{\xi}) = (2.527, -0.271)$ | AI |
| HW | 0.072 | $(\hat{\delta}, \hat{\omega}) = (0.125, 0.621)$ | AI |
| E1 | $2.352 \times 10^{-7}$ | $(\hat{\alpha}, \hat{\beta}, \hat{\mu}) = (9.546, 2.171, 0.855)$ | AI |
| E2 | **0.900** | $(\hat{\alpha}, \hat{\xi}) = (2.316 - 0.791)$ | AI |
| Model | BIC | $\hat{\boldsymbol{\theta}}_{\text{CMLE}}$ | Extremal dependence |
| W | 393.877 | $(\hat{\alpha}, \hat{\xi}) = (2.364, -0.187)$ | AI |
| HW | 393.629 | $(\hat{\delta}, \hat{\omega}) = (0.242, 0.584)$ | AI |
| E1 | **393.138** | $(\hat{\alpha}, \hat{\beta}, \hat{\mu}) = (2.634, 1.112, 0.735)$ | AI |
| E2 | 396.562 | $(\hat{\alpha}, \hat{\xi}) = (2.115, -0.645)$ | AI |

Table 5.5.4: Model selection procedure obtained through the neural classifier for censoring level $\tau = 0.85$, and parameter estimates given by the trained NBE for pair (STF, STJ). The results through censoring MLE and BIC are given in the bottom table. All the values are rounded to 3 decimal places.

| Model | $\hat{p}$ | $\hat{\boldsymbol{\theta}}_{\text{NBE}}$ | Extremal dependence |
|---|---|---|---|
| W | 0.308 | $(\hat{\alpha}, \hat{\xi}) = (2.670, -0.317)$ | AI |
| HW | 0.019 | $(\hat{\delta}, \hat{\omega}) = (0.111, 0.632)$ | AI |
| E1 | $5.876 \times 10^{-5}$ | $(\hat{\alpha}, \hat{\beta}, \hat{\mu}) = (10.393, 3.228, 0.996)$ | AI |
| E2 | **0.672** | $(\hat{\alpha}, \hat{\xi}) = (2.420, -0.849)$ | AI |
| Model | BIC | $\hat{\boldsymbol{\theta}}_{\text{CMLE}}$ | Extremal dependence |
| W | 401.813 | $(\hat{\alpha}, \hat{\xi}) = (2.222, -0.153)$ | AI |
| HW | 400.142 | $(\hat{\delta}, \hat{\omega}) = (0.025, 0.601)$ | AI |
| E1 | **398.434** | $(\hat{\alpha}, \hat{\beta}, \hat{\mu}) = (13.559, 3.023, 1.000)$ | AI |
| E2 | 412.893 | $(\hat{\alpha}, \hat{\xi}) = (2.072, -0.680)$ | AI |

From the results obtained, a few conclusions can be drawn. For instance, although the selected model may not always be the best to capture the extremal dependence of the data, a good representation can still be achieved if the model with the second highest probability is considered instead. Additionally, with the proposed toolbox, we are not only able to infer about some model characteristics, but also able to assess the sensitivity to the censoring threshold by considering a range of different levels $\tau$. In particular, from the results obtained for the range of $\tau$ considered (see Section C.5 of the Supplementary

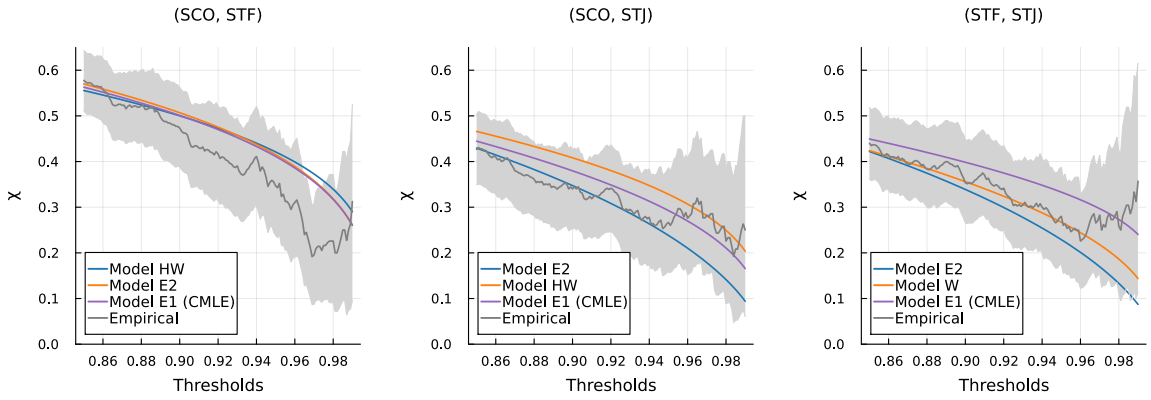



Figure 5.5.2: Empirical (in grey) and model $\chi(y)$ estimated via the NBE for $y \in [0.85, 0.99]$ for the models with the two highest posterior probabilities. For each pair, the estimated $\chi(y)$ for the selected model is given by the blue line, followed by its estimate obtained with the model with the second highest probability in orange. Estimated model $\chi(y)$ for the selected model through AIC is given by the purple line. The 95% confidence bands were obtained by block boostrapping.

Material), we can see that, for lower censoring levels, Model HW is clearly the preferred one to fit the three pairs, with posterior probabilities above 0.9 in all cases. This changes for censoring levels $\tau \geq 0.85$, or $\tau \geq 0.9$ for pair (SCO, STF), for which Model E2 is the most suitable to fit the data according to the neural classifier. In all cases, the selected models consistently indicate the presence of asymptotic independence in each pair. This sensitivity analysis to the censoring level is computationally expensive in a likelihood-based inference method, where the likelihood needs to be evaluated for all considered models across various censoring levels; this leads to the need for choosing a censoring level before the analysis, which might end up not being the most suitable to the underlying application.

## 5.6 Conclusion and discussion

In some situations, while the likelihood function of a model may be available, its evaluation is computationally costly; this might be due to the need for inversion of functions, numerical integration or even a combination of the two. This computational burden in

the inference process poses limitations on the use of certain models in practice since modelling normally entails consideration of different candidate models and threshold levels. In this paper, we exploited the use of neural Bayes estimation, a likelihood-free approach which uses neural networks, to perform inference on the vector of model parameters. In particular, we focused on two types of models which are available in the multivariate extremes literature: a weighted copula model, which is able to represent both the body and tail regions of a data set, and models based on random scale constructions, which are flexible enough to capture the two regimes of extremal dependence, interpolating between the two in the interior of the parameter space. When likelihood evaluation is no longer required, model selection criteria such as the AIC or BIC are now unavailable for determining the best-fitting model. We overcame this by proposing a model selection neural classifier which allows us to choose the most suitable model from the candidates in a fast and effective manner.

For the models where it is feasible, albeit relatively slow, to evaluate the likelihood function, we compared the performance of the neural Bayes estimators and neural classifier with (censored) maximum likelihood inference and BIC, respectively. Through simulation studies, we have shown that NBEs tend to be more biased than the (censored) MLE, though they do not always exhibit higher variability. Moreover, when estimating quantities of interest, such as the extremal measures $\chi(y)$ from equation (5.3.1), we did not find the bias significant, as the dependence structure appeared to be well captured with the trained NBEs. However, the bias of the estimates might be reflected in the lower coverage shown by the bootstrap-based credible intervals for some parameters in Section 5.4.2. Whilst the coverage probability significantly improved when we trained a neural interval estimator (obtaining a value of 0.96 or higher) it would be ideal to also obtain such values with bootstrap-based intervals as we would only need to train the estimator once in a single loss function. Despite this, since the neural Bayes estimates are considerably faster to obtain than the (censored) MLE and the extremal

dependence structure is generally well captured, the methodology may be considered preferable overall.

When comparing the neural classifier with the BIC, both procedures performed similarly for the case when $M = 2$. However, when $M = 4$, the neural classifier was able to correctly categorise more data sets than the BIC for all considered models, except Model W. We have also shown that the proposed toolbox for model selection and inference performs well — though not always as well as classical likelihood inference — for scenarios where the data set does not originate from one of the models considered in this paper, with the extremal dependence structure being generally well captured. In particular, through repeated simulation, we showed that the neural classifier and NBE captured the true extremal dependence structure the majority of the times, especially for asymptotically independent data. A general consideration of the performance of NBEs in misspecified scenarios is an important future line of research.

In all our examples, we assumed the parameters to be uniformly distributed a priori. However, we note this implies non-uniform priors in alternative parameterisations and we did not assess the effect of this choice on the performance of the trained NBE. Despite this, for studies involving the weighted copula model from Section 5.3.4, we found that reparameterising some of the model parameters helped the neural network to learn about them in the training step of the NBE. More specifically, the assessment of the NBE showed less variability and higher coverage probabilities (with narrower range) of the 95% uncertainty intervals for each model parameters obtained with bootstrap-based intervals. The performance of the NBE is inherently influenced by the choice of the prior distributions. While an informative prior might reduce the volume of the parameter space, and allow lower values for $K$ and $J$ in equation (5.2.2), this restricts the use of the trained estimators across several applications; when the latter is desirable, a vague prior is advisable (Sainsbury-Dale et al., 2024a). On the other hand, when the likelihood is available, and feasible, informative prior distributions can be constructed

using likelihood-based estimates (Lenzi et al., 2023). Finally, we restricted our analysis to the bivariate setting. Each of the random scale models described in Section 5.3.3 could be expanded to higher dimensions; however, this may not be as useful since the models considered are only suitable when all the variables are asymptotically dependent or asymptotically independent. On the other hand, the mixture model could usefully be extended to higher dimensions, and the benefits of NBEs may be even clearer in this setting as the complexity of the likelihood increases with dimension for this model.

Neural network-based approaches have be a growing subject in the extremes literature. For instance, in a univariate setting, Cannon (2010), Cannon (2011), Carreau and Vrac (2011), Ceresetti et al. (2012), Vasiliades et al. (2015), Bennett et al. (2015) and Shrestha et al. (2017) leverage neural networks to estimate the parameters, risk measures or to build mixture models. In the multivariate framework, this has been mainly predominant in, but not restricted to, the spatial setting. Ahmed et al. (2022) and Wixson and Cooley (2024) propose using neural networks as classification tools for testing the extremal dependence type of a data set. Whilst the former only focus on spatial processes, the latter applies the classifier to both spatial and non spatial data sets. In the approach of Murphy-Barltrop et al. (2024), neural networks are used to learn about the extremal dependence structure of higher-dimensional data, by considering a geometric approach whereby the joint behaviour can be inferred from a star-shaped limit set. In a regression context, Cisneros et al. (2024) use neural networks to model the full distribution of wildfire spread, whereas Pasche and Engelke (2024) and Richards and Huser (2022) propose neural-based approaches to extreme quantile regression problems.

# Chapter 6

# Extreme value methods for estimating rare events in Utopia

## 6.1 Introduction

This paper details an approach to the data challenge organised for the Extreme Value Analysis (EVA) 2023 Conference. The objective of the challenge was to estimate extremal probabilities, or their associated quantiles, for simulated environmental data sets for various locations in a fictitious country called Utopia. The data challenge is split into 4 challenges; challenges C1 and C2 focus on a setting where data is obtained from a single location while challenges C3 and C4 concern multivariate data sets, where data is obtained simultaneously from multiple locations.

Challenge C1 requires estimation of the 0.9999-quantile of the distribution of the environmental response variable $Y$ conditional on a covariate vector $\boldsymbol{X}$, for 100 realisations of covariates. To do so, we model the tail of $Y \mid \boldsymbol{X} = \boldsymbol{x}$ using a generalised Pareto distribution (GPD; Pickands, 1975) and employ the extreme value generalised additive modelling (EVGAM) framework, first introduced by Youngman (2019), to account for the non-stationary data structure. We consider a variety of model formulations and

select our final model using cross-validation. Furthermore, central 50% confidence intervals are estimated via a non-stationary bootstrapping technique, and the final model performance is assessed using the number of times the true conditional quantile lies in the confidence intervals (Rohrbeck et al., 2024). For Challenge C2, we are interested in estimating the value of $q$ that satisfies $\Pr(Y > q) = 1/(300T)$, where $T = 200$.

Challenges C3 and C4 concern the estimation of probabilities for extreme multivariate regions, subsets of $\mathbb{R}^d$, where some or all of the components are so large that we seldom observe any data in them. Such estimates require techniques for modelling and extrapolating within the joint tail. For challenge C3, we want to estimate two joint tail probabilities for three unknown non-stationary environmental variables. To achieve this, we propose a non-stationary extension of the model introduced by Wadsworth and Tawn (2013). Lastly, for challenge C4, we wish to estimate the probability that 50 variables (locations) jointly exceed prespecified extreme thresholds. Based on an initial analysis, we separate the variables into five independent groups, and obtain distinct probability estimates for each group using the conditional extremes approach of Heffernan and Tawn (2004).

The remainder of the paper is structured as follows. A suitable background to EVA is provided in Section 6.2, introducing concepts required throughout our work. Section 6.3 covers our approach to the univariate challenges C1 and C2, and the multivariate challenges C3 and C4 are considered in Sections 6.4 and 6.5, respectively. The paper ends with a discussion of the results of all challenges in Section 6.6.

## 6.2 EVA background

### 6.2.1 Univariate modelling

Univariate EVA methods are concerned with capturing the behaviour of the tail of a distribution which allows for extreme quantities to be estimated. A common univariate

approach is the peaks-over-threshold framework. Consider a continuous, independent and identically distributed (IID) random variable $Y$ with distribution function $F$ and upper endpoint $y^F := \sup\{y : F(y) < 1\}$. Pickands (1975) shows that, for some high threshold $v < y^F$, the excesses $(Y - v) \mid Y > v$, after suitable rescaling, converge in distribution to a GPD as $v \to y^F$. Davison and Smith (1990) provide an overview of the properties of the GPD, and also propose an extension of this framework to the non-stationary setting: given a non-stationary process $Y$ with associated covariate(s) $\boldsymbol{X}$, the authors propose the following model

$$\Pr(Y > y + v \mid Y > v, \boldsymbol{X} = \boldsymbol{x}) = \left(1 + \frac{y\xi(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)_+^{-1/\xi(\boldsymbol{x})}, \tag{6.2.1}$$

for $y > 0$, where $\sigma(\cdot), \xi(\cdot)$ are the covariate-dependent scale and shape parameters, respectively. Recent extensions of the Davison and Smith (1990) framework include allowing the threshold to be covariate-dependent, i.e., $v(\boldsymbol{x})$ (Kyselý et al., 2010; Northrop and Jonathan, 2011), and using generalised additive models (GAMs; Chavez-Demoulin and Davison, 2005; Youngman, 2019) to capture the functions $\sigma(\cdot)$ and $\xi(\cdot)$ in a flexible manner.

### 6.2.2 Extremal dependence measures

In addition to analysing marginal tail behaviours, multivariate EVA methods are concerned with quantifying the dependence between extremes of the individual components. An important classification of this dependence is obtained through the measure $\chi$ (Joe, 1997): given a $d$-dimensional random vector $\boldsymbol{Z}$, with $d \geq 2$ and $Z_i \sim F$ for all $i \in \{1, \ldots, d\}$,

$$\chi(u) := \left(\frac{1}{1-u}\right) \Pr(F(Z_1) > u, \ldots, F(Z_d) > u), \tag{6.2.2}$$

with $u \in [0, 1)$. Where the limit exists, we set $\chi := \lim_{u \to 1} \chi(u) \in [0, 1]$. When $\chi > 0$, we say that the variables in $\boldsymbol{Z}$ exhibit asymptotic dependence, i.e., can take their largest

values simultaneously, with the strength of dependence increasing as $\chi$ approaches 1. If $\chi = 0$, the variables cannot all take their largest values together. In particular, for $d = 2$, we refer to the case $\chi = 0$ as asymptotic independence.

We also consider the coefficient of tail dependence proposed by Ledford and Tawn (1996). Using the formulation given in Resnick (2002), let

$$\eta(u) := \frac{\log(1-u)}{\log \Pr\left(F(Z_1) > u, \ldots, F(Z_d) > u\right)},$$

with $u \in [0, 1)$. When the limit exists, we set $\eta := \lim_{u \to 1} \eta(u) \in (0, 1]$. The cases $\eta = 1$ and $\eta < 1$, correspond to $\chi > 0$ and $\chi = 0$, respectively. For $\eta < 1$, this coefficient quantifies the form of dependence for random vectors that do not take their largest values simultaneously.

Since $\chi$ and $\eta$ are limiting values, they are unknown in practice and must be approximated using numerical techniques. Therefore, when quantifying extremal dependence, we approximate $\chi$ $(\eta)$ using empirical estimates of $\chi(u)$ $\big(\eta(u)\big)$ for some high threshold $u$.

## 6.3 Challenges C1 and C2

Both challenges concern 70 years of daily data for the capital city of Amaurot. Each year has 12 months of 25 days and two seasons (season 1 for months 1-6, and season 2 for months 7-12). Suppose $Y$ is an unknown response variable, and $\boldsymbol{X} = (V_1, \ldots, V_8)$ is a vector of covariates, $(V_1, V_2, V_3, V_4)$ denoting unknown environmental variables and $(V_5, V_6, V_7, V_8)$ denoting season, wind direction (radians), wind speed (unknown scale), and atmosphere (recorded monthly), respectively.

For C1, we build a model for $Y \mid \boldsymbol{X}$ and estimate the 0.9999-quantile, with associated 50% confidence intervals, for 100 different covariate combinations denoted $\tilde{\boldsymbol{x}}_i$ for $i \in \{1, \ldots, 100\}$. Note $\tilde{\boldsymbol{x}}_i$ are not covariates observed within the data set, but new

observations provided by the challenge organisers.

For C2, we estimate the marginal quantile $q$ such that $\Pr(Y > q) = (6 \times 10)^{-4}$, which corresponds to a once in 200-year event in the IID setting; in particular, $q$ is obtained subject to a predefined loss function. We first estimate the marginal distribution $F_Y(y)$ using Monte-Carlo techniques; see for instance, Eastoe and Tawn (2009). Since we have a large sample size, $n = 21,000$, it is reasonable to assume that the observed covariate sample is representative of $\boldsymbol{X}$. Thus, we can approximate the marginal distribution $F_Y(y)$ as follows,

$$\hat{F}_Y(y) = \int_{\boldsymbol{X}} F_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n} \sum_{t=1}^{n} F_{Y_t|\boldsymbol{X}_t}(y_t \mid \boldsymbol{x}_t). \tag{6.3.1}$$

where $F_{Y|\boldsymbol{X}}(\cdot)$ is the conditional distribution function of $Y \mid \boldsymbol{X}$ and $f_{\boldsymbol{X}}(\cdot)$ denotes the joint probability density of the covariates $\boldsymbol{X}$.

We incorporate the following loss function provided by the challenge organisers,

$$\mathcal{L}(q, \hat{q}) = \begin{cases} 0.9(0.99q - \hat{q}) & \text{if } 0.99q > \hat{q}, \\ 0 & \text{if } |q - \hat{q}| \leq 0.01q, \\ 0.1(\hat{q} - 1.01q) & \text{if } 1.01q < \hat{q}, \end{cases} \tag{6.3.2}$$

where $q$ and $\hat{q}$ are the true and estimated marginal quantiles, respectively. This loss function penalises under-estimation more heavily than an over-estimation.

We conduct the same exploratory data analysis for both challenges given the same covariates are used; this is outlined in Section 6.3.1. In Section 6.3.2 we introduce our techniques for modelling $Y \mid \boldsymbol{X}$, which is then used for modelling $Y$ via (6.3.1). Our approach for uncertainty quantification is outlined in Section 6.3.3, and we give our results for both challenges in Section 6.3.4.

### 6.3.1 Exploratory data analysis

Given the covariate vector $\boldsymbol{X_t} = \{V_{1,t}, \ldots, V_{8,t}\}$, the environmental response variable $Y_t$, $t \in \{1, \ldots, n\}$, is temporally independent (Rohrbeck et al., 2024). However, it is not clear which covariates affect $Y$, and what form these covariate-response relationships take. In what follows, we aim to explore these relationships so we can account for them in our modelling framework.

To begin, we explore the dependence between all variables to understand the relationships between covariates, as well as the relationships between individual covariates and the response variable. We investigate dependence in the main body of the data using Kendall's $\tau$ measure, while for the joint tails, we use the pairwise extremal dependence coefficients $\chi$ and $\eta$ defined in Section 6.2; values for all pairs are shown in Figure 6.3.1, with the threshold $u$ set at the empirical 0.95-quantile for the extremal measures.

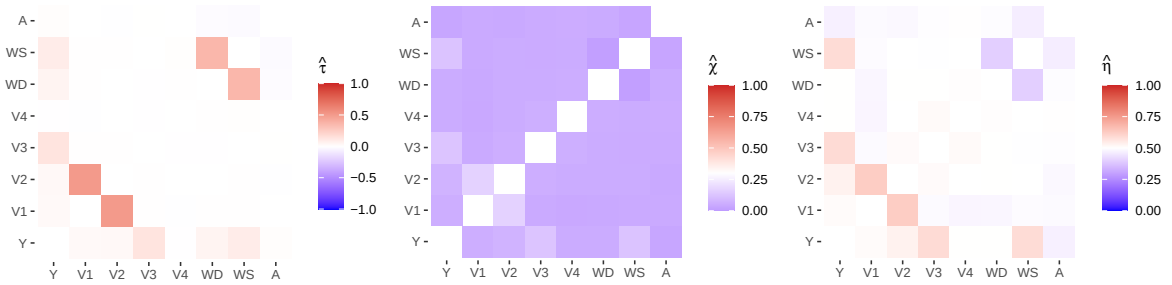



Figure 6.3.1: Heat maps for dependence measures for each pair of variables: Kendall's $\tau$ (left), $\chi$ (middle) and $\eta$ (right). Note the scale in each plot varies, depending on the support of the measure, and the diagonals are left blank, where each variable is compared against itself.

The response variable $Y$ has the strongest dependence with $V_3$ in the body of the distribution (see $\hat{\tau}$ in Figure 6.3.1), followed by $V_6$ (wind speed) then $V_7$ (wind direction). For the tail of the distribution, $Y$ has strongest dependence with $V_2$, $V_3$ and $V_6$ (see $\hat{\chi}$ and $\hat{\eta}$ in Figure 6.3.1). We also find strong dependence between $V_6$ and $V_7$ in the

body, but evidence of weak dependence in the tail (dark blue for $\hat{\chi}$ and $\hat{\eta}$). There is also strong dependence between $V_1$ and $V_2$ in both the body and tail (see dark red for $\hat{\eta}$). We find very similar dependence relationships when the data are split into seasons. In the Supplementary Material, we show scatter plots of each covariate against the response variable; these demonstrate a highly non-linear relationship for each explanatory variable with $Y$.

Next, we explore temporal relationships for the response variable $Y$. We first find temporal non-stationarity as the distribution of $Y$ varies significantly with $V_5$ (season); see the Supplementary Material for more detail. The mean and range of $Y$ is higher in season 1 than season 2, with greater extreme values observed in season 1. However, within each season, across months, there is little temporal variation in the distribution of $Y$. We also find that $Y$ exhibits temporal independence at all lags, with auto-correlation function (acf) values close to zero; see the Supplementary Material.

As noted in Rohrbeck et al. (2024), 11.7% of the observations have at least one value missing completely at random (MCAR). A detailed breakdown of the pattern of missing predictor observations is provided in the Supplementary Material. Since we can assume the data are MCAR, ignoring the observations that have a missing predictor covariate will not bias our inference, however, a complete case analysis is undesirable due to the amount of data loss. To mitigate against this, we attempted to impute the observations where predictors are missing but ultimately could not find an imputation method that satisfactorily retained the dependence structure between the response and covariates, particularly in the tails of the distribution. Therefore, we use a case analysis approach, whereby an observation is only removed if a predictor covariate of interest is missing. This results in only 4% of observations being removed for our final model.

### 6.3.2 Methods

Due to the complex nature of the data, we consider various non-stationary GPD models, as in equation (6.2.1), that are formulated as GAMs to fit $Y \mid \boldsymbol{X}$. For threshold selection, we extend the method proposed by Murphy et al. (2024) to select a threshold for non-stationary, covariate-dependent GPD models; the details are provided in Section 6.3.2. Our inference and model selection procedures are then provided in Sections 6.3.2 and 6.3.2, respectively. We note that the same model formulation is used for both C1 and C2 with a small adjustment to the parameter estimation procedure for C2 to incorporate the provided loss function given in (6.3.2). We utilise equation (6.3.1) to obtain the marginal distribution of $Y$.

#### General model formulation

Let $\tilde{\boldsymbol{X}}_t$ denote the set of predictor covariates with $t \in \{1, \ldots, n\}$. Then $y_t$ and $\tilde{\boldsymbol{x}}_t$ denote the observations of the response variable and predictive covariates, respectively. We consider models with the following form,

$$F_{Y_t|\tilde{\boldsymbol{X}}_t}(y_t|\tilde{\boldsymbol{X}}_t = \tilde{\boldsymbol{x}}_t) = 1 - \zeta(\tilde{\boldsymbol{x}}_t)\left[1 + \xi(\tilde{\boldsymbol{x}}_t)\left(\frac{y_t - v(\tilde{\boldsymbol{x}}_t)}{\sigma(\tilde{\boldsymbol{x}}_t)}\right)\right]_+^{-1/\xi(\tilde{\boldsymbol{x}}_t)}, \tag{6.3.3}$$

where $v(\tilde{\boldsymbol{x}}_t)$ and $\zeta(\tilde{\boldsymbol{x}}_t)$ are a covariate-dependent threshold and rate parameter, respectively, such that the rate parameter corresponds to the probability of exceeding the threshold.

Our analysis in Section 6.3.1 indicates that $V_3$, $V_5$ (season), and $V_6$ (wind speed) exhibit non-trivial dependence relationships with the response variable. Therefore we assume these variables can be used as predictor variables for modelling $Y$, and set $\tilde{\boldsymbol{x}} := (\boldsymbol{V}_j)_{j\in\{3,5,6\}}$. Although $V_7$ (wind direction) also exhibits strong dependence with $Y$, we do not consider it here since it is highly correlated with wind speed so would involve adding complex interaction terms to the model formulation, and $V_6$ has a stronger

relationship with $Y$ compared to $V_7$ (see Figure 6.3.1).

Owing to the complex covariate structure observed in the data, as described in Section 6.3.1, we employ the flexible EVGAM framework proposed in Youngman (2019) for modelling tail behaviour. Under this framework, GAM formulations are used to capture non-stationarity in the threshold, scale and shape functions given in equation (6.3.3). Without loss of generality, consider the scale function $\sigma(\cdot)$. We assume that

$$h(\sigma(\tilde{\boldsymbol{x}})) = \psi_\sigma(\tilde{\boldsymbol{x}}), \quad \text{with} \quad \psi_\sigma(\tilde{\boldsymbol{x}}) = \beta_0 + \sum_{\kappa=1}^{K} \sum_{p=1}^{P_\kappa} \beta_{\kappa p} b_{\kappa p}(\tilde{\boldsymbol{x}}), \tag{6.3.4}$$

where $h(x) := \log(x)$ denotes the link function which ensures the correct support, with coefficients $\beta_0, \beta_{\kappa p} \in \mathbb{R}$ and basis functions $b_{\kappa p}$ for $p \in \{1, \ldots, P_\kappa\}, \kappa \in \{1, \ldots, K\}$, where $K$ is the number of splines in the GAM formulation and $P_\kappa$ is the basis dimension relating to spline $\kappa$. The basis functions can be in terms of individual covariates, i.e., $b_{\kappa p} : \mathbb{R} \mapsto \mathbb{R}$, or multiple covariates, i.e., $b_{\kappa p} : \mathbb{R}^m \mapsto \mathbb{R}$, $1 < m \leq 8$. Analogous forms can be taken for $v(\cdot)$ and $\xi(\cdot)$, adjusting the link function $h(\cdot)$ as appropriate, although these are not considered here for reasons detailed below.

To select an appropriate threshold, we employ the threshold selection method of Murphy et al. (2024) and extend this approach to select a threshold for non-stationary, covariate-dependent GPD models. The method selects a threshold based on minimising the expected quantile discrepancy (EQD) between the sample quantiles and fitted GPD model quantiles. When fitting a non-stationary model, the excesses will not be identically distributed across covariates. Thus, to utilise the EQD method in this case, we use the fitted non-stationary GPD parameter estimates to transform the excesses to common standard exponential margins and compare sample quantiles against theoretical quantiles from the standard exponential distribution. This transformation is a common approach for checking the model fit of a non-stationary GPD (Coles, 2001).

We use a stepped-threshold according to season as there is clear variation in the

distribution, and thereby the extremes, of $Y$ between seasons; see the Supplementary Material for more details. Specifically, we set $v(\tilde{\boldsymbol{x}}_t) := \mathbb{1}(\tilde{x}_{2,t} = 1)v_1 + \mathbb{1}(\tilde{x}_{2,t} = 2)v_2$, $v_1, v_2 \in \mathbb{R}$, with corresponding rate parameter $\zeta(\tilde{\boldsymbol{x}}_t) := \mathbb{1}(\tilde{x}_{2,t} = 1)\zeta_1 + \mathbb{1}(\tilde{x}_{2,t} = 2)\zeta_2$, where $\zeta_1, \zeta_2 \in [0, 1]$ denote the probabilities of exceeding the threshold for seasons 1 and 2, respectively, and $\tilde{x}_{r,t}$ are realisations of the $r^{\text{th}}$ component of $\tilde{\boldsymbol{x}}$ for $r \in \{1, 2, 3\}$. This seasonal threshold significantly improves model fits; see the Supplementary Material for further details. GAM forms for the threshold were also explored, but did not offer significant improvement. Furthermore, the smooth GAM formulation of the GPD scale parameter adequately captures any residual variation in the response arising due to covariate dependence.

### Inference

For all GAM formulations, we only consider basis functions of singular covariates, since specifying basis functions of multiple variables requires a detailed understanding of covariate interactions and can significantly increase the computational complexity of the modelling procedure (Wood, 2017). We keep the shape function $\xi(\boldsymbol{x}) := \xi \in \mathbb{R}$ constant across covariates; this is common in non-stationary analyses, since this parameter is difficult to estimate (Chavez-Demoulin and Davison, 2005). Within the GAM formulation, we consider several parametric forms to account for the predictive covariates in the scale parameter using linear models, indicator functions and splines.

When using splines, we are required to select a basis dimension $P_\kappa \in \mathbb{N}$; this determines the number of coefficients to be estimated. Basis dimension is the most important choice within spline modelling procedures and directly corresponds with the flexibility of the framework (Wood, 2017). We only consider splines for $V_3$ and $V_6$. For each $\tilde{X}_r$, $r \in \{1, 3\}$, we determine the basis dimension $P_1$ and $P_2$, respectively, by first building a model for $Y_t \mid \tilde{X}_{r,t}$, to allow us to consider the effect of this predictor on the response directly. We vary the basis dimension and compare the resulting models using cross

validation (CV), detailed in the following section. We set $P_1 = 4$ and $P_2 = 3$ for $V_3$ and $V_6$, respectively.

For C2, we incorporate the loss function of equation (6.3.2) into the estimation procedure. Let $\mathcal{I}_v := \{t \in \{1, \ldots, n\} \mid y_t > v(\tilde{\boldsymbol{x}}_t)\}$ denote the set of temporal indices corresponding to threshold exceedances and $n_v := |\mathcal{I}_v|$. We consider the objective function

$$S(\boldsymbol{\theta}) := -l_R(\boldsymbol{\theta}) + \sum_{i \in \mathcal{I}_v} \mathcal{L}(q_i^*, \hat{q}_i)/n_v, \tag{6.3.5}$$

where $l_R(\boldsymbol{\theta})$ denotes the penalised log-likelihood function of the restricted maximum likelihood estimation (REML) approach (Wood, 2017), $\boldsymbol{\theta}$ denotes the parameter vector associated with the GPD formulation of equation (6.3.4), and $\sum_{i \in \mathcal{I}_v} \mathcal{L}(q_i^*, \hat{q}_i)/n_v$ denotes the average loss between the sample quantiles of the transformed excesses and the theoretical standard exponential quantiles. Specifically, we transform the excesses, $(y_t - v(\tilde{\boldsymbol{x}}_t))_{t \in \mathcal{I}_v}$, to standard exponential margins using the fitted non-stationary GPD parameter estimates and compare the ordered excesses, $\boldsymbol{q}^*$, to the theoretical quantiles, $\hat{\boldsymbol{q}}$, from a standard exponential distribution evaluated at probabilities $\{p_i = i/(n_v + 1), i = 1, \ldots, n_v\}$. Minimising the objective function $S(\boldsymbol{\theta})$ ensures that the parameter estimates also account for and minimise the loss function, $\mathcal{L}$. We use this formulation to adjust the GPD parameters for challenge C2 once a threshold is selected.

### Model selection

To determine the best-fitting model, we use a forward selection process and aim to minimise the model's CV score. For each model, we apply $k$-fold CV (Hastie et al., 2001, Ch 7.) utilising the continuous ranked probability score (CRPS, Gneiting and Katzfuss, 2014) as our goodness-of-fit metric. CRPS describes the discrepancy between the predicted distribution function and observed values without the specification of empirical quantiles. We explore model ranking by taking both $k = 10$ and 50, and find that both give an equivalent ranking; we present results for the latter. We also

provide the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values to aid in model selection. A subset of models used in the forward selection process are detailed in Table 6.3.1 where, for each model, we provide the change in the CRPS, AIC and BIC relative to model 1. The parameterisation of model 7 achieves the largest reduction for all three metrics relative to the baseline model.

Table 6.3.1: Table of selected models considered for challenge C1. $\mathbb{1}(\cdot)$ denotes an indicator function, $s_i(\cdot)$ for $i \in \{1, 2\}$ denote thin-plate regression splines, $\beta_0, \beta_1$ are coefficients to be estimated, and $\tilde{x}_{r,t}$ is defined as in the text. All values have been given to one decimal place.

| Model | $\sigma(\tilde{\boldsymbol{x}}_t)$ | $\Delta$CRPS | $\Delta$AIC | $\Delta$BIC |
|---|---|---|---|---|
| 1 | $\beta_0$ | 0 | 0 | 0 |
| 2 | $\beta_0 + \beta_1 \mathbb{1}(\tilde{x}_{2,t} = 1)$ | -0.5 | -33.4 | -26.1 |
| 3 | $\beta_0 + s_1(\tilde{x}_{1,t})$ | -0.9 | -408.5 | -379.2 |
| 4 | $\beta_0 + s_2(\tilde{x}_{3,t})$ | -0.5 | -284.3 | -276.8 |
| 5 | $\beta_0 + \beta_1 \mathbb{1}(\tilde{x}_{2,t} = 1) + s_1(\tilde{x}_{1,t})$ | -0.9 | -425.8 | -388.1 |
| 6 | $\beta_0 + s_1(\tilde{x}_{1,t}) + s_2(\tilde{x}_{3,t})$ | -1.0 | -752.7 | -717.2 |
| 7 | $\beta_0 + \beta_1 \mathbb{1}(\tilde{x}_{2,t} = 1) + s_1(\tilde{x}_{1,t}) + s_2(\tilde{x}_{3,t})$ | **-1.1** | **-780.0** | **-735.3** |

### 6.3.3 Uncertainty

For each of the 100 different covariate combinations, $\tilde{\boldsymbol{x}}_i$ for $i \in \{1, \ldots, 100\}$, we need to construct central 50% confidence intervals. We use a bootstrapping procedure to avoid making potentially inaccurate assumptions such as the asymptotic normality approximation of maximum likelihood estimates, for example. Traditional bootstrap approaches are non-parametric and randomly resample the data with replacement. However, in Section 6.3.1 we find that the response variable is dependent on covariates, and these covariates exhibit temporal dependence. A standard bootstrap procedure would therefore not retain this dependence. Instead, we preserve the temporal dependence structure of covariates and their relationship with the response variable by approximating our confidence intervals using the stationary, semi-parametric bootstrapping procedure adopted by D'Arcy et al. (2023).

First, the response variable $Y_t$ is transformed to Uniform(0,1) margins to preserve its non-stationary behaviour; denote this sequence $U_t^Y = F_{Y_t|\tilde{\boldsymbol{X}}_t}(Y_t|\tilde{\boldsymbol{X}}_t = \tilde{x}_t)$ where $F_{Y_t|\tilde{\boldsymbol{X}}_t}$ is the estimated model given in equation (6.3.3). We then adopt the stationary bootstrap procedure of Politis and Romano (1994) to retain the temporal dependence in the response and explanatory variables by sampling blocks of consecutive observations. The block length $L$ is random and simulated from a Geometric$(1/l)$ distribution, where the mean block length $l \in \mathbb{N}$ is carefully selected based on the autocorrelation function. This was selected at 50 days, the maximum lag for which the autocorrelation was significant across all variables; see the Supplementary Material. Denote this bootstrapped sequence on Uniform margins by $U_t^B$. We transform $U_t^B$ back to the original scale using our fitted model, preserving the original structure of $Y_t$; we denote this series $Y_t^B$. Then we fit our model to $Y_t^B$ to re-estimate all of the parameters and thus the quantile of interest. We repeat this procedure to obtain 200 bootstrap samples.

### 6.3.4 Results

For C1, we use our final model of Section 6.3.2 to estimate the 0.9999-quantile of $Y \mid \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_i$, $i \in \{1, \ldots, 100\}$, for the set of 100 covariate combinations. The left panel of Figure 6.3.2 shows the quantile-quantile (QQ) plot for our model. There is general alignment between the model and empirical quantiles; however, there is some over-estimation in the upper tail, and our 95% tolerance bounds do not contain some of the most extreme response values. The right panel of Figure 6.3.2 shows our predicted quantiles, and their associated confidence intervals, compared to their true quantiles. As expected, our predictions tend to over-estimate the true quantiles. We note this figure is different from the one presented by Rohrbeck et al. (2024) due to an error in our code being fixed after submission. In this scenario, our estimated confidence intervals lead to a 14% coverage of the true quantiles, which does not alter our ranking for this challenge. Our performance and model improvements are discussed in Section

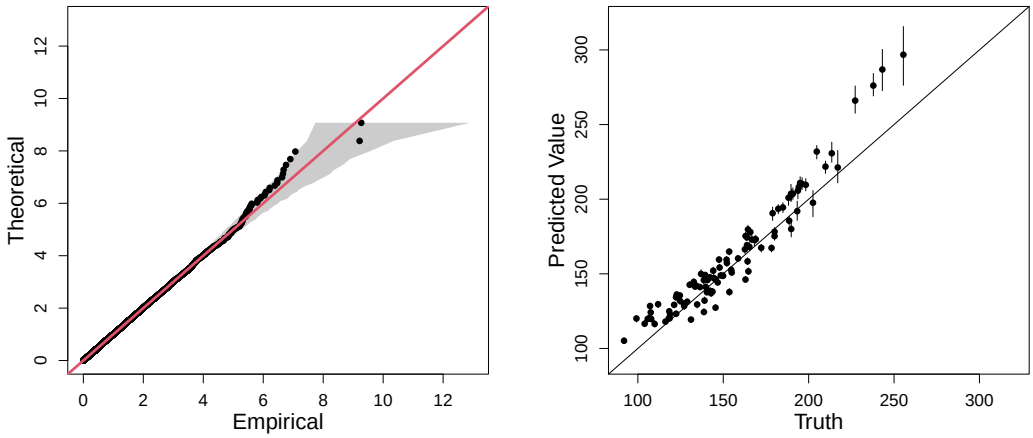



Figure 6.3.2: QQ plot for our final model (model 7 in Table 6.3.1) on standard exponential margins. The $y = x$ line is given in red and the grey region represents the 95% tolerance bounds (left). Predicted 0.9999−quantiles against true quantiles for the 100 covariate combinations. The points are the median predicted quantile over 200 bootstrapped samples and the vertical error bars are the corresponding 50% confidence intervals. The $y = x$ line is also shown (right).

6.6.

For challenge C2, we estimate the quantile of interest as $\hat{q} = 213.1$ $(209.3, 242.1)$. A 95% confidence interval for the estimate is given in parentheses based on the bootstrapping procedure outlined in Section 6.3.2. Due to a coding error, this value differs from the original estimate submitted for the EVA (2023) Conference Data Challenge. The updated value over-estimates compared to the truth ($q = 196.6$).

## 6.4 Challenge C3

### 6.4.1 Exploratory data analysis

For challenge C3, we are provided with 70 years of daily data of an environmental variable for three towns on the island of Coputopia. These data are denoted by $Y_{i,t}$, $i \in \{1, 2, 3\}$, $t \in \{1, \ldots, n\}$, where $i$ is the index of each town and $t$ is the point in

time. Each year consists of 12 months, each lasting 25 days, resulting in $n = 21,000$ observations for each location.

We are also provided with daily covariate observations $\boldsymbol{X}_t = (S_t, A_t)$, where $S_t$ and $A_t$ denote seasonal and atmospheric conditions, respectively. Season is a binary variable, taking values in the set $\{1, 2\}$, with each year of observations exhibiting both seasons for exactly 150 consecutive days. Atmospheric conditions are piecewise constant over months, with large variation in the observed values between months. A descriptive figure of both covariates is given in the Supplementary Material.

In Rohrbeck et al. (2024), we are informed that $Y_{i,t}$ are distributed identically across all sites and over time, with standard Gumbel margins. However, it is not known whether the covariates $\boldsymbol{X}_t$ influence the dependence structure of $\boldsymbol{Y}_t := (Y_{1,t}, Y_{2,t}, Y_{3,t})$. We are also informed that, conditioned on covariates, the process is independent over time, i.e., $(\boldsymbol{Y}_t \mid \boldsymbol{X}_t) \perp\!\!\!\perp (\boldsymbol{Y}_{t'} \mid \boldsymbol{X}_{t'})$ for any $t \neq t'$. In this section, we examine what influence, if any, the covariate process $\boldsymbol{X}_t$ may have on the dependence structure of $\boldsymbol{Y}_t$.

We begin by transforming the time series $Y_{i,t}$ to standard exponential margins, denoted by $\boldsymbol{Z}_{i,t}$, via the probability integral transform. This transformation is common in the study of multivariate extremes and can simplify the description of extremal dependence (Keef et al., 2013a). To explore the extremal dependence in the Coputopia time series, we consider all 2- and 3-dimensional subvectors of the process, i.e., $\{Z_{i,t}, i \in I, t \in \{1, \ldots, n\}\},\ I \in \mathcal{I} := \{\{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$. This separation is important to ensure the overall dependence structure is fully understood, since intermediate scenarios can exist where a random vector exhibits $\chi = 0$, but $\chi > 0$ for some 2-dimensional subvector(s) (Simpson et al., 2020).

Furthermore, to explore the impact of covariates on the dependence structure, we partition the time series into subsets using the covariates. For the seasonal covariate, let $G^S_{I,j} := \{Z_{i,t}, i \in I, S_t = j\}$ for $j = 1, 2$, and for the atmospheric covariate, let $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ denote the permutation associated with the order statistics

of $A_t$, defined so that ties in the data are accounted for. We then split the data into 10 equally sized subsets corresponding to the atmospheric order statistics, i.e., $G^A_{I,k} := \left\{Z_{i,t}, i \in I, t \in \Sigma^k\right\}$ for $k = 1, 2, \ldots, 10$, where $\Sigma^k := \{t \mid (k-1)n/10 + 1 \leq \pi(t) \leq kn/10\}$. Thus, the atmospheric values associated with each subset $G^A_{I,k}$ will increase over $k$.

The idea behind these subsets is to examine whether altering the values of either covariate impacts the extremal dependence structure. Consequently, we set $u = 0.9$ and estimate $\chi(u)$ using the techniques outlined in Section 6.2, with uncertainty quantified through bootstrapping with 200 samples. The bootstrapped $\chi$ estimates for $G^A_{I,k}$ with $I = \{1, 2, 3\}$ are given in Figure 6.4.1. The plots for the remaining index sets in $\mathcal{I}$, along with the subsets associated with the seasonal covariate, are given in the Supplementary Material. The estimates of $\chi$ appear to vary, in the majority of cases, across both subset types (seasonal and atmospheric), suggesting both covariates have an impact on the dependence structure. For the atmospheric process in particular, the values of $\chi$ tend to decrease for higher atmospheric values, suggesting a negative association between the strength of positive extremal dependence and the atmospheric covariate. We also observe that across all subsets, $\chi$ appears consistently low in magnitude, suggesting the extremes of some, if not all, of the sub-vectors are unlikely to occur simultaneously. As such, for modelling the Coputopia time series, we require a framework that can capture such forms of dependence. We also consider pointwise estimates of the function $\lambda(\cdot)$, as defined later in equation (6.4.1), over $G^S_{I,j}$ and $G^A_{I,k}$ for fixed simplex points; these results are given in the Supplementary Material. Similar to $\chi$, estimates of $\lambda(\cdot)$ vary significantly across subsets, providing additional evidence of non-stationarity within extremal dependence structure.

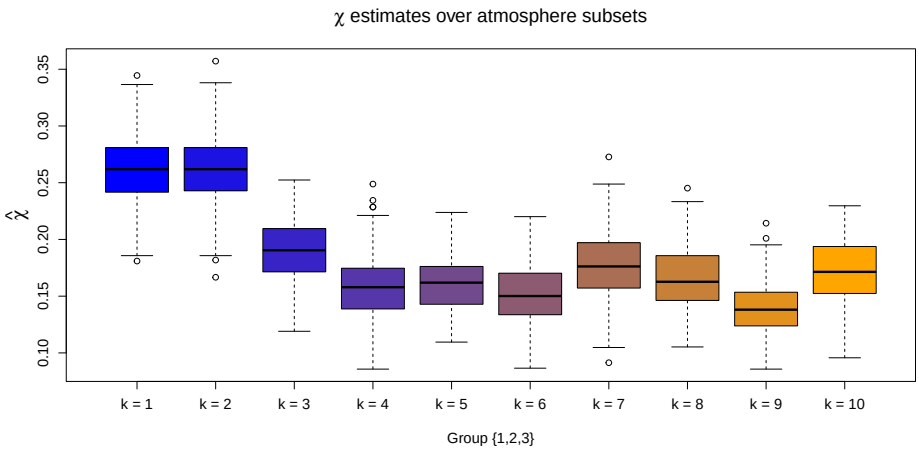



Figure 6.4.1: Boxplots of empirical $\chi$ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2, 3\}$. The colour transition (from blue to orange) over $k$ illustrates the trend in $\chi$ estimates as the atmospheric values are increased.

### 6.4.2 Modelling of joint tail probabilities under asymptotic independence

For challenge C3, we are required to estimate probabilities $p_1 := \Pr(Y_1 > y, Y_2 > y, Y_3 > y)$ and $p_2 := \Pr(Y_1 > v, Y_2 > v, Y_3 < m)$, with $y = 6$, $v = 7$ and $m = -\log(\log(2))$. Note that $p_1$ and $p_2$ are independent of the covariate process and correspond to different extremal regions in $\mathbb{R}^3$; we refer to $p_1$ and $p_2$ as parts 1 and 2 of the challenge, respectively. For the remainder of this section we will consider the transformed exponential variables $(Z_1, Z_2, Z_3)$, omitting the subscript $t$ for ease of notation. Observe that $F_{(-Z_3)}(z) = e^z$, for $z < 0$; setting $\tilde{Z}_3 := -\log(1 - \exp(-Z_3))$, we have

$$p_2 = \Pr(Z_1 > \tilde{v}, Z_2 > \tilde{v}, Z_3 < \tilde{m}) = \Pr\left(Z_1 > \tilde{v}, Z_2 > \tilde{v}, \tilde{Z}_3 > \tilde{m}\right),$$

where $\tilde{v}$ and $\tilde{m}$ denote the values $v$ and $m$ transformed to the standard exponential scale, e.g., $\tilde{v} := -\log(1 - \exp(-\exp(-v)))$. Similarly, we have $p_1 = \Pr(Z_1 > \tilde{y}, Z_2 > \tilde{y}, Z_3 > \tilde{y})$. Consequently, both $p_1$ and $p_2$ can be considered as joint survivor probabilities.

Since not all extremes of $Z_1$, $Z_2$ and $Z_3$ are observed simultaneously, we employ the framework by Wadsworth and Tawn (2013), which is a generalisation of the approach

proposed in Ledford and Tawn (1996). The model of Wadsworth and Tawn (2013) assumes that for any ray $\boldsymbol{\omega} \in \boldsymbol{S}^2 := \{(\omega_1, \omega_2, \omega_3) \in [0,1]^3 : \omega_1 + \omega_2 + \omega_3 = 1\}$, where $\boldsymbol{S}^2$ denotes the standard 2-dimensional simplex,

$$\begin{aligned}\Pr\left(Z_1 > \omega_1 r,\, Z_2 > \omega_2 r,\, Z_3 > \omega_3 r\right) =& \Pr\left(\min\{Z_1/\omega_1,\, Z_2/\omega_2,\, Z_3/\omega_3\} > r\right) \\ =& \mathcal{L}(e^r; \boldsymbol{\omega}) e^{-r\lambda(\boldsymbol{\omega})}, \end{aligned} \tag{6.4.1}$$

as $r \to \infty$, where $\lambda(\boldsymbol{\omega}) \geq \max(\boldsymbol{\omega})$ is known as the angular dependence function (ADF). Asymptotic dependence occurs at the lower bound, i.e., $\lambda(\boldsymbol{\omega}) = \max(\boldsymbol{\omega})$ for all $\boldsymbol{\omega} \in \boldsymbol{S}^2$, and the coefficient of tail dependence is related to the ADF via $\eta = 1/\{3\lambda(1/3, 1/3, 1/3)\}$. In practice, equation (6.4.1) can be used to evaluate extreme joint survivor probabilities; in particular, probabilities $p_1$ and $p_2$ can be identified with the rays $\boldsymbol{\omega}^{(1)} := (\tilde{y}, \tilde{y}, \tilde{y})/r^{(1)}$ and $\boldsymbol{\omega}^{(2)} := (\tilde{v}, \tilde{v}, \tilde{m})/r^{(2)}$ in $\boldsymbol{S}^2$, respectively, where $r^{(1)} := \tilde{y} + \tilde{y} + \tilde{y}$ and $r^{(2)} := \tilde{v} + \tilde{v} + \tilde{m}$. See Section 6.4.4 for further details.

### 6.4.3 Accounting for non-stationary dependence

In the stationary setting, pointwise estimates of $\lambda(\cdot)$ can be obtained via the Hill estimator (Hill, 1975), from which tail probabilities can be approximated. However, alternative procedures are required for data exhibiting trends in dependence, such as the Coputopia data set. Existing approaches for capturing non-stationary dependence structures are sparse in the extremes literature, and most approaches are limited to asymptotically dependent data structures. For the case when data are not asymptotically dependent, Mhalla et al. (2019) and Murphy-Barltrop and Wadsworth (2024) propose non-stationary extensions of the Wadsworth and Tawn (2013) framework, while Jonathan et al. (2014) and Guerrero et al. (2023) propose non-stationary extensions of the Heffernan and Tawn (2004) model (see Murphy-Barltrop and Wadsworth (2024) for a detailed review).

To account for non-stationary dependence in C3, we propose an extension of the Wadsworth and Tawn (2013) framework. With $\boldsymbol{Z}_t = (Z_{1,t}, Z_{2,t}, Z_{3,t})$ and $\boldsymbol{X}_t$, defined as in Section 6.4.1, we define the structure variable $T_{\boldsymbol{\omega},t} := \min\{Z_{1,t}/\omega_1, Z_{2,t}/\omega_2, Z_{3,t}/\omega_3\}$, for any $\boldsymbol{\omega} \in \boldsymbol{S}^2$; we refer to $T_{\boldsymbol{\omega},t}$ as the min-projection variable at time $t$. From Section 6.4.1, we know that the joint distribution of $\boldsymbol{Z}_t$ is not identically distributed over $t$; which implies non-stationarity in the distribution of $T_{\boldsymbol{\omega},t}$. To account for this, Mhalla et al. (2019) and Murphy-Barltrop and Wadsworth (2024) assume the following model given the vector of covariates $\boldsymbol{x}_t$:

$$\Pr\left(T_{\boldsymbol{\omega},t} > u \mid \boldsymbol{X}_t = \boldsymbol{x}_t\right) = \mathcal{L}\left(e^u \mid \boldsymbol{\omega}, \boldsymbol{x}_t\right) e^{-\lambda(\boldsymbol{\omega};\boldsymbol{x}_t)u} \text{ as } u \to \infty, \tag{6.4.2}$$

for all $t$, where $\lambda\left(\cdot; \boldsymbol{x}_t\right)$ denotes the non-stationary ADF. Note that this assumption is very similar in form to equation (6.4.1), with the primary difference being the function $\lambda(\cdot; \boldsymbol{x}_t)$ is non-stationary over $t$. From equation (6.4.2), we have

$$\Pr\left(T_{\boldsymbol{\omega},t} - u > z \mid T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\right) = e^{-\lambda(\boldsymbol{\omega};\boldsymbol{x}_t)z} \text{ as } u \to \infty, \tag{6.4.3}$$

for $z > 0$. Consequently, equation (6.4.2) is equivalent to assuming $(T_{\boldsymbol{\omega},t} - u) \mid \{T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\} \sim \text{Exp}(\lambda\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right))$ as $u \to \infty$.

We found that equation (6.4.2) was not flexible enough to capture the tail of $T_{\boldsymbol{\omega},t}$ for the Coputopia data; see Section 6.4.3 for further discussion. Thus, we propose the following model: given any $z > 0$ and a fixed $\boldsymbol{\omega} \in \boldsymbol{S}^2$, we assume

$$\Pr\left(T_{\boldsymbol{\omega},t} - u > z \mid T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\right) = \left(1 + \frac{\xi\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right) z}{\sigma\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right)}\right)^{-1/\xi(\boldsymbol{\omega};\boldsymbol{x}_t)} \text{ as } u \to \infty, \tag{6.4.4}$$

where $\sigma(\cdot; \boldsymbol{x}_t), \xi(\cdot; \boldsymbol{x}_t)$ are non-stationary scale and shape parameter functions, respectively. This is equivalent to assuming $(T_{\boldsymbol{\omega},t} - u) \mid \{T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\} \sim \text{GPD}(\sigma\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right), \xi\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right))$ as $u \to \infty$, and equation (6.4.3) is recovered by taking the limit as

$\xi(\boldsymbol{\omega}; \boldsymbol{x}_t) \to 0$ for all $t$.

Our proposed formulation in equation (6.4.4) allows for additional flexibility within the modelling framework by including a GPD shape parameter $\xi(\boldsymbol{\omega}; \boldsymbol{x}_t)$, which quantifies the tail behaviour of $T_{\boldsymbol{\omega},t}$. Given the wide range of distributions in the domain of attraction of a GPD (Pickands, 1975), it is reasonable to assume that the tail of $T_{\boldsymbol{\omega},t}$ can be approximated by equation (6.4.4). For the Coputopia time series, this assumption appears valid, as demonstrated by the diagnostics in Section 6.4.3.

### Model fitting

To apply equation (6.4.4), we first fix $\boldsymbol{\omega} \in \boldsymbol{S}^2$ and assume that the formulation holds approximately for some sufficiently high threshold level from the distribution of $T_{\boldsymbol{\omega},t}$; we denote the corresponding quantile level by $\tau \in (0, 1)$. For simplicity, the same quantile level is considered across all $t$. Further, let $v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)$ denote the corresponding threshold function, i.e., $\Pr(T_{\boldsymbol{\omega},t} \leq v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t) \mid \boldsymbol{X}_t = \boldsymbol{x}_t) = \tau$ for all $t$. Under our assumption, we have $(T_{\boldsymbol{\omega},t} - v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)) \mid \{T_{\boldsymbol{\omega},t} > v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t), \boldsymbol{X}_t = \boldsymbol{x}_t\} \sim \text{GPD}(\sigma(\boldsymbol{\omega}; \boldsymbol{x}_t), \xi(\boldsymbol{\omega}; \boldsymbol{x}_t))$. We emphasise that $v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)$ is not constant in $t$, and we would generally expect $v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t) \neq v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_{t'})$ for $t \neq t'$.

As detailed in Section 6.4.2, both $p_1$ and $p_2$ can be associated with points on the simplex $\boldsymbol{S}^2$, denoted by $\boldsymbol{\omega}^{(1)}$ and $\boldsymbol{\omega}^{(2)}$, respectively. Letting $\boldsymbol{\omega} \in \{\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}\}$, our estimation procedure consists of two stages: estimation of the threshold function $v_\tau(\boldsymbol{\omega}, \boldsymbol{z}_t)$ for a fixed $\tau \in (0, 1)$, followed by estimation of GPD parameter functions $\sigma(\boldsymbol{\omega}; \boldsymbol{x}_t), \xi(\boldsymbol{\omega}; \boldsymbol{x}_t)$. For both steps, we take a similar approach to Section 6.3.2 and use GAMs to capture these covariate relationships. To simplify our approach, we falsely assume that the atmospheric covariate $A_t$ is continuous over $t$; this step allows us to utilise GAM formulations containing smooth basis functions. Given the significant variability in $A_t$ between months, discrete formulations for this covariate would significantly increase the number of model parameters and result in higher variability.

Let $\log(v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)) = \psi_v(\boldsymbol{x}_t)$, $\log(\sigma(\boldsymbol{\omega}; \boldsymbol{x}_t)) = \psi_\sigma(\boldsymbol{x}_t)$ and $\xi(\boldsymbol{\omega}; \boldsymbol{x}_t) = \psi_\xi(\boldsymbol{x}_t)$ denote the GAM formulations of each function, where $\psi_-$ denotes the basis representation of equation (6.3.4). Exact forms of basis functions are specified in Section 6.4.3. As in Section 6.3.2, model fitting is carried out using the `evgam` software package (Youngman, 2022). For the first stage, $v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)$ is estimated by exploiting a link between the loss function typically used for quantile regression and the asymmetric Laplace distribution (Yu and Moyeed, 2001). The spline coefficients associated with $\psi_\sigma$ and $\psi_\xi$ are estimated subsequently using the obtained threshold exceedances.

### Selection of GAM formulations and diagnostics

Prior to estimation of the threshold and parameter functions, we specify a quantile level $\tau$ and formulations for each of the GAMs. To begin, we fix $\tau = 0.9$ and consider a variety of formulations for each $\psi_v, \psi_\sigma$ and $\psi_\xi$. By comparing metrics for model selection, namely AIC, BIC and CRPS, we found the following formulations to be sufficient

$$\psi_v(\boldsymbol{x}_t) = \beta_v + s_v(a_t) + \beta_s \mathbb{1}(s_t = 2), \quad \psi_\sigma(\boldsymbol{x}_t) = \beta_\sigma + s_\sigma(a_t) \quad \text{and} \quad \psi_\xi(\boldsymbol{x}_t) = \beta_\xi, \tag{6.4.5}$$

for parts 1 and 2, where $\beta_v, \beta_\sigma, \beta_\xi \in \mathbb{R}$ denote constant intercept terms, $\mathbb{1}$ denotes the indicator function with corresponding coefficient $\beta_s \in \mathbb{R}$, and $s_v, s_\sigma$ denote cubic regression splines of dimension 10. The shape parameter is set to constant for the reasons outlined in Section 6.3.2. Cubic basis functions are used for $\psi_v$ and $\psi_\sigma$ since they have several desirable properties, including continuity and smoothness (Wood, 2017). A dimension of size 10 appears more than sufficient to capture the trends relating to the atmosphere variable. Alternative formulations were tested for both parts, but this made little difference to the resulting model fits.

We remark that the seasonal covariate is only present with the formulation for $\psi_v$. Once accounted for in the non-stationary threshold, the seasonal covariate appeared to

have little influence on the fitted GPD parameters. More complex GAM formulations were tested involving interaction terms between the seasonal and atmospheric covariates, which showed little to no improvement in model fits. Thus, we prefer the simpler formulations on the basis of parsimony.

With GAM formulations selected, we now consider the quantile level $\tau \in (0, 1)$. To assess sensitivity in our formulation, we set $\mathrm{T} := \{0.8, 0.81, \ldots, 0.99\}$ and fit the GAMs outlined in equation (6.4.5) for each $\tau \in \mathrm{T}$. Letting $\delta_{\boldsymbol{\omega},t}$ and $\mathcal{T}_\tau := \{t \in \{1, \ldots, n\} \mid \delta_{\boldsymbol{\omega},t} > v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)\}$ denote the min-projection observations and indices of threshold-exceeding observations, respectively, we expect the set $\mathcal{E} := \{-\log\{1 - F_{GPD}(\delta_{\boldsymbol{\omega},t} - v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)) \mid \sigma(\boldsymbol{\omega}; \boldsymbol{x}_t), \xi(\boldsymbol{\omega}; \boldsymbol{x}_t)\} \mid t \in \mathcal{T}_\tau\}$ to follow a standard exponential distribution.

With all exceedances transformed to a unified scale, we compare the empirical and model exponential quantiles using QQ plots, through which we assess the relative performance of each $\tau \in \mathrm{T}$. We selected $\tau$ values for which the empirical and theoretical quantiles appeared most similar in magnitude. From this analysis, we set $\tau = 0.83$ and $\tau = 0.85$ for parts 1 and 2, respectively. The corresponding QQ plots are given in Figure 6.4.2, where we observe reasonable agreement between the empirical and theoretical quantiles. However, whilst these values appeared optimal within T, we stress that adequate model fits were also obtained for other quantile levels, suggesting our modelling procedure is not particularly sensitive to the exact choice of quantile. Furthermore, we also tested a range of quantile levels below the 0.8-level, but were unable to improve the quality of model fits.

Plots illustrating the estimated GPD scale parameter functions are given in the Supplementary Material, with the resulting dependence trends in agreement with the observed trends from Section 6.4.1. We also remark that the estimated GPD shape parameters obtained for parts 1 and 2 were $0.042\,(0.01, 0.075)$ and $0.094\,(0.059, 0.128)$, respectively, where the brackets denote 95% confidence intervals obtained using pos-

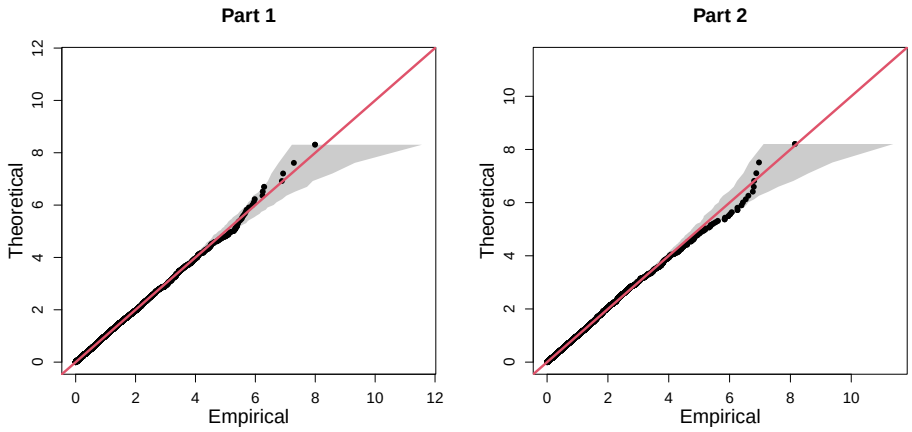



Figure 6.4.2: Final QQ plots for parts 1 (left) and 2 (right) of C3, with the $y = x$ line given in red. In both cases, the grey regions represent the 95% bootstrapped tolerance bounds.

terior sampling (Wood, 2017). These estimates, which indicate slightly heavy-tailed behaviour within the min-projection variable, provide insight into why the original exponential modelling framework is not appropriate for C3.

Overall, these results suggest different extremal dependence trends exist for the two simplex points $\boldsymbol{\omega}^{(1)}$ and $\boldsymbol{\omega}^{(2)}$, illustrating the importance of the flexibility in our model. These findings are also in agreement with empirical trends observed in Section 6.4.1, suggesting our modelling framework is successfully capturing the underlying extremal dependence structures.

### 6.4.4 Results

Given estimates of threshold and parameter functions, probability estimates can be obtained via Monte Carlo techniques. Taking $p_1$, for instance, we have

$$
\begin{aligned}
p_1 =& \Pr(Z_1 > \tilde{y}, Z_2 > \tilde{y}, Z_3 > \tilde{y}) \\
=& \Pr\left(\min\left(Z_1/\omega_1^{(1)}, Z_2/\omega_2^{(1)}, Z_3/\omega_3^{(1)}\right) > r^{(1)}\right) \\
=& \int_{\boldsymbol{X}_t} \Pr\left(T_{\boldsymbol{\omega}^{(1)},\, t} > r^{(1)} \mid \boldsymbol{X}_t = \boldsymbol{x}_t\right) f_{\boldsymbol{X}_t}(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t \\
=& (1-\tau) \int_{\boldsymbol{X}_t} \Pr(T_{\boldsymbol{\omega}^{(1)},\, t} > r^{(1)} \mid T_{\boldsymbol{\omega}^{(1)},\, t} > v_\tau(\boldsymbol{\omega}^{(1)}, \boldsymbol{x}_t), \boldsymbol{X}_t = \boldsymbol{x}_t) f_{\boldsymbol{X}_t}(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t \\
\approx& \frac{1-\tau}{n} \sum_{t=1}^{n} \left(1 + \frac{\xi(\boldsymbol{\omega}^{(1)}; \boldsymbol{x}_t)\left(r^{(1)} - v_\tau(\boldsymbol{\omega}^{(1)}, \boldsymbol{x}_t)\right)}{\sigma\left(\boldsymbol{\omega}^{(1)}; \boldsymbol{x}_t\right)}\right)^{-1/\xi\left(\boldsymbol{\omega}^{(1)}; \boldsymbol{x}_t\right)},
\end{aligned}
$$

assuming $\{\boldsymbol{x}_t : t \in \{1, \ldots, n\}\}$ is a representative sample from $\boldsymbol{X}_t$. The procedure for $p_2$ is analogous. We note that this estimation procedure is only valid when $r^{(1)} > v_\tau(\boldsymbol{\omega}^{(1)}, \boldsymbol{x}_t)$, or $r^{(2)} > v_\tau(\boldsymbol{\omega}^{(2)}, \boldsymbol{x}_t)$, for all $t$: however, for each $\tau \in \mathrm{T}$, this inequality is always satisfied, owing to the very extreme nature of the probabilities in question. Through this approximation, we obtain $\hat{p}_1 = 1.480 \times 10^{-5}$ and $\hat{p}_2 = 2.461 \times 10^{-5}$.

## 6.5 Challenge C4

### 6.5.1 Exploratory data analysis

Challenge C4 entails estimating survival probabilities across 50 locations on the island of Utopula. As stated in Rohrbeck et al. (2024), the Utopula island is split in two administrative areas, for which the respective regional governments 1 and 2 have collected data concerning the variables $Y_{i,t}$, $i \in I = \{1, \ldots, 50\}$, $t \in \{1, \ldots, 10,000\}$. Index $i$ denotes the $i^{\text{th}}$ location, with locations $i \in \{1, \ldots, 25\}$ and $i \in \{26, \ldots, 50\}$ belonging to the administrative areas of governments 1 and 2, respectively. Index $t$ denotes the time point in days; however, since $Y_{i,t}$ are IID for all $i$, we drop the subscript $t$ for the remainder of this section.

Since many multivariate extreme value models are only applicable in low-to-moderate dimensions, we consider dimension reduction based on an exploration of the extremal

dependence structure of the data. In particular, we analyse pairwise estimates of the extremal dependence coefficient $\chi(u)$, introduced in equation (6.2.2), for all possible pairwise combinations of sites; the resulting estimates, using $u = 0.95$, are presented in the heat map of Figure 6.5.1. Identification of any dependence clusters is achieved through visual investigation, which seems appropriate for this data. We note, however, that should visual considerations not suffice, alternative more sophisticated clustering methods are available and can be applied; see for example Bernard et al. (2013).

Figure 6.5.1 suggests the existence of 5 distinct subgroups where all variables within each subgroup have similar extremal dependence characteristics, while variables in different subgroups appear to be approximately independent of each other in the extremes. It is worth mentioning that the same clusters are identified when we analyse pairwise estimates of the extremal dependence coefficient $\eta(u)$; the resulting estimates can be found in the Supplementary Material. Moreover, examining the magnitudes of $\chi(\cdot)$ and $\eta(\cdot)$ estimates, it does not appear reasonable to assume asymptotic dependence between variables in the same group. We therefore consider models that can be applied to data structures that do not take their extreme values simultaneously. The indices of the five aforementioned subgroups are $G_1 = \{4, 14, 19, 28, 30, 38, 43, 44\}$, $G_2 = \{3, 10, 15, 18, 22, 29, 45, 47\}$, $G_3 = \{8, 21, 25, 26, 32, 33, 34, 40, 41, 42, 48, 49, 50\}$, $G_4 = \{1, 2, 5, 7, 9, 17, 20, 31, 46\}$ and $G_5 = \{6, 11, 12, 13, 16, 23, 24, 27, 35, 36, 37, 39\}$. Groups $G_1$ and $G_2$ include the most strongly dependent variables (shown by the darkest color blocks in Figure 6.5.1), followed by group $G_3$, while groups $G_4$ and $G_5$ contain the most weakly dependent variables. We henceforth assume independence between these groups of variables, i.e., $\Pr((Y_i)_{i\in G_k} \in A_k, (Y_i)_{i\in G_{k'}} \in A_{k'}) = \Pr((Y_i)_{i\in G_k} \in A_k) \Pr((Y_i)_{i\in G_{k'}} \in A_{k'})$, $A_k \subset \mathbb{R}^{|G_k|}, A_{k'} \subset \mathbb{R}^{|G_{k'}|}$, for any $k \neq k' \in \{1, \ldots, 5\}$.

Challenge C4 requires us to estimate the probabilities $p_1 = \Pr(Y_i > s_i;\, i \in I)$ and $p_2 = \Pr(Y_i > s_1;\, i \in I)$, where $s_i := \mathbb{1}(i \in \{1, 2, \ldots, 25\})s_1 + \mathbb{1}(i \in \{26, 27, \ldots, 50\})s_2$ and $s_1$ ($s_2$) denotes the marginal level exceeded once every year (month) on average.

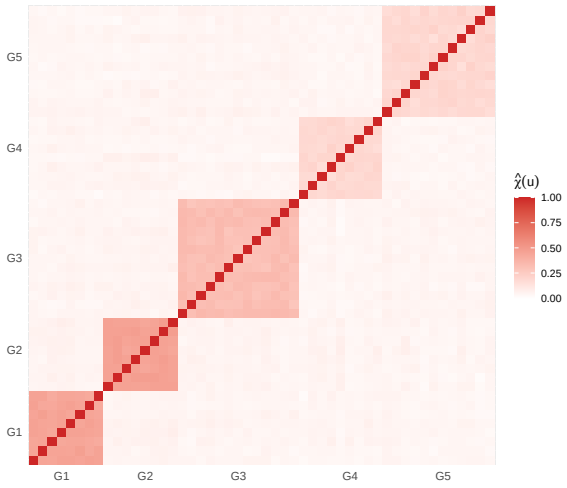



Figure 6.5.1: Heat map of estimated empirical pairwise $\chi(u)$ extremal dependence coefficients with $u = 0.95$.

Under the assumption of independence between groups, the challenge can be broken down to 5 lower-dimensional challenges involving the estimation of joint tail probabilities for each $G_k$, $k \in \{1, \ldots, 5\}$. These can then be multiplied together to obtain the required overall probabilities due to (assumed) between-group independence; specifically, we have $p_1 = \prod_{k=1}^{5} \Pr(Y_i > s_i;\ i \in G_k)$ and $p_2 = \prod_{k=1}^{5} \Pr(Y_i > s_1;\ i \in G_k)$.

### 6.5.2 Conditional extremes

The conditional multivariate extreme value model (CMEVM) of Heffernan and Tawn (2004) provides a flexible multivariate extreme value framework capable of capturing a range of extremal dependence forms without making assumptions about the specific form of joint dependence structure. Consider a $d$-dimensional random variable $\boldsymbol{W} = (W_1, \ldots, W_d)$ on standard Laplace margins. For $i \in \{1, \ldots, d\}$, the CMEVM approach assumes the existence of parameter vectors $\boldsymbol{\alpha}_{-|i} \in [-1, 1]^{d-1}$ and $\boldsymbol{\beta}_{-|i} \in (-\infty, 1]^{d-1}$ such that

$$\lim_{u_i \to \infty} \Pr\left\{ \boldsymbol{W}_{-i} \leq \boldsymbol{\alpha}_{-|i} W_i + W_i^{\boldsymbol{\beta}_{-|i}} \boldsymbol{z}_{|i}, W_i - u_i > w \mid W_i > u_i \right\} = e^{-w} H_{|i}\left(\boldsymbol{z}_{|i}\right), \quad w > 0,$$

with non-degenerate distribution function $H_{|i}(\cdot)$, vector operations being applied componentwise, and conditional threshold $u_i$. The vector $\boldsymbol{W}_{-i}$ denotes $\boldsymbol{W}$ excluding its $i^{\text{th}}$ component and $\boldsymbol{z}_{|i}$ is within the support of the residual random vector $\boldsymbol{Z}_{|i} = (\boldsymbol{W}_{-i} - \boldsymbol{\alpha}_{-|i} w_i)/w_i^{\boldsymbol{\beta}_{-|i}} \sim H_{|i}(\cdot)$. We apply this model to data where $W_i > u_i$, for some finite conditioning threshold $u_i$, to estimate the probabilities $p_1$ and $p_2$ defined in Section 6.5.1, using the inference procedure of Keef et al. (2013a).

### 6.5.3 Results

Let $\boldsymbol{W} := (W_1, \ldots, W_{50})$ denote the random vector after transformation to standard Laplace margins. This vector is divided into the five subgroups identified in Section 6.5.1, and the subgroup probabilities are estimated using predictions obtained from the sampling method of Heffernan and Tawn (2004). We condition on the first variable of each subgroup being extreme, and simulate $10^8$ predictions from each of the resulting fitted conditional extremes models. To account for uncertainty in the estimates, we perform a parametric bootstrapping procedure with 100 samples.

Sensitivity analyses of the estimated probabilities to the choice of conditioning variable suggest no significant effect. Furthermore, we consider a range of conditioning thresholds; the corresponding estimates of subgroup probabilities defined in Section 6.5.1 appear relatively stable with respect to the conditioning threshold quantile. We ultimately select 0.85-quantiles for the conditioning thresholds of our final probability estimates. These are given by $\hat{p}_1 = 1.094 \times 10^{-26}$ $(2.150 \times 10^{-36}, 1.359 \times 10^{-24})$ and $\hat{p}_2 = 1.076 \times 10^{-31}$ $(1.596 \times 10^{-46}, 1.850 \times 10^{-29})$, with 95% confidence intervals obtained from parametric bootstrapping given in parentheses.

## 6.6 Discussion

In this paper, we have proposed a range of statistical methods for estimating extreme quantities for challenges C1-C4. For the univariate challenge C1, we estimated the 0.9999-quantile, and the associated 50% confidence intervals, of $Y \mid \boldsymbol{X} = \boldsymbol{x_i}$, $i \in \{1, \ldots, n\}$. For challenge C2, we estimated a quantile, corresponding to a once in 200 year level, of the marginal distribution $Y$ whilst incorporating the loss function in equation (6.3.2). Overall we ranked $6^{\text{th}}$ and and $4^{\text{th}}$ for challenges C1 and C2, respectively.

For challenge C1, our final model (model 7 in Table 6.3.1) was chosen to minimise the model selection criteria; however, QQ plots showed over-estimation of the most extreme values of the response (see Figure 6.3.2). As a result, the conditional quantiles calculated for C1 are generally over-estimated when compared with the true quantiles. If we ignored the model selection criteria and chose the model based on a visual assessment of QQ plots, we would have chosen model 5 in Table 6.3.1 and this would have covered the true quantile on fewer occasions than our chosen model. Therefore, the main issue with our results concerns the width of the confidence intervals.

Narrow confidence intervals are an indication of over-fitting and this could have arisen in several places. For instance, Rohrbeck et al. (2024) suggested all the seasonality is captured in the threshold, while our model includes a seasonal threshold and a covariate for seasonality in the scale parameter of the GPD model. As well as over-fitting, the model may not have been flexible enough; this could be, in part, due to our model missing covariates. For instance, the true model contained $V_2$ as a covariate (Rohrbeck et al., 2024) whilst our model did not. In addition, the basis dimensions for our splines are low. In practice, a higher dimension than we would expect should be considered and, although we chose the dimension using a model-based approach, it may have resulted in the splines not being flexible enough to capture all of the trends in the data.

Narrow confidence intervals may have also resulted from the choice of uncertainty quantification procedure. Changing the average block length $l$ in our stationary bootstrap procedure would alter the confidence interval widths, although this was carefully chosen to reflect the temporal dependence in the data. Alternative methods, such as the standard bootstrap procedure or the delta method, could be implemented to investigate how this affects the confidence interval widths. We expect that such confidence intervals will be wider than those presented here since the dependence in the data is not accounted for, but assuming temporal independence would be inaccurate. Therefore, whilst adopting an alternative procedure may widen confidence intervals, thus improving our performance, such intervals may not be well calibrated for this data set.

The over-fitting and over-estimation issues encountered in C1 are carried through to C2 since the same model is used for both challenges. However, one aspect specific to C2 is the choice of quantile evaluation within the loss function. Many methods exist for evaluating the non-stationary quantiles which feed into the loss function term of the objective function $S(\boldsymbol{\theta})$ in equation (6.3.5). As the loss function will be dominated by the log-likelihood in $S(\boldsymbol{\theta})$, we choose to transform to standard exponential margins when evaluating the quantiles in order to give more importance to the loss function. Since the data is light tailed ($\xi < 0$) this transformation elongates the tail and therefore inflates any deviations between the model and theoretical quantiles which in turn, inflates the contribution of the average loss function to $S(\boldsymbol{\theta})$. However, this approach means that the objective function will have a preference to minimise the deviations in the upper-tail of the distribution, leading to potential over-fitting to the upper-tail and possibly, a poor fit in the rest of the tail. This may not necessarily be undesirable since the loss function penalises under-estimation more than over-estimation, however, since the model in C1 already over-fits, this method may only exacerbate the problem for C2.

For the first multivariate challenge C3, we employed an extension of the method proposed by Wadsworth and Tawn (2013) to estimate probabilities of three variables

lying in extremal sets. Our extension accounts for non-stationarity in the extremal dependence structure, with GAMs used to represent covariate relationships. The QQ plots for the resulting model suggested reasonable fits. For this challenge, we ranked 5$^{\text{th}}$ and our estimates are on the same order of magnitude as the truth (Rohrbeck et al., 2024).

We note similarities in the methodologies presented for the challenges C1, C2, and C3. Specifically, each of the proposed methods used the EVGAM framework for capturing non-stationary tail behaviour via a generalised Pareto distribution. We acknowledge that the model selection tool proposed for C1 and C2 could also be applied for C3. However, we opted not to use this tool for several reasons. Firstly, unlike the univariate setting, there is no guarantee of convergence to a GPD in the limit, and the GPD tail assumption thereby needs to be tested. Moreover, in exploratory analysis, we tested the model selection tool for C3 but found the selected models and quantiles to not be satisfactory, particularly in the upper tail of the min-projection variable. We therefore selected a model manually, using QQ plots to evaluate performance. Exploring threshold and model selection techniques for multivariate extremes represents an important area of research.

In the final multivariate challenge C4, we estimated very high-dimensional joint survival probabilities. To do so, we split the probability into 5 lower-dimensional components which are assumed independent of each other, then estimated each using the CMEVM of Heffernan and Tawn (2004). In the final rankings of Rohrbeck et al. (2024), we ranked 3$^{\text{rd}}$ for this challenge. A more prudent method could have been implemented, as groups of variables were never truly independent. Alternatively, although we achieve relatively stable probability estimates with respect to threshold in Section 6.5.2 (see Supplementary Material for details), our approach could potentially have been improved by estimating individual group probabilities across varying thresholds and taking an average value as our final result. We also do not report the effect of

the choice of the conditioning variable on our estimates. Preliminary analysis suggested this to be negligible. However, conditioning on each site in a given subgroup and then taking a weighted sum of the resulting probabilities (e.g., Keef et al., 2013b) may have resulted in more robust estimates.

# Chapter 7

# Conclusions and further work

The main goal of this thesis was to develop dependence models which are able to jointly capture the body and tail regions of multivariate data, while ensuring an accurate representation across both regions. To address the computational challenges associated with intensive likelihood evaluation, we also explored the neural Bayes estimation methodology proposed by Sainsbury-Dale et al. (2024a), facilitating the inference process for the proposed models and other available models from the bivariate extremes literature.

In Chapter 3, we introduced a dependence model that jointly captures the body and tail regions of bivariate data, providing a smooth transition between them. A key advantage of the proposed model is its ability to bypass the need for selecting, or estimating, a threshold vector above which the observations are deemed extreme. We have also demonstrated that, even under misspecified dependence structures, the copula model proved sufficiently flexible to capture different behaviours. Finally, when applied to model the relationship between temperature and ozone concentrations, the weighted copula model significantly outperformed the fit provided by a single copula model. Furthermore, we showed that the copula model is able to distinguish between contrasting behaviours in the body and tail regions. More specifically, it identified negative associations in the body, and positive dependence in the tail.

The weighted copula model, however, is computationally expensive due to the need for inversion and numerical integration of functions. As a result, it does not scale well to higher dimensions. While extending the model beyond the bivariate case is theoretically possible, it leads to significant computational challenges or even infeasibility. This issue became evident when we tried to incorporate one of the bivariate models based on random scale constructions from Chapter 2.2.7 as the tail component. Given that these models also rely on inversion of functions and/or numerical integration, they are themselves computationally expensive. This led to the computational time when fitting the weighted copula model being infeasible. Additionally, we need to choose a priori which copula families to include in the weighted copula model; this results in the need for comparing various model specifications to identify the most suitable one for each data set. Given the computational time required to evaluate one likelihood, this process is inevitably intensive.

It may be the case that we are not able to draw conclusions about the extremal dependence with the weighted copula model, i.e., when $\chi(r) > 0$ and $\eta(r) < 1$ as $r \to 1$, which can result in misrepresentation of the extremal region when extrapolating beyond the observed data. This could be addressed by incorporating one of the copulas from Chapter 2.2.7 as the tail component, since these models are capable of representing both regimes of extremal dependence. As this approach proved computationally infeasible, studying the extremal dependence measures for the proposed model is crucial. While we have numerically investigated these properties for specific model configurations, and theoretically derived the measures for one specific case, further exploration is needed to fully describe the extremal region across various configurations.

In Chapter 4, we proposed an alternative copula model based on a mixture of multivariate Gaussian distributions. Similarly to the weighted copula model from Chapter 3, this model is able to represent the body and tail regions of multivariate data without requiring the definition of an extremal region. In contrast to the weighted

copula model from Chapter 3, this model does not require the selection a priori of the copula or distributions to be included in the mixture. Owing to its construction, the model intrinsically exhibits asymptotic independence. Nevertheless, we showed that it is sufficiently flexible to accommodate a wide range of complex extremal dependence structures at near-asymptotic levels, including scenarios with asymptotically dependent or non-exchangeable data. Contrarily to the model from Chapter 3, the Gaussian mixture copula scales effectively to dimensions beyond the bivariate case. More specifically, through simulation studies, we showed that the proposed model is identifiable in a 5-dimensional setting. We illustrated its performance using a 5-dimensional air pollution data set analysed previously by Heffernan and Tawn (2004), and our results showed that the Gaussian mixture copula reasonably characterised the joint behaviour of the variables.

Despite the model scaling well up to $d = 5$, the evaluation of the likelihood of the Gaussian mixture copula becomes increasingly computationally expensive with the dimension. This is primarily due to the high parameterisation of the Gaussian mixture copula. This issue leads to complications in the inference procedure, particularly when results suggest that adding an extra mixture component would be beneficial. As discussed in Chapter 4, exploring the graphical structure of the mixture components presents an interesting avenue for further work, as it might aid the inference procedure. More specifically, identifying potential conditional independence between pairs of variables (within the same mixture component) could be incorporated into the likelihood function, resulting in a reduction in the dimensionality of the Gaussian mixture copula model. The high number of parameters in the model may reduce its interpretability. Further imposing an ordering on all the means of the mixture components, rather than just the first element of each component, i.e., $\mu^i_{j-1} < \mu^i_j$ for all $i \in D$ and $j = 2, \ldots, k$, could improve the interpretability of the Gaussian mixture copula. In this approach, it would be clearer that each mixture component is progressively further into the tails.

In Chapters 3 and 4, we have assumed stationarity, which is a common assumption in practice but may often not be the case in real world applications. Therefore, extending both models to accommodate for non-stationarity in the dependence structure due to covariates across the entire distribution is of interest, and an important avenue for future research. While the typical strategy in multivariate extremes is to focus on trends that might occur in the extreme observations, it is also essential to consider non-stationarity in the body region as only a subset of components may be extreme. In a univariate setting, capturing trends in the data by also considering the non-extreme observations has been studied by Eastoe and Tawn (2009) or by de Carvalho et al. (2022) in a regression context. Although the conditional extremes method proposed by Heffernan and Tawn (2004) (recall Chapter 2.2.6) could be applied in situations where only a subset of variables is extreme, defining a threshold above which the conditioning variables are deemed extreme may introduce discontinuities at the threshold. As we would expect similarities in the trends of the body and tail regions, this approach may not be realistic. Therefore, adapting the weighted copula and the Gaussian mixture copula models to capture changes in the dependence across both regions may be advantageous.

In Chapter 5, we introduced an amortised statistical toolbox for model selection and inference, which leverages neural networks and bypasses the need for likelihood evaluation. In particular, we exploited the utility of neural Bayes estimation (Sainsbury-Dale et al., 2024a) to perform inference on the weighted copula model from Chapter 3, as well as on bivariate models based on random scale constructions, which are able to interpolate between asymptotic independence and dependence at the interior of their parameter space. Through simulation studies, we have shown that the inference process is considerably faster with neural Bayes estimation than maximum likelihood estimation (when this is feasible). Additionally, the derived extremal dependence measures are well calibrated, indicating that the neural Bayes estimator effectively captures the tail be-

haviour. As previously mentioned, incorporating one of the models from Chapter 2.2.7 as the tail component in the weighted copula model from Chapter 3 is computationally infeasible with a likelihood-based approach. With neural Bayes estimation, this is no longer the case, as demonstrated in Chapter 5.4.2 and C.2.1 where we performed inference within such model configurations.

We have also proposed a model selection neural classifier that enables fast and effective selection of the most suitable model from a set of candidates, performing comparably to the BIC in simulation studies. Additionally, we demonstrated that the proposed toolbox for model selection and inference is robust even when the data is not generated by any of the models considered. Finally, when applying the toolbox to study the pairwise extremal dependence of the changes in horizontal geomagnetic field fluctuation across three locations, we showed that it allows for sensitivity analysis to the threshold used to censor non-extremal observations, when the focus is solely on extreme observations. By applying the methodology across a range of threshold levels, we assessed the impact of this choice on the results. Notably, this step would be computationally expensive with a likelihood-based approach.

Simulation studies showed, however, that the estimates given by the NBE were generally more biased compared to those obtained through maximum likelihood inference, when likelihood evaluation was feasible. Such bias led to poor coverage (and wider) bootstrap-based uncertainty intervals. Whilst the coverage probabilities significantly improved when training a second estimator to target specific posterior marginal quantiles, rather than targeting posterior quantities such as the median, doing so results in extra computational cost. Further investigation into reducing the bias and improving coverage rates is of importance; to the best of our knowledge, exploring coverage rates in the NBE context has not been done yet.

One possible approach is to explore different neural network architectures, such as adding more layers or initialising the weights and biases at different values. Addition-

ally, exploring different optimiser algorithms and adjusting their learning rates may lead to different, potentially better, results. Another possibility could be to implement an ensemble approach, where multiple models are trained, each with different weights and biases, and their estimates combined. However, an initial exploration of these strategies did not yield improvements when applied to the models considered in Chapter 5. Finally, the impact of the prior choice on the results can be further explored. Specifically, the simulation studies shown in Chapter C.2.1 suggest that reparameterising certain model parameters improved the inference procedure, reducing the bias observed throughout the analysis of Chapter 5. However, uniform priors do not necessarily lead to uniform priors in alternative parameterisations, and thus more careful consideration of priors is required.

Similarly to Chapter 3, we have restricted our analysis to the bivariate setting. However, given that we no longer rely on likelihood evaluations, applying the NBE to higher dimensions is now available. While this might not be useful for the random scale models from Chapter 2.2.7, as they are only meaningful when all variables exhibit joint asymptotic dependence or independence, the same is not true for the weighted copula model from Chapter 3 and the Gaussian mixture copula from Chapter 4. For these models, the advantages of neural Bayes estimation would be even more pronounced, especially for the model from Chapter 3, since likelihood evaluation becomes computationally infeasible with triple integrals and beyond. In addition, neural Bayes estimation may be useful for incorporating covariates into the models to account for non-stationarity, especially if doing so demands extra computational resources. In this case, we would require a model for the covariates from which we could simulate.

The weighted copula model from Chapter 3 and the Gaussian mixture copula from Chapter 4 both require making a few choices in advance. As mentioned before, for the former, we need to decide which copula families to include and which weighting function to use, while for the latter, the number of mixture components to incorporate into

the model need to be specified. Inevitably, if a likelihood-based inference procedure is adopted, model selection for both copula models becomes computationally expensive, as all possible options have to be fitted for comparison. Furthermore, if no prior knowledge of the data structure can be leveraged into the analysis, this leads to hundreds of possible model specifications — particularly for the weighted copula model — resulting in a tedious and time-consuming process.

On the other hand, if asymptotic (in)dependence is sought, then an asymptotically (in)dependent copula should be chosen for the tail component, possibly along with a monotonic increasing weighting function, such as those used in Chapter 3. Similar considerations can be made for the Gaussian mixture copula from Chapter 4; in a higher dimensional setting and/or when the data exhibits a more complex dependence structure (which can be explored visually, for example), additional mixture components are needed, as shown by the simulation and case studies. Instead, the model selection can be facilitated by adopting a likelihood-free framework, such as the one discussed in Chapter 5. In this case, a neural network would need to be trained on all possible model specifications. However, depending on the number of candidate models, a different — perhaps more flexible — neural network architecture than the one used in Section 5.2.4 may need to be considered.

Lastly, Chapter 6 detailed the contribution of a wider team for the 2023 EVA conference data challenge. This challenge was comprised of four challenges C1-C4, for which we proposed a range of statistical methods to estimate extreme quantiles. For challenge C1, the 0.9999-quantile and its 50% confidence intervals, were estimated for a response variable conditioned on a set of environmental covariates. For challenge C2, we estimated the marginal 200-year return level; to do so we have incorporated a specific loss function. For challenge C3, we extended the method of Wadsworth and Tawn (2013) to account for non-stationarity, whilst using generalised additive models, to estimate probabilities of three variables lying in two different extremal regions. For

the final challenge, C4, we estimated two 50-dimensional joint survival probabilities; we do so by splitting the probabilities into five lower-dimensional components, assumed independent of each other, and estimating each using the conditional extremes approach of Heffernan and Tawn (2004).

For challenge C3 (recall Section 6.4), the threshold above which a GP distribution is assumed for the tail was selected through exploratory analysis. In particular, by considering a range of thresholds, we have assessed the fit of the GP distribution through QQ plots. A similar approach was taken to select the threshold above which the conditioning variable is assumed large in challenge C4 (recall Section 6.5). More specifically, by performing a sensitivity analysis to the choice of conditioning threshold, we showed that the threshold appeared to not have a significant effect on the estimated probabilities, obtaining relatively stable results. Nevertheless, exploring threshold selection methods for multivariate extremes is of interest as, in general, different choices may lead to different conclusions.

To address the high dimensionality presented by challenge C4, we have identified subgroups of variables, which we assumed independent of each other, through visual inspection of pairwise estimates of the extremal dependence coefficient. While this proved sufficient for the Utopula data, considering clustering methods for multivariate extremes (e.g., Bernard et al., 2013, Chautru, 2015 or Janßen and Wan, 2020) to identify lower-dimensional components is an alternative. Finally, we found the choice of conditioning variable to be insignificant for C4. However, since the conditional extremes method (recall Section 2.2.6) is known to not be self-consistent (Liu and Tawn, 2014), and thus different conclusions may be obtained when considering a different conditioning site, more robust estimates might have been obtained by considering a weighted sum of the probabilities considering each site as the conditioning variable instead.

# Appendix A

# Supplementary material for Chapter 3

## A.1 Copula densities

In this appendix we give the copula distribution function $C$ and density function $c$ for all copulas used in the paper.

### Gaussian copula

The Gaussian copula with correlation parameter $\rho \in (-1, 1)$ is given by

$$C(u, v; \rho) = \Phi_2\left(\Phi_1^{-1}(u), \Phi_1^{-1}(v); \rho\right), \quad u,\, v \in (0, 1),$$

where $\Phi_2(\cdot, \cdot; \rho)$ is the bivariate standard normal distribution function with correlation $\rho$ and $\Phi_1^{-1}(\cdot)$ is the inverse of the univariate standard normal distribution function. The Gaussian copula density can be written as

$$c(u, v; \rho) = \frac{1}{\sqrt{1-\rho^2}} \exp\left\{-\frac{\rho^2 x^2 + \rho^2 y^2 - 2\rho xy}{2(1-\rho^2)}\right\}, \quad u,\, v \in (0, 1),$$

where $x = \Phi_1^{-1}(u)$ and $y = \Phi_1^{-1}(v)$.

## Student t copula

The Student t copula with correlation parameter $\rho \in (-1, 1)$ and $\nu > 0$ degrees of freedom is given by

$$C(u, v; \rho, \nu) = T_{2,\nu}\left(T_\nu^{-1}(u), T_\nu^{-1}(v); \rho\right), \quad u,\, v \in (0, 1),$$

where $T_{2,\nu}(\cdot, \cdot; \rho)$ is the bivariate t distribution function with correlation parameter $\rho$ and $T_\nu^{-1}(\cdot)$ is the inverse of the univariate t distribution function. The Student t copula density can be written as

$$c(u, v; \rho, \nu) = \frac{1}{\sqrt{1-\rho^2}} \frac{\Gamma\left(\frac{\nu+2}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)^2} \frac{\left[\left(1+\frac{x^2}{\nu}\right)\left(1+\frac{y^2}{\nu}\right)\right]^{(\nu+1)/2}}{\left[1+\frac{(x^2+y^2-2\rho sr)}{\nu(1-\rho)^2}\right]^{(\nu+2)/2}}, \quad u,\, v \in (0, 1),$$

where $x = T^{-1}(u)$ and $y = T^{-1}(v)$.

## Frank copula

The Frank copula with parameter $\alpha \in \mathbb{R} \setminus \{0\}$ is given by

$$C(u, v; \alpha) = -\frac{1}{\alpha} \log\left(1 - \frac{(1-e^{-\alpha u})(1-e^{-\alpha v})}{1-e^{-\alpha}}\right), \quad u,\, v \in (0, 1),$$

and its density can be written as

$$c(u, v; \alpha) = \frac{\alpha(1-e^{-\alpha})e^{-\alpha(u+v)}}{\left[1-e^{-\alpha}-(1-e^{-\alpha u})(1-e^{-\alpha v})\right]^2}, \quad u,\, v \in (0, 1).$$

## Clayton copula

The Clayton copula with parameter $\alpha \in \mathbb{R}^+$ is given by

$$C(u, v; \alpha) = \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}, \quad u,\, v \in (0, 1),$$

and its density can be written as

$$c(u, v; \alpha) = \frac{(\alpha + 1)(uv)^{\alpha}}{\left(u^{\alpha} + v^{\alpha} - (uv)^{\alpha}\right)^{1/\alpha + 2}}, \quad u,\, v \in (0, 1).$$

## Joe copula

The Joe copula with parameter $\alpha > 1$ is given by

$$C(u, v; \alpha) = 1 - \left[(1-u)^{\alpha} + (1-v)^{\alpha} - (1-u)^{\alpha}(1-v)^{\alpha}\right]^{1/\alpha}, \quad u,\, v \in (0, 1),$$

and its density can be written as

$$c(u, v; \alpha) = \left(x^{\alpha} + y^{\alpha} - (xy)^{\alpha}\right)^{1/\alpha - 2}(xy)^{\alpha - 1}\left(\alpha - 1 + x^{\alpha} + y^{\alpha} - (xy)^{\alpha}\right), \quad u,\, v \in (0, 1),$$

where $x = 1 - u$ and $y = 1 - v$.

## Gumbel copula

The Gumbel copula with parameter $\alpha > 1$ is given by

$$C(u, v; \alpha) = \exp\left\{-\left(x^{\alpha} + y^{\alpha}\right)^{1/\alpha}\right\}, \quad u,\, v \in (0, 1),$$

where $x = -\log(u)$ and $y = -\log(v)$. The Gumbel copula density can be written as

$$c(u, v; \alpha) = \frac{C(u, v; \alpha)}{uv}(xy)^{\alpha - 1}\left(x^{\alpha} + y^{\alpha}\right)^{1/\alpha - 2}\left[\left(x^{\alpha} + y^{\alpha}\right)^{1/\alpha} + \alpha - 1\right], \quad u,\, v \in (0, 1).$$

The Inverted Gumbel copula density is obtained if we substitute $u$ and $v$ by $(1-u)$ and $(1-v)$, respectively.

## Hüsler-Reiss copula

The Hüsler-Reiss copula with parameter $\alpha \in \mathbb{R}^+$ is given by

$$C(u,v;\alpha) = \exp\left\{-x\Phi\left(\frac{1}{\alpha}+\frac{\alpha}{2}\log\left(\frac{x}{y}\right)\right) - y\Phi\left(\frac{1}{\alpha}+\frac{\alpha}{2}\log\left(\frac{y}{x}\right)\right)\right\}, \quad u,\, v \in (0,1),$$

where $x = -\log(u)$ and $y = -\log(v)$. The Hüsler-Reiss copula density can be written as

$$c(u,v;\alpha) = \frac{C(u,v;\alpha)}{uv}\left[\Phi\left(\frac{1}{\alpha}+\frac{\alpha}{2}\log\left(\frac{x}{y}\right)\right)\Phi\left(\frac{1}{\alpha}+\frac{\alpha}{2}\log\left(\frac{y}{x}\right)\right) + \frac{\alpha}{2y}\phi\left(\frac{1}{\alpha}+\frac{\alpha}{2}\log\left(\frac{x}{y}\right)\right)\right], \quad u,\, v \in (0,1).$$

## Galambos copula

The Galambos copula with parameter $\alpha \in \mathbb{R}^+$ is given by

$$C(u,v;\alpha) = \exp\left\{-x-y+\left(x^{-\alpha}+y^{-\alpha}\right)^{-1/\alpha}\right\}, \quad u,\, v \in (0,1),$$

where $x = -\log(u)$ and $y = -\log(v)$. For $u,\, v \in (0,1)$, the Galambos copula density can be written as

$$c(u,v;\alpha) = \frac{C(u,v;\alpha)}{uv}\left[1-(x^{-\alpha}+y^{-\alpha})^{-1-1/\alpha}(x^{-\alpha-1}+y^{-\alpha-1}) + (x^{-\alpha}+y^{-\alpha})^{-2-1/\alpha}(xy)^{-\alpha-1}\left(1+\alpha+(x^{-\alpha}+y^{-\alpha})^{-1/\alpha}\right)\right].$$

### Coles-Tawn copula

The Coles-Tawn copula with parameters $\alpha, \beta \in \mathbb{R}^+$ is given by

$$C(u, v; \alpha, \beta) = \exp\left\{-x\left(1 - \mathrm{Be}(q; \alpha + 1, \beta)\right) - y\mathrm{Be}(q; \alpha, \beta + 1)\right\}, \quad u,\, v \in (0, 1),$$

where $x = -\log(u)$, $y = -\log(v)$, $q = \dfrac{\alpha x}{\alpha y + \beta x}$ and $\mathrm{Be}(q; a, b)$ represents the Beta distribution function with shape parameters $a > 0$ and $b > 0$. The Coles-Tawn copula density can be written as

$$c(u, v; \alpha, \beta) = \frac{C(u, v; \alpha, \beta)}{uvx^2y^2}\left[x^2y^2\left(1 - \mathrm{Be}\left(q; \alpha + 1, \beta\right)\right)\mathrm{Be}\left(q; \alpha, \beta + 1\right) + \frac{\alpha\beta\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha)\Gamma(\beta)}\frac{q^{\alpha-1}(1 - q)^{\beta-1}}{(\alpha/x + \beta/y)^3}\right], \quad u,\, v \in (0, 1).$$

## A.2 Extremal dependence properties

The extremal dependence measures $\chi$ and $\eta$ of the weighted copula model presented in Section 2 of the main text were derived for the case where $c_b$ is a Frank copula and $c_t$ a Gumbel copula, with two different weighting functions, and are presented in this Section. From equation (2) of Section 1.3 of the main text, we have

$$\begin{aligned}
\chi &= \lim_{r\to 1}\chi(r) = \lim_{r\to 1}\frac{P[U^* > r, V^* > r]}{P[U^* > r]} \\
&= \lim_{r\to 1}\frac{(1/K)\int_r^1\int_r^1 f_{c_t}(u^*, v^*; \boldsymbol{\alpha}, \theta)\mathrm{d}v^*\,\mathrm{d}u^* + (1/K)\int_r^1\int_r^1 f_{c_b}(u^*, v^*; \boldsymbol{\beta}, \theta)\mathrm{d}v^*\,\mathrm{d}u^*}{(1/K)\int_r^1\int_0^1 f_{c_t}(u^*, v^*; \boldsymbol{\alpha}, \theta)\mathrm{d}v^*\,\mathrm{d}u^* + (1/K)\int_r^1\int_0^1 f_{c_b}(u^*, v^*; \boldsymbol{\beta}, \theta)\mathrm{d}v^*\,\mathrm{d}u^*} \\
&= \lim_{r\to 1}\frac{\int_r^1\int_r^1 f_{c_t}(u^*, v^*; \boldsymbol{\alpha}, \theta)\mathrm{d}v^*\,\mathrm{d}u^* + \int_r^1\int_r^1 f_{c_b}(u^*, v^*; \boldsymbol{\beta}, \theta)\mathrm{d}v^*\,\mathrm{d}u^*}{\int_r^1\int_0^1 f_{c_t}(u^*, v^*; \boldsymbol{\alpha}, \theta)\mathrm{d}v^*\,\mathrm{d}u^* + \int_r^1\int_0^1 f_{c_b}(u^*, v^*; \boldsymbol{\beta}, \theta)\mathrm{d}v^*\,\mathrm{d}u^*},
\end{aligned}$$

where $f_{c_t} = K_t f_t$ and $f_{c_b} = K_b f_b$ with $K_t$, $K_b$, $f_t$, $f_b$ and $K$ as defined in Section 2.2 of the main text.

### Case 2: $c_b$ is a Frank copula, $c_t$ is a Gumbel copula and $\pi(u^*, v^*; \theta) = (uv)^\theta$

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

$$f_{c_b}(u^*, v^*; \beta, \theta) = [1 - (u^* v^*)^\theta] \frac{\beta(1 - \exp\{-\beta\}) \exp\{-\beta(u^* + v^*)\}}{[1 - \exp\{-\beta\} - (1 - \exp\{-\beta u^*\})(1 - \exp\{-\beta v^*\})]^2}$$

and

$$\begin{aligned} f_{c_t}(u^*, v^*; \alpha, \theta) &= (u^* v^*)^\theta \frac{C_t(u^*, v^*; \alpha)}{u^* v^*} (xy)^{\alpha-1} \left(x^\alpha + y^\alpha\right)^{1/\alpha - 2} \left[\left(x^\alpha + y^\alpha\right)^{1/\alpha} + \alpha - 1\right] \\ &= (u^* v^*)^{\theta-1} C_t(u^*, v^*; \alpha)(xy)^{\alpha-1} \left(x^\alpha + y^\alpha\right)^{1/\alpha - 2} \left[\left(x^\alpha + y^\alpha\right)^{1/\alpha} + \alpha - 1\right], \end{aligned}$$

with $x = -\log(u^*)$, $y = -\log(v^*)$ and $C_t(u^*, v^*; \alpha) = \exp\left\{-\left(x^\alpha + y^\alpha\right)^{1/\alpha}\right\}$.

#### Effect of the body copula $c_b$

Since the interest is on the limit when $u^*$ and $v^*$ are very near (1,1) and $f_{c_b}(u^*, v^*; \beta, \theta)$ is defined at (1,1), a Taylor approximation of order 1 can be used about (1,1) with point $(1 - s, 1 - t)$ for $\int_r^1 \int_r^1 f_{c_b}(u^*, v^*) \mathrm{d}v^* \, \mathrm{d}u^*$, where $s, t \to 0$. Therefore, for some norm $\|\cdot\|$ near 0, we have

$$f_{c_b}(1 - s, 1 - t; \beta, \theta) = f_{c_b}(1, 1) - s \frac{\partial f_{c_b}}{\partial s}(1, 1) - t \frac{\partial f_{c_b}}{\partial t}(1, 1) + \mathcal{O}\left(\|(s, t)\|^2\right),$$

where

$$\begin{aligned} \frac{\partial f_{c_b}}{\partial s} =& \frac{2\beta^2[1 - (st)^\theta](1 - \exp\{-\beta\})(1 - \exp\{-\beta t\}) \exp\{-\beta(2s + t)\}}{[1 - \exp\{-\beta\} - (1 - \exp\{-\beta s\})(1 - \exp\{-\beta t\})]^3} \\ &- \frac{\beta\theta s^{\theta-1} t^\theta (1 - \exp\{-\beta\}) \exp\{-\beta(s + t)\}}{[1 - \exp\{-\beta\} - (1 - \exp\{-\beta s\})(1 - \exp\{-\beta t\})]^2} \\ &- \frac{\beta^2[1 - (st)^\theta](1 - \exp\{-\beta\}) \exp\{-\beta(s + t)\}}{[1 - \exp\{-\beta\} - (1 - \exp\{-\beta s\})(1 - \exp\{-\beta t\})]^2}. \end{aligned}$$

At the point (1,1), $f_{c_b}(1,1) = 0$ and

$$\frac{\partial f_{c_b}}{\partial s}(1,1) = \frac{\partial f_{c_b}}{\partial t}(1,1) = -\beta\theta\left(1-\exp\{-\beta\}\right)^{-1}.$$

So,

$$f_{c_b}(1-s,1-t;\beta,\theta) = \beta\theta\left(1-\exp\{-\beta\}\right)^{-1}(s+t) + \mathcal{O}\left(\|(s,t)\|^2\right).$$

Taking $s = 1-u^*$ and $t^* = 1-v^*$, we have

$$\begin{aligned}
&\int_r^1\int_r^1 f_{c_b}(u^*,v^*)\mathrm{d}v^*\,\mathrm{d}u^* \\
&\quad= \int_0^{1-r}\int_0^{1-r} \beta\theta\left(1-\exp\{-\beta\}\right)^{-1}(s+t)\mathrm{d}t\,\mathrm{d}s + \mathcal{O}\left((1-r)^4\right) \\
&\quad= \beta\theta\left(1-\exp\{-\beta\}\right)^{-1}\int_0^{1-r}\int_0^{1-r}(s+t)\mathrm{d}t\,\mathrm{d}s + \mathcal{O}\left((1-r)^4\right) \\
&\quad= \beta\theta\left(1-\exp\{-\beta\}\right)^{-1}(1-r)^3 + \mathcal{O}\left((1-r)^4\right).
\end{aligned}$$

Similarly, for $\int_r^1\int_0^1 f_{c_b}(u^*,v^*)\mathrm{d}v^*\,\mathrm{d}u^*$, a Taylor approximation of order 1 can be used about $(1,v^*)$ with point $(u^*,v^*)$. Thus, we have

$$f_{c_b}(u^*,v^*;\beta,\theta) = f_{c_b}(1,v^*) + (u^*-1)\frac{\partial f_{c_b}}{\partial u^*}(1,v^*) + \mathcal{O}\left((u^*-1)^2\right),$$

where

$$f_{c_b}(1,v^*) = \frac{(1-(v^*)^\theta)\beta\exp\{-\beta(1-v^*)\}}{1-\exp\{-\beta\}} = A_{v^*,\beta,\theta}$$

and

$$\begin{aligned}\frac{\partial f_{c_b}}{\partial u^*}(1,v^*) =& \frac{2\beta^2(1-(v^*)^\theta)(1-\exp\{-\beta v^*\})\exp\{-2\beta(1-v^*)\}}{(1-\exp\{-\beta\})^2} \\ & - \frac{\beta\theta(v^*)^\theta\exp\{-\beta(1-v^*)\}}{1-\exp\{-\beta\}} - \frac{\beta^2(1-(v^*)^\theta)\exp\{-\beta(1-v^*)\}}{1-\exp\{-\beta\}} \\ =& B_{v^*,\beta,\theta}.\end{aligned}$$

So, $f_{c_b}(u^*,v^*) = A_{v^*,\beta,\theta} + B_{v^*,\beta,\theta}(u^*-1) + \mathcal{O}\left((u^*-1)^2\right)$, and we obtain

$$\begin{aligned}&\int_r^1\int_0^1 f_{c_b}(u^*,v^*)\mathrm{d}v^*\,\mathrm{d}u^* \\ &\quad= \int_r^1\int_0^1 [A_{v^*,\beta,\theta} + B_{v^*,\beta,\theta}(u^*-1)]\mathrm{d}v^*\,\mathrm{d}u^* + \mathcal{O}\left((1-r)^3\right) \\ &\quad= \int_0^1 A_{v^*,\beta,\theta}\int_r^1 \mathrm{d}u^*\,\mathrm{d}v^* + \int_0^1 B_{v^*,\beta,\theta}\int_r^1 (u^*-1)\mathrm{d}u^*\,\mathrm{d}v^* + \mathcal{O}\left((1-r)^3\right) \\ &\quad= (1-r)\underbrace{\int_0^1 A_{v^*,\beta,\theta}\mathrm{d}v^*}_{C_{\beta,\theta}} - \frac{1}{2}(1-r)^2\underbrace{\int_0^1 B_{v^*,\beta,\theta}\mathrm{d}v^*}_{D_{\beta,\theta}} + \mathcal{O}\left((u^*)^2\right) \\ &\quad= C_{\beta,\theta}(1-r) - \frac{D_{\beta,\theta}}{2}(1-r)^2 + \mathcal{O}\left((1-r)^3\right)\end{aligned}$$

**Effect of the tail copula $c_t$**

Contrarily to $f_{c_b}(\cdot)$, $f_{c_t}(u^*,v^*;\alpha,\theta)$ is not finite at (1,1). For this reason, it is not possible to use a Taylor approximation about (1,1). Instead, we use asymptotics near this point. Specifically, we now write $u^*$ and $v^*$ in terms of $s$ and $t$, where $s,t>0$ and $u^* = 1-s+o(s)$ and $v^* = 1-t+o(t)$ as $s,t\to 0$. This describes the behaviour of $u^*$ and $v^*$ as they tend to 1. Thus, for the first term of $f_{c_t}$, we have

$$\begin{aligned}(u^*v^*)^{\theta-1} =& (1-s)^{\theta-1}(1-t)^{\theta-1} + o(s) + o(t) \\ =& [1-(\theta-1)s][1-(\theta-1)t] + o(s) + o(t),\end{aligned}$$

as $s,t\to 0$.

Let us first consider the case when $x = -\log(u^*) > y = -\log(v^*)$. For $(u^*, v^*) \to (1,1)$, i.e., $s \to 0$ and $t \to 0$, with $t/s \to c$ for $c \in (0,1)$, the copula density term follows asymptotically

$$c_t(u^*, v^*; \alpha) \sim (\alpha - 1)x^{-\alpha}y^{\alpha-1}\left[1 + \left(\frac{y}{x}\right)^{\alpha}\right]^{1/\alpha-2}.$$

Analogously, when $x < y$, i.e., $s \to 0$ and $t \to 0$, with $t/s \to c$ for $c \in (1, \infty)$,

$$c_t(u^*, v^*; \alpha) \sim (\alpha - 1)y^{-\alpha}x^{\alpha-1}\left[1 + \left(\frac{x}{y}\right)^{\alpha}\right]^{1/\alpha-2}.$$

Moreover, $x = s + o(s)$ and $y = t + o(t)$ as $s, t \to 0$. So, considering the symmetry between cases $x > y$ and $x < y$, and recalling $u^* = 1 - s + o(s)$ and $v^* = 1 - t + o(t)$,

$$\int_r^1 \int_r^1 f_{c_t}(u^*, v^*)\mathrm{d}v^* \,\mathrm{d}u^* = P[1 - S > r, 1 - T > r] = 2P[S < 1 - r, T < S].$$

So, we have

$$\begin{aligned}
P[S < 1 - r, T < S] &= \int_0^{1-r} \int_0^s f^*_{c_t}(s, t; \alpha, \theta)\mathrm{d}t \,\mathrm{d}s \\
&= \int_0^{1-r} \int_0^s [1 - (\theta - 1)s][1 - (\theta - 1)t](\alpha - 1)s^{-\alpha}t^{\alpha-1} \\
&\qquad \times \left[1 + \left(\frac{t}{s}\right)^{\alpha}\right]^{1/\alpha-2} \mathrm{d}t \,\mathrm{d}s + o\left((1-r)^2\right) \\
&= (\alpha - 1)\int_0^{1-r} [1 - (\theta - 1)s]s^{-\alpha} \\
&\qquad \underbrace{\int_0^s [1 - (\theta - 1)t]t^{\alpha-1}\left[1 + \left(\frac{t}{s}\right)^{\alpha}\right]^{1/\alpha-2} \mathrm{d}t}_{A(s)} \,\mathrm{d}s + o\left((1-r)^2\right)
\end{aligned}$$

as $r \to 1$. Evaluating $A(s)$ by parts, we get

$$\int_0^s [1-(\theta-1)t]t^{\alpha-1}\left[1+\left(\frac{t}{s}\right)^{\alpha}\right]^{1/\alpha-2}\mathrm{d}t$$
$$= \frac{2^{1/\alpha-1}s^{\alpha}}{1-\alpha} - \frac{2^{1/\alpha-1}(\theta-1)s^{\alpha+1}}{1-\alpha} - \frac{s^{\alpha}}{1-\alpha} - \frac{(1-\theta)s^{\alpha+1}}{1-\alpha}C_{\alpha},$$

with $C_{\alpha} = \int_0^1 (1+q^{\alpha})^{1/\alpha-1}\mathrm{d}q$. And, by substituting $A(s)$ in the outer integral, we obtain

$$P[S<1-r,T<s] = (1-2^{1/\alpha-1})(1-r) + \left[(2^{1/\alpha}-1-C_{\alpha})(\theta-1)/2\right](1-r)^2 + o\left((1-r)^2\right), \quad \text{as } r \to 1.$$

Then, as $r \to 1$,

$$\begin{aligned}
&\int_r^1\int_r^1 f_{c_t}(u^*,v^*)\mathrm{d}v^*\,\mathrm{d}u^* \\
&= 2(1-2^{1/\alpha-1})(1-r) + 2\left[(2^{1/\alpha}-1-C_{\alpha})(\theta-1)/2\right](1-r)^2 + o\left((1-r)^2\right) \\
&= (2-2^{1/\alpha})(1-r) + (2^{1/\alpha}-1-C_{\alpha})(\theta-1)(1-r)^2 + o\left((1-r)^2\right),
\end{aligned}$$

Since for $\int_r^1\int_0^1 f_{c_t}(u^*,v^*)\mathrm{d}v^*\,\mathrm{d}u^*$ we need to integrate over the support for $v^*$, it is not possible to approximate $f_{c_t}(\cdot)$ as above. Instead, we take the change of variable $y = xz$, with $z = y/x \in \mathbb{R}^+$, so we have $u^* = \exp\{-x\}$ and $v^* = \exp\{-xz\}$. Thus, we obtain

$$
\begin{aligned}
&\int_r^1 \int_0^1 f_{c_t}(u^*, v^*; \alpha, \theta) \mathrm{d}v^* \, \mathrm{d}u^* \\
&= \int_r^1 \int_0^1 (u^* v^*)^{\theta-1} C_t(u^*, v^*; \alpha)(xy)^{\alpha-1} \left(x^\alpha + y^\alpha\right)^{1/\alpha-2} \\
&\qquad \times \left[\left(x^\alpha + y^\alpha\right)^{1/\alpha} + \alpha - 1\right] \mathrm{d}v^* \, \mathrm{d}u^* \\
&= \int_0^{-\log(r)} \int_0^\infty \exp\left\{-x\left[\theta(1+z) + (1+z^\alpha)^{1/\alpha}\right]\right\} z^{\alpha-1} \left(1+z^\alpha\right)^{1/\alpha-2} \\
&\qquad \times \left[x\left(1+z^\alpha\right)^{1/\alpha} + \alpha - 1\right] \mathrm{d}z \, \mathrm{d}x \\
&= \int_0^{-\log(r)} \int_0^\infty x \underbrace{z^{\alpha-1}\left(1+z^\alpha\right)^{2/\alpha-2}}_{g(z)} \exp\left\{-x \underbrace{\left[\theta(1+z) + (1+z^\alpha)^{1/\alpha}\right]}_{h(z)}\right\} \mathrm{d}z \, \mathrm{d}x \\
&\quad + (\alpha-1) \int_0^{-\log(r)} \int_0^\infty \underbrace{z^{\alpha-1}\left(1+z^\alpha\right)^{1/\alpha-2}}_{f(z)} \\
&\qquad \times \exp\left\{-x \underbrace{\left[\theta(1+z) + (1+z^\alpha)^{1/\alpha}\right]}_{h(z)}\right\} \mathrm{d}z \, \mathrm{d}x \\
&= \int_0^\infty g(z) \underbrace{\int_0^{-\log(r)} x \exp\{-xh(z)\} \mathrm{d}x}_{B(z,r)} \, \mathrm{d}z \\
&\qquad + (\alpha-1) \int_0^\infty f(z) \underbrace{\int_0^{-\log(r)} \exp\{-xh(z)\} \mathrm{d}x}_{C(z,r)} \, \mathrm{d}z.
\end{aligned}
$$

Evaluating $B(z,r)$ by parts, we get

$$
\begin{aligned}
&\int_0^{-\log(r)} x \exp\{-xh(z)\} \mathrm{d}x \\
&\qquad = \left[-\frac{x}{h(z)} \exp\{-xh(z)\}\right]_{x=0}^{x=-\log(r)} - \left[\frac{1}{h^2(z)} \exp\{-xh(z)\}\right]_{x=0}^{x=-\log(r)} \\
&\qquad = \frac{\log(r)}{h(z)} r^{h(z)} - \frac{1}{h^2(z)} r^{h(z)} + \frac{1}{h^2(z)}.
\end{aligned}
$$

Analogously, by evaluating $C(z,r)$, we have

$$\begin{aligned}\int_0^{-\log(r)} \exp\{-xh(z)\}\mathrm{d}x &= \left[-\frac{1}{h(z)}\exp\{-h(z)x\}\right]_{x=0}^{x=-\log(r)} \\ &= -\frac{1}{h(z)}r^{h(z)} + \frac{1}{h(z)}.\end{aligned}$$

Substituting $B(z,r)$ and $C(z,r)$ in the outer integral, we obtain

$$\begin{aligned}\int_r^1\int_0^1 f_{c_t}(u^*,v^*;\alpha,\theta)\mathrm{d}v^*\,\mathrm{d}u^* =& \log(r)\int_0^\infty \frac{g(z)}{h(z)}r^{h(z)}\mathrm{d}z + \int_0^\infty \frac{g(z)}{h^2(z)}\left(1-r^{h(z)}\right)\mathrm{d}z \\ &+ (\alpha-1)\int_0^\infty \frac{f(z)}{h(z)}\left(1-r^{h(z)}\right)\mathrm{d}z.\end{aligned}$$

Evaluating $\int_0^\infty \frac{f(z)}{h(z)}\left(1-r^{h(z)}\right)\mathrm{d}z$ by parts, we have

$$\begin{aligned}&\int_0^\infty \frac{f(z)}{h(z)}\left(1-r^{h(z)}\right)\mathrm{d}z \\ &\quad= \left[\frac{1}{1-\alpha}\left(1+z^\alpha\right)^{1/\alpha-1}\frac{1-r^{h(z)}}{h(z)}\right]_0^\infty \\ &\qquad \underbrace{-\int_0^\infty \frac{1}{1-\alpha}\left(1+z^\alpha\right)^{1/\alpha-1}\left(\frac{h'(z)\left(r^{h(z)}-1\right)}{h^2(z)} - \frac{\log(r)h'(z)r^{h(z)}}{h(z)}\right)\mathrm{d}z}_{D(r)} \\ &\quad= \frac{1}{1-\alpha}\lim_{z\to\infty}\left(1+z^\alpha\right)^{1/\alpha-1}\frac{1-r^{\theta(1+z)+(1+z^\alpha)^{1/\alpha}}}{\theta(1+z)+(1+z^\alpha)^{1/\alpha}} \\ &\qquad -\frac{1}{1-\alpha}\frac{1-r^{\theta+1}}{\theta+1} - D(r) \\ &\quad= \frac{1}{1-\alpha}\lim_{z\to\infty} z^{1-\alpha}\frac{r^{z(\theta+1)}}{z(1+\theta)} + \frac{1-r^{\theta+1}}{(\alpha-1)(\theta+1)} - D(r) \\ &\quad= \frac{1}{1-\alpha}\lim_{z\to\infty} z^{-\alpha}\frac{r^{z(\theta+1)}}{1+\theta} + \frac{1-r^{\theta+1}}{(\alpha-1)(\theta+1)} - D(r) \\ &\quad= \frac{1-r^{\theta+1}}{(\alpha-1)(\theta+1)} - D(r).\end{aligned}$$

Noting that $h'(z) = \theta + z^{\alpha-1}(1+z^\alpha)^{1/\alpha-1}$, and recalling that $g(z) = z^{\alpha-1}(1+z^\alpha)^{2/\alpha-2}$, $D(r)$ can be simplified as below

$$
\begin{aligned}
D(r) =& \int_0^\infty \frac{1}{1-\alpha}\left(1+z^\alpha\right)^{1/\alpha-1}\left(\frac{h'(z)\left(r^{h(z)}-1\right)}{h^2(z)} - \frac{\log(r)h'(z)r^{h(z)}}{h(z)}\right)\mathrm{d}z \\
=& \int_0^\infty \frac{1}{1-\alpha}\left(1+z^\alpha\right)^{1/\alpha-1}\left[\theta + z^{\alpha-1}(1+z^\alpha)^{1/\alpha-1}\right]\frac{\left(r^{h(z)}-1\right)}{h^2(z)}\mathrm{d}z \\
& - \log(r)\int_0^\infty \frac{1}{1-\alpha}\left(1+z^\alpha\right)^{1/\alpha-1}\left[\theta + z^{\alpha-1}(1+z^\alpha)^{1/\alpha-1}\right]\frac{r^{h(z)}}{h(z)}\mathrm{d}z \\
=& -\theta\int_0^\infty \frac{1}{1-\alpha}\left(1+z^\alpha\right)^{1/\alpha-1}\frac{\left(1-r^{h(z)}\right)}{h^2(z)}\mathrm{d}z \\
& - \int_0^\infty \frac{1}{1-\alpha}z^{\alpha-1}\left(1+z^\alpha\right)^{2/\alpha-2}\frac{\left(1-r^{h(z)}\right)}{h^2(z)}\mathrm{d}z \\
& - \theta\log(r)\int_0^\infty \frac{1}{1-\alpha}\left(1+z^\alpha\right)^{1/\alpha-1}\frac{r^{h(z)}}{h(z)}\mathrm{d}z \\
& - \log(r)\int_0^\infty \frac{1}{1-\alpha}z^{\alpha-1}\left(1+z^\alpha\right)^{2/\alpha-2}\frac{r^{h(z)}}{h(z)}\mathrm{d}z \\
=& \frac{\theta}{\alpha-1}\int_0^\infty \left(1+z^\alpha\right)^{1/\alpha-1}\frac{\left(1-r^{h(z)}\right)}{h^2(z)}\mathrm{d}z + \frac{1}{\alpha-1}\int_0^\infty \frac{g(z)}{h^2(z)}\left(1-r^{h(z)}\right)\mathrm{d}z \\
& + \frac{\theta\log(r)}{\alpha-1}\int_0^\infty \left(1+z^\alpha\right)^{1/\alpha-1}\frac{r^{h(z)}}{h(z)}\mathrm{d}z + \frac{\log(r)}{\alpha-1}\int_0^\infty \frac{g(z)}{h(z)}r^{h(z)}\mathrm{d}z.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\int_r^1\int_0^1 f_{c_t}(u^*,v^*;\alpha,\theta)\mathrm{d}v^*\,\mathrm{d}u^* =& \log(r)\int_0^\infty \frac{g(z)}{h(z)}r^{h(z)}\mathrm{d}z + \int_0^\infty \frac{g(z)}{h^2(z)}\left(1-r^{h(z)}\right)\mathrm{d}z \\
& + (\alpha-1)\frac{1-r^{\theta+1}}{(\alpha-1)(\theta+1)} - (\alpha-1)\frac{\theta}{\alpha-1}\int_0^\infty \left(1+z^\alpha\right)^{1/\alpha-1}\frac{\left(1-r^{h(z)}\right)}{h^2(z)}\mathrm{d}z \\
& - (\alpha-1)\frac{1}{\alpha-1}\int_0^\infty \frac{g(z)}{h^2(z)}\left(1-r^{h(z)}\right)\mathrm{d}z \\
& - (\alpha-1)\frac{\theta\log(r)}{\alpha-1}\int_0^\infty \left(1+z^\alpha\right)^{1/\alpha-1}\frac{r^{h(z)}}{h(z)}\mathrm{d}z \\
& - (\alpha-1)\frac{\log(r)}{\alpha-1}\int_0^\infty \frac{g(z)}{h(z)}r^{h(z)}\mathrm{d}z \\
=& \frac{1-r^{\theta+1}}{\theta+1} - \theta\int_0^\infty \left(1+z^\alpha\right)^{1/\alpha-1}\frac{\left(1-r^{h(z)}\right)}{h^2(z)}\mathrm{d}z \\
& - \theta\log(r)\int_0^\infty \left(1+z^\alpha\right)^{1/\alpha-1}\frac{r^{h(z)}}{h(z)}\mathrm{d}z \\
=& 1-r-\frac{\theta}{2}(1-r)^2 + o\left((1-r)^2\right),
\end{aligned}
$$

where $-\theta \int_0^\infty (1+z^\alpha)^{1/\alpha-1} \frac{\left(1-r^{h(z)}\right)}{h^2(z)} \mathrm{d}z - \theta \log(r) \int_0^\infty (1+z^\alpha)^{1/\alpha-1} \frac{r^{h(z)}}{h(z)} \mathrm{d}z = o((1-r)^2)$ as $r \to 1$. Additionally, $r^{\theta+1} = 1-(\theta+1)(1-r)+[\theta(\theta+1)/2](1-r)^2+o\left((1-r)^2\right)$ as $r \to 1$ by the Binomial expansion.

**Extremal dependence $\chi$ for this case**

Let

$$
\begin{aligned}
c_1 &= 2-2^{1/\alpha} = \chi_{Gumbel}, & c_5 &= -\theta/2 + o\left((1-r)^2\right), \\
c_2 &= (2^{1/\alpha}-1-C_\alpha)(\theta-1), & c_6 &= C_{\beta,\theta} = \beta\left(1-\exp\{-\beta\}\right)^{-1} \\
c_3 &= \beta\theta\left(1-\exp\{-\beta\}\right)^{-1}, & &\quad \times \int_0^1 (1-(v^*)^\theta) e^{-\beta(1-v^*)} \mathrm{d}v^*, \\
c_4 &= 1, & c_7 &= -D_{\beta,\theta}/2.
\end{aligned}
$$

We then have

$$
\begin{aligned}
\chi =& \lim_{r\to 1} \frac{c_1(1-r)+c_2(1-r)^2+c_3(1-r)^3+o\left((1-r)^3\right)}{c_4(1-r)+c_5(1-r)^2+c_6(1-r)+c_7(1-r)^2+o\left((1-r)^2\right)} \\
=& \lim_{r\to 1} \left( \frac{c_1}{c_4+c_6} + \left[ \frac{c_2-c_1(c_5+c_7)}{(c_4+c_6)^2} \right](1-r) + \mathcal{O}\left((1-r)^2\right) \right) \\
=& \frac{c_1}{c_4+c_6} = \frac{2-2^{1/\alpha}}{1+\beta\left(1-\exp\{-\beta\}\right)^{-1} \int_0^1 (1-(v^*)^\theta) e^{-\beta(1-v^*)} \mathrm{d}v^*}
\end{aligned} \tag{A.2.1}
$$

For the vector of parameters $\boldsymbol{\gamma} = (3, 1, 1.844444)$, $c_1 \approx 0.740079$, $c_5 = 1$ and $c_7 \approx 0.5630892$. Thus, from equation (A.2.1), we have $\chi \approx 0.473472$. Moreover, from the numerical investigation, $\chi(r) \approx 0.4699556$ with $r = 0.9998779$. Figure A.2.1 shows this comparison.

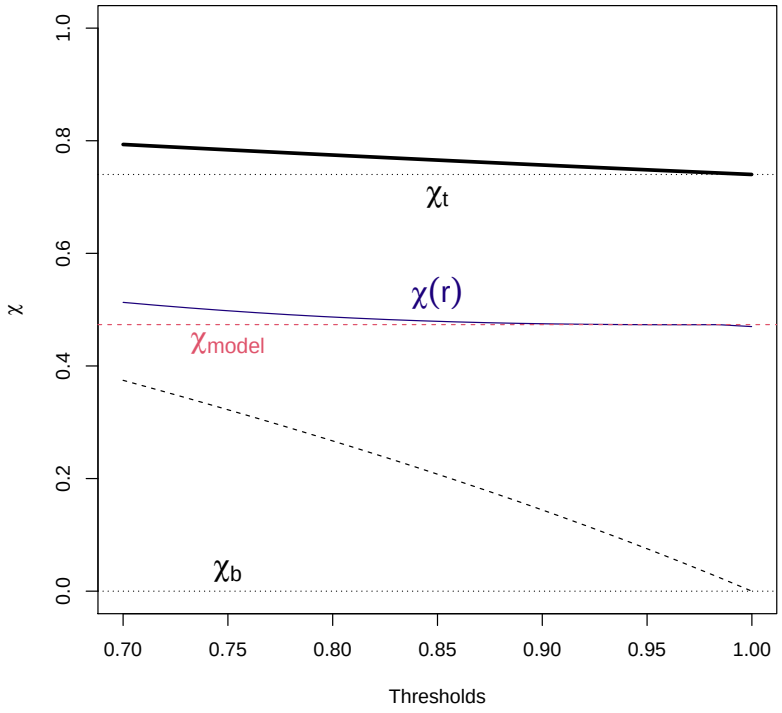



Figure A.2.1: The blue line represents $\chi(r)$ for $r \in [0.7, 1)$ with weighting function $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$ and $\theta = 1.84444$. The thick black lines represent the single copula models - Frank (dashed) and Gumbel (solid). The theoretical values for the Frank and Gumbel copulas based on Table 2 of Section 2.3 from the main text are represented by the horizontal dashed lines, and the value derived for the model is represented by the pink dashed line.

For the vector of parameters $\boldsymbol{\gamma} = (1.5, 3, 3.488889)$, $c_1 \approx 0.4125989$, $c_5 = 1$ and $c_7 \approx 0.5555462$. Thus, from equation (A.2.1), we have $\chi \approx 0.2652438$. Moreover, from the numerical investigation, $\chi(r) \approx 0.2842924$ with $r = 0.9998779$. Figure A.2.2 shows this comparison.

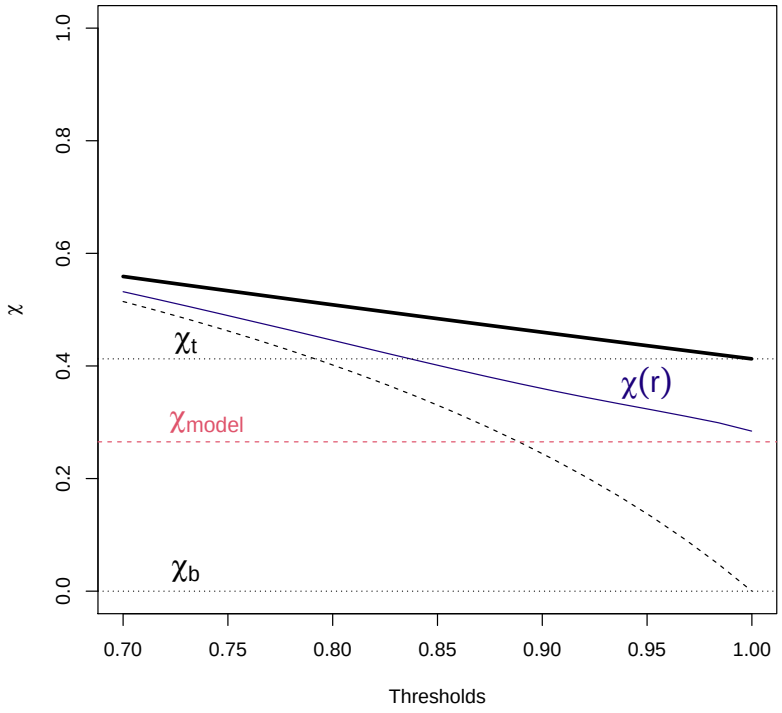



Figure A.2.2: The blue line represents $\chi(r)$ for $r \in [0.7, 1)$ with weighting function $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$ and $\theta = 3.488889$. The thick black lines represent the single copula models - Frank (dashed) and Gumbel (solid). The theoretical values for the Frank and Gumbel copulas based on Table 2 of Section 2.3 from the main text are represented by the horizontal dashed lines, and the value derived for the model is represented by the pink dashed line.

**Extremal dependence $\eta$ for this case**

As $\chi > 0$, we should expect $\eta = 1$. Following equation (4) of Section 1.3 from the main text, we have

$$\begin{aligned}
\eta =& \lim_{r\to 1} \frac{\log\left(P[U^* > r]\right)}{\log\left(P[U^* > r, V^* > r]\right)} \\
=& \lim_{r\to 1} \frac{\log\left[c_4(1-r) + c_5(1-r)^2 + c_6(1-r) + c_7(1-r)^2 + o\left((1-r)^2\right)\right]}{\log\left[c_1(1-r) + c_2(1-r)^2 + c_3(1-r)^3 + o\left((1-r)^3\right)\right]} \\
\overset{\left(\frac{\infty}{\infty}\right)}{=}& \lim_{r\to 1} \frac{-c_4 - c_6 - 2(c_5+c_7)(1-r) + o(1-r)}{-c_1 - 2c_2(1-r) - 3c_3(1-r)^2 + o\left((1-r)^2\right)} \\
& \times \frac{c_1 + c_2(1-r) + c_3(1-r)^2 + o((1-r)^2)}{c_4 + c_6 + (c_5+c_7)(1-r) + o\left(1-r\right)} \\
=& \frac{c_4 + c_6}{c_1} \frac{c_1}{c_4 + c_6} = 1,
\end{aligned}$$

by L'Hôpital's Rule.

## Case 2.1: $c_b$ is a Frank copula, $c_t$ is a Gumbel copula and $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$

Let us now assume a different weighting function $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$. We have

$$\begin{aligned}
f_{c_b}(u^*, v^*; \beta, \theta) =& [1 - \exp\{-\theta(1-u^*)(1-v^*)\}] \\
& \times \frac{\beta(1-\exp\{-\beta\})\exp\{-\beta(u^*+v^*)\}}{\left[1-\exp\{-\beta\} - (1-\exp\{-\beta u^*\})(1-\exp\{-\beta v^*\})\right]^2}
\end{aligned}$$

and

$$\begin{aligned}
f_{c_t}(u^*, v^*; \alpha, \theta) =& \exp\{-\theta(1-u^*)(1-v^*)\} \frac{C_t(u^*, v^*; \alpha)}{u^* v^*} (xy)^{\alpha-1} \left(x^\alpha + y^\alpha\right)^{1/\alpha - 2} \\
& \times \left[\left(x^\alpha + y^\alpha\right)^{1/\alpha} + \alpha - 1\right],
\end{aligned}$$

with $x = -\log(u^*)$, $y = -\log(v^*)$ and $C_t(u^*, v^*; \alpha) = \exp\left\{-\left(x^\alpha + y^\alpha\right)^{1/\alpha}\right\}$.

**Effect of the body copula $c_b$**

As the above case, a Taylor approximation of order 1 can be used about (1,1) with point $(1-s, 1-t)$ for $\int_r^1 \int_r^1 f_{c_b}(u^*, v^*)\mathrm{d}v^*\,\mathrm{d}u^*$, where $s, t \to 0$. Therefore, for some norm $\|\cdot\|$ near 0, we have

$$f_{c_b}(1-s, 1-t; \beta, \theta) = f_{c_b}(1,1) - s\frac{\partial f_{c_b}}{\partial s}(1,1) - t\frac{\partial f_{c_b}}{\partial t}(1,1) + \mathcal{O}\left(\|(s,t)\|^2\right),$$

where

$$\begin{aligned}\frac{\partial f_{c_b}}{\partial s} =& -\exp\{-\theta(1-s)(1-t)\} \\ &\times \frac{2\beta^2(1-\exp\{-\beta\})(1-\exp\{-\beta t\})\exp\{-\beta(2s+t)\}}{[1-\exp\{-\beta\}-(1-\exp\{-\beta s\})(1-\exp\{-\beta t\})]^3} \\ &-\exp\{-\theta(1-s)(1-t)\} \\ &\times \frac{\beta(1-\exp\{-\beta\})[\theta(1-t)-\beta]\exp\{-\beta(s+t)\}}{[1-\exp\{-\beta\}-(1-\exp\{-\beta s\})(1-\exp\{-\beta t\})]^2}.\end{aligned}$$

At the point (1,1), $f_{c_b}(1,1) = 0$ and

$$\frac{\partial f_{c_b}}{\partial s}(1,1) = \frac{\partial f_{c_b}}{\partial t}(1,1) = -\beta^2\left(1-\exp\{-\beta\}\right)^{-1}.$$

So,

$$f_{c_b}(1-s, 1-t; \beta, \theta) = \beta^2\left(1-\exp\{-\beta\}\right)^{-1}(s+t) + \mathcal{O}\left(\|(s,t)\|^2\right),$$

and we obtain

$$\begin{aligned}\int_r^1 \int_r^1 & f_{c_b}(u^*, v^*)\mathrm{d}v^* \,\mathrm{d}u^* \\ &= \int_r^1 \int_r^1 \beta^2 \left(1 - \exp\{-\beta\}\right)^{-1} (s+t)\mathrm{d}t \,\mathrm{d}s + \mathcal{O}\left((1-r)^4\right) \\ &= \beta^2 \left(1 - \exp\{-\beta\}\right)^{-1} (1-r)^3 + \mathcal{O}\left((1-r)^4\right).\end{aligned}$$

Similarly, for $\int_r^1 \int_0^1 f_{c_b}(u^*, v^*)\mathrm{d}v^* \,\mathrm{d}u^*$, a Taylor approximation of order 1 can be used about $(1, v^*)$ with point $(u^*, v^*)$. Thus, we have

$$f_{c_b}(u^*, v^*; \beta, \theta) = f_{c_b}(1, v^*) + (u^* - 1)\frac{\partial f_{c_b}}{\partial u^*}(1, v^*) + \mathcal{O}\left((u^* - 1)^2\right),$$

where $f_{c_b}(1, v^*) = 0$ and

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

So, $f_{c_b}(u^*, v^*) = A_{v^*,\beta,\theta} + \mathcal{O}\left((u^* - 1)^2\right)$, and we obtain

$$\begin{aligned}\int_r^1 \int_0^1 f_{c_b}(u^*, v^*)\mathrm{d}v^* \,\mathrm{d}u^* &= \int_r^1 \int_0^1 A_{v^*,\beta,\theta}(u^* - 1)\mathrm{d}v^* \,\mathrm{d}u^* + \mathcal{O}\left((1-r)^3\right) \\ &= \int_0^1 A_{v^*,\beta,\theta} \int_r^1 (u^* - 1)\mathrm{d}u^* \,\mathrm{d}v^* + \mathcal{O}\left((1-r)^3\right) \\ &= -\frac{1}{2}(1-r)^2 \underbrace{\int_0^1 A_{v^*,\beta,\theta}\mathrm{d}v^*}_{B_{\beta,\theta}} + \mathcal{O}\left((1-r)^3\right) \\ &= -\frac{B_{\beta,\theta}}{2}(1-r)^2 + \mathcal{O}\left((1-r)^3\right)\end{aligned}$$

**Effect of the tail copula $c_t$**

Let us again write $u^*$ and $v^*$ in terms of $s$ and $t$, where $s, t > 0$ and $u^* = 1 - s + o(s)$ and $v^* = 1 - t + o(t)$ as $s, t \to 0$. As before, this describes the behaviour of $u^*$ and $v^*$ as they tend to 1. For the weighting function term of $f_{c_t}$, we have

$$\exp\{-\theta(1-u^*)(1-v^*)\} = \exp\{-\theta st\} + o(s) + o(t),$$

as $s, t \to 0$.

Similarly to the previous case, we consider $x = -\log(u^*) > y = -\log(v^*)$. For $(u^*, v^*) \to (1,1)$, i.e, $s \to 0$ and $t \to 0$, with $t/s \to c$ for $c \in (0,1)$, the copula density term follows asymptotically

$$c_t(u^*, v^*; \alpha) \sim (\alpha - 1)x^{-\alpha}y^{\alpha-1}\left[1 + \left(\frac{y}{x}\right)^{\alpha}\right]^{1/\alpha - 2}.$$

And, when $x < y$, i.e, $s \to 0$ and $t \to 0$, with $t/s \to c$ for $c \in (1, \infty)$,

$$c_t(u^*, v^*; \alpha) \sim (\alpha - 1)y^{-\alpha}x^{\alpha-1}\left[1 + \left(\frac{x}{y}\right)^{\alpha}\right]^{1/\alpha - 2}.$$

Finally, $x = s + o(s)$ and $y = t + o(t)$ as $s, t \to 0$. Thus, considering the symmetry between cases $x > y$ and $x < y$, and recalling $u^* = 1 - s + o(s)$ and $v^* = 1 - t + o(t)$,

$$\int_r^1 \int_r^1 f_{c_t}(u^*, v^*) \mathrm{d}v^* \, \mathrm{d}u^* = P[1 - S > r, 1 - T > r] = 2P[S < 1 - r, T < s].$$

So, we have

$$
\begin{aligned}
P[S < 1-r, T < s] &= \int_0^{1-r} \int_0^s f^*_{c_t}(s,t;\alpha,\theta)\mathrm{d}t\,\mathrm{d}s \\
&= (\alpha-1)\int_0^{1-r}\int_0^s \exp\{-\theta s t\} s^{-\alpha} t^{\alpha-1}\left[1+\left(\frac{t}{s}\right)^{\alpha}\right]^{1/\alpha-2}\mathrm{d}t\,\mathrm{d}s + o\left((1-r)^2\right) \\
&= (\alpha-1)\int_0^{1-r} s^{-\alpha} \underbrace{\int_0^s \exp\{-\theta s t\} t^{\alpha-1}\left[1+\left(\frac{t}{s}\right)^{\alpha}\right]^{1/\alpha-2}\mathrm{d}t}_{A(s)}\,\mathrm{d}s + o\left((1-r)^2\right)
\end{aligned}
$$

as $r \to 1$. Evaluating $A(s)$ by parts, we get

$$
\begin{aligned}
\int_0^s \exp\{-\theta s t\} t^{\alpha-1}\left[1+\left(\frac{t}{s}\right)^{\alpha}\right]^{1/\alpha-2}\mathrm{d}t =& \frac{2^{1/\alpha-1}\exp\{-\theta s^2\}s^{\alpha}}{1-\alpha} - \frac{s^{\alpha}}{1-\alpha} \\
&+ \frac{\theta s^{\alpha+2}}{1-\alpha}C_{\alpha} - \frac{\theta^2 s^{\alpha+4}}{1-\alpha}C^*_{\alpha},
\end{aligned}
$$

with $C_{\alpha} = \int_0^1 (1+q^{\alpha})^{1/\alpha-1}\mathrm{d}q$ and $C^*_{\alpha} = \int_0^1 q(1+q^{\alpha})^{1/\alpha-1}\mathrm{d}q$. By substituting $A(s)$ in the outer integral, we obtain

$$
\begin{aligned}
P[S < 1-r, T < s] =& -2^{1/\alpha-1}\int_0^{1-r} e^{-\theta s^2}\mathrm{d}s + \int_0^{1-r}\mathrm{d}s \\
&- \theta C_{\alpha}\int_0^{1-r} s^2\mathrm{d}s + \theta^2 C^*_{\alpha}\int_0^{1-r} s^4\mathrm{d}s + o\left((1-r)^2\right) \\
=& -2^{1/\alpha-1}\int_0^{1-r}(1-\theta s^2)\mathrm{d}s + (1-r) + o\left((1-r)^2\right) \\
=& (1-2^{1/\alpha-1})(1-r) + o\left((1-r)^2\right),
\end{aligned}
$$

as $r \to 1$ and where $\exp\{-\theta s^2\} = 1 - \theta s^2 + \mathcal{O}\left((1-r)^4\right)$ as $s \to 0$. Thus,

$$
\begin{aligned}
\int_r^1\int_r^1 f_{c_t}(u^*,v^*)\mathrm{d}v^*\,\mathrm{d}u^* =& 2(1-2^{1/\alpha-1})(1-r) + o\left((1-r)^2\right) \\
=& (2-2^{1/\alpha})(1-r) + o\left((1-r)^2\right),
\end{aligned}
$$

as $r \to 1$.

As before, for $\int_r^1 \int_0^1 f_{c_t}(u^*, v^*)\mathrm{d}v^* \,\mathrm{d}u^*$, we take the change of variable $y = xz$, with $z = y/x \in \mathbb{R}^+$, so we have $u^* = \exp\{-x\}$ and $v^* = \exp\{-xz\}$. Thus, we obtain

$$\begin{aligned}
&\int_r^1 \int_0^1 f_{c_t}(u^*, v^*)\mathrm{d}v^* \,\mathrm{d}u^* \\
&= \int_r^1 \int_0^1 \exp\{-\theta(1-u^*)(1-v^*)\} \frac{C_t(u^*, v^*; \alpha)}{u^* v^*} (xy)^{\alpha-1} \left(x^\alpha + y^\alpha\right)^{1/\alpha - 2} \\
&\quad \times \left[\left(x^\alpha + y^\alpha\right)^{1/\alpha} + \alpha - 1\right] \mathrm{d}v^* \,\mathrm{d}u^* \\
&= \int_0^{-\log(r)} \int_0^\infty \exp\{-\theta(1 - \exp\{-x\} - \exp\{-xz\} + \exp\{-x - xz\}) - x(1+z^\alpha)^{1/\alpha}\} \\
&\quad \times z^{\alpha-1} \left(1+z^\alpha\right)^{1/\alpha - 2} \left[x\left(1+z^\alpha\right)^{1/\alpha} + \alpha - 1\right] \mathrm{d}z \,\mathrm{d}x
\end{aligned}$$

We have $\exp\{-x\} = 1 - x + \frac{x^2}{2} + \mathcal{O}\left(x^3\right)$, $\exp\{-xz\} = 1 - xz + \frac{x^2 z^2}{2} + \mathcal{O}\left(x^3\right)$ and $\exp\{-x(1+z)\} = 1 - x(1+z) + \frac{x^2(1+z)^2}{2} + \mathcal{O}\left(x^3\right)$ as $x \to 0$. So, the exponential term

$$\begin{aligned}
&\exp\{-\theta(1 - \exp\{-x\} - \exp\{-xz\} + \exp\{-x(1+z)\}) - x(1+z^\alpha)^{1/\alpha}\} \\
&= \exp\left\{-\theta\left[1 - \left(1 - x + \frac{x^2}{2}\right) - \left(1 - xz + \frac{x^2 z^2}{2}\right)\right.\right. \\
&\qquad \left.\left. + \left(1 - x(1+z) + \frac{x^2(1+z)^2}{2}\right)\right] - x(1+z^\alpha)^{1/\alpha}\right\} + \mathcal{O}\left(x^3\right) \\
&= \exp\left\{-\theta x^2 z - x(1+z^\alpha)^{1/\alpha}\right\} + \mathcal{O}\left(x^3\right) = \exp\{-x(1+z^\alpha)^{1/\alpha}\} + \mathcal{O}\left(x^2\right)
\end{aligned}$$

as $x \to 0$.

So, we have

$$
\begin{aligned}
&\int_r^1 \int_0^1 f_{c_t}(u^*, v^*)\mathrm{d}v^*\,\mathrm{d}u^* \\
&\quad= \int_0^{-\log(r)} \int_0^\infty \exp\{-x(1+z^\alpha)^{1/\alpha}\} z^{\alpha-1}\,(1+z^\alpha)^{1/\alpha-2} \\
&\qquad \times \left[x\,(1+z^\alpha)^{1/\alpha} + \alpha - 1\right] \mathrm{d}z\,\mathrm{d}x \\
&\quad= \int_0^{-\log(r)} \int_x^\infty \exp\{-w\} \left(\frac{x}{w}\right)^\alpha (w+\alpha-1)\frac{1}{x}\mathrm{d}w\,\mathrm{d}x \\
&\quad= \int_0^{-\log(r)} \int_0^w x^{\alpha-1} \exp\{-w\} w^{-\alpha}(w+\alpha-1)\mathrm{d}x\,\mathrm{d}w \\
&\qquad + \int_{-\log(r)}^\infty \int_0^{-\log(r)} x^{\alpha-1} \exp\{-w\} w^{-\alpha}(w+\alpha-1)\mathrm{d}x\,\mathrm{d}w \\
&\quad= \int_0^{-\log(r)} \exp\{-w\} w^{-\alpha}(w+\alpha-1) \left[\frac{x^\alpha}{\alpha}\right]_0^w \mathrm{d}w \\
&\qquad + \int_{-\log(r)}^\infty \exp\{-w\} w^{-\alpha}(w+\alpha-1) \left[\frac{x^\alpha}{\alpha}\right]_0^{-\log(r)} \mathrm{d}w
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{\alpha} \int_0^{-\log(r)} \exp\{-w\}(w+\alpha-1)\mathrm{d}w \\
&+ \frac{[-\log(r)]^\alpha}{\alpha} \int_{-\log(r)}^\infty \exp\{-w\} w^{-\alpha}(w+\alpha-1)\mathrm{d}w \\
&= \frac{1}{\alpha} \int_0^{-\log(r)} w\exp\{-w\}\mathrm{d}w + \frac{\alpha-1}{\alpha} \int_0^{-\log(r)} \exp\{-w\}\mathrm{d}w \\
&+ \frac{[-\log(r)]^\alpha}{\alpha} \int_1^\infty r^t[-\log(r)t]^{-\alpha}(-\log(r)t+\alpha-1)(-\log(r)\mathrm{d}t \\
=&\frac{1}{\alpha}\,(r\log(r) - r + 1) + \frac{\alpha-1}{\alpha}\,(-r+1) \\
&+ \frac{-\log(r)[-\log(r)]^{-\alpha}[-\log(r)]^\alpha}{\alpha} \int_1^\infty r^t t^{-\alpha}(-\log(r)t+\alpha-1)\mathrm{d}t \\
=&\frac{r\log(r)}{\alpha} + 1 - r - \frac{\log(r)}{\alpha} \int_1^\infty r^t t^{-\alpha}(-\log(r)t+\alpha-1)\mathrm{d}t,
\end{aligned}
$$

where $w = x(1+z^\alpha)^{1/\alpha}$ and $t = \frac{w}{-\log(r)}$. As $r \to 1$, we have

$$\begin{aligned}
&\int_r^1 \int_0^1 f_{c_t}(u^*, v^*)\mathrm{d}v^*\,\mathrm{d}u^* \\
&\quad = \frac{r\log(r)}{\alpha} + 1 - r - \frac{\log(r)}{\alpha}\int_1^\infty t^{-\alpha}(\alpha-1)\mathrm{d}t \\
&\quad = \frac{r\log(r)}{\alpha} + 1 - r - \frac{\log(r)}{\alpha} \\
&\quad = \left(1 - \frac{\log(r)}{\alpha}\right)(1-r) = \left(1 - \frac{-(1-r) + \mathcal{O}\left((1-r)^2\right)}{\alpha}\right)(1-r) \\
&\quad = (1-r) + \frac{1}{\alpha}(1-r)^2 + \mathcal{O}\left((1-r)^3\right).
\end{aligned}$$

**Extremal dependence $\chi$ for this case**

Let

$$\begin{aligned}
c_1 &= 2 - 2^{1/\alpha} = \chi_{Gumbel}, & c_3 &= 1/\alpha, \\
c_2 &= 1, & c_4 &= -B_{\beta,\theta}/2.
\end{aligned}$$

we then have

$$\begin{aligned}
\chi &= \lim_{r\to 1} \frac{c_1(1-r) + o\left((1-r)^2\right)}{c_2(1-r) + c_3(1-r)^2 + c_4(1-r)^2 + o\left((1-r)^2\right)} \\
&= \lim_{r\to 1}\left(\frac{c_1}{c_2} - \frac{c_3+c_4}{c_2^2}(1-r) + \mathcal{O}\left((1-r)^2\right)\right) = \frac{c_1}{c_2} = 2 - 2^{1/\alpha}
\end{aligned} \tag{A.2.2}$$

For the vector of parameters $\boldsymbol{\gamma} = (3, 1, 1.844444)$, $c_1 \approx 0.740079$ and $c_5 = 1$. Thus, from equation (A.2.2), we have $\chi \approx 0.740079$. Moreover, from the numerical investigation, $\chi(r) \approx 0.7350891$ with $r = 0.9998779$. Figure A.2.3 shows this comparison.

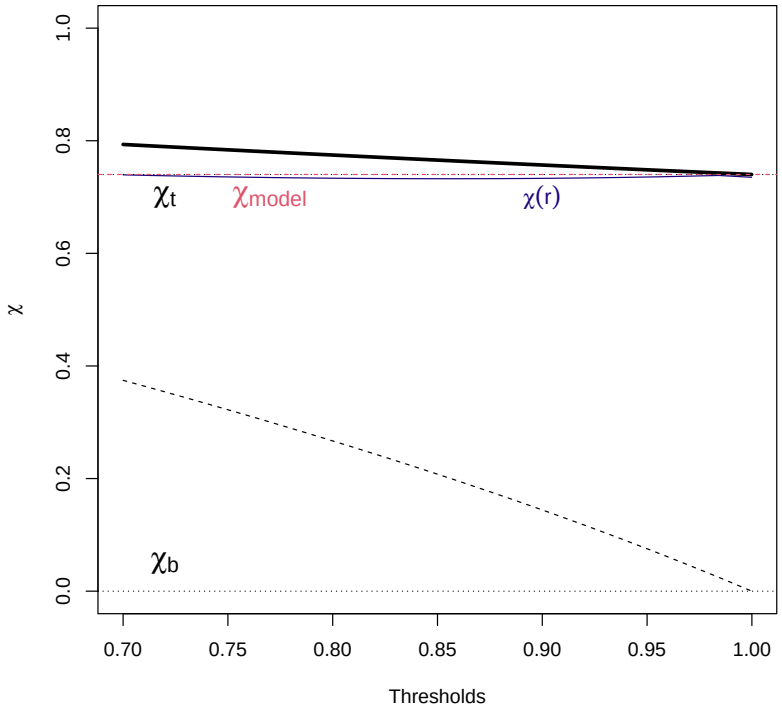



Figure A.2.3: The blue line represents $\chi(r)$ for $r \in [0.7, 1)$ with weighting function $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$ and $\theta = 1.84444$. The thick black lines represent the single copula models - Frank (dashed) and Gumbel (solid). The theoretical values for the Frank and Gumbel copulas based on Table 2 of Section 2.3 from the main text are represented by the horizontal dashed lines, and the value derived for the model is represented by the pink dashed line. Note that the theoretical value for the Gumbel copula, $\chi_t$, is the same as the one derived for the model, $\chi_{\text{Model}}$.

For the vector of parameters $\boldsymbol{\gamma} = (1.5, 2, 3.488889)$, $c_1 \approx 0.4125989$ and $c_5 = 1$. Thus, from equation (A.2.1), we have $\chi \approx 0.4125989$. Moreover, from the numerical investigation, $\chi(r) \approx 0.4093587$ with $r = 0.9998779$. Figure A.2.4 shows this comparison.

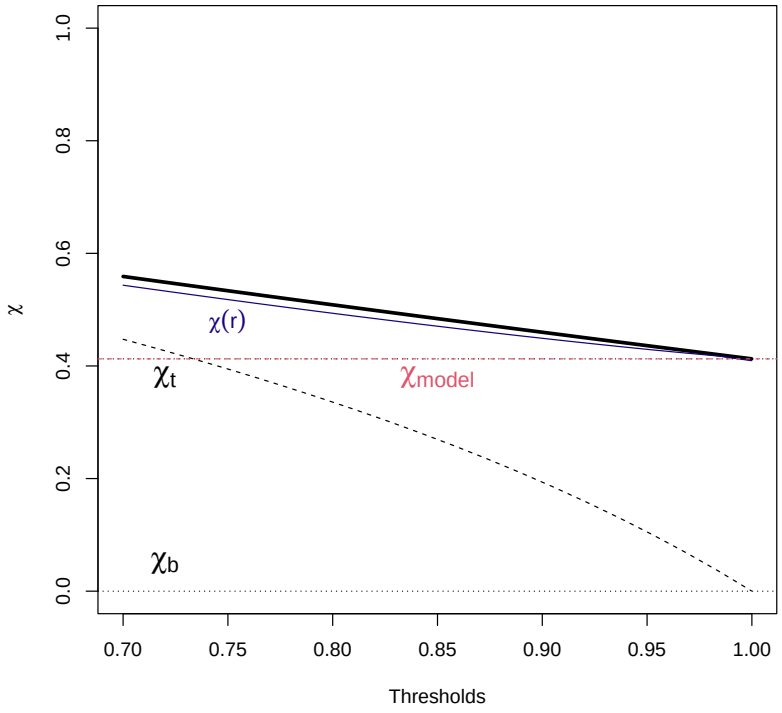



Figure A.2.4: The blue line represents $\chi(r)$ for $r \in [0.7, 1)$ with weighting function $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$ and $\theta = 3.488889$. The thick black lines represent the single copula models - Frank (dashed) and Gumbel (solid). The theoretical values for the Frank and Gumbel copulas based on Table 2 of Section 2.3 from the main text are represented by the horizontal dashed lines, and the value derived for the model is represented by the pink dashed line. Note that the theoretical value for the Gumbel copula, $\chi_t$, is the same as the one derived for the model, $\chi_{\text{Model}}$.

### Extremal dependence $\eta$ for this case

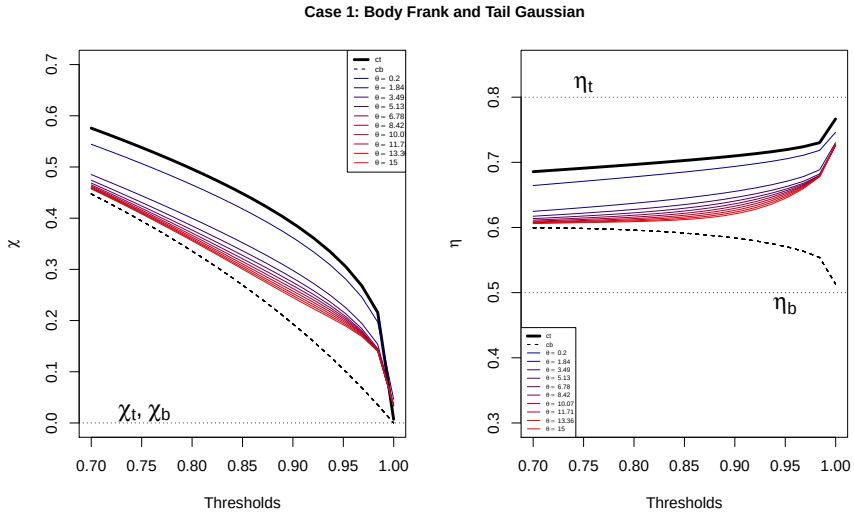
As $\chi > 0$, we should expect $\eta = 1$. Following equation (4) of Section 1.3 from the main text, we have

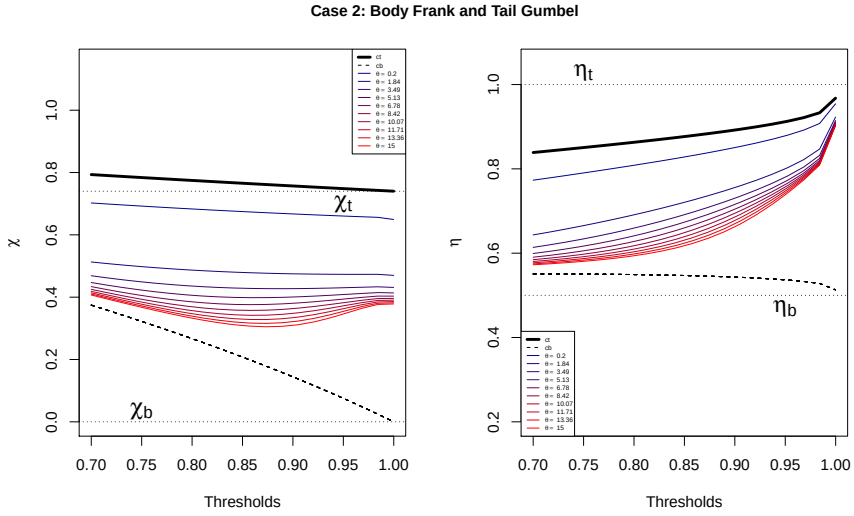
$$
\begin{aligned}
\eta &= \lim_{r\to 1} \frac{\log\left(P[U^* > r]\right)}{\log\left(P[U^* > r, V^* > r]\right)} \\
&= \lim_{r\to 1} \frac{\log\left[c_2(1-r) + c_3(1-r)^2 + c_4(1-r)^2 + o\left((1-r)^2\right)\right]}{\log\left[c_1(1-r) + o\left((1-r)^2\right)\right]} \\
&\overset{(\frac{\infty}{\infty})}{=} \lim_{r\to 1} \frac{-c_2 - 2(c_3+c_4)(1-r) + o\left(1-r\right)}{-c_1 + o\left(1-r\right)} \frac{c_1 + o\left((1-r)^2\right)}{c_2 + (c_3+c_4)(1-r) + o\left((1-r)^2\right)} \\
&= \frac{c_2}{c_1}\frac{c_1}{c_2} = 1
\end{aligned}
$$

by L'Hôpital's Rule.

## A.3 Extremal dependence properties: numerical investigation

Figures A.3.1 and A.3.2 show the results of the numerical study presented in Section 2.3 of the main text for the remaining three models considered.



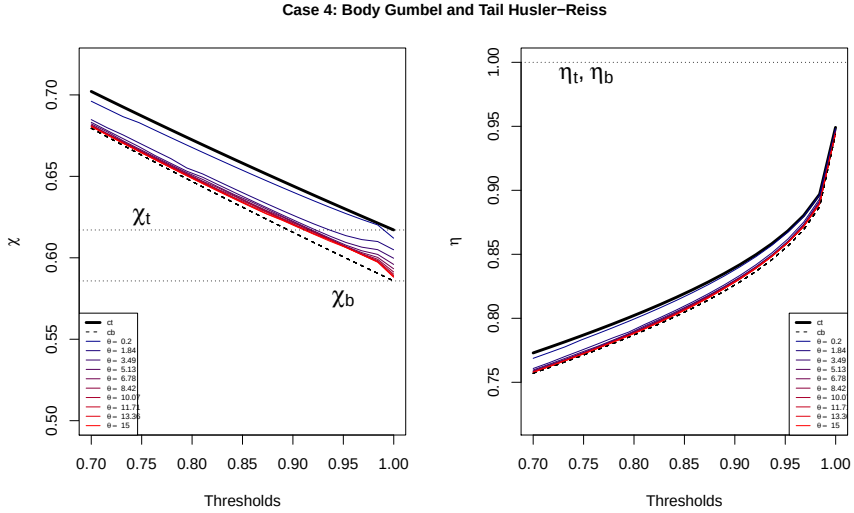
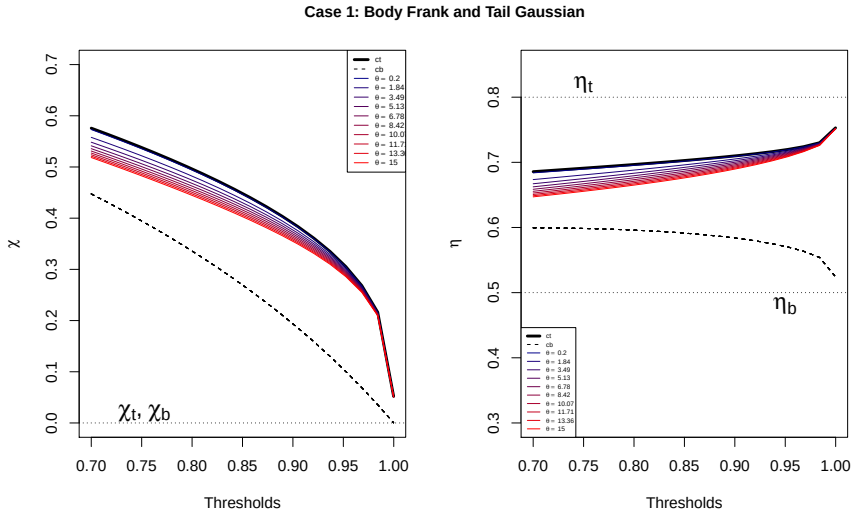
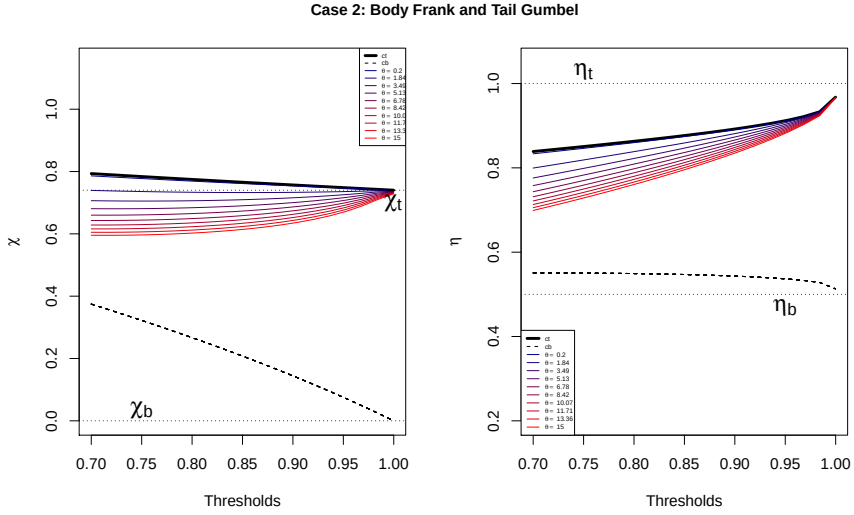




Figure A.3.1: $\chi(r)$ and $\eta(r)$ for $r \in [0.7, 1)$ with weighting function $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$.



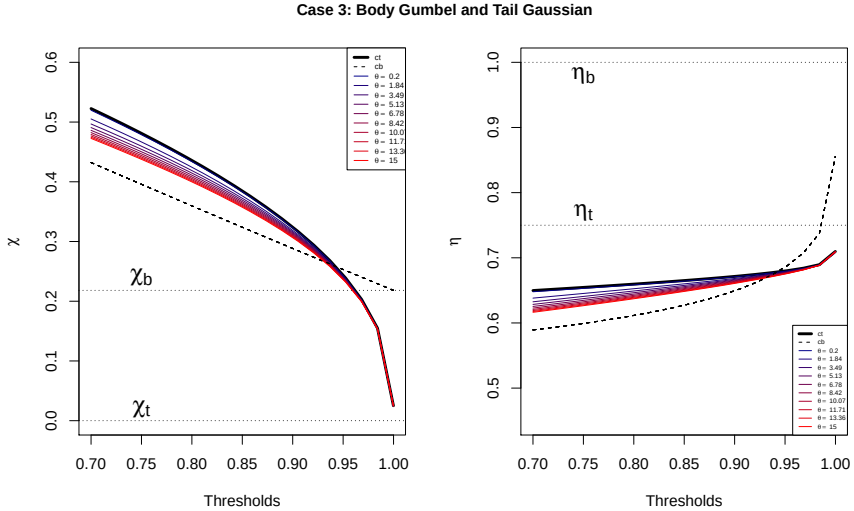
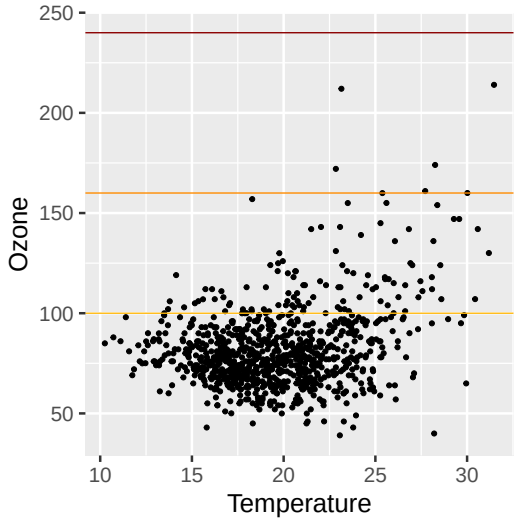



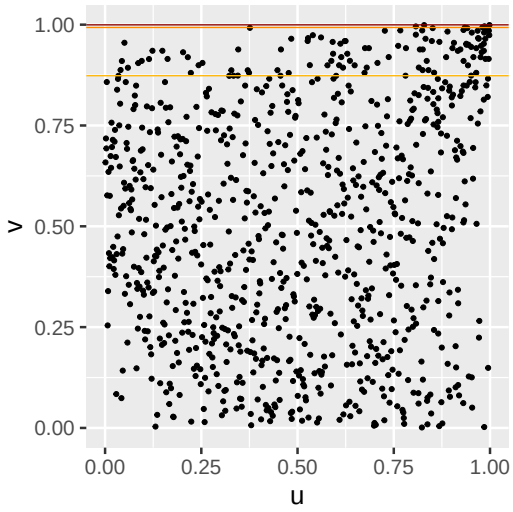

Figure A.3.2: $\chi(r)$ and $\eta(r)$ with weighting function $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$, for $r \in [0.7, 1)$.

## A.4 Ozone and temperature analysis for Weybourne, UK

Following the same structure as the case study in Section 4 in the main paper, the analysis for the summers of 2010 to 2019 of Weybourne, UK, is presented here. Figures A.4.1a and A.4.1b show the scatterplots of the daily maxima of temperature and the daily maxima of ozone on the original scale and on uniform margins, respectively.




(a) Daily maxima of temperature and ozone. The moderate, high and very high DAQI are represented by the yellow, orange and red lines, respectively.




(b) Daily maxima of temperature (u) and ozone (v) on uniform margins. The corresponding moderate, high and very high DAQI are represented by the yellow, orange and red lines, respectively.

Figure A.4.1: Summer data from 2010 to 2019 for Weybourne, UK.

### Model fitting

Table A.4.1 shows the MLEs obtained by fitting a range of single copulas and the corresponding AIC values, whereas Figure A.4.2 illustrates the comparison between the empirical extremal dependence measure $\eta(r)$ for $r \in (0, 1)$ and the model-derived ones.

Table A.4.1: MLEs for ten copulas and their AIC values. Lower AIC values are preferred.

| Copula | Parameter | AIC |
|---|---|---|
| Clayton | $7.21 \times 10^{-9}$ | 2.0 |
| Frank | 0.94 | -19.2 |
| Gumbel | 1.18 | -81.7 |
| Inverted Gumbel | 1.03 | 0.9 |
| Galambos | 0.43 | -82.9 |
| Gaussian | 0.18 | -27.6 |
| Joe | 1.34 | -113.8 |
| Student t | 0.17 8.95 | -34.9 |
| Hüsler-Reiss | 0.82 | -99.1 |
| Coles-Tawn | 0.16 0.24 | -80.4 |

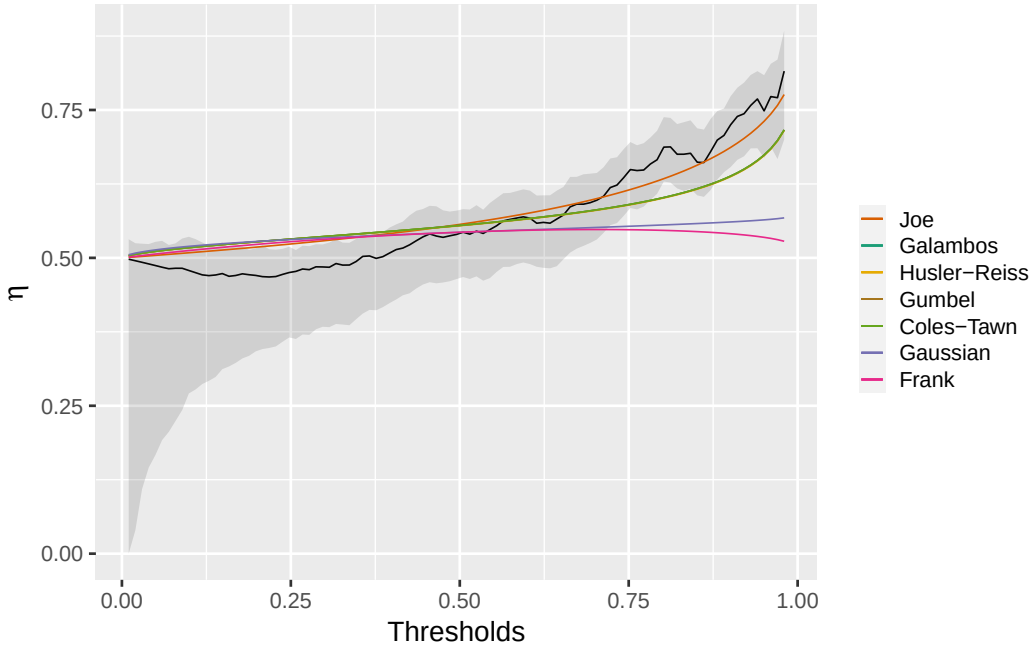



Figure A.4.2: Empirical $\eta(r)$ (in black) and $\eta(r)$ for seven copulas (in colour) for $r \in (0, 1)$. The 95% confidence bands were obtained by block bootstrapping. Note that the $\eta(r)$ for the Galambos, the Hüsler-Reiss, the Gumbel and the Coles-Tawn copulas overlap.

Table A.4.2 shows the MLEs when fitting a range of weighted copula models with $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$ and their AIC values. Table A.4.3 shows the MLEs of the five best models according to AIC when the weighting function is $\pi(u^*, v^*; \theta) = \exp\{-\theta(1 - u^*)(1 - v^*)\}$.

Table A.4.2: MLEs for different weighted copula models and their AIC values when the weighting function used is $\pi(u^*, v^*; \theta) = (u^* v^*)^\theta$. Lower AIC values are preferred.

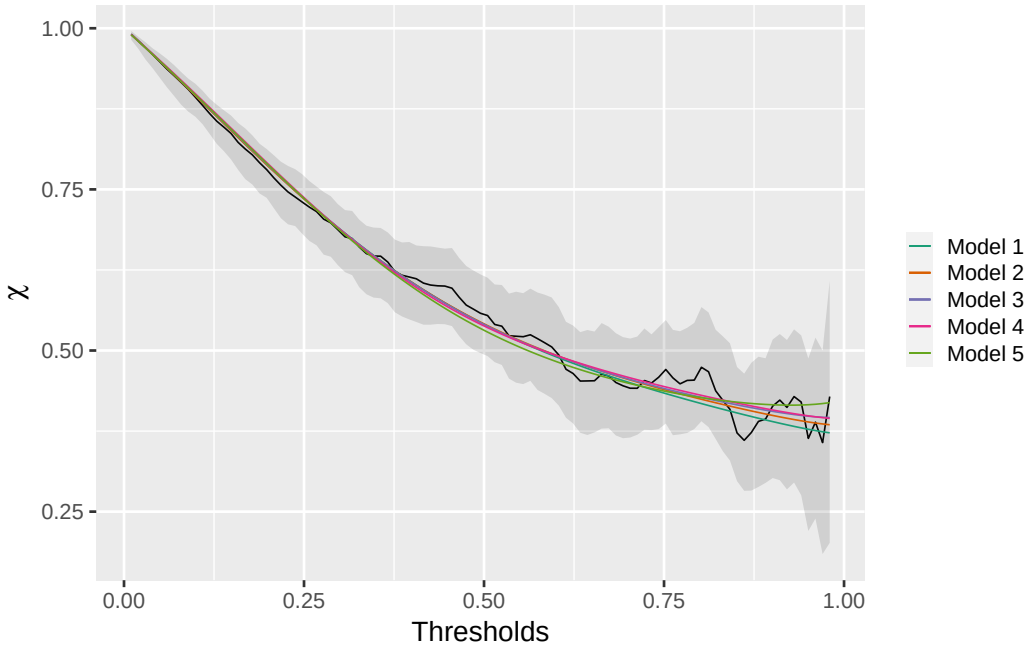
| Model | $c_t$ | $c_b$ | $\hat{\boldsymbol{\alpha}}$ | $\hat{\beta}$ | $\hat{\theta}$ | AIC |
|---|---|---|---|---|---|---|
| Model 1 | Hüsler-Reiss | Gaussian | 1.08 | -0.23 | 0.34 | -124.2 |
| Model 2 | Galambos | Gaussian | 0.66 | -0.23 | 0.33 | -121.9 |
| Model 3 | Coles-Tawn | Gaussian | 0.29 1.10 | -0.22 | 0.34 | -122.5 |
| Model 4 | Coles-Tawn | Frank | 0.30 1.22 | -1.59 | 0.32 | -123.8 |
| Model 5 | Joe | Frank | 1.46 | -1.95 | 0.16 | -126.7 |
| Model 6 | Clayton | Gaussian | 14.99 | -0.05 | 4.33 | -92.8 |
| Model 7 | Inverted Gumbel | Gaussian | 2.33 | -0.15 | 0.96 | -105.4 |
| Model 8 | Hüsler-Reiss | Joe | 1.19 | 1.26 | 4.93 | -112.2 |
| Model 9 | Student t | Galambos | 0.69 4.82 | 0.27 | 2.71 | -98.0 |
| Model 10 | Gaussian | Clayton | 0.75 | $1.16 \times 10^{-5}$ | 2.45 | -99.1 |
| Model 11 | Gumbel | Joe | 1.47 | 1.26 | 4.27 | -111.8 |

Table A.4.3: MLEs for five weighted copula models and their AIC values when the weighting function used is $\pi(u^*, v^*; \theta) = \exp\{-\theta(1-u^*)(1-v^*)\}$. Lower AIC values are preferred.

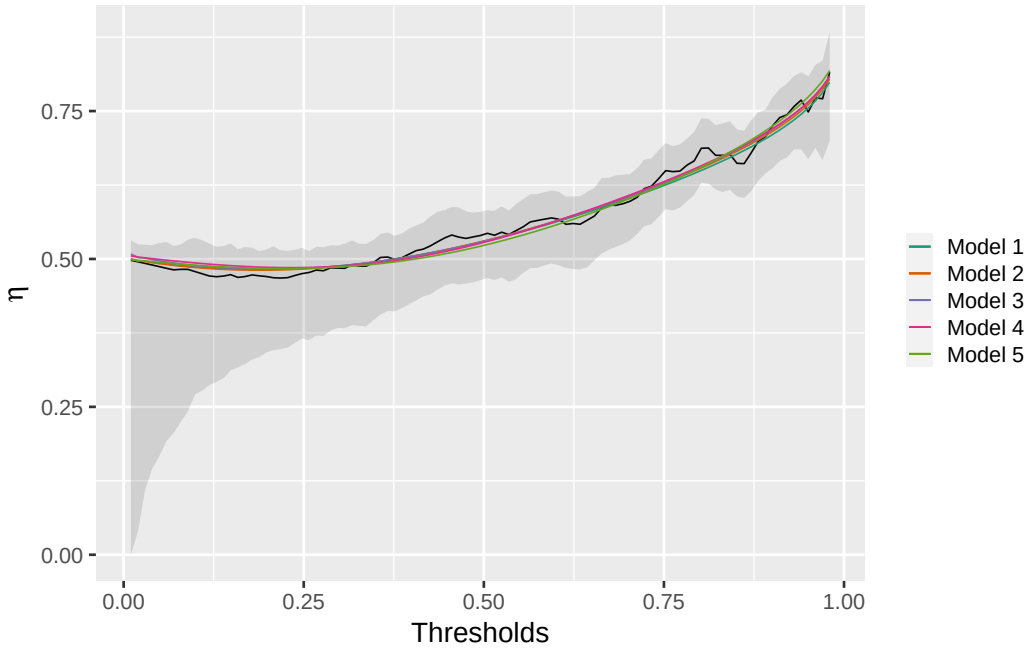
| Model | $c_t$ | $c_b$ | $\hat{\boldsymbol{\alpha}}$ | $\hat{\beta}$ | $\hat{\theta}$ | AIC |
|---|---|---|---|---|---|---|
| Model 1 | Hüsler-Reiss | Gaussian | 1.12 | -0.52 | 3.21 | -158.5 |
| Model 2 | Galambos | Gaussian | 0.72 | -0.51 | 3.48 | -159.2 |
| Model 3 | Coles-Tawn | Gaussian | 0.46 0.82 | -0.48 | 4.13 | -158.1 |
| Model 4 | Coles-Tawn | Frank | 0.48 0.74 | -3.05 | 3.61 | -150.0 |
| Model 5 | Joe | Frank | 1.52 | -2.63 | 2.85 | -147.1 |

## Diagnostics

Figure A.4.3 displays $\chi(r)$ and $\eta(r)$ for $r \in (0, 1)$ for the five models considered. A clear improvement from the single copula models shown in Figure A.4.2 can be seen as now all five models offer a reasonable fit throughout the whole support of the data. In summer, the average temperature in Weybourne is between 18°C and 22°C and the observed 90th, 95th and 99th percentiles of the temperature are around 24°C, 26°C and 29°C, respectively. Table A.4.4 shows Kendall's $\tau$ and some probabilities of interest.




(a) Empirical $\chi(r)$ (in black) and $\chi(r)$ for the five models (in colour) for $r \in (0, 1)$. The 95% confidence bands were obtained by block bootstrapping.




(b) Empirical $\eta(r)$ (in black) and $\eta(r)$ for the five models (in colour) for $r \in (0, 1)$. The 95% confidence bands were obtained by block bootstrapping.

Figure A.4.3: Dependence measures $\chi(r)$ and $\eta(r)$.

Table A.4.4: Diagnostics for the best five models according to their AIC values. The 95% confidence intervals for the empirical values were obtained by block bootstrapping. The empirical probability $P[O_3 \geq 160 \mid 29 \leq T \leq 30]$ and its 95% confidence interval are explained by the low number of observations present in the data set.

| Model | Kendall's $\tau$ | $P[T \leq 15, O_3 \geq 100]$ | $P[T \geq 24, O_3 \geq 100]$ |
|---|---|---|---|
| Empirical | 0.0966 | 0.0045 | 0.0460 |
| (95% CI) | (0.0555 , 0.1934) | (0.0000 , 0.0050) | (0.0338 , 0.0667) |
| Model 1 | 0.0881 | 0.0072 | 0.0491 |
| Model 2 | 0.0900 | 0.0076 | 0.0502 |
| Model 3 | 0.0853 | 0.0084 | 0.0509 |
| Model 4 | 0.0944 | 0.0069 | 0.0512 |
| Model 5 | 0.0882 | 0.0068 | 0.0517 |

| Model | $P[T \geq 26, O_3 \geq 100]$ | $P[O_3 \geq 100 \mid 24 \leq T \leq 25]$ | $P[O_3 \geq 160 \mid 29 \leq T \leq 30]$ |
|---|---|---|---|
| Empirical | 0.0291 | 0.1520 | 0.0000 |
| (95% CI) | (0.0189 , 0.0438) | (0.0488 , 0.2800) | (0.0000 , 0.0000) |
| Model 1 | 0.0283 | 0.2557 | 0.1912 |
| Model 2 | 0.0287 | 0.2617 | 0.1982 |
| Model 3 | 0.0300 | 0.2516 | 0.1894 |
| Model 4 | 0.0298 | 0.2573 | 0.1921 |
| Model 5 | 0.0297 | 0.2646 | 0.2176 |

# Appendix B

# Supplementary material for Chapter 4

## B.1 Formulation of set $A_w$ from Section 4.3.2

Here we present only the bivariate case, with the general $d$-dimensional case following similarly. Consider $w \in \mathcal{S}_1$, and standard exponential random variables, $X_1^E$ and $X_2^E$, with marginal distribution function $F_E(x) = 1 - \exp\{-x\}$ for $x > 0$ and $i = 1, 2$. Following Wadsworth and Tawn (2013), we are interested in regions of the form

$$A(x, y) = \{X_1^E > x, X_2^E > y\}, \tag{B.1.1}$$

for $x > 0$ and $y > 0$.

Let us now assume that $w := x/(x + y)$ and $\max\{x, y\} = u^E$, where $u^E$ is some threshold level in exponential margins. Two examples of such sets are shown by the shaded regions in Figure B.1.1.

By the definition of $w$, we have that

$$\max\{x, y\} = x \Leftrightarrow \max\left\{x, \frac{1-w}{w}x\right\} = x \Rightarrow 1 > \frac{1-w}{w} \Leftrightarrow w > \frac{1}{2}.$$

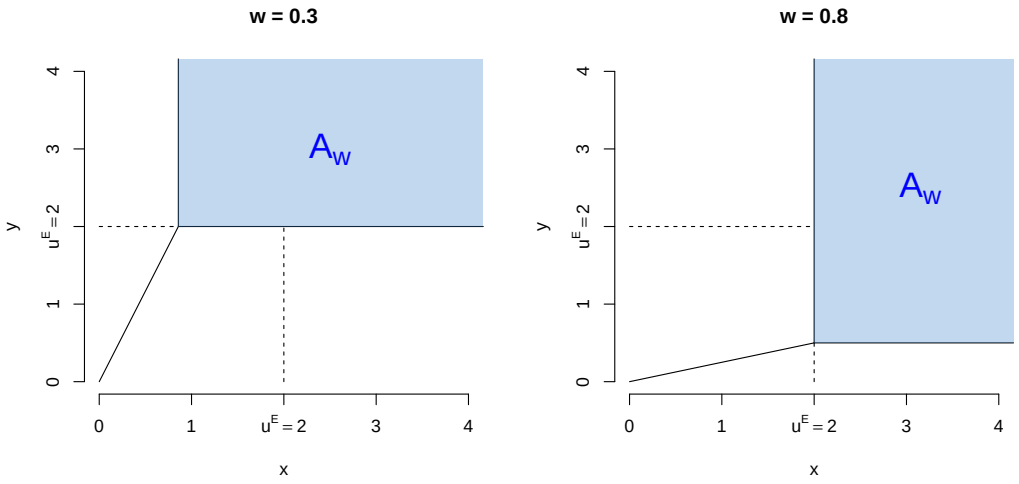



Figure B.1.1: Example of regions $A_w$ for $w = \{0.3, 0.8\}$ and $u^E = 2$.

Thus, we have $x = u^E$ and set $A(x, y)$ from expression (B.1.1) can be rewritten as

$$A_w = \left\{ X_1^E > u^E, X_2^E > \frac{1-w}{w} u^E \right\},$$

when $w > 1/2$.

Similarly, we have $w \leq 1/2$ when $\max\{x, y\} = y = u^E$. Therefore,

$$A_w = \left\{ X_1^E > \frac{w}{1-w} u^E, X_2^E > u^E \right\},$$

for $w \leq 1/2$. Combining the two, we arrive to region given in Section 4.3.2.

$$A_w = \left\{ X_1^E > \max\left\{ \frac{w}{1-w}, 1 \right\} u^E, X_2^E > \max\left\{ \frac{1-w}{w}, 1 \right\} u^E \right\}.$$

## B.2 Simulation Studies

### B.2.1 Model inference

For Case III, we take $p = 0.71$, $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = (5, 3, 2, 3, 5)$, $\boldsymbol{\sigma}_{\Sigma_1} = (1.00, 0.60, 1.60, 0.80, 1.80)$, $\boldsymbol{\sigma}_{\Sigma_2} = (6.26, 4.31, 3.23, 4.01, 1.34)$, $\boldsymbol{\rho}_{\Sigma_1} = (0.26, -0.08, 0.34, 0.37, -0.41, 0.14, 0.19, -0.27, 0.35, -0.17)$ and $\boldsymbol{\rho}_{\Sigma_2} = (-0.14, -0.06, -0.02, -0.04, 0.06, 0.08, -0.23, 0.68, -0.56, 0.19)$.

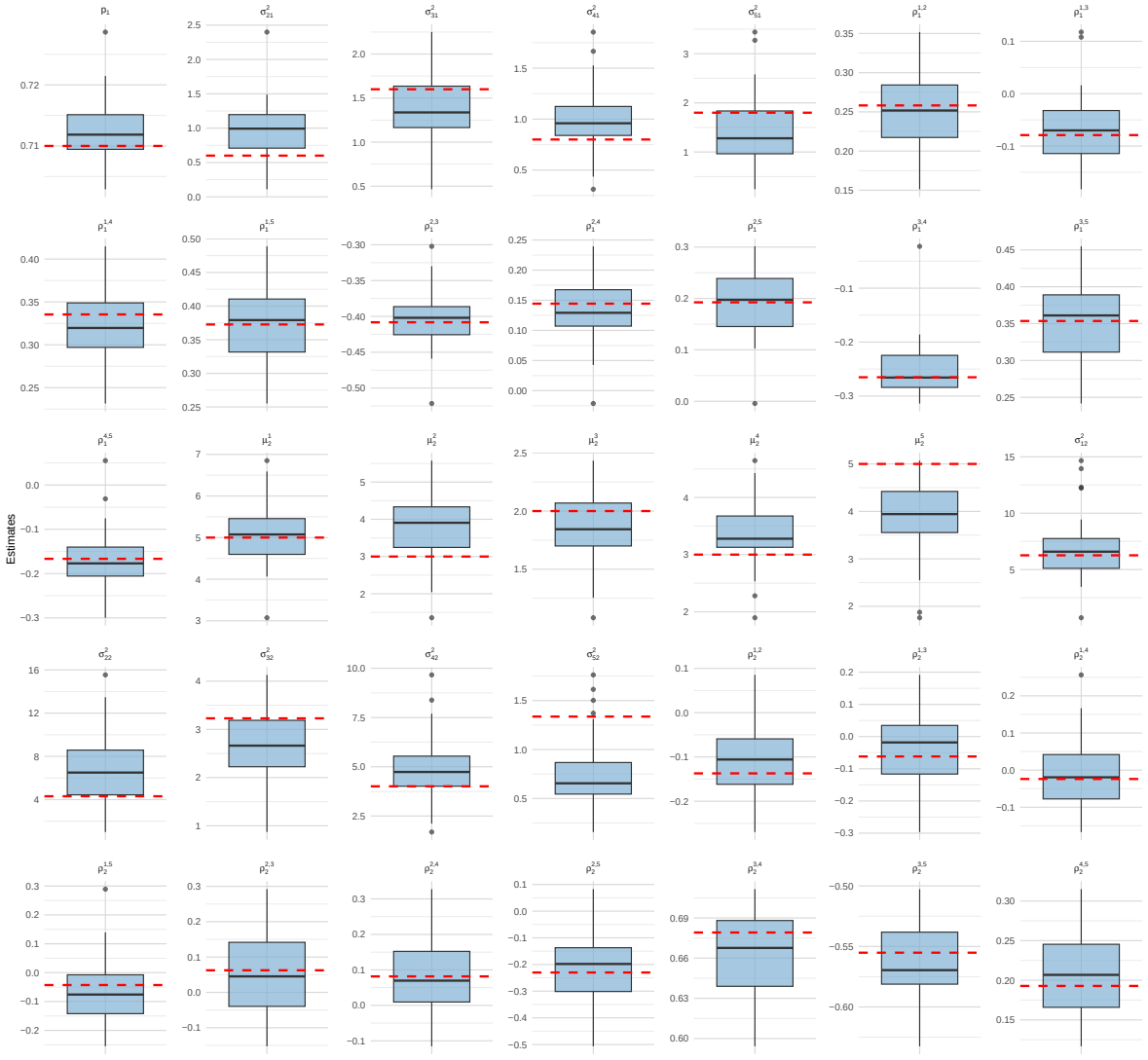



Figure B.2.1: Boxplots of estimates of the Gaussian mixture copula model based on 50 replicated data sets for Case III. The true parameter values are indicated by the red lines.

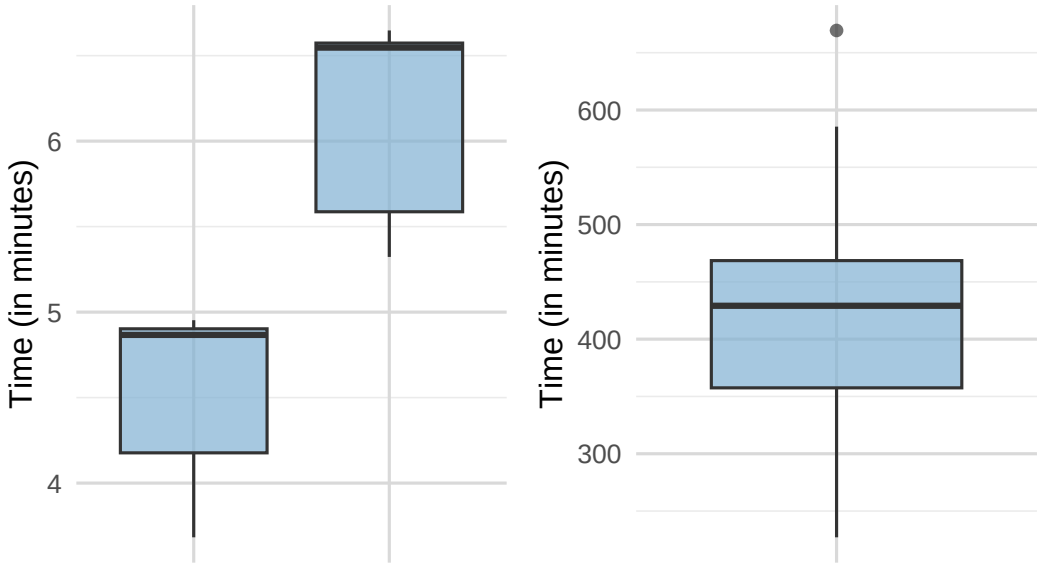



Figure B.2.2: Time (in minutes) taken to optimise the log-likelihood 4.2.4 of a model with $d = 2$ and $k = 2$ or $d = 2$ and $k = 3$ (left), and $d = 5$ and $k = 2$ (right).

**Pairwise exchangeability**

We present now examples of simplified model specifications for Cases I-III defined in Section 4.3.1 from the main paper. More specifically, we consider pairwise exchangeability where for each mixture component $\boldsymbol{Z}_j$, $\mu_j^1 = \ldots = \mu_j^d$ and $\sigma_{1j}^2 = \ldots = \sigma_{dj}^2$ for $j \in K$, $d \in D$.

For Case I, we set $p_1 = 0.30$, $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = \mathbf{3}$, $\boldsymbol{\sigma}_{\Sigma_1} = \mathbf{1}$, $\boldsymbol{\sigma}_{\Sigma_2} = \mathbf{1}.62$, $\rho_{\Sigma_1} = 0.29$ and $\rho_{\Sigma_2} = 0.20$. In Case II, when an extra mixture component is added, we retain the models for the $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ mixture components, and for the extra mixture component we take $(p_1, p_2) = (0.20, 0.53)$, $\boldsymbol{\mu}_3 = \mathbf{5}$, $\sigma_{\Sigma_3} = \mathbf{2}.51$ and $\rho_{\Sigma_3} = 0.02$. For Case III, we take $p = 0.27$, $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = \mathbf{2}$, $\boldsymbol{\sigma}_{\Sigma_1} = \mathbf{1}$, $\boldsymbol{\sigma}_{\Sigma_2} = \mathbf{0}.6$, $\boldsymbol{\rho}_{\Sigma_1} = (-0.12, 0.79, 0.03, 0.11, -0.39, -0.20, -0.24, 0.03, -0.30, -0.23)$ and $\boldsymbol{\rho}_{\Sigma_2} = (-0.14, -0.06, -0.02, -0.04, 0.06, 0.08, -0.23, 0.68, -0.56, 0.19)$. The results are shown in Figure B.2.3.

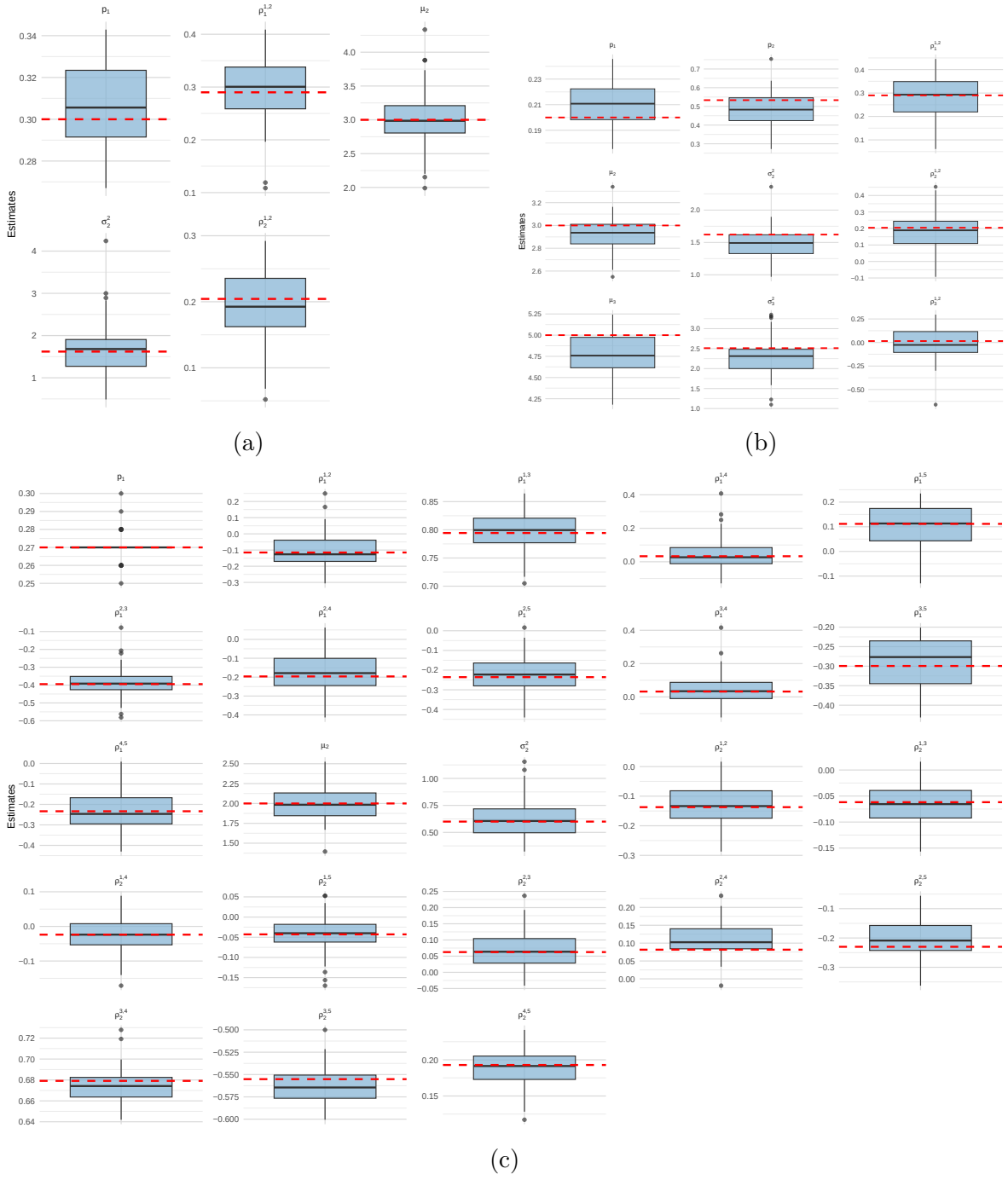



Figure B.2.3: Boxplots of estimates of the Gaussian mixture copula model when assuming pairwise exchangeability based on 50 replicated data sets: (a) Case I, (b) Case II and (c) Case III. The true parameter values are indicated by the red lines.

## B.2.2 Model fit and diagnostics

### Asymptotically independent data

Figure B.2.4 shows the results for $\eta_D(r)$ for the case where the underlying data is AI given in the main paper.

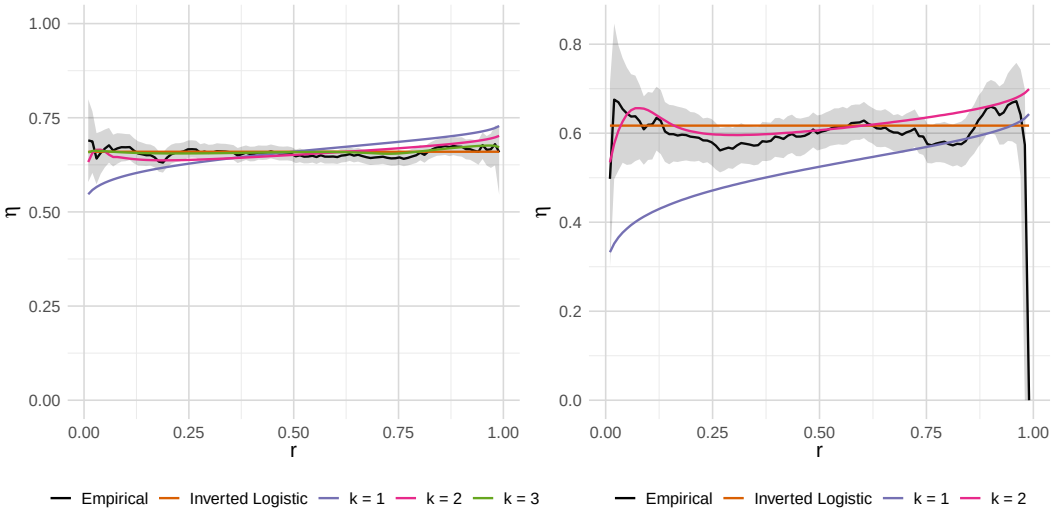



Figure B.2.4: Estimates of $\eta_D(r)$ for $r \in (0.1)$ with true (in orange) and empirical (in black) values also shown. The pointwise 95% confidence intervals for the empirical $\eta_D(r)$ are obtained through bootstrap. When $d = 2$ (left), models with $k = 1 - 3$ mixture components are considered, whereas when $d = 5$ only models with $k = 1 - 2$ mixture components are studied.

When considering a smaller sample size ($n = 1000$), the decrease in AIC with $k = 3$ in relation to when $k = 1$ is of $-32.15$, with $k = 2$ of $-29.47$ relative to $k = 1$. These results indicate that either the $k = 2$ or the $k = 3$ model is suitable to model the data, with a slight preference for the $k = 3$ model. Figure B.2.5 shows a comparison between model-based $\chi_2(r)$ and $\eta_2(r)$ with their true and empirical counterparts. Although small, there are differences between the three fits, especially for the $k = 3$ model.

### Asymptotically dependent data

Figure B.2.6 shows the results for $\eta_D(r)$ for the case where the underlying data is AD given in the main paper.

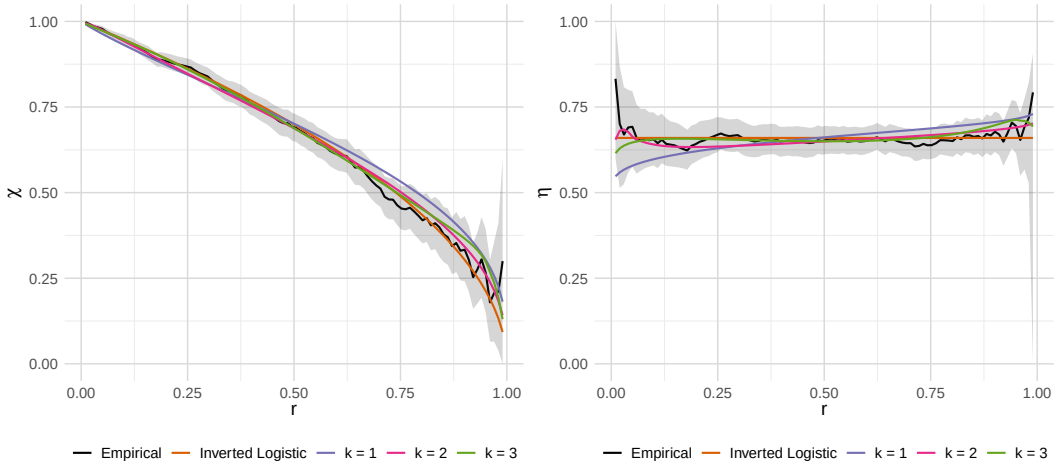



Figure B.2.5: Estimates of $\chi_2(r)$ (left) and of $\eta_2(r)$ (right) for $r \in (0.1)$ with true (in orange) and empirical (in black) values also shown. The pointwise 95% confidence intervals for the empirical $\chi_2(r)$ and $\eta_2(r)$ are obtained through bootstrap.

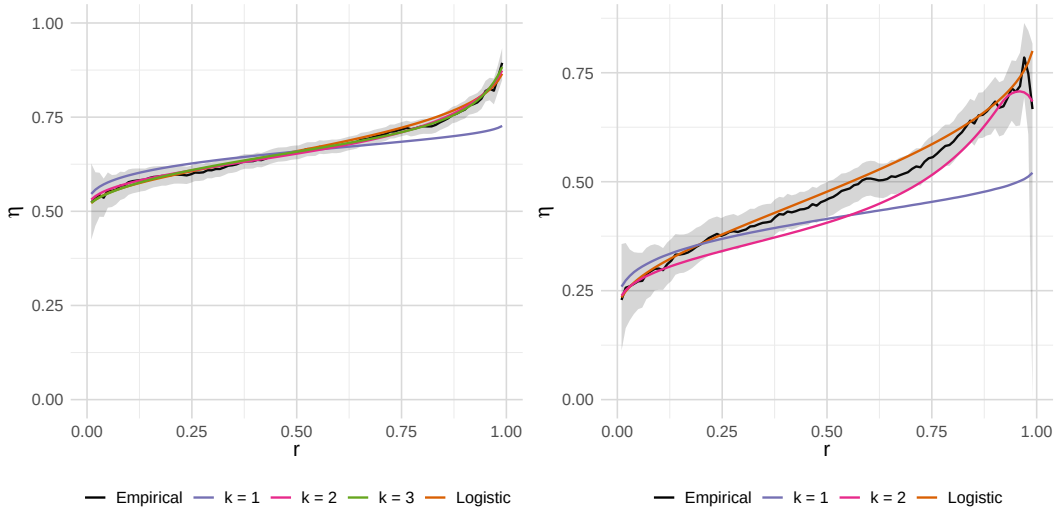



Figure B.2.6: Estimates of $\eta_D(r)$ for $r \in (0.1)$ with true (in orange) and empirical (in black) values also shown. The pointwise 95% confidence intervals for the empirical $\eta_D(r)$ are obtained through bootstrap. When $d = 2$ (left), models with $k = 1 - 3$ mixture components are considered, whereas when $d = 5$ only models with $k = 1 - 2$ mixture components are studied.

When considering a smaller sample size ($n = 1000$), the decrease in AIC with $k = 3$ in relation to when $k = 1$ is $-43.18$, whereas there is an increase in AIC of 11.69 with $k = 2$ relative to $k = 1$. These results indicate that the $k = 3$ model is the most suitable for the underlying data. Figure B.2.7 shows a comparison between model-based $\chi_2(r)$

and $\eta_2(r)$ with their true and empirical counterparts. Only the $k = 3$ is able to capture the extremal behaviour of the underlying data.

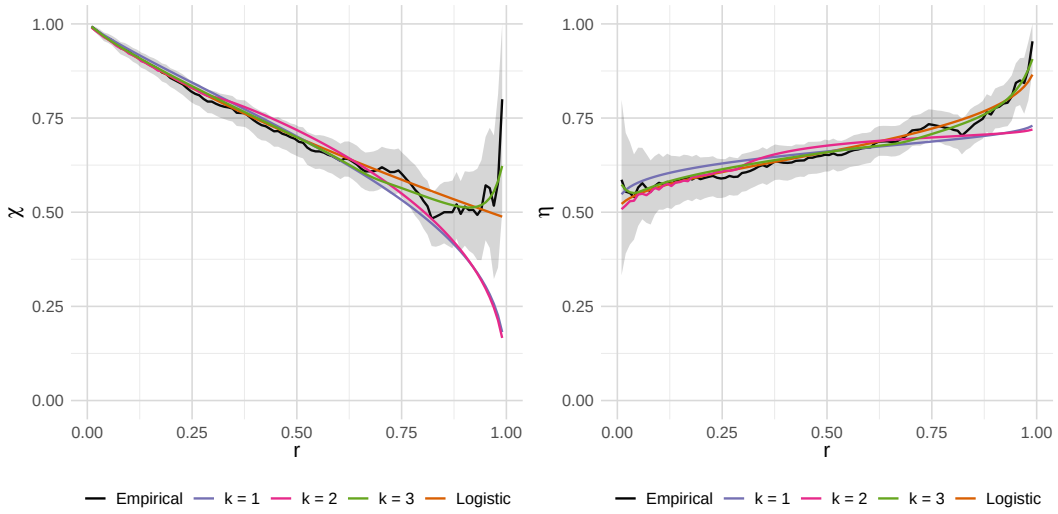



Figure B.2.7: Estimates of $\chi_2(r)$ (left) and of $\eta_2(r)$ (right) for $r \in (0.1)$ with true (in orange) and empirical (in black) values also shown. The pointwise 95% confidence intervals for the empirical $\chi_2(r)$ and $\eta_2(r)$ are obtained through bootstrap.

### Non-exchangeable data

Figure B.2.8 shows the results for $\eta_2(r)$ in the case where the underlying data exhibits asymmetry patterns given in the corresponding section of the main paper.

### Weighted copula model

Figure B.2.9 shows the results for $\eta_2(r)$ for the case where the underlying data is generated from the WCM, which is given in the corresponding section of the main paper.

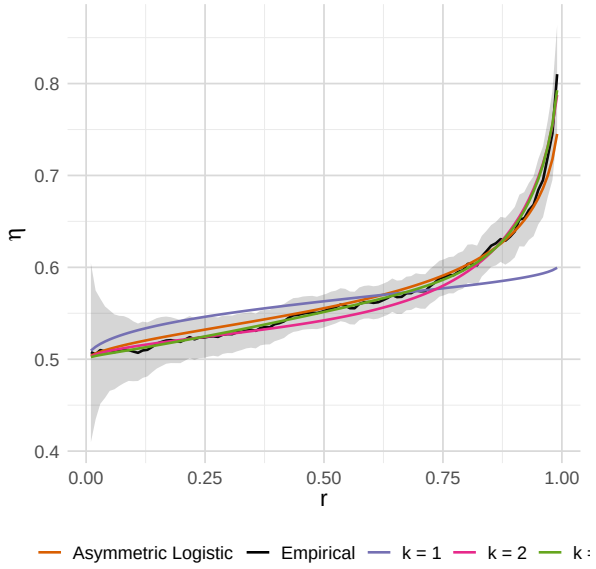



Figure B.2.8: Estimates of $\eta_2(r)$ for $r \in (0.1)$ with true (in orange) and empirical (in black) values also shown. The pointwise 95% confidence intervals for the empirical $\eta_2(r)$ are obtained through bootstrap.

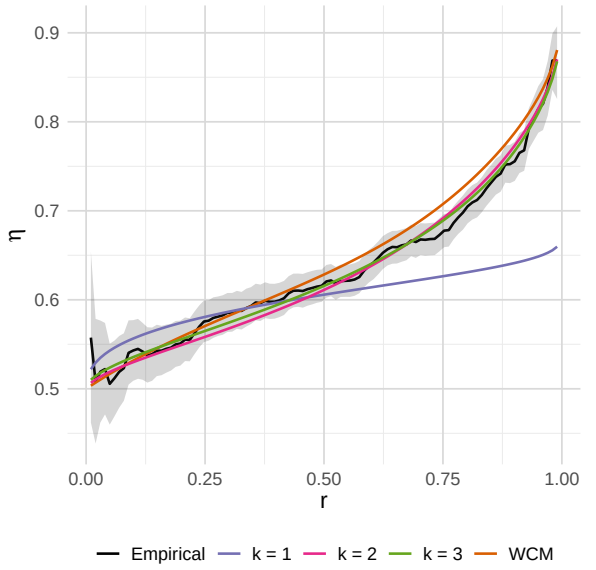



Figure B.2.9: Estimates of $\eta_2(r)$ for $r \in (0.1)$ with true (in orange) and empirical (in black) values also shown. The pointwise 95% confidence intervals for the empirical $\eta_2(r)$ are obtained through bootstrap.

## B.3 Case study: air pollution data

### Pairwise analysis

Figure B.3.1 shows the results for $\eta_2(r)$ for the pairwise analysis presented in the main paper.

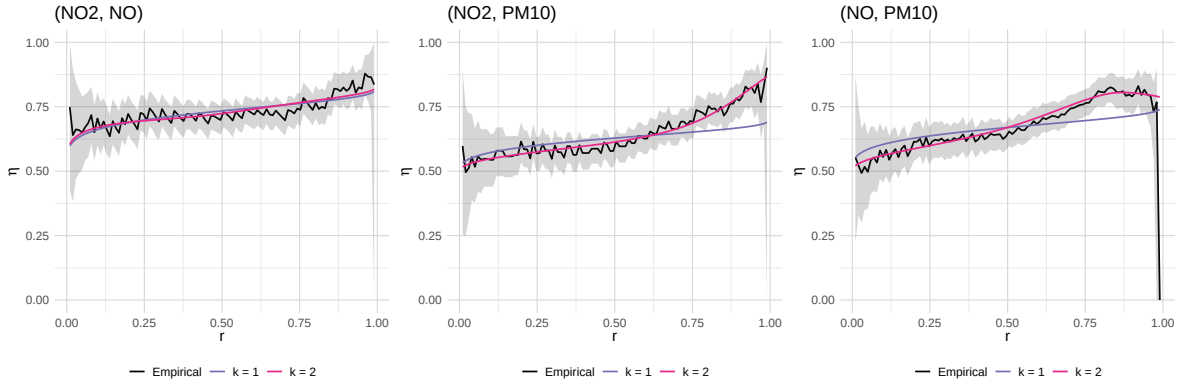



Figure B.3.1: Estimates of $\eta_2(r)$ for $r \in (0.1)$ with empirical (in black) values also shown for pairs $(NO_2, NO)$ (left), $(NO_2, PM_{10})$ (middle) and $(NO, PM_{10})$ (right). The pointwise 95% confidence intervals for the empirical $\chi_2(r)$ are obtained through bootstrap.

### Trivariate analysis

Figure B.3.2 shows the results for $\eta_3(r)$ for the trivariate analysis presented in the corresponding section of the main paper.

### Higher dimensional analysis

Figure B.3.3 shows the results for $\eta_5(r)$ for the full analysis presented in the main paper.

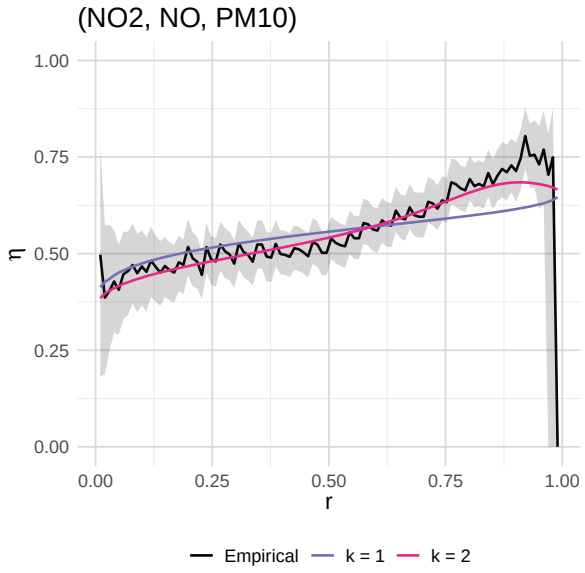



Figure B.3.2: Estimates of $\chi_3(r)$ for $r \in (0.1)$ with empirical (in black) values also shown for the triple $(NO_2, NO, PM_{10})$. The pointwise 95% confidence intervals for the empirical $\chi_2(r)$ are obtained through bootstrap.

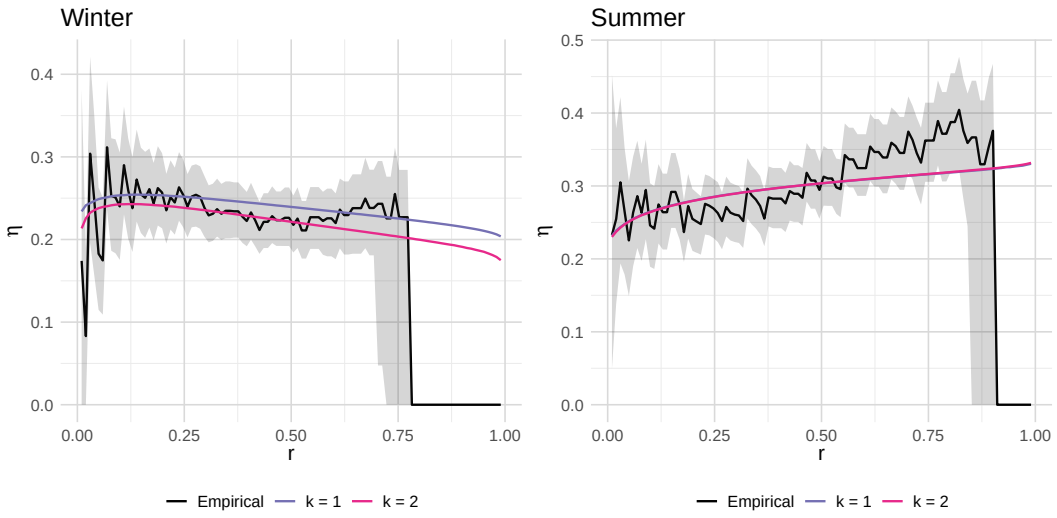



Figure B.3.3: Estimates of $\eta_5(r)$ for $r \in (0.1)$ with empirical (in black) values also shown for $(O_3, NO_2, NO, SO_2, PM_{10})$ in the winter season (left) and the summer season (right). The pointwise 95% confidence intervals for the empirical $\chi_5(r)$ are obtained through bootstrap. Note that $\eta_5(r)$ for $k = 1$ and $k = 2$ overlap in the right panel.

# Appendix C

# Supplementary material for Chapter 5

## C.1 DeepSets architecture

A schematic of the DeepSets architecture (recall Section 5.2.1 of the main paper) used is shown in Figure C.1.1.

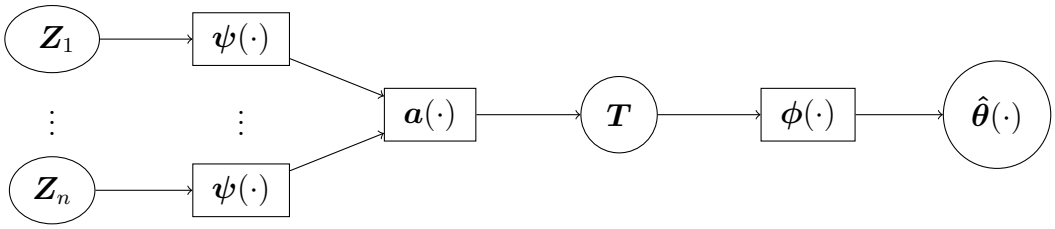



Figure C.1.1: In the first step, the data inputs $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are transformed independently through neural network $\boldsymbol{\psi}(\cdot)$, They are then aggregated through a permutation-invariant function $\boldsymbol{a}(\cdot)$, obtaining the summary statistic $\boldsymbol{T}$. In the last step, neural network $\boldsymbol{\phi}(\cdot)$ maps the summary statistic $\boldsymbol{T}$ to an estimate of the vector of model parameters $\hat{\boldsymbol{\theta}}(\cdot)$.

## C.2 Parameter estimation

In this section, we present the simulation studies done for the remaining models considered in this work. In Section C.2.1, we show the performance of the NBEs for uncensored data in five configurations of the WCM from Section 5.3.4 of the main paper. Where feasible, a comparison with maximum likelihood inference is presented. In Section C.2.2, we show the performance of the NBEs when the sample size and censoring level are kept fixed, and when the sample size is assumed variable but the censoring level is still fixed for Model W. Finally, in Sections C.2.3, C.2.4 and C.2.5, we present the results for the remaining three models from Section 5.3.3 of the main paper. In all these cases, a comparison with censored maximum likelihood estimation is given.

The neural network architecture used for parameter estimation (recall Section 5.4.2 of the main paper) is given in Table C.2.1.

Table C.2.1: Summary of the neural network architecture used to train the NBE. The input array to the first layer represents the dimension $d = 2$ of data set $\boldsymbol{Z}$; this differs for uncensored and censored data. For the censored case, a dense Bilinear layer is used instead, and an extra dimension for the indicator vector $\boldsymbol{I}$ is needed. In addition, the input layer of $\boldsymbol{\phi}(\cdot)$ has an extra dimension in the case of censored data with random censoring level $\tau$. The output array $|\boldsymbol{\theta}|$ to the last layer represents the number of parameters in the model.

| Neural network | Layer type | Input dimension | Output dimension |
|---|---|---|---|
| $\boldsymbol{\psi}(\cdot)$ | Dense | [2] or [2, 2] | [128] |
| | Dense | [128] | [128] |
| | Dense | [128] | [256] |
| $\boldsymbol{\phi}(\cdot)$ | Dense | [256] or [257] | [128] |
| | Dense | [128] | [ $\lvert\boldsymbol{\theta}\rvert$ ] |

### C.2.1 Weighted copula model

We consider now five additional configurations of the WCM. For the first two models, we assume $c_b$ and $c_t$ to be one-parameter copulas, while for the remaining three configurations $c_b$ is assumed to be a Gaussian copula, and $c_t$ is one of the flexible copulas

mentioned in Section 5.3.3 of the main article. For these three models configurations, the likelihood is infeasible and hence no comparison with MLE is provided. In all the models, we take $\pi(x_1, x_2; \gamma) = (x_1 x_2)^\gamma$ as the weighting function. Since preliminary analysis indicated that the neural network was struggling to learn $\gamma$, we set $\kappa = \log \gamma$ and estimate $\kappa$ instead. Lastly, the model-based $\chi(y)$ estimates of the WCM are obtained using a Monte Carlo approximation with 500 000 samples.

**Model 1: $c_b$ is a Gaussian copula and $c_t$ is a logistic copula**

For the first model, we consider the copula tailored to the body $c_b$ to be a Gaussian copula with correlation parameter $\rho \in (-1, 1)$, and the copula tailored to the tail $c_t$ to be a logistic copula with $\alpha_L \in (0, 1]$. Similarly to the weighting function parameter $\gamma$, we take an alternative parameterisation and set $\tau_L = \text{logit}(\alpha_L)$. Additionally, we set $\rho \sim \text{Unif}(-1, 1)$, $\tau_L \sim \text{Unif}(-3, 3)$, which results in $\alpha_L \in (0.05, 0.95)$, and $\kappa \sim \text{Unif}(-3.51, 1.95)$, which leads to $\gamma \in (-0.03, 7.03)$, as the priors for the parameters. The performance of the NBE is assessed in Figure C.2.1 where the true values of the parameters are compared with their estimated values. It can be seen that parameter $\kappa$ exhibits a bit of variability, while parameters $\rho$ and $\tau_L$ are estimated quite well via the NBE. The coverage probabilities and average length of the 95% uncertainty intervals obtained via a non-parametric bootstrap procedure shown in Table C.2.2. Similarly to the main paper, we compute the coverage probabilities of 95% uncertainty intervals, and their average length, for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$; the results are shown on the right of Table C.2.2. According to these results, the true $\chi(y)$ is within the confidence intervals in more than 77% of the time, which suggest that this measure is well derived from the NBE.

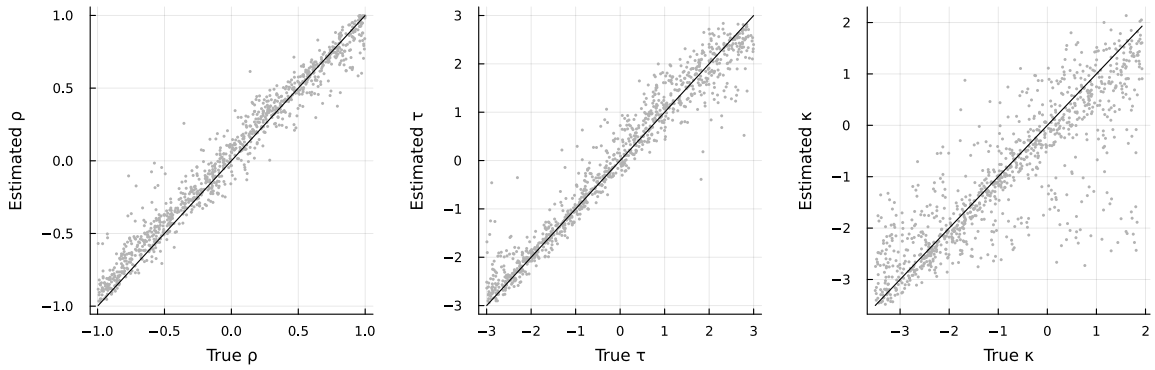



Figure C.2.1: Assessment of the NBE when $c_b$ is a Gaussian copula with correlation parameter $\rho$, $c_t$ is a logistic copula with parameter $\tau_L = \text{logit}(\alpha_L)$, and with weighting function $\pi(x_1, x_2; \kappa) = (x_1 x_2)^{\exp\{\kappa\}}$, $x_1, x_2 \in (0, 1)$ for a sample size of $n = 1000$.

Table C.2.2: Coverage probability and average length of the 95% uncertainty intervals for the parameters (left) and for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$ (right) obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Coverage | Length |
|---|---|---|
| $\rho$ | 0.71 | 0.26 |
| $\tau_L$ | 0.75 | 0.76 |
| $\kappa$ | 0.69 | 1.33 |

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.50)$ | 0.77 | 0.05 |
| $\chi(0.80)$ | 0.79 | 0.08 |
| $\chi(0.95)$ | 0.78 | 0.09 |

**Comparison with maximum likelihood estimation**

Since the likelihood of this model is feasible, though computationally intensive, we compare the estimations obtained by the NBE to the MLEs. With the assigned priors, we generate five different parameter vectors $\boldsymbol{\theta} = (\rho, \tau_L, \kappa)'$ and the corresponding data sets, each of which has $n = 1000$. Additionally, each data set is simulated 100 times. The results are shown in Figure C.2.2; it can be seen that the NBE estimates are generally more biased, and sometimes more variable, than the MLEs. However, they are less likely to have big outliers as the neural network is trained in a bounded interval. Despite slightly more biased estimates, the estimates obtained with the NBE are generally good. Furthermore, it is substantially faster to obtain an estimate through NBE than through maximum likelihood. In particular, on average, the MLE took 3 hours and 12 minutes to evaluate, while the NBE took 0.653 seconds; this means that the NBE is about

17,663 times faster, which is a substantial improvement in computational time.

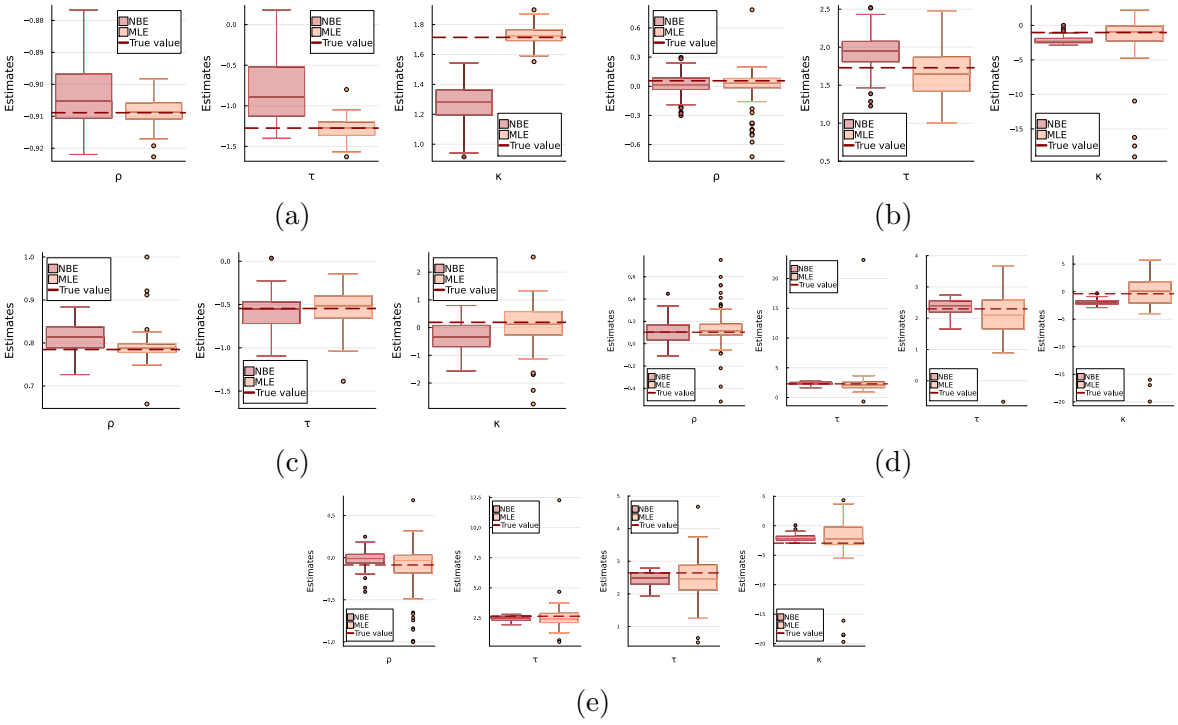



Figure C.2.2: Comparison between parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\rho}, \hat{\tau}_L, \hat{\kappa})'$ given by MLE (orange) and by NBE (red) for 100 samples with $n = 1000$. The true parameter values are given by the red line. (a) $\boldsymbol{\theta} = (0.91, -1.27, 1.71)'$, (b) $\boldsymbol{\theta} = (0.91, 1.73, -1.03)'$, (c) $\boldsymbol{\theta} = (0.91, -0.55, 0.19)'$, (d) $\boldsymbol{\theta} = (0.91, 2.30, -0.38)'$ and (e) $\boldsymbol{\theta} = (0.91, 2.64, -2.95)'$. For better visualisation, the larger outliers obtained through MLE were removed for $\hat{\tau}_L$ in (d) and (e).

### Model 2: $c_b$ is a Frank copula and $c_t$ is a Joe copula

For the second model, we consider $c_b$ to be a Frank copula (Frank, 1979) with parameter $\beta_F \in \mathbb{R}$, and $c_t$ to be a Joe copula (Joe, 1996) with $\alpha_J > 1$. As priors for the model parameters, we take $\beta_F \sim \text{Unif}(-15, 15)$, $\alpha_J \sim \text{Unif}(1, 15)$ and $\kappa \sim \text{Unif}(-3.51, 1.95)$. The performance of the NBE is assessed in Figure C.2.3 where the true values of the parameters are compared with their estimated values. It can be seen that all the parameters are estimated quite well with the NBE, with $\beta_F$ and $\alpha_J$ showing a bit of variability for lower and higher values, respectively. The coverage probabilities and average length of the 95% uncertainty intervals obtained via a non-parametric bootstrap

procedure for the parameter estimates and for $\chi(y)$ at levels $y \in \{0.50, 0.80, 0.95\}$ are shown in Table C.2.3. The lower coverage rates given on the left table reflect the bias shown by the parameter estimates. However, the results for $\chi(y)$ suggest that the NBE is able to capture the dependence structure of the data, especially for higher $y$, with the true $\chi(y)$ being within the confidence intervals in more than 59% of the time.

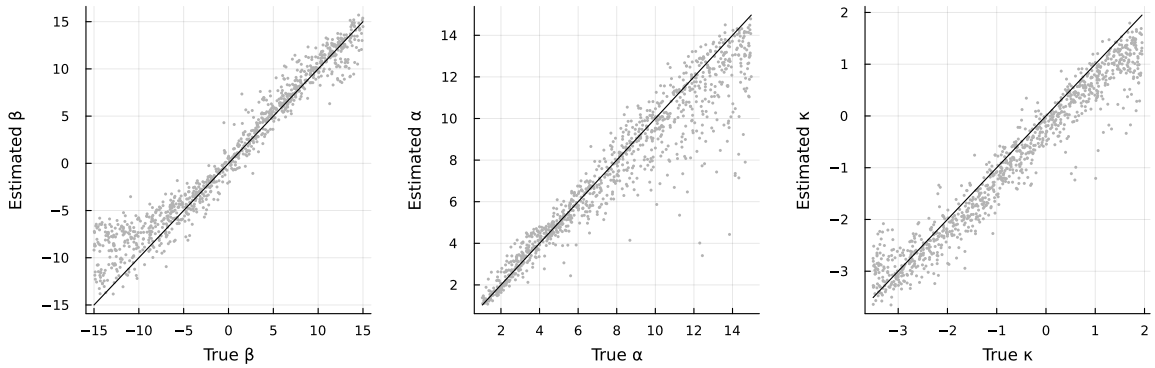



Figure C.2.3: Assessment of the NBE when $c_b$ is a Frank copula with parameter $\beta$, $c_t$ is a Joe copula with parameter $\alpha_J$, and with weighting function $\pi(x_1, x_2; \kappa) = (x_1 x_2)^{\exp\{\kappa\}}$, $x_1, x_2 \in (0, 1)$ for a sample size of $n = 1000$.

Table C.2.3: Coverage probability and average length of the 95% uncertainty intervals for the parameters (left) and for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$ (right) obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Coverage | Length |
|---|---|---|
| $\beta_F$ | 0.68 | 3.03 |
| $\alpha_J$ | 0.72 | 1.95 |
| $\kappa$ | 0.62 | 0.92 |

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.50)$ | 0.61 | 0.05 |
| $\chi(0.80)$ | 0.59 | 0.06 |
| $\chi(0.95)$ | 0.64 | 0.06 |

**Comparison with maximum likelihood estimation**

For this model the likelihood is feasible, though computational intensive. Therefore, we compare the estimations obtained by the NBE and by the MLE for five different parameter vectors $\boldsymbol{\theta} = (\beta_F, \alpha_J, \kappa)'$ generated with the pre-specified priors. Additionally, each data set with $n = 1000$ is simulated 100 times. The results are shown in Figure C.2.4. Similarly to the first model, the NBE estimates are generally more biased

than the MLEs, are less prone to have large outliers, and are generally good. While, on average, the MLE took 52 minutes to evaluate, the NBE took 0.203 seconds, which is about 15,339 times faster.

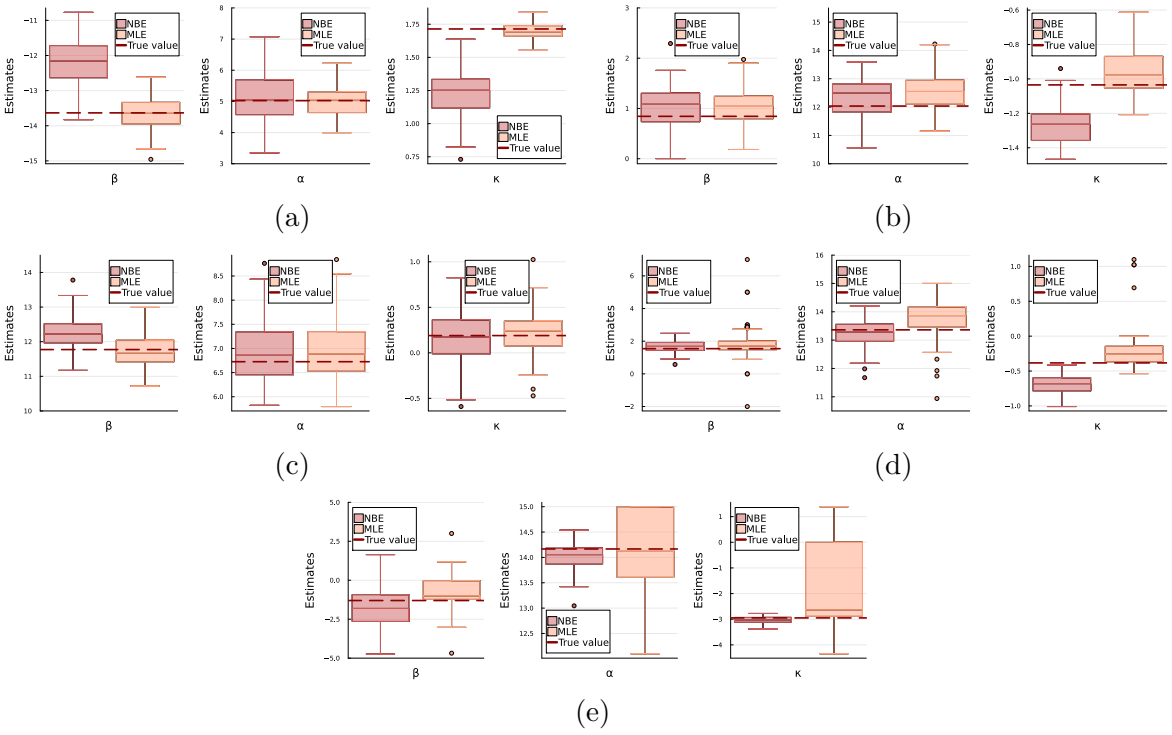



Figure C.2.4: Comparison between parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\beta}_F, \hat{\alpha}_J, \hat{\kappa})'$ given by MLE (orange) and by NBE (red) for 100 samples with $n = 1000$. The true parameter values are given by the red line. (a) $\boldsymbol{\theta} = (-13.63, 5.02, 1.71)'$, (b) $\boldsymbol{\theta} = (0.84, 12.04, -1.03)'$, (c) $\boldsymbol{\theta} = (11.77, 6.73, 0.19)'$, (d) $\boldsymbol{\theta} = (1.54, 13.36, -0.38)'$ and (e) $\boldsymbol{\theta} = (-1.30, 14.17, -2.95)'$.

### Model 3: $c_b$ is a Gaussian copula and $c_t$ is Model W

For the third model, we consider $c_t$ to be Model W, for which the priors for the parameters are those mentioned in Section 5.4.1 from the main paper. Figure C.2.5 displays the performance of the NBE. Despite the variability shown, especially by $\alpha$ and $\kappa$, the NBE provides good estimates overall. The coverage probabilities and average length of the 95% uncertainty intervals for the parameters and for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$ obtained via a non-parametric bootstrap procedure are given in Table C.2.4 on the left and right, respectively. The results for the parameter uncertainty are in agreement

with Figure C.2.5 with the coverage probability for $\alpha$ being the lowest and its average length the highest. However, as shown by the coverage probability for $\chi(y)$, this bias does not affect this dependence quantity. More specifically, the true value is within the confidence intervals in more than 85% of the time.

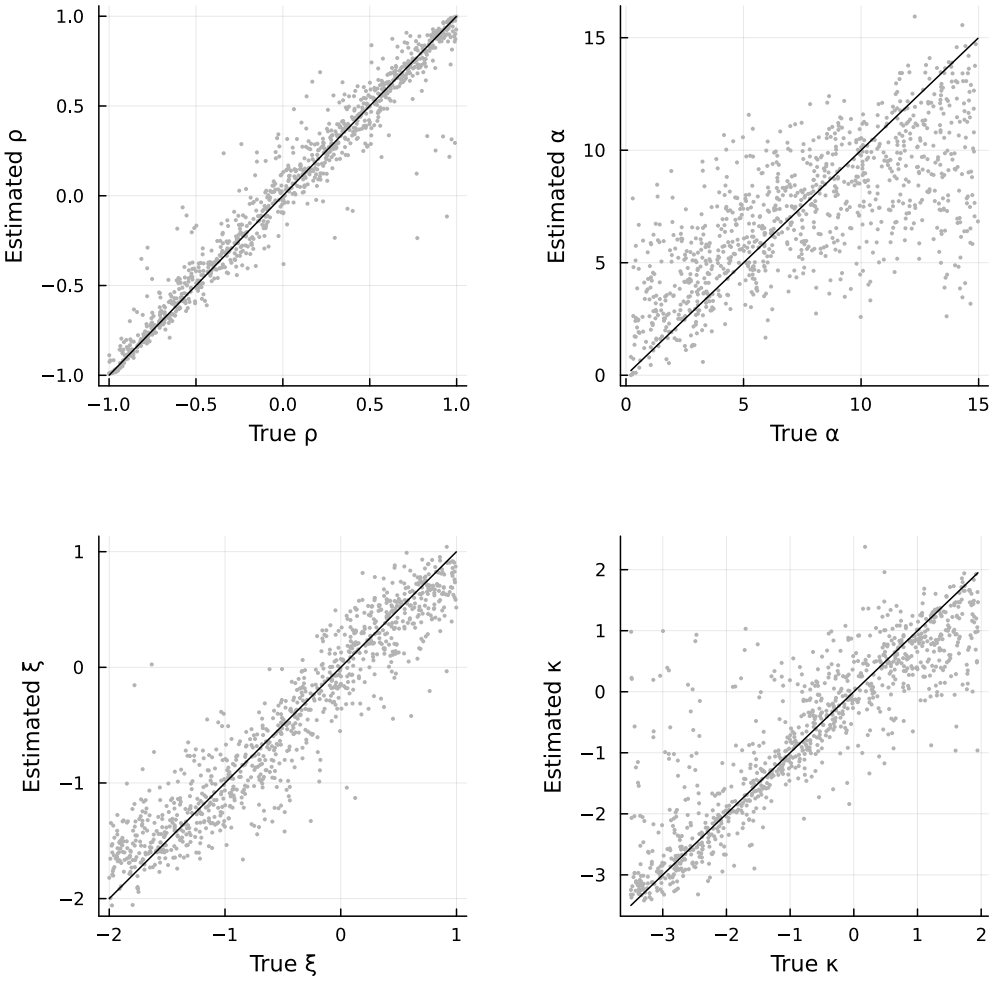



Figure C.2.5: Assessment of the NBE when $c_b$ is a Gaussian copula with correlation parameter $\rho$, $c_t$ is Model W with parameters $(\alpha, \xi)'$ and with weighting function $\pi(x_1, x_2; \kappa) = (x_1 x_2)^{\exp\{\kappa\}}$, $x_1, x_2 \in (0, 1)$ for a sample size of $n = 1000$.

### Model 4: $c_b$ is a Gaussian copula and $c_t$ is Model HW

For the forth model, we consider $c_t$ to be Model HW with the priors for the model parameters mentioned in Section 5.4.1 from the main paper. Figure C.2.6 displays the

Table C.2.4: Coverage probability and average length of the 95% uncertainty intervals for the parameters (left) and for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$ (right) obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Coverage | Length |
|---|---|---|
| $\rho$ | 0.85 | 0.24 |
| $\alpha$ | 0.60 | 4.28 |
| $\xi$ | 0.71 | 0.56 |
| $\kappa$ | 0.73 | 1.16 |

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.50)$ | 0.91 | 0.06 |
| $\chi(0.80)$ | 0.89 | 0.09 |
| $\chi(0.95)$ | 0.85 | 0.11 |

performance of the NBE, showing that $\delta$ and $\omega$ seem to be over-estimated by the NBE for lower values. Table C.2.5 shows the coverage probabilities and average length of the 95% uncertainty intervals obtained via a non-parametric bootstrap procedure for the parameters on the left, and for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$ on the right. The results for the parameter estimates mirror the variability shown in Figure C.2.6, where the coverage probability for $\omega$ and $\delta$ are the lowest. The coverage probabilities of 95% uncertainty intervals for $\chi(y)$ show that the true value is within the confidence intervals in more than 86% of the time, indicating that despite the bias shown by the estimation, this dependence measure is well calibrated.

Table C.2.5: Coverage probability and average length of the 95% uncertainty intervals for the parameters (left) and for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$ (right) obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Coverage | Length |
|---|---|---|
| $\rho$ | 0.80 | 0.28 |
| $\delta$ | 0.58 | 0.15 |
| $\omega$ | 0.44 | 0.47 |
| $\kappa$ | 0.71 | 1.33 |

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.50)$ | 0.91 | 0.07 |
| $\chi(0.80)$ | 0.88 | 0.10 |
| $\chi(0.95)$ | 0.86 | 0.12 |

### Model 5: $c_b$ is a Gaussian copula and $c_t$ is Model E2

For the final model, we take $c_t$ to be Model E2 with the priors for the model parameters mentioned in Section 5.4.1 from the main paper. Figure C.2.7 shows the performance of the NBE. Similarly to Model 3, there is some variability in the NBEs, especially for

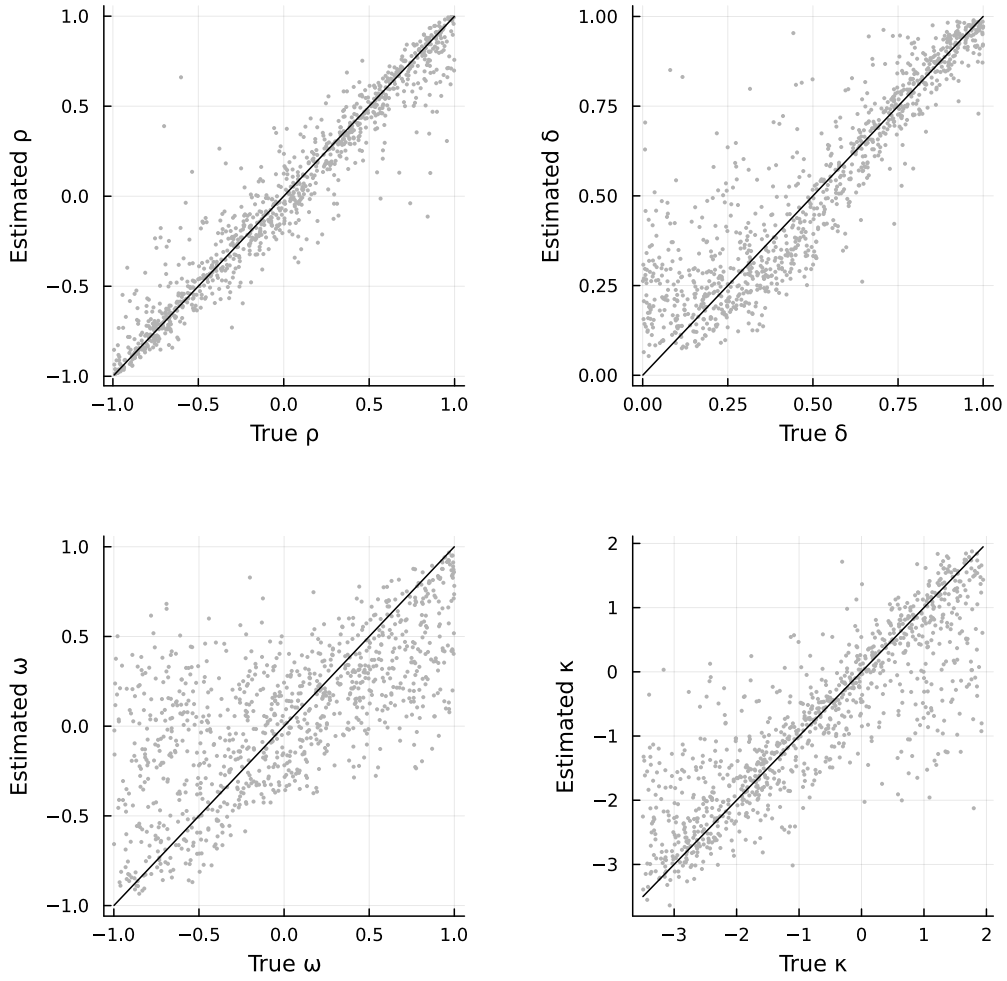



Figure C.2.6: Assessment of the NBE when $c_b$ is a Gaussian copula with correlation parameter $\rho$, $c_t$ is Model HW with parameters $(\delta, \omega)'$, and with weighting function $\pi(x_1, x_2; \kappa) = (x_1 x_2)^{\exp\{\kappa\}}$, $x_1, x_2 \in (0, 1)$ for a sample size of $n = 1000$.

$\alpha$. This parameter is also the one with lowest coverage probability and wider intervals for the parameters estimation, as shown in left of Table C.2.6. Similarly to the previous models, the coverage probability for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$, shown in the right of Table C.2.6, indicate that this measure is well captured by the NBE, with the true value lying within the confidence intervals in at least 83% of the time.

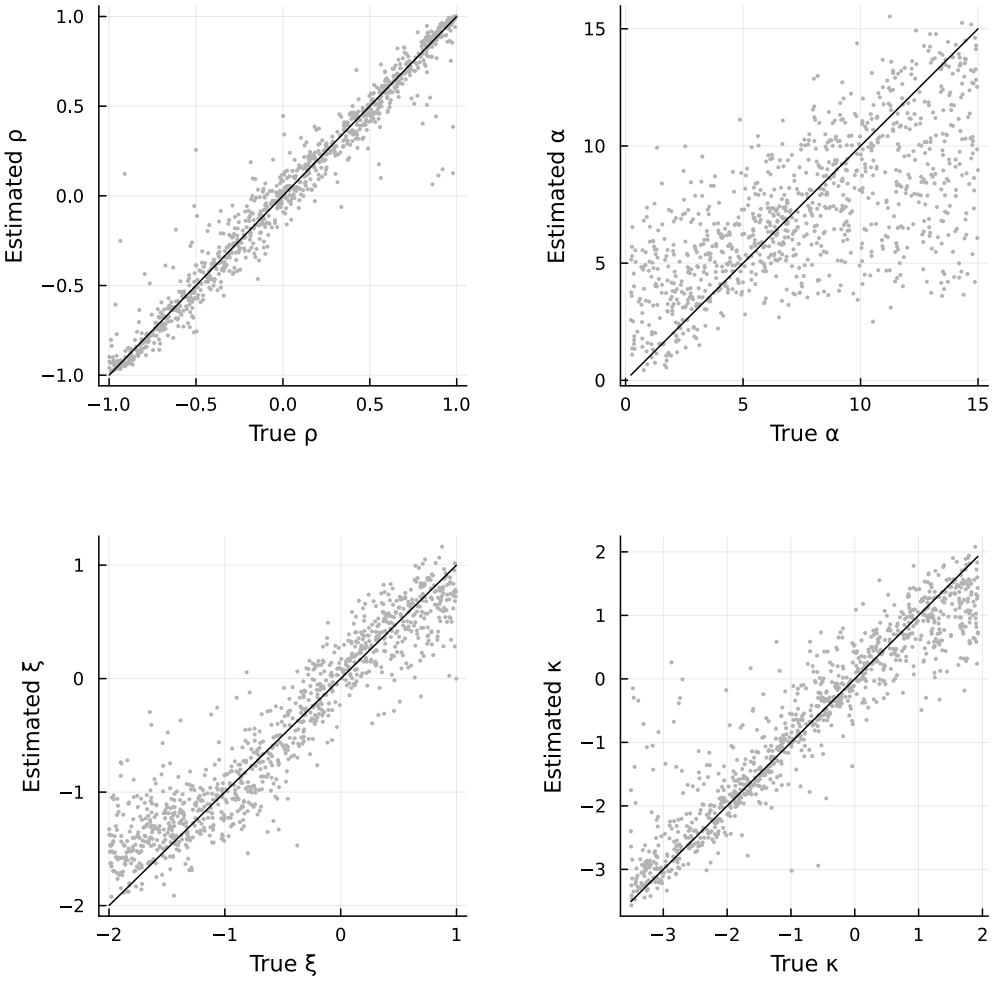



Figure C.2.7: Assessment of the NBE when $c_b$ is a Gaussian copula with correlation parameter $\rho$, $c_t$ is Model E2 with parameters $(\alpha, \xi)'$ and with weighting function $\pi(x_1, x_2; \kappa) = (x_1 x_2)^{\exp\{\kappa\}}$, $x_1, x_2 \in (0, 1)$ for a sample size of $n = 1000$.

Table C.2.6: Coverage probability and average length of the 95% uncertainty intervals for the parameters (left) and for $\chi(y)$ at levels $y = \{0.50, 0.80, 0.95\}$ (right) obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Coverage | Length |
|---|---|---|
| $\rho$ | 0.83 | 0.22 |
| $\alpha$ | 0.57 | 4.48 |
| $\xi$ | 0.72 | 0.59 |
| $\kappa$ | 0.75 | 0.96 |

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.50)$ | 0.88 | 0.05 |
| $\chi(0.80)$ | 0.84 | 0.07 |
| $\chi(0.95)$ | 0.83 | 0.10 |

## C.2.2 Model W

### Variable sample size and censoring level

### Comparison with censored maximum likelihood estimation

The comparison between the NBE and CMLE estimators for the remaining three parameter vectors considered in the simulation study of the main paper is given in Figure C.2.8.

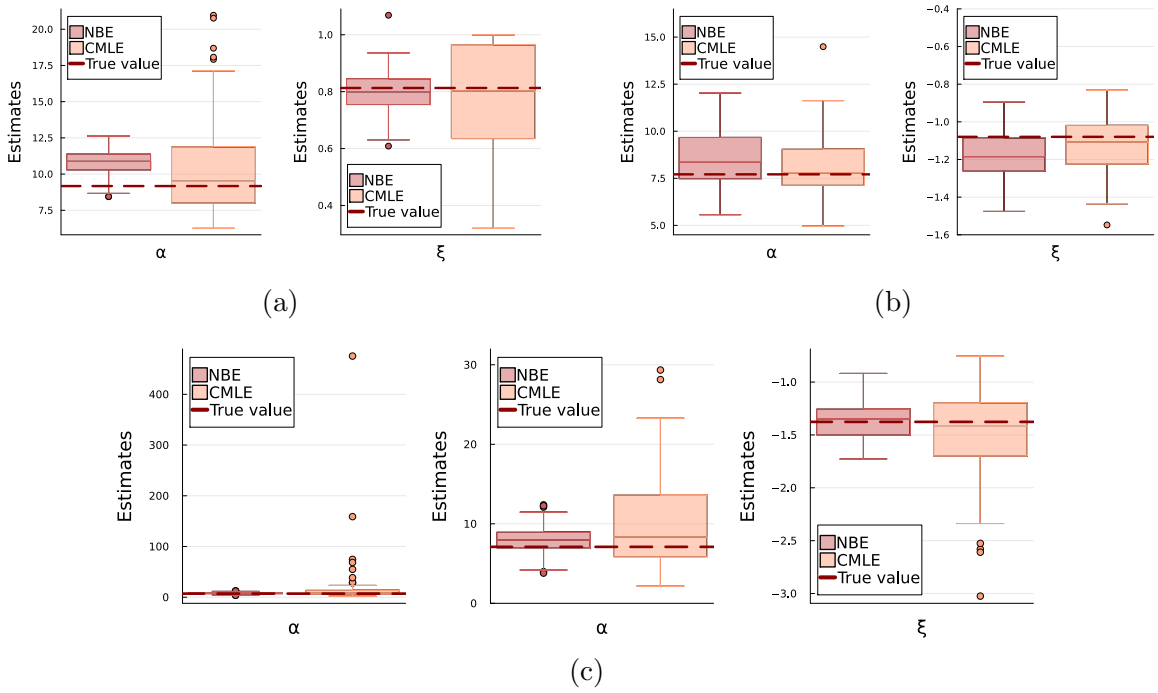



Figure C.2.8: Comparison between parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\xi})'$ given by CMLE (orange) and by NBE (red) for 100 samples with $n = 1000$. The true parameters are given by the red line. (a) $\boldsymbol{\theta}_3 = (9.17, 0.81)'$ with $\tau_3 = 0.80$, (b) $\boldsymbol{\theta}_4 = (7.71, -1.08)'$ with $\tau_4 = 0.73$ and (c) $\boldsymbol{\theta}_5 = (7.10, -1.38)'$ with $\tau_5 = 0.98$. For better visualisation, the larger outliers obtained through MLE are removed for $\hat{\alpha}$ in (e).

For comparison with the simulation study of Model W given in Section 5.4.2, we now present the results for when both sample size $n$ and censoring level $\tau$ are kept fixed, and for when the sample size is assumed variable but the censoring level is kept fixed.

### Fixed sample size and censoring level

We first assume that the sample size and the censoring level are kept fixed at $n = 1000$ and $\tau = 0.8$, respectively, and the performance of the NBE is shown in Figure C.2.9. Similarly to the case presented in Section 5.4.2 of the main paper, the NBE exhibits some bias for larger values of $\alpha$. This is also noticeable with the average length of the 95% uncertainty intervals obtained via a non-parametric bootstrap procedure given in Table C.2.7. It can also be seen that the coverage probabilities are slightly higher than the ones from Section 5.4.2. This might be due to the fact that there are less unknown variables in this configuration. Finally, the coverage probabilities of 95% uncertainty intervals, and their average length, for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$ are shown on the right of Table C.2.7. The results are similar to the ones presented in the main paper, with a slightly higher coverage for larger $y$.

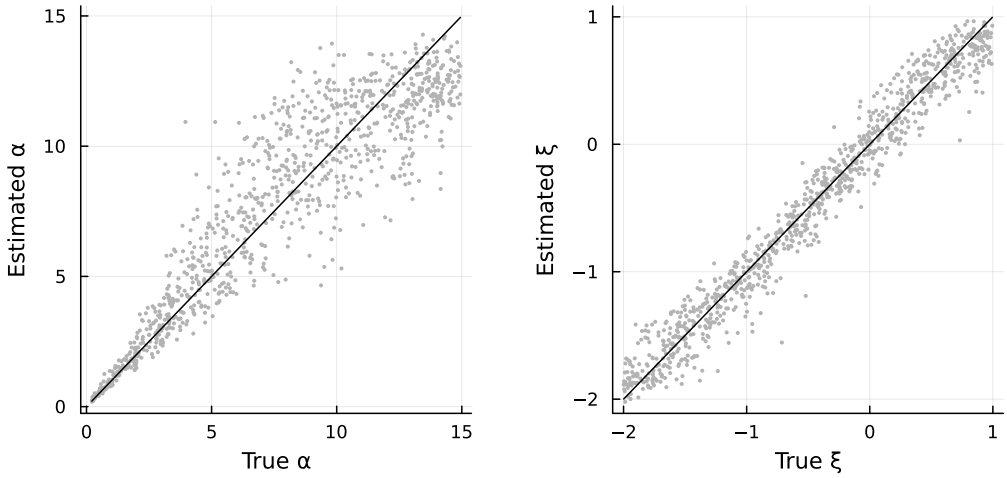



Figure C.2.9: Assessment of the NBE for Model W with parameters $\boldsymbol{\theta} = (\alpha, \xi)'$ for a sample size of $n = 1000$ and fixed censoring level $\tau = 0.8$.

### Comparison with censored maximum likelihood estimation

We compare the estimations obtained by the NBE and by the MLE for the five parameter vectors $\boldsymbol{\theta} = (\alpha, \xi)'$ considered in Section 5.4.2 with now fixed $\tau = 0.8$. Likewise

Table C.2.7: Coverage probability and average length of the 95% uncertainty intervals for the parameters (left) and for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$ (right) obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Coverage | Length |
|---|---|---|
| $\alpha$ | 0.79 | 4.06 |
| $\xi$ | 0.88 | 0.51 |

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.80)$ | 0.91 | 0.07 |
| $\chi(0.95)$ | 0.92 | 0.09 |
| $\chi(0.99)$ | 0.92 | 0.10 |

before, each data set is simulated 100 times and has a sample size of $n = 1000$. The results are shown in Figure C.2.10; these are fairly similar to those obtained when $n$ and $\tau$ are assumed unknown, and given in the main paper. This is also the configuration for which censored MLE is faster; for instance, on average, the CMLE took 77.550 seconds, while the NBE was 356 times faster with an average time of 0.218 seconds.

### Variable sample size and fixed censoring level

We now assume the sample size is unknown but we keep the censoring level fixed at $\tau = 0.8$. The performance of the NBE is given in Figure C.2.11, where a similar behaviour to the results obtained either when $n$ is assumed fixed or when $\tau$ is also assumed unknown. The coverage probability of the 95% uncertainty intervals obtained via (non-parametric) bootstrap shown in Table C.2.8 are now slightly lower than the ones from the case when $n$ is assumed fixed at 1000. However, these are still slightly higher that those from Section 5.4.2. The coverage probabilities of 95% uncertainty intervals, and their average length, for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$ are shown on the right of Table C.2.8, and are similar in magnitude to the corresponding results presented in the main paper.

### Comparison with censored maximum likelihood estimation

The same five parameter vectors $\boldsymbol{\theta} = (\alpha, \xi)'$ considered for the cases where $n$ is fixed at 1000 and the one presented in Section 5.4.2 are used to compare the NBE and CMLE

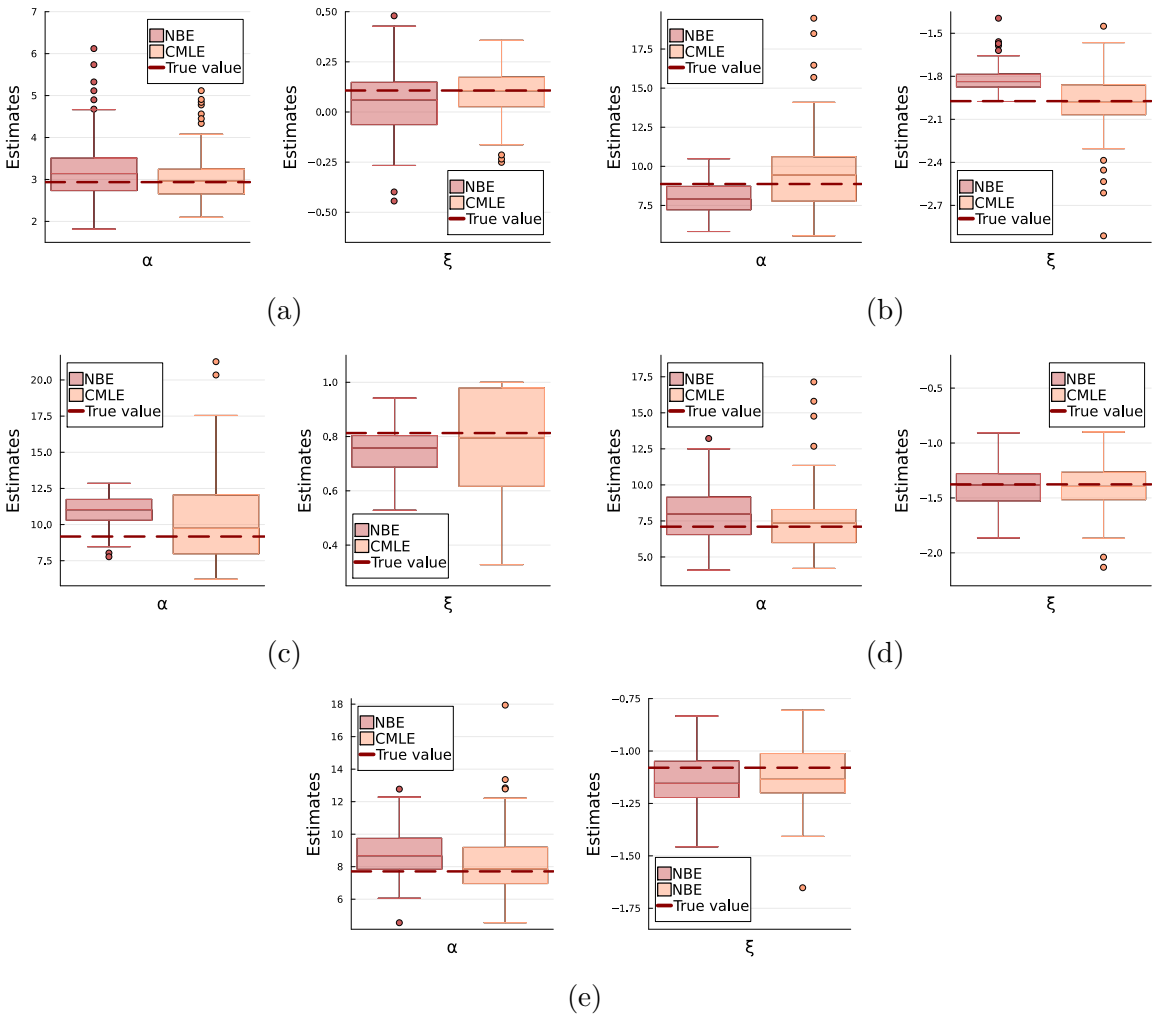



Figure C.2.10: Comparison between parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\xi})'$ given by CMLE (orange) and by NBE (red) for 100 samples with $n = 1000$. The true parameters are given by the red line. (a) $\boldsymbol{\theta} = (2.94, 0.11)'$, (b) $\boldsymbol{\theta} = (8.87, -1.97)'$, (c) $\boldsymbol{\theta} = (9.17, 0.81)'$, (d) $\boldsymbol{\theta} = (7.10, -1.38)'$ and (e) $\boldsymbol{\theta} = (7.71, -1.08)'$.

Table C.2.8: Coverage probability and average length of the 95% uncertainty intervals for the parameters (left) and for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$ (right) obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Coverage | Length |
|---|---|---|
| $\alpha$ | 0.72 | 3.53 |
| $\xi$ | 0.80 | 0.46 |

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.80)$ | 0.90 | 0.06 |
| $\chi(0.95)$ | 0.88 | 0.08 |
| $\chi(0.99)$ | 0.85 | 0.09 |

estimates. Similarly to the previous case, we fix $\tau = 0.8$, and each data set has a sample size of $n = 1000$ and is simulated 100 times. No evident key differences to the estimates

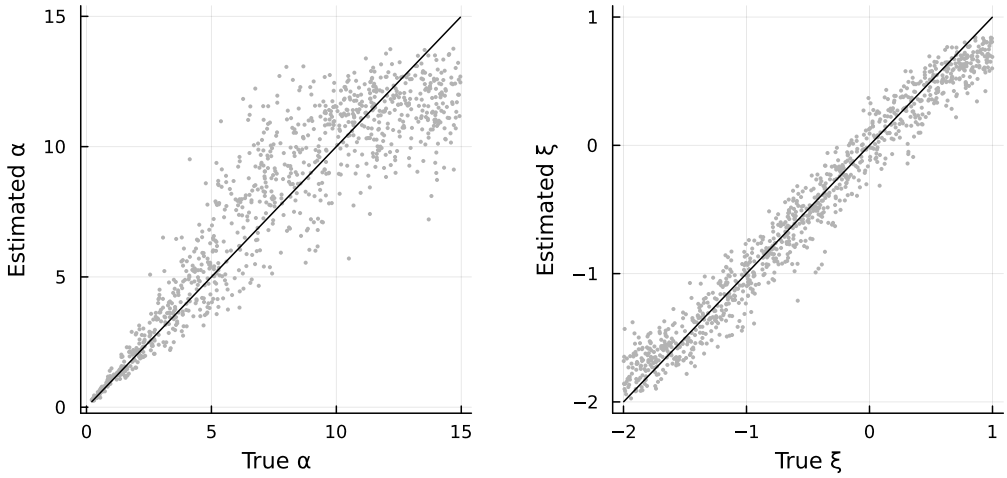



Figure C.2.11: Assessment of the NBE for Model W with parameters $\boldsymbol{\theta} = (\alpha, \xi)'$ for a sample size of $n = 1000$ and fixed censoring level $\tau = 0.8$.

obtained when $n$ is assumed fixed and when $n$ and $\tau$ are assumed unknown are shown by the results in Figure C.2.12. For this case, the average time to get a NBE is of 0.470 seconds, which is about 165 faster than CMLE on average.

### General conclusions

The results with fixed censoring level ($\tau = 0.8$) with fixed ($n = 1000$) and variable sample size exhibit similar findings. In the case where both $\tau$ and $n$ are fixed, the obtained bootstrap-based intervals have better coverage. Although not as evident, this is also the case when we assume fixed $\tau = 0.8$ with a variable sample size. When comparing the estimates given by the NBEs with the ones obtained by classical inference techniques, fixing one or both $n$ and $\tau$ did not improve the performance of the estimators.

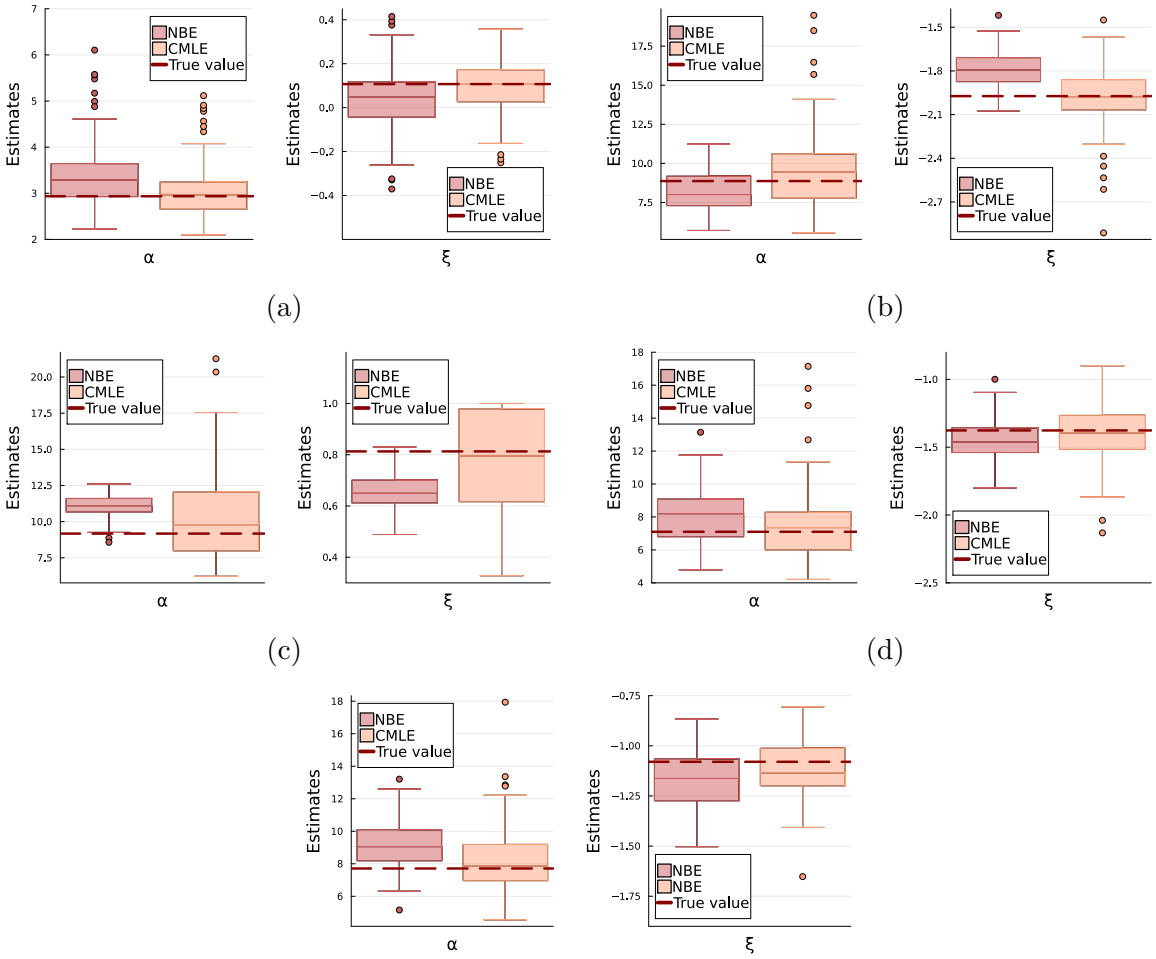



Figure C.2.12: Comparison between parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\xi})'$ given by CMLE (orange) and by NBE (red) for 100 samples with $n = 1000$. The true parameters are given by the red line. (a) $\boldsymbol{\theta} = (2.94, 0.11)'$, (b) $\boldsymbol{\theta} = (8.87, -1.97)'$, (c) $\boldsymbol{\theta} = (9.17, 0.81)'$, (d) $\boldsymbol{\theta} = (7.10, -1.38)'$ and (e) $\boldsymbol{\theta} = (7.71, -1.08)'$.

### C.2.3 Model HW

We assess the performance of the NBE for Model HW. The results are shown in Figure C.2.13 and Table C.2.9. Similarly to Model W, there is some variability in the estimates, in particular for lower values of $\delta$ and $\omega$. The coverage probabilities of 95% uncertainty intervals, and their average length, for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$, shown on the right of Table C.2.8, indicate that even with biased results, the NBE is able to characterise the extremal dependence at high levels of $y$. We note that, similarly to the study involving Model W given in the main paper, the coverage probabilities for $\chi(y)$ are achieved with new data sets for 1000 parameter configurations, each generated with a fixed censoring level $\tau = 0.8$.

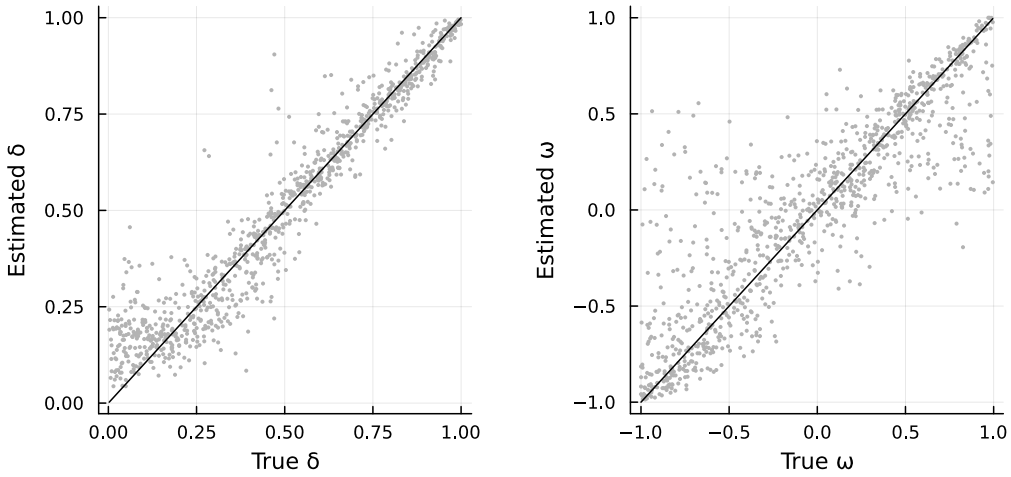



Figure C.2.13: Assessment of the NBE for Model HW, where $\boldsymbol{V}$ follows a bivariate Gaussian copula, with parameters $\boldsymbol{\theta} = (\delta, \omega)'$ for a sample size of $n = 1000$.

Table C.2.9: Coverage probability and average length of the 95% uncertainty intervals for the parameters (left) and for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$ (right) obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Coverage | Length |
|---|---|---|
| $\delta$ | 0.64 | 0.12 |
| $\omega$ | 0.69 | 0.41 |

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.80)$ | 0.90 | 0.08 |
| $\chi(0.95)$ | 0.90 | 0.09 |
| $\chi(0.99)$ | 0.90 | 0.10 |

### Comparison with censored maximum likelihood estimation

Similarly to the previous cases, we generate five parameter vectors from the priors considered with the corresponding data sets with $n = 1000$, and we simulate each data set 100 times. The comparison between the NBE and CMLE is shown in Figure C.2.14; the estimates given by the NBE are quite good, particularly for lower censoring levels. As for computational times, on average the CMLE took 198.489 seconds, while the NBE was 732 times faster with an average of 0.271 seconds.

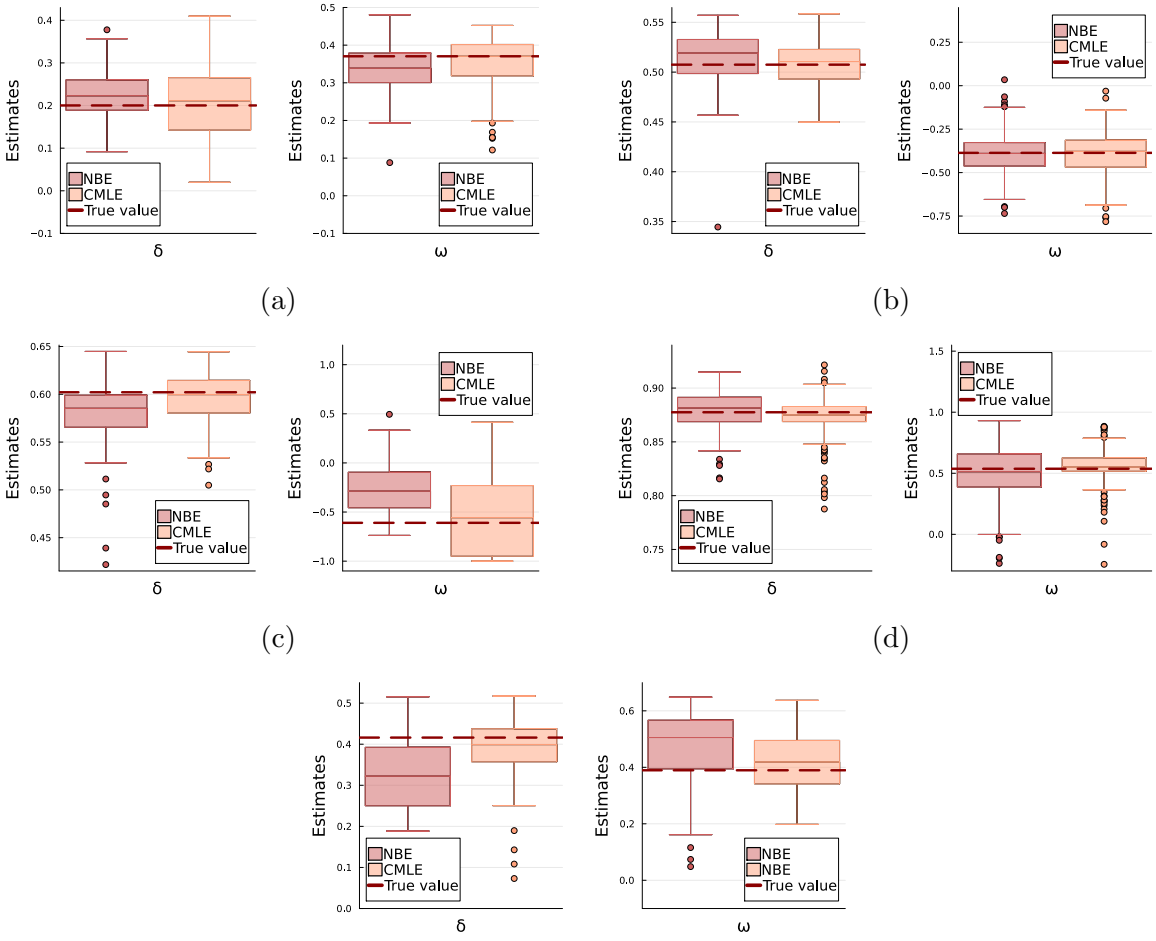



Figure C.2.14: Comparison between parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\delta}, \hat{\omega})'$ given by CMLE (orange) and by NBE (red) for 100 samples with $n = 1000$. The true parameters are given by the red line. (a) $\boldsymbol{\theta} = (0.20, 0.37)'$ with $\tau = 0.65$, (b) $\boldsymbol{\theta} = (0.51, -0.39)'$ with $\tau = 0.76$, (c) $\boldsymbol{\theta} = (0.60, -0.61)'$ with $\tau = 0.95$, (d) $\boldsymbol{\theta} = (0.88, 0.54)'$ with $\tau = 0.57$ and (e) $\boldsymbol{\theta} = (0.42, 0.39)'$ with $\tau = 0.91$.

### C.2.4 Model E1

Figure C.2.15 and Table C.2.10 show the performance of the NBE for Model E1. As can be seen, parameters $\alpha$ and $\beta$ have the lowest coverage probability and higher average length of their 95% uncertainty intervals; this is in agreement with the variability shown when comparing the true values with their estimated values in Figure C.2.15. As before, we compute the coverage probabilities of the 95% confidence intervals for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$ by considering new data sets for 1000 parameter configurations, each generated with a fixed censoring level $\tau = 0.8$. The results, given on the right of Table C.2.10, indicate that the bias shown by the NBE does not seem to influence the estimation of $\chi(y)$. In particular, the true value is within the confidence intervals in more than 81% of the time.

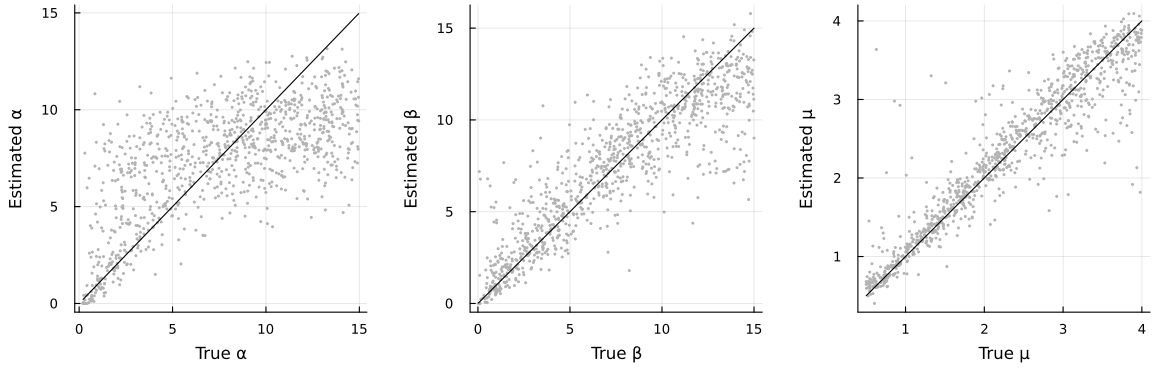



Figure C.2.15: Assessment of the NBE for Model E1 with parameters $\boldsymbol{\theta} = (\alpha, \beta, \mu)'$ for a sample size of $n = 1000$.

Table C.2.10: Coverage probability and average length of the 95% uncertainty intervals for the parameters (left) and for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$ (right) obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Coverage | Length |
|---|---|---|
| $\alpha$ | 0.38 | 3.04 |
| $\beta$ | 0.61 | 2.84 |
| $\mu$ | 0.70 | 0.51 |

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.80)$ | 0.82 | 0.11 |
| $\chi(0.95)$ | 0.82 | 0.11 |
| $\chi(0.99)$ | 0.81 | 0.12 |

### Comparison with censored maximum likelihood estimation

We compare the estimations obtained by the NBE and by the CMLE for five parameter vectors $\boldsymbol{\theta} = (\alpha, \beta, \mu)'$ generated from the priors considered. Each corresponding data set has $n = 1000$ and is simulated 100 times. Similarly to the other models considered, the NBE is more biased than the CMLE and, in some cases, can be more variable than the CMLE. Despite that, on average, the CMLE took 28 minutes, whereas the NBE took 0.159 seconds, meaning that the NBE is about $10,420$ times faster.

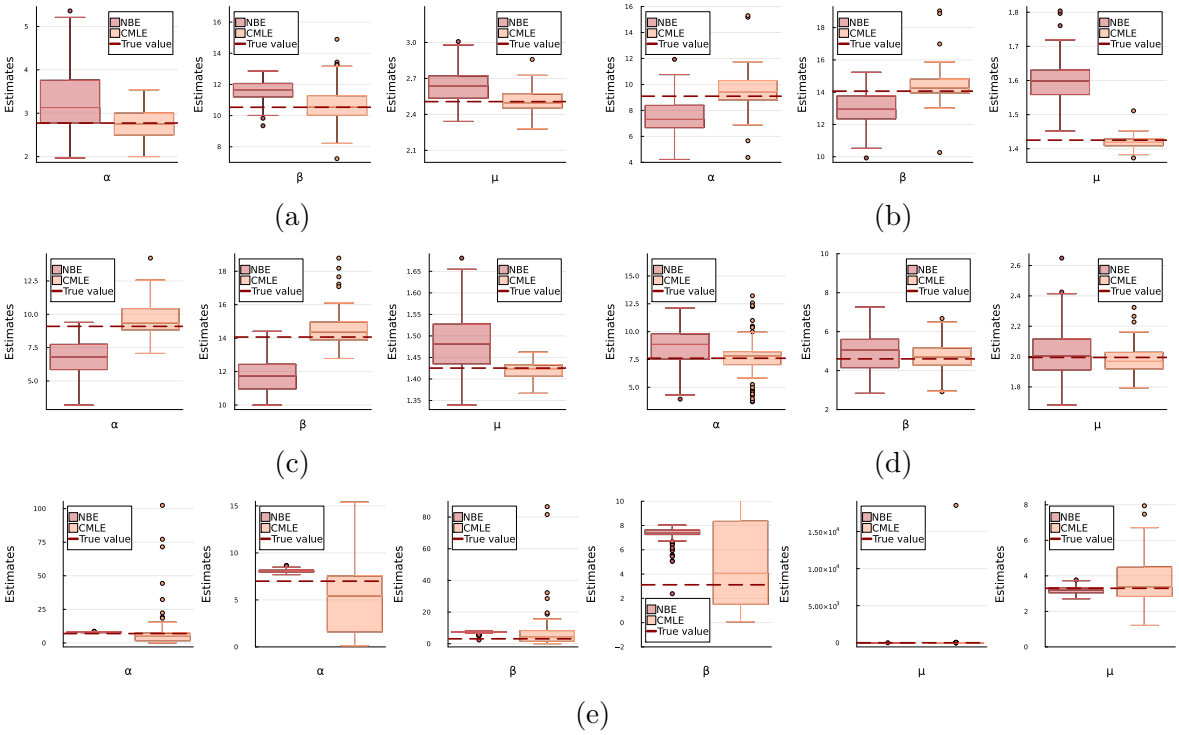



Figure C.2.16: Comparison between parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}, \hat{\mu})'$ given by CMLE (orange) and by NBE (red) for 100 samples with $n = 1000$. The true parameter values are given by the red line. (a) $\boldsymbol{\theta} = (2.77, 10.54, 2.51)'$ with $\tau = 0.79$, (b) $\boldsymbol{\theta} = (9.09, 14.06, 1.43)'$ with $\tau = 0.60$, (c) $\boldsymbol{\theta} = (9.09, 14.06, 1.43)'$ with $\tau = 0.80$, (d) $\boldsymbol{\theta} = (7.61, 4.60, 1.99)'$ with $\tau = 0.73$ and (e) $\boldsymbol{\theta} = (6.99, 3.12, 3.30)'$ with $\tau = 0.98$. For better visualisation, the larger values obtained through MLE were removed for $\boldsymbol{\theta}$ in (e).

### C.2.5 Model E2

For the final model, we consider Model E2, for which the performance of the NBE is shown in Figure C.2.17 and Table C.2.11. Parameter $\alpha$ shows the higher variability, especially for larger values, with its 95% uncertainty intervals being wider and having lower coverage probabilities. The coverage probabilities of 95% uncertainty intervals of $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$, shown on the right of Table C.2.11, indicate that the this measure is well calibrated, with the true $\chi(y)$ lying within the intervals at least 87% of the time in spite of the bias shown by the NBE. As before, the results for $\chi(y)$ are obtained with ne data sets for 1000 parameter configurations with a fixed censoring level $\tau = 0.8$.

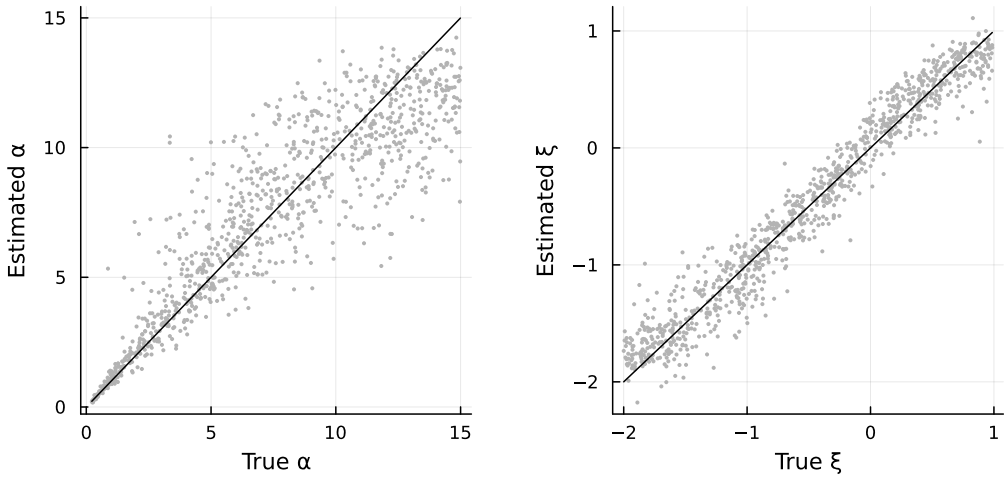



Figure C.2.17: Assessment of the NBE for Model E1 with parameters $\boldsymbol{\theta} = (\alpha, \xi)'$ for a sample size of $n = 1000$.

Table C.2.11: Coverage probability and average length of the 95% uncertainty intervals for the parameters (left) and for $\chi(y)$ at levels $y = \{0.80, 0.95, 0.99\}$ (right) obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

| Parameter | Coverage | Length |
|---|---|---|
| $\alpha$ | 0.65 | 3.10 |
| $\xi$ | 0.75 | 0.44 |

| $\chi(y)$ | Coverage | Length |
|---|---|---|
| $\chi(0.80)$ | 0.94 | 0.06 |
| $\chi(0.95)$ | 0.90 | 0.10 |
| $\chi(0.99)$ | 0.87 | 0.12 |

**Comparison with censored maximum likelihood estimation**

The estimation obtained by the NBE and the CMLE is assessed for five parameter vectors $\boldsymbol{\theta} = (\alpha, \xi)'$ and their corresponding data sets with $n = 1000$, which are simulated 100 times each. The results, shown in Figure C.2.18, indicate that the NBE is more biased than the CMLE. However, as with the previous models, the NBE is about $2,314$ times faster than CMLE; in particular, on average, the CMLE took 21 minutes to compute, whilst the NBE took 0.557 seconds.

## C.3 Model selection

The neural network architecture used for model selection (recall Section 5.4.3 of the main paper) is given in Table C.3.1.

Table C.3.1: Summary of the neural network architecture used for the model selection classifier. The input array to the first layer represents the dimension $d$ of data set $\boldsymbol{Z}$ and the one-hot encoded vector $\boldsymbol{I}$; see Section 5.2.3. The output array of the last layer of neural network $\psi$ differ based on the number of models $M$ : for $M = 2$, we have $w_\psi = 128$, while for $M = 4$, $w_\psi = 256$. The output array of the last layer of neural network $\phi$ represents the output class probabilities $\hat{\boldsymbol{p}}$.

| Neural network | Layer type | Input dimension | Output dimension |
|---|---|---|---|
| $\boldsymbol{\psi}(\cdot)$ | Dense | $[2, 2]$ | $[128]$ |
| | Dense | $[128]$ | $[128]$ |
| | Dense | $[128]$ | $[w_\psi]$ |
| $\boldsymbol{\phi}(\cdot)$ | Dense | $[d_\psi + 1]$ | $[128]$ |
| | Dense | $[128]$ | $[M]$ |

## C.4 Misspecified scenarios

We present now two examples, one for each study performed in Section 5.4.4 from the main paper. For each case, the best model is selected through the trained neural classifier, and the vector of parameters is estimated using the NBE trained for inference on the selected model. In addition, a comparison with classical model selection tools and

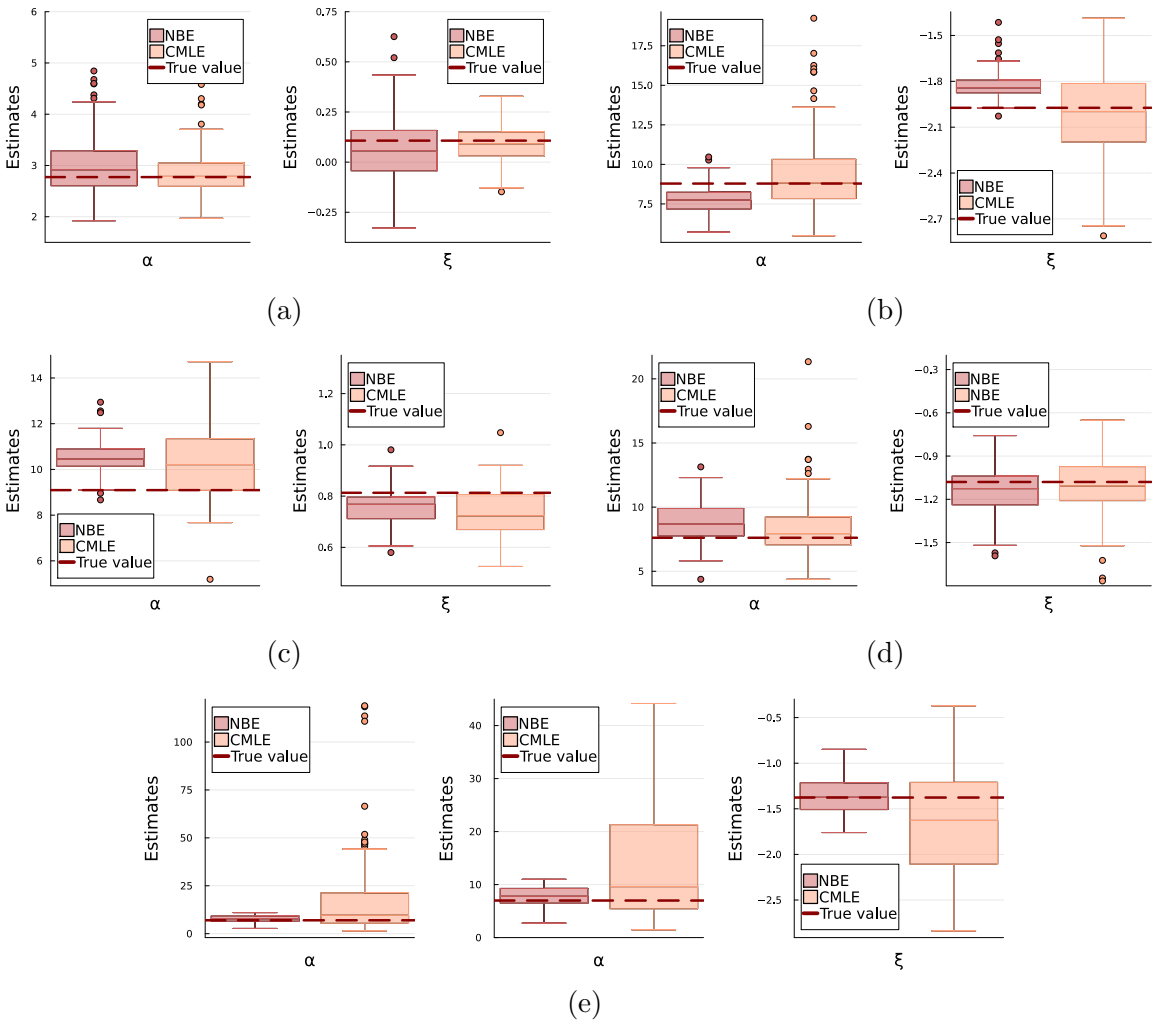



Figure C.2.18: Comparison between parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\xi})'$ given by CMLE (orange) and by NBE (red) for 100 samples with $n = 1000$. The true parameter values are given by the red line. (a) $\boldsymbol{\theta} = (2.77, 0.11)'$ with $\tau = 0.79$, (b) $\boldsymbol{\theta} = (8.79, -1.97)'$ with $\tau = 0.60$, (c) $\boldsymbol{\theta} = (9.09, 0.81)'$ with $\tau = 0.80$, (d) $\boldsymbol{\theta} = (7.61, -1.08)'$ with $\tau = 0.73$ and (e) $\boldsymbol{\theta} = (6.99, -1.38)'$ with $\tau = 0.98$. For better visualisation, the larger outliers obtained through MLE were removed for $\hat{\alpha}$ in (e).

inference is given. As a further diagnostic, we compare $\chi(y)$ for $y \in (0, 1)$ obtained with the NBE for the estimated model with their empirical counterparts, the true values and with the ones obtained by the model selected and estimated through BIC and CMLE.

Results for model selection through the neural classifier and BIC for the Gaussian data case can be seen on the left of Table C.4.1, and the estimates for the vector of parameters obtained by the NBE and CMLE are on the right. The neural classifier

Table C.4.1: Model selection procedure obtained through the probabilities given by the neural classifier and through BIC (left), and parameter estimates given by the NBE and by the CMLE (right) for the selected model (in bold). All the values are rounded up to 3 decimal places.

| Model | $\hat{\boldsymbol{p}}$ | BIC |
|---|---|---|
| Model W | $4.609 \times 10^{-5}$ | **569.257** |
| Model HW | **0.987** | 570.297 |
| Model E1 | $2.392 \times 10^{-8}$ | 576.295 |
| Model E2 | 0.013 | 576.452 |

| Method | Model parameters |
|---|---|
| NBE (Model HW) | $(\hat{\delta}, \hat{\omega}) = (0.201, 0.400)$ |
| CMLE (Model W) | $(\hat{\alpha}, \hat{\xi}) = (1.155, -0.092)$ |

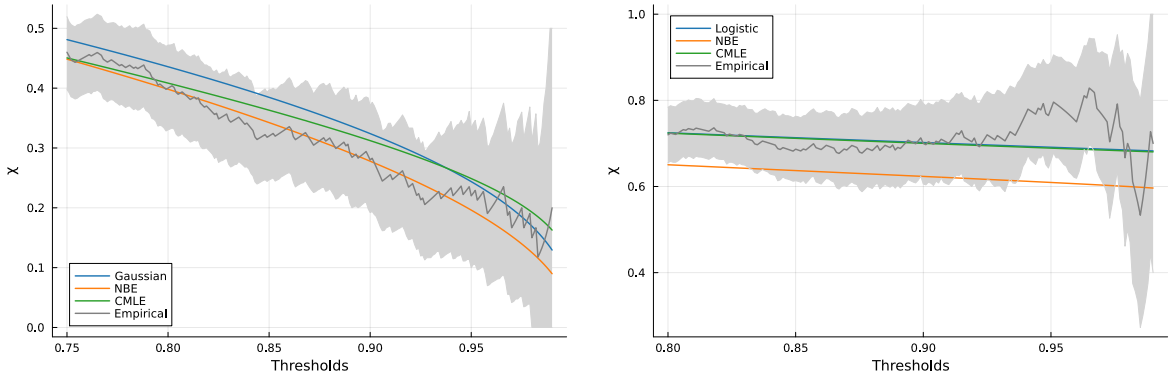
selected Model HW as the most suitable one for the data set, whilst according to the BIC, Model W is the best one. In spite of this difference, both models indicate the presence of asymptotically independent data since $\hat{\delta} \leq 0.5$ and $\hat{\xi} < 0$. This is in agreement with the underlying Gaussian data being AI. The comparison between $\chi(y)$ obtained by the models estimated through the NBE and the CMLE, with the true values of $\chi(y)$ based on the Gaussian copula, and their empirical estimates, for $y \in [0.75, 0.99]$ are shown in the left panel of Figure C.4.1. The model estimates obtained through CMLE inference are overall closer to the truth than the ones given by the NBE; however, the NBE estimates are closer to the empirical estimates. Overall, the extremal dependence behaviour of the data is well captured by the trained NBE.

Table C.4.1 gives the results for model selection and parameter estimation for the logistic data case. For the model selection, the neural classifier selected Model HW, whereas BIC preferred Model W, though the difference with Model E2 is very small. Despite this difference, looking at the parameters for each model that indicate the extremal dependence structure, we have $\hat{\delta} > 0.5$ and $\hat{\xi} > 0$, respectively. Thus, both parameters indicate correctly the presence of asymptotically dependent data. The comparison between $\chi(y)$ obtained by the models estimated through the NBE and the CMLE, with the true values $\chi(y)$ for the logistic data, and their empirical estimates, for $y \in [0.8, 0.99]$ is shown in right panel of Figure C.4.1. For this case, the estimated model $\chi(y)$ given by the CMLE overlap with the true values for the logistic data. On

Table C.4.2: Model selection procedure obtained through the probabilities given by the neural classifier and through BIC (left), and parameter estimates given by the NBE and by the CMLE (right) for the selected model (in bold). All the values are rounded to 3 decimal places.

| Model | $\hat{\boldsymbol{p}}$ | BIC |
|---|---|---|
| Model W | $7.774 \times 10^{-5}$ | $\mathbf{-53.077}$ |
| Model HW | $\mathbf{0.999}$ | $-46.030$ |
| Model E1 | $2.104 \times 10^{-7}$ | $-38.450$ |
| Model E2 | 0.001 | $-52.987$ |

| Method | Model parameters |
|---|---|
| NBE (Model HW) | $(\hat{\delta}, \hat{\omega}) = (0.640, -0.147)$ |
| CMLE (Model W) | $(\hat{\alpha}, \hat{\xi}) = (2.173, 1.000)$ |




(a) (b)

Figure C.4.1: Model-based $\chi(y)$ given by the NBE (in orange) and by the CMLE (in green), and empirical $\chi(y)$ (in grey) for $y \in [\tau, 0.99]$. The 95% confidence bands were obtained by boostrapping. (a) $\chi(y)$ for a Gaussian copula with correlation parameter $\rho = 0.5$ (in blue) and censoring level $\tau = 0.75$, and (b) $\chi(y)$ for a logistic distribution with dependence parameter $\alpha_L = 0.4$ (in blue) and censoring level $\tau = 0.8$. Note that $\chi(y)$ for the logistic data and for the model given by the CMLE almost overlap (right).

the other hand, the model $\chi(y)$ estimated by the NBE seems to under-estimate the truth. However, as before, the extremal dependence structure is still reasonably well captured with the trained NBE.

## C.5 Case study: changes in horizontal geomagnetic field fluctuations

In this section, we summarise the results for the remaining censoring levels for each pair of locations. Contrarily to the main paper, we only show the selected model for each censoring level and the type of extremal dependence estimated with it.

### Pair (SCO, STF)

Table C.5.1: Model selected by the neural classifier for censoring levels $\tau = \{0.60, 0.65, \ldots, 0.95\}$ and parameter estimates given by the trained NBE for pair (SCO, STF). All the values are rounded up to 3 decimal places.

| $\tau$ | Model | $\hat{p}$ | $\hat{\boldsymbol{\theta}}_{\text{NBE}}$ | Extremal dependence |
|---|---|---|---|---|
| 0.60 | Model HW | 0.999 | $(\hat{\delta}, \hat{\omega}) = (0.170, 0.743)$ | AI |
| 0.65 | Model HW | 0.998 | $(\hat{\delta}, \hat{\omega}) = (0.178, 0.767)$ | AI |
| 0.70 | Model HW | 0.954 | $(\hat{\delta}, \hat{\omega}) = (0.178, 0.800)$ | AI |
| 0.75 | Model HW | 0.906 | $(\hat{\delta}, \hat{\omega}) = (0.195, 0.767)$ | AI |
| 0.80 | Model HW | 0.917 | $(\hat{\delta}, \hat{\omega}) = (0.228, 0.742)$ | AI |
| 0.85 | Model HW | 0.922 | $(\hat{\delta}, \hat{\omega}) = (0.258, 0.714)$ | AI |
| 0.90 | Model E2 | 0.935 | $(\hat{\alpha}, \hat{\xi}) = (3.512, -0.368)$ | AI |
| 0.95 | Model E2 | 0.640 | $(\hat{\alpha}, \hat{\xi}) = (3.616, -0.399)$ | AI |

**Pair (SCO, STJ)**

Table C.5.2: Model selected by the neural classifier for censoring levels $\tau = \{0.60, 0.65, \ldots, 0.95\}$ and parameter estimates given by the trained NBE for pair (SCO, STJ). All the values are rounded up to 3 decimal places.

| $\tau$ | Model | $\hat{p}$ | $\hat{\boldsymbol{\theta}}_{\text{NBE}}$ | Extremal dependence |
|---|---|---|---|---|
| 0.60 | Model HW | 1.000 | $(\hat{\delta}, \hat{\omega}) = (0.085, 0.560)$ | AI |
| 0.65 | Model HW | 1.000 | $(\hat{\delta}, \hat{\omega}) = (0.093, 0.580)$ | AI |
| 0.70 | Model HW | 1.000 | $(\hat{\delta}, \hat{\omega}) = (0.109, 0.586)$ | AI |
| 0.75 | Model HW | 1.000 | $(\hat{\delta}, \hat{\omega}) = (0.104, 0.591)$ | AI |
| 0.80 | Model HW | 0.958 | $(\hat{\delta}, \hat{\omega}) = (0.105, 0.626)$ | AI |
| 0.85 | Model E2 | 0.900 | $(\hat{\alpha}, \hat{\xi}) = (2.316\ -0.791)$ | AI |
| 0.90 | Model E2 | 0.940 | $(\hat{\alpha}, \hat{\xi}) = (2.748, -0.834)$ | AI |
| 0.95 | Model E2 | 0.875 | $(\hat{\alpha}, \hat{\xi}) = (3.223, -0.782)$ | AI |

**Pair (STF, STJ)**

Table C.5.3: Model selected by the neural classifier for censoring levels $\tau = \{0.60, 0.65, \ldots, 0.95\}$ and parameter estimates given by the trained NBE for pair (STF, STJ). All the values are rounded up to 3 decimal places.

| $\tau$ | Model | $\hat{p}$ | $\hat{\boldsymbol{\theta}}_{\text{NBE}}$ | Extremal dependence |
|---|---|---|---|---|
| 0.60 | Model HW | 1.000 | $(\hat{\delta}, \hat{\omega}) = (0.106, 0.558)$ | AI |
| 0.65 | Model HW | 1.000 | $(\hat{\delta}, \hat{\omega}) = (0.113, 0.571)$ | AI |
| 0.70 | Model HW | 1.000 | $(\hat{\delta}, \hat{\omega}) = (0.117, 0.588)$ | AI |
| 0.75 | Model HW | 0.996 | $(\hat{\delta}, \hat{\omega}) = (0.134, 0.585)$ | AI |
| 0.80 | Model HW | 0.920 | $(\hat{\delta}, \hat{\omega}) = (0.125, 0.610)$ | AI |
| 0.85 | Model E2 | 0.672 | $(\hat{\alpha}, \hat{\xi}) = (2.420, -0.849)$ | AI |
| 0.90 | Model E2 | 0.727 | $(\hat{\alpha}, \hat{\xi}) = (2.573, -0.846)$ | AI |
| 0.95 | Model E2 | 0.832 | $(\hat{\alpha}, \hat{\xi}) = (3.711, -0.772)$ | AI |

# Appendix D

# Supplementary material for Chapter 6

## D.1 Additional figures for Section 6.3

In this section, we present additional figures for Section 6.3 of the main paper, concerned with challenges C1 and C2. Figures D.1.1-D.1.3 support the exploratory analysis for challenges C1 and C2. We explore the within-year seasonality of the response variable $Y$ in Figure D.1.1, looking at the distribution of $Y$ per month and across the two seasons. This shows that there is a significant difference in the distribution of $Y$ between seasons 1 and 2, but within each season there is little difference across months.

Figure D.1.2 shows a scatter plot of $Y$ against each covariate $V_1, \ldots, V_8$, excluding $V_6$ which corresponds to season. Covariates $V_1, V_2$ and $V_8$ do not seem to have a relationship with $Y$, whilst there seems to be dependence for the remaining covariates. These observed relationships appear complex and non-linear.

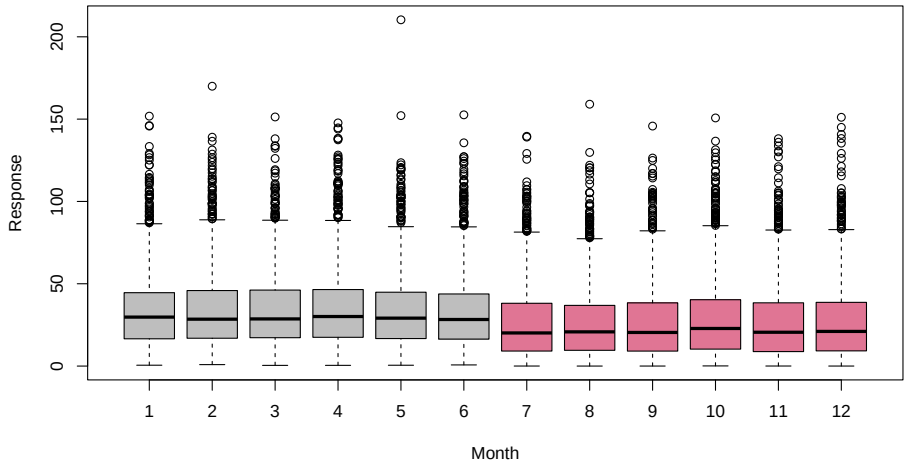



Figure D.1.1: Box plot of the response variable $Y$ with each month and season (season 1 in grey and season 2 in red).

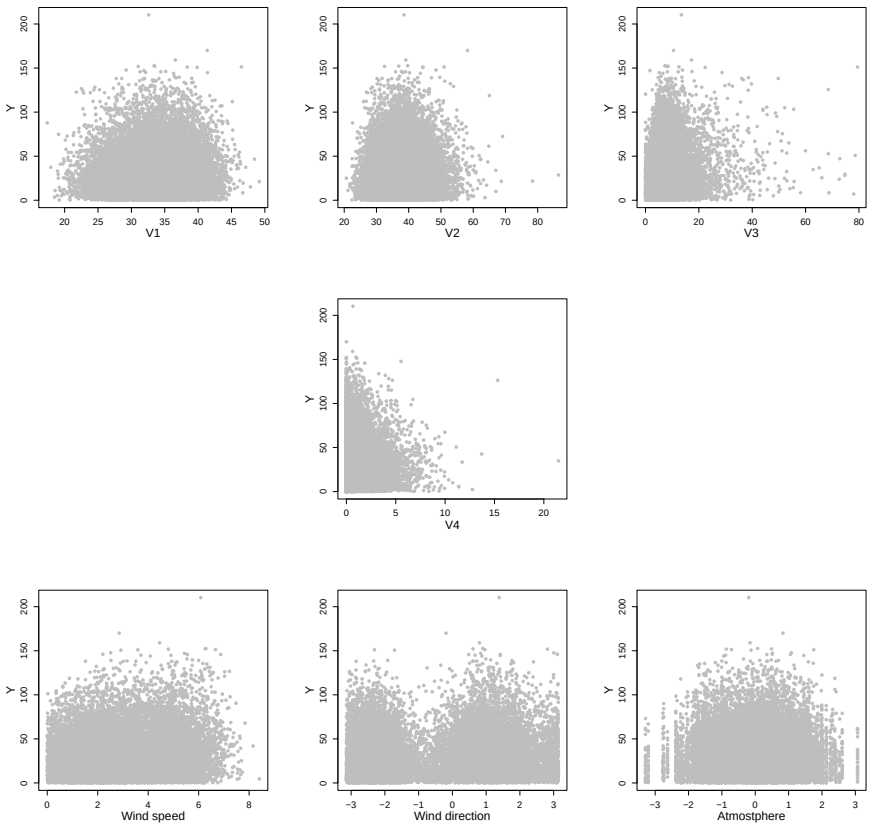



Figure D.1.2: Scatter plots of explanatory variables $V_1, \ldots, V_4$, wind speed ($V_6$), wind direction ($V_7$) and atmosphere ($V_8$), from top-left to bottom-right (by row), against the response variable $Y$.

We also explore temporal dependence in Figure D.1.3 that details the auto-correlation function (acf) values for the response $Y$ and explanatory variables $V_1, \ldots, V_4$, $V_6, \ldots, V_8$, up to a lag of 60. All variables have negligible acf values beyond lag 0, except $V_6$ (wind speed), $V_7$ (wind direction) and $V_8$ (atmosphere).

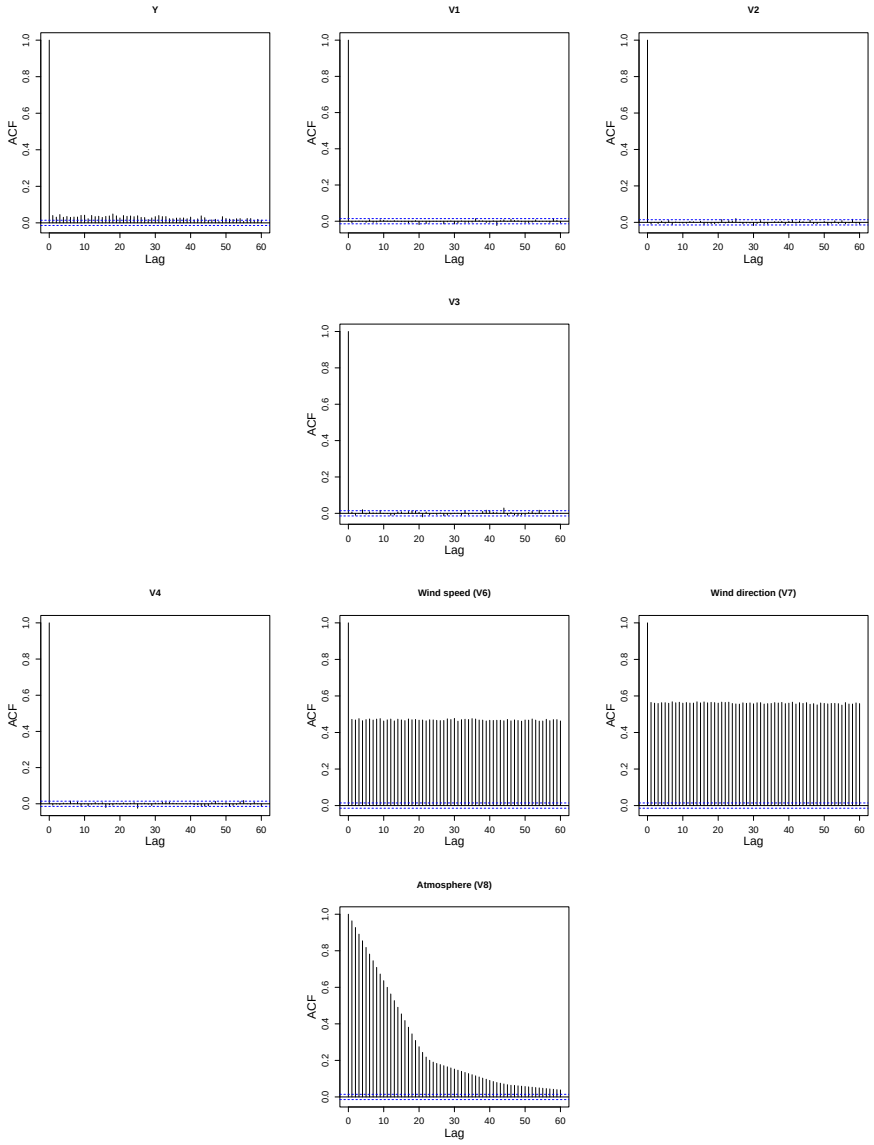



Figure D.1.3: Autocorrelation function plots for the response variable $Y$ and explanatory variables $V1, \ldots, V4$, wind speed ($V6$), wind direction ($V7$) and atmosphere ($V8$), from top-left to bottom-right (by row).

Figure D.1.4 shows the QQ-plots corresponding to a standard GPD model fitted to the excesses of $Y$ above a constant (left) and seasonally-varying threshold (right).

95% tolerance bounds (grey) show a lack of agreement between observations and the standard GPD model above a constant threshold. The second plot demonstrates a significant improvement in model fit.

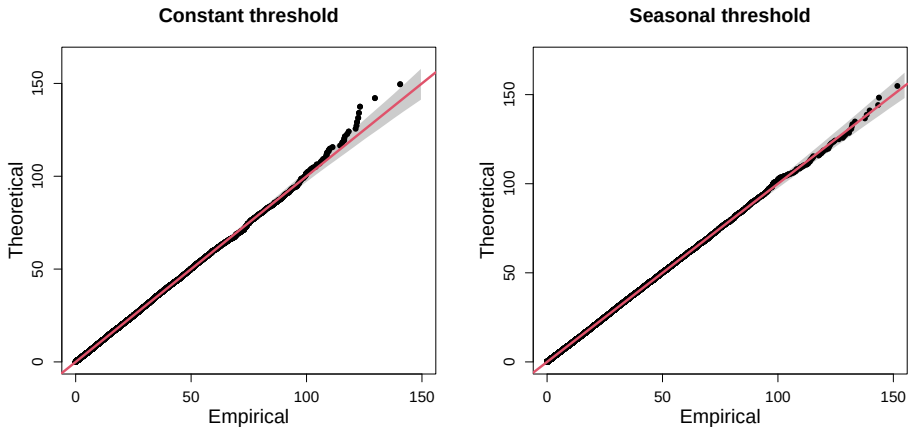



Figure D.1.4: QQ-plots showing standard GPD model fits with 95% tolerance bounds (grey) above a constant (left) and stepped-seasonal (right) threshold.

Figure D.1.5 shows a detailed summary of the pattern of missing data in the data and can be produced using the `missing_pattern` function in the `finalfit` package in `R` (Harrison et al., 2024). To interpret the figure note that blue and red squares represent observed and missing variables, respectively. The number on the right indicates the number of missing predictor variables (i.e., the number of red squares in the row), while the number on the left is the number of observations that fall into the row category. On the bottom, we have the number of observations that fall into the column category. For example, 18,545 observations are fully observed (denoted by the first row); there are 407 observations where only $V4$ is missing (denoted by the second row), 13 observations where both $V4$ and $V6$ are missing (denoted by the fourth row), 456 observations where $V4$ and at least one other predictor is missing (denoted by the last column in the table), etc. It can be seen that there are very few observations where more than one predictor is missing.

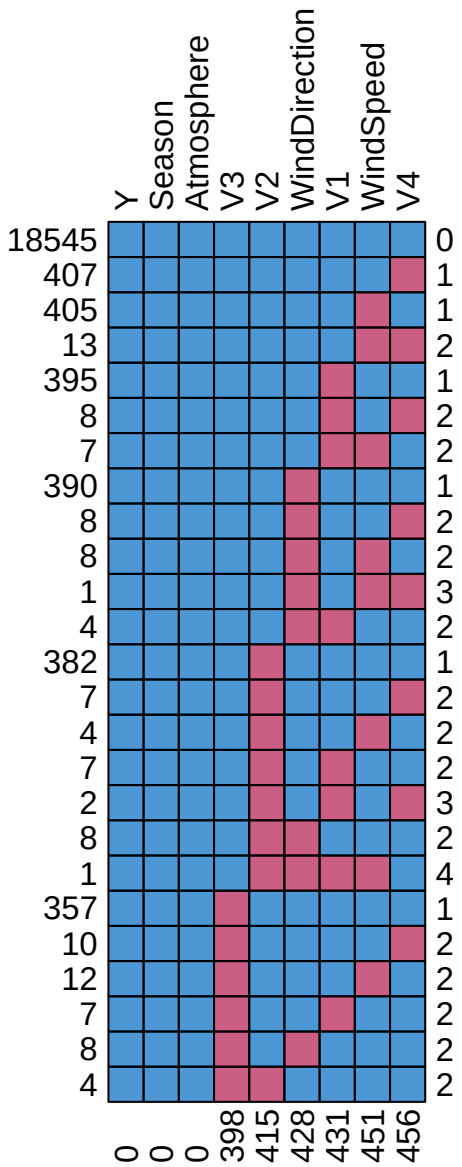



Figure D.1.5: Detailed pattern of missing predictor variables in the Amaurot data set.

## D.2 Additional figures for Section 6.4

In this section, we present additional plots related to Section 6.4 of the main article. Figure D.2.1 illustrates the time series of both covariates for the first 3 years of the observation period. It can be seen how the seasons vary periodically over each year, as well as the discrete nature of the atmospheric covariate.

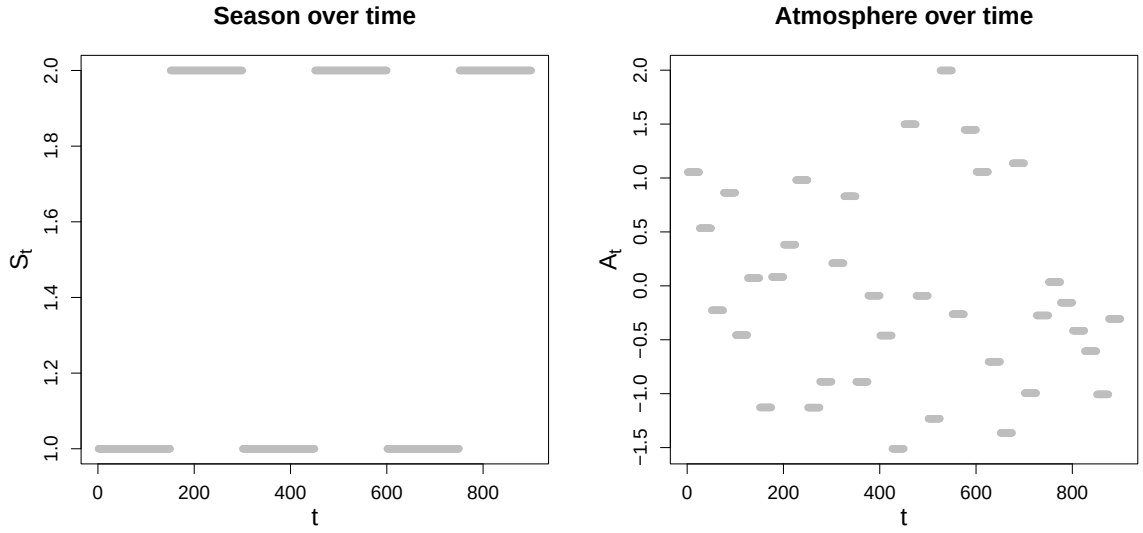



Figure D.2.1: Plots of $S_t$ (left) and $A_t$ (right) against $t$ for the first 3 years of the observation period.

Bootstrapped $\chi$ estimates for the groups $G^A_{I,k}, k \in \{1, \ldots, 10\}$, $I \in \mathcal{I} \setminus \{1, 2, 3\}$ and $G^S_{I,k}, k \in \{1, 2\}$, $I \in \mathcal{I}$ are given in Figures D.2.2 - D.2.5. These estimates illustrate the impact of atmosphere on the dependence structure.

For a 3-dimensional random vector, the angular dependence function, denoted $\lambda(\cdot)$, is defined on the unit-simplex $\boldsymbol{S}^2$ and describes extremal dependence along different rays $\boldsymbol{\omega} \in \boldsymbol{S}^2$. As noted in Section 4.2 of the main manuscript, we can associate each of the probabilities from C3, $p_1$ and $p_2$, with points on $\boldsymbol{S}^2$, denoted $\boldsymbol{\omega}^1$ and $\boldsymbol{\omega}^2$ respectively. With $I = \{1, 2, 3\}$, we consider $\lambda(\boldsymbol{\omega}^1)$ and $\lambda(\boldsymbol{\omega}^2)$ over the subsets $G^S_{I,k}$, $k \in \{1, 2\}$ and $G^A_{I,k}$, $k \in \{1, \ldots, 10\}$. We note that $\lambda(\boldsymbol{\omega}^1)$ is analogous with the coefficient of tail dependence $\eta \in (0, 1]$ (Ledford and Tawn, 1996), with $\eta = 1/3\lambda(\boldsymbol{\omega}^1)$; this corresponds

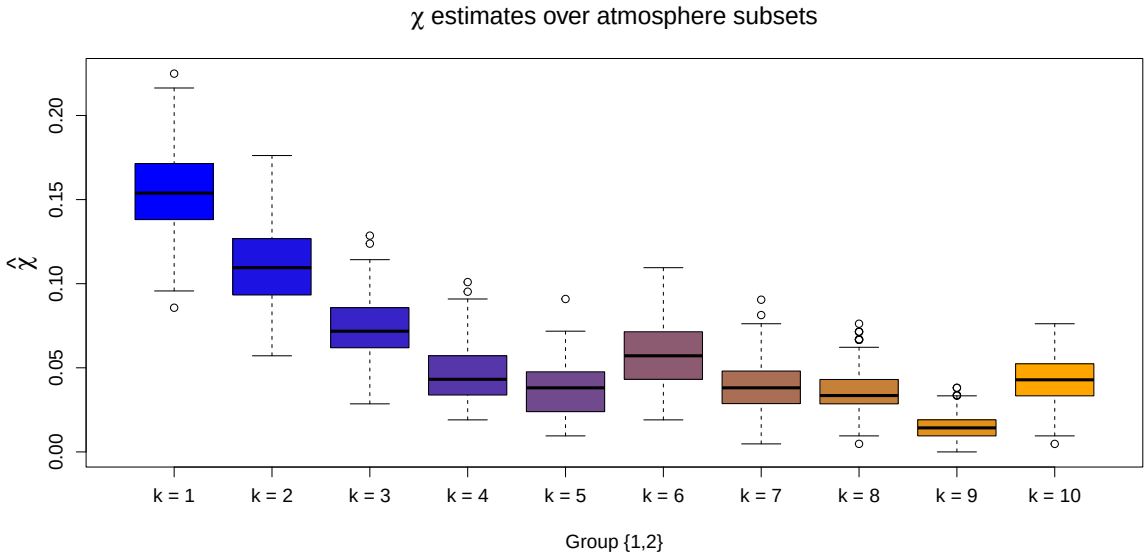



Figure D.2.2: Boxplots of empirical $\chi$ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2\}$. The colour transition (from blue to orange) over $k$ illustrates the trend in $\chi$ estimates as the atmospheric values are increased.

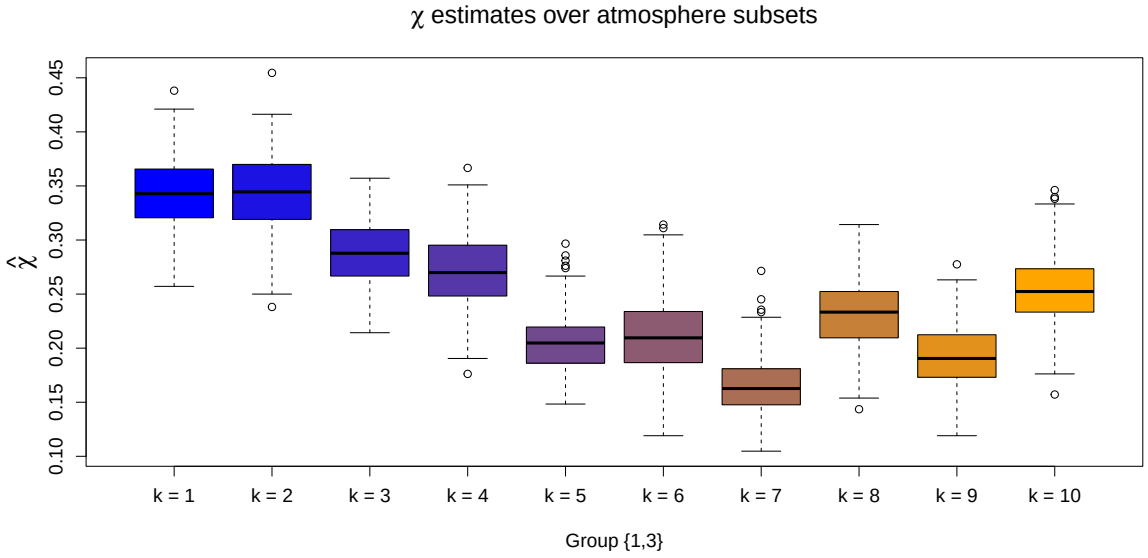



Figure D.2.3: Boxplots of empirical $\chi$ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 3\}$. The colour transition (from blue to orange) over $k$ illustrates the trend in $\chi$ estimates as the atmospheric values are increased.

with the region where all variables are simultaneously extreme. Furthermore, $\lambda(\boldsymbol{\omega}^2)$, which corresponds to a region where only two variables are extreme, is only evaluated after an additional marginal transformation of the third Coputopia time series; see

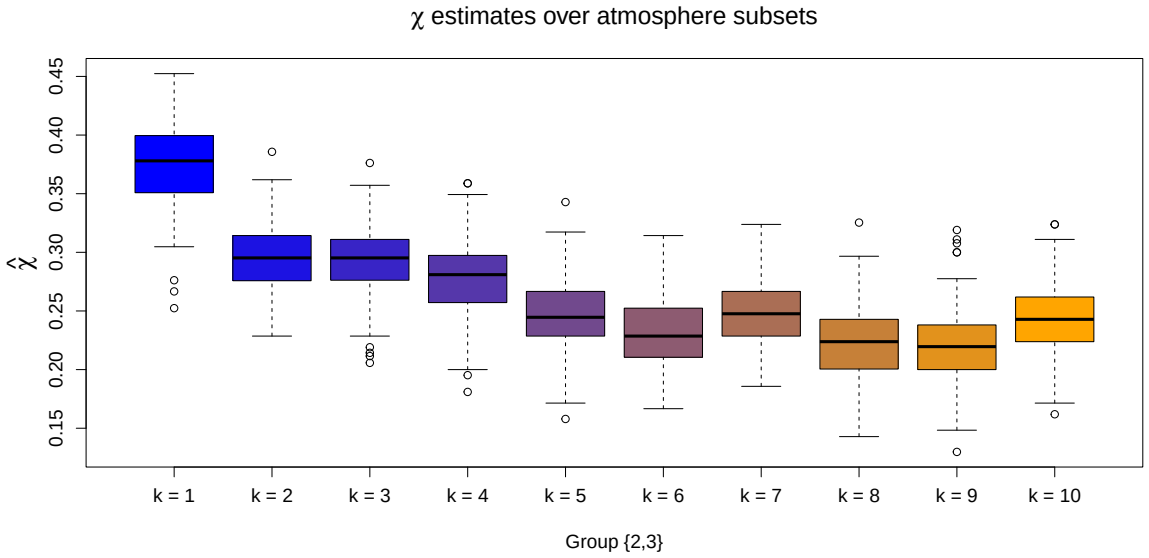



Figure D.2.4: Boxplots of empirical $\chi$ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{2, 3\}$. The colour transition (from blue to orange) over $k$ illustrates the trend in $\chi$ estimates as the atmospheric values are increased.

Section 4.2 of the main manuscript.

Estimation of $\lambda(\cdot)$ for each simplex point and subset was achieved using the Hill estimator (Hill, 1975) at the 90% level, with uncertainty subsequently quantified via bootstrapping. These results are given in Figures D.2.6 - D.2.9. These plots provide further evidence of a relationship between the extremal dependence structure and the covariates.

To illustrate the estimated trend in dependence, Figure D.2.10 shows the estimated scale functions, $\sigma(\boldsymbol{\omega}; \boldsymbol{x}_t)$, over atmosphere for parts 1 and 2. Under the assumption of asymptotic normality in the spline coefficients, 95% confidence intervals are obtained via posterior sampling; see Wood (2017) for more details. We observe that $\sigma$ tends to increase and decrease over atmosphere for parts 1 and 2, respectively, although the trend is less pronounced for the latter. Under our modelling framework, we note that higher values of $\sigma$ are associated with less positive extremal dependence in the direction $\boldsymbol{\omega}$ of interest; to see this, observe that the survivor function of the GPD with fixed $\xi$ is negatively associated with $\sigma$. Considering the trend in $\sigma(\boldsymbol{\omega}; \boldsymbol{x}_t)$, our results indicate a

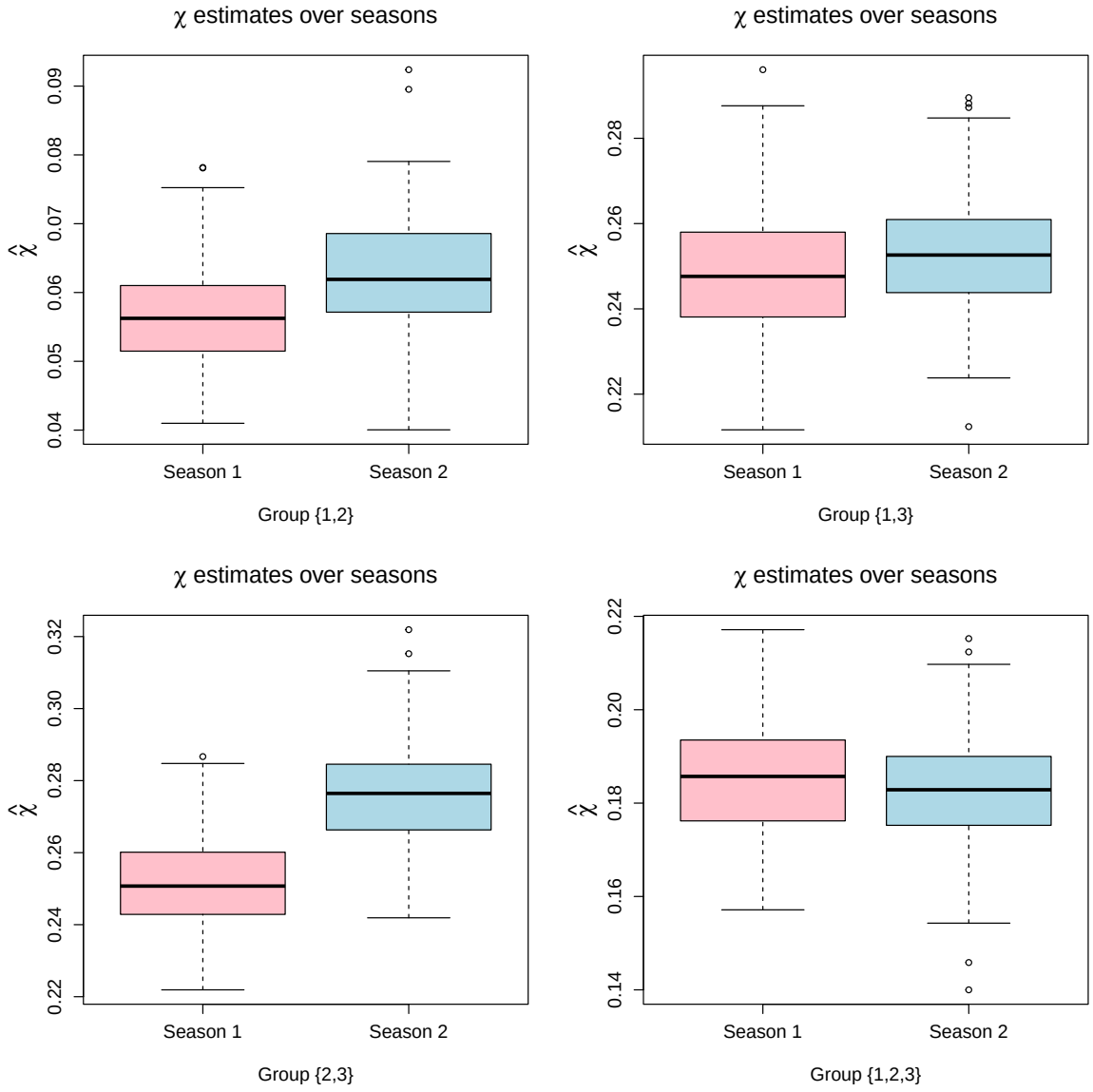



Figure D.2.5: Boxplots of empirical $\chi$ estimates obtained for the subsets $G^S_{I,k}$, with $k = 1, 2$. In each case, pink and blue colours illustrate estimates for seasons 1 and 2, respectively. From top left to bottom right: $I = \{1, 2, 3\}$, $I = \{1, 2\}$, $I = \{1, 3\}$, $I = \{2, 3\}$.

decrease in dependence in the region where all variables are extreme.

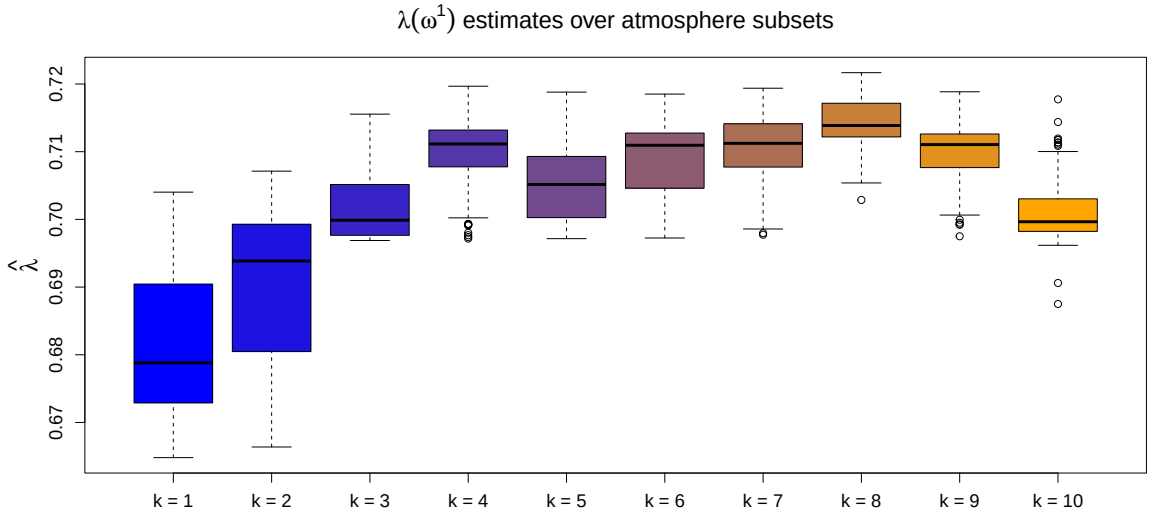



Figure D.2.6: Boxplots of empirical $\lambda(\boldsymbol{\omega}^1)$ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2, 3\}$. The colour transition (from blue to orange) over $k$ illustrates the trend in $\lambda$ estimates as the atmospheric values are increased.

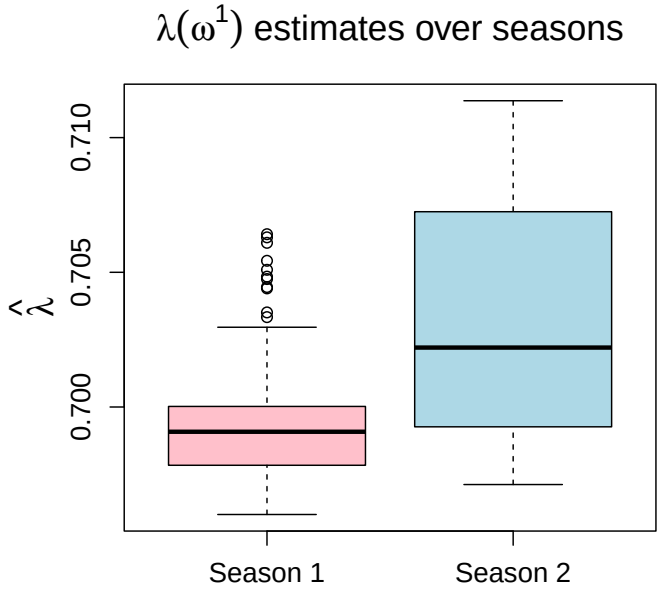



Figure D.2.7: Boxplots of empirical $\lambda(\boldsymbol{\omega}^1)$ estimates obtained for the subsets $G^S_{I,k}$, with $k = 1, 2$ and $I = \{1, 2, 3\}$. In each case, pink and blue colours illustrate estimates for seasons 1 and 2, respectively.

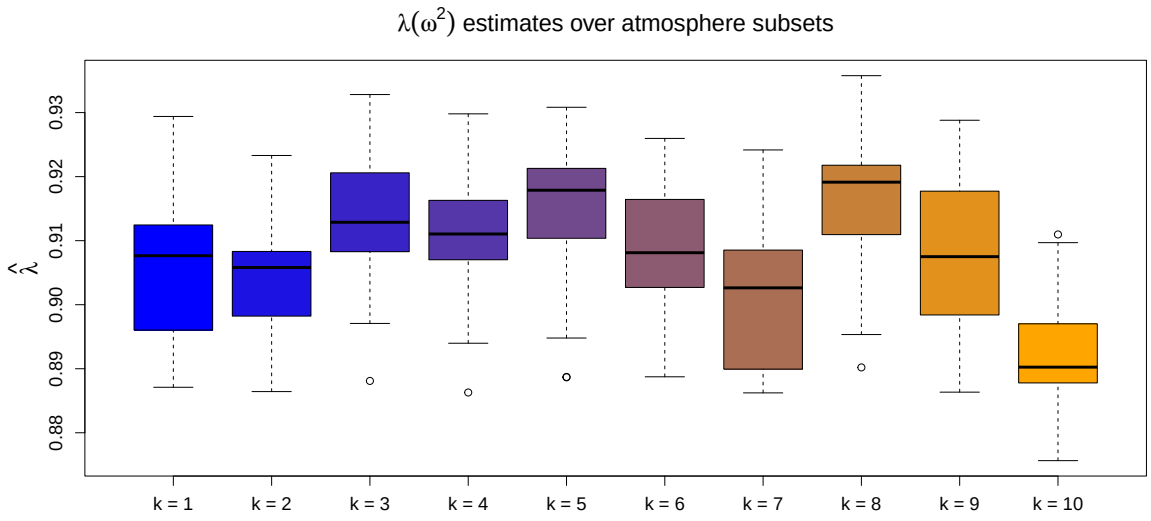



Figure D.2.8: Boxplots of empirical $\lambda(\boldsymbol{\omega}^2)$ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2, 3\}$. The colour transition (from blue to orange) over $k$ illustrates the trend in $\lambda$ estimates as the atmospheric values are increased.

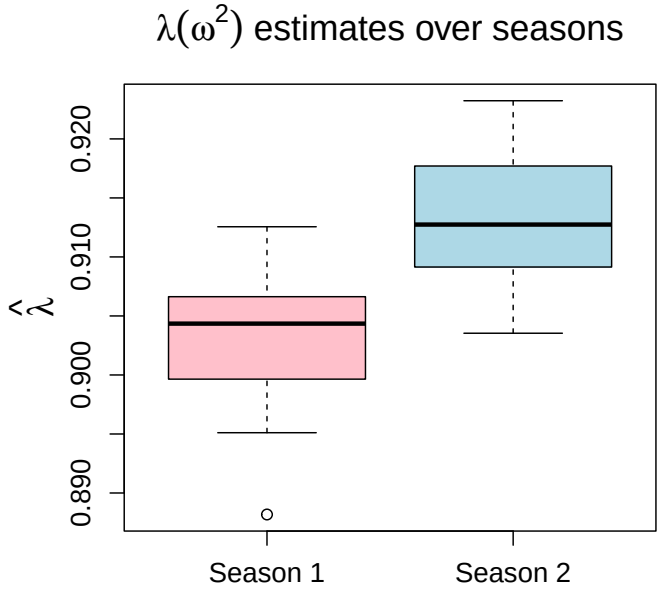



Figure D.2.9: Boxplots of empirical $\lambda(\boldsymbol{\omega}^2)$ estimates obtained for the subsets $G^S_{I,k}$, with $k = 1, 2$ and $I = \{1, 2, 3\}$. In each case, pink and blue colours illustrate estimates for seasons 1 and 2, respectively.

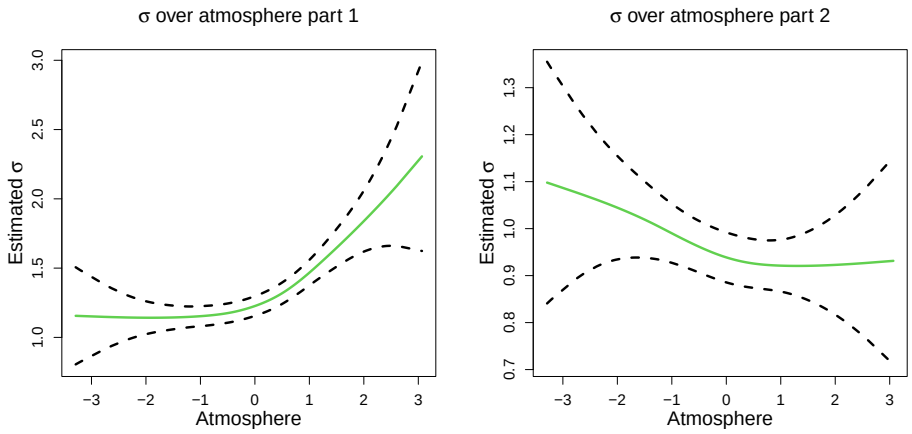



Figure D.2.10: Estimated $\sigma$ functions (green) over atmosphere for part 1 (left) and 2 (right). In both cases, the regions defined by the black dotted lines represent 95% confidence intervals obtained using posterior sampling.

## D.3 Additional figures for Section 6.5

In this section, we present additional plots related to Section 6.5 of the main article an we refer to $p_1$ and $p_2$ as parts 1 and 2 of C4, respectively. Figure D.3.1 shows a heat map of empirically estimated $\eta(\cdot)$ dependence coefficients and provides further evidence of the existence of the 5 dependence subgroups identified in our exploratory analysis for challenge C4. It also suggests that our modelling assumptions are reasonable; specifically that there is in-between group independence, and that the extremes within each group do not occur simultaneously.

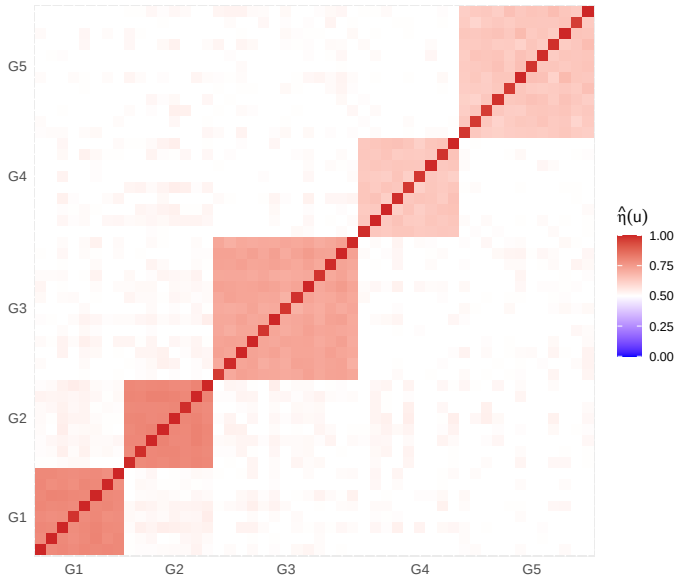



Figure D.3.1: Heat map of estimated empirical pairwise $\eta(u)$ extremal dependence coefficients with $u = 0.95$.

Figure D.3.2 shows the bootstrapped estimated individual group and overall probabilities with respect to conditioning threshold quantile for part 1 of challenge C4. Similarly, Figure D.3.3 shows the bootstrapped estimated individual group and overall probabilities with respect to conditioning threshold quantile for part 2 of challenge C4.

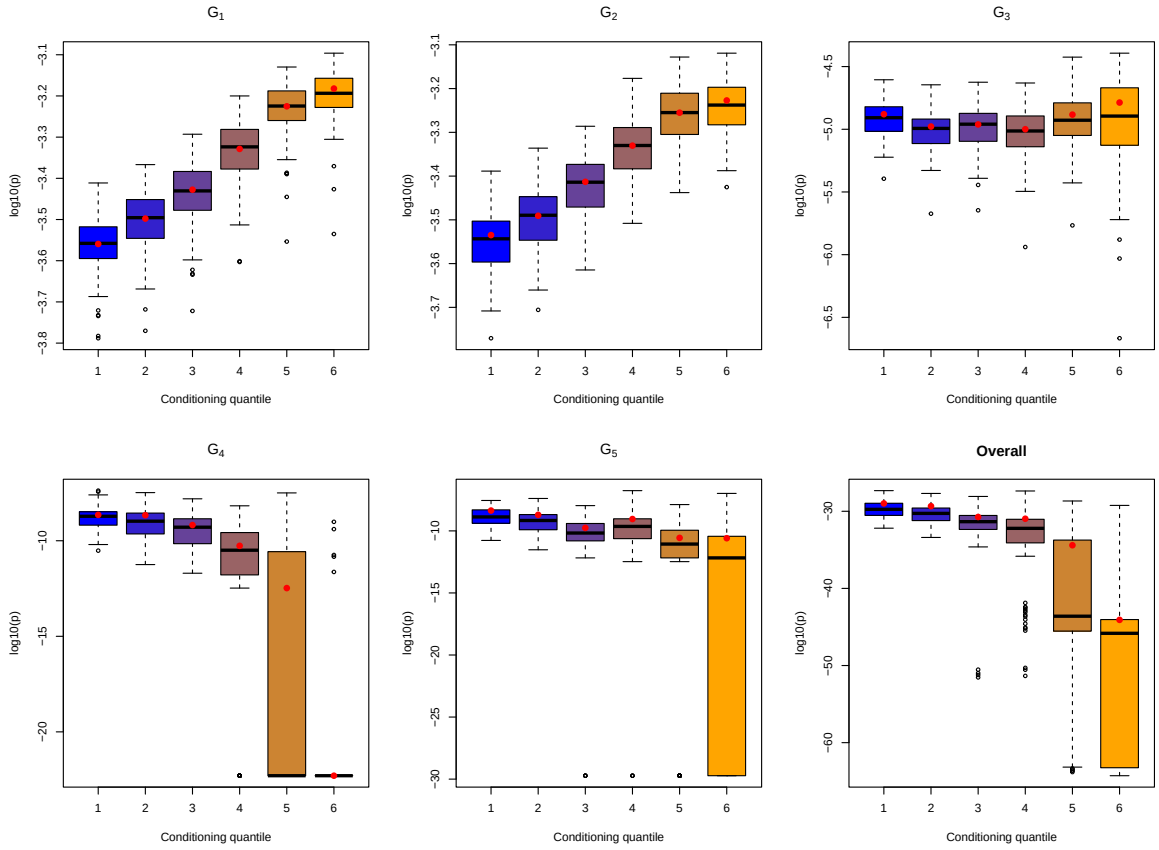



Figure D.3.2: Part 1 subgroup and overall bootstrapped probability estimates on the log scale. The red points indicate the original sample estimates and the colouring of the boxplots indicates the choice of conditioning threshold, with the conditioning quantile indices 1-6 referring to the quantile levels $\{0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$, respectively.

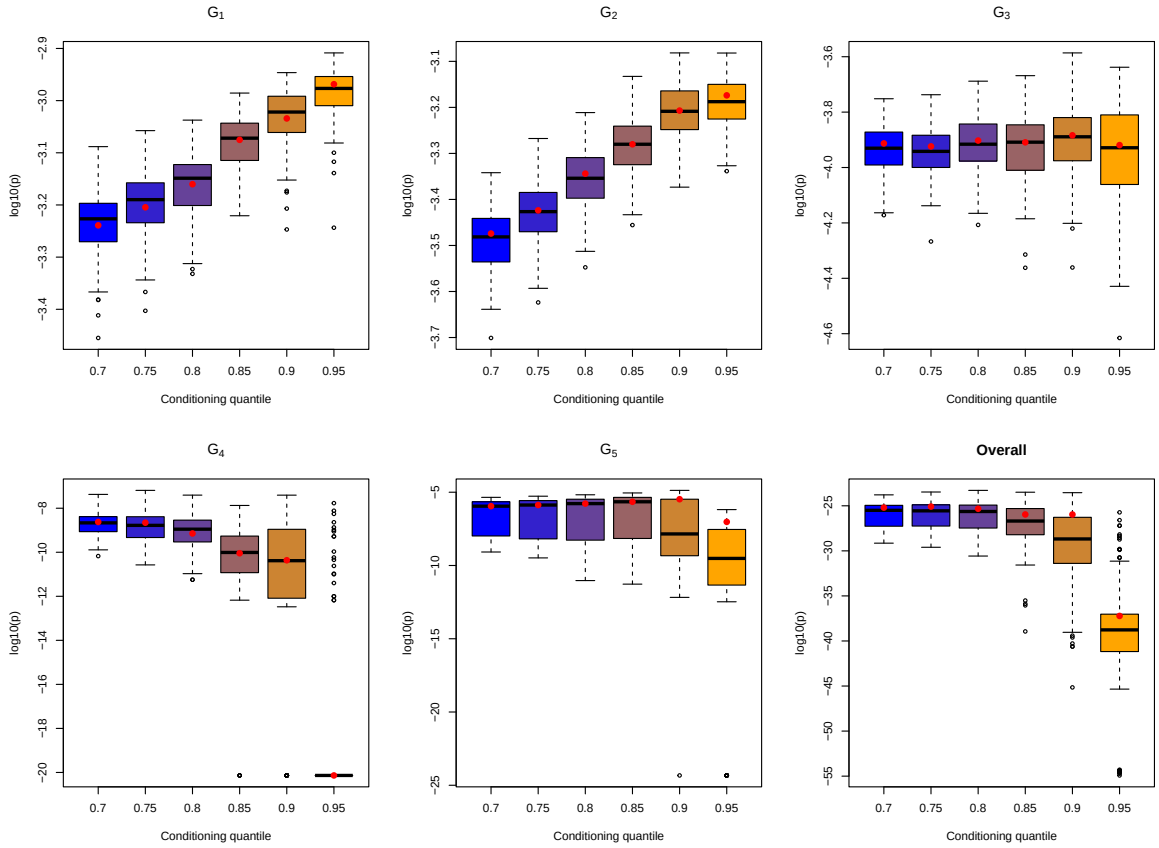



Figure D.3.3: Part 2 subgroup and overall bootstrapped probability estimates on the log scale for C4. The red points indicate the original sample estimates and the colouring of the boxplots indicates the choice of conditioning threshold, with the conditioning quantile indices 1-6 referring to the quantile levels $\{0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$, respectively.

# Bibliography


Ahmed, M., Maume-Deschamps, V., and Ribereau, P. (2022). Recognizing a spatial extreme dependence structure: A deep learning approach. *Environmetrics*, 33(4):e2714.

Ahmed, M., Maume-Deschamps, V., and Ribereau, P. (2024). Model selection for extremal dependence structures using deep learning: Application to environmental data. *arXiv*, 2409.13276.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alsina, C., Frank, M. J., and Schweizer, B. (2006). *Associative functions: Triangular norms and copulas*. World Scientific Publishing Co., Singapore.

André, L., Wadsworth, J., and O'Hagan, A. (2024). Joint modelling of the body and tail of bivariate data. *Computational Statistics and Data Analysis*, 189:107841.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37(2):697–725.

Aulbach, S., Bayer, V., and Falk, M. (2012a). A multivariate piecing-together approach with an application to operational loss data. *Bernoulli*, 18(2):455–475.

Aulbach, S., Falk, M., and Hofmann, M. (2012b). The multivariate piecing-together approach revisited. *Journal of Multivariate Analysis*, 110:161–170.

Bacigál, T., Juráňová, M., and Mesiar, R. (2010). On some new constructions of Archimedean copulas and applications to fitting problems. *Neural network world*, 20(1):81–90.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society: Series A (General)*, 139(3):318–344.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160.

Behrens, C. N., Lopes, H. F., and Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4(3):227–244.

Bennett, K. E., Cannon, A. J., and Hinzman, L. (2015). Historical trends and extremes in boreal Alaska river basins. *Journal of Hydrology*, 527:590–607.

Bernard, E., Naveau, P., Vrac, M., and Mestre, O. (2013). Clustering of maxima: Spatial dependencies among heavy rainfall in France. *Journal of Climate*, 26(20):7929–7937.

Bopp, G. P. and Shaby, B. A. (2017). An exponential-gamma mixture model for extreme Santa Ana winds. *Environmetrics*, 28(8).

Bottolo, L., Consonni, G., Dellaportas, P., and Lijoi, A. (2003). Bayesian analysis of extreme values by mixture modeling. *Extremes*, 6:25–47.

Cabras, S. and Castellanos, M. (2010). An objective Bayesian approach for threshold estimation in the peaks over the threshold model. Technical report, Análisis de riesgo.

Cannon, A. J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes*, 24:673–685.

Cannon, A. J. (2011). Gevcdn: An R package for nonstationary extreme value analysis by generalized extreme value conditional density estimation network. *Computers and Geosciences*, 37:1532–1533.

Carreau, J. and Bengio, Y. (2009). A hybrid Pareto model for asymmetric fat-tailed data: The univariate case. *Extremes*, 12(1):53–76.

Carreau, J. and Vrac, M. (2011). Stochastic downscaling of precipitation with neural network conditional mixture models. *Water Resources Research*, 47:10502.

Castro-Camilo, D., Huser, R., and Rue, H. (2019). A spliced Gamma-Generalized Pareto model for short-term extreme wind speed probabilistic forecasting. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):517–534.

Ceresetti, D., Ursu, E., Carreau, J., Anquetin, S., Creutin, J. D., Gardes, L., Girard, S., and Molinié, G. (2012). Evaluation of classical spatial-analysis schemes of extreme rainfall. *Hazards Earth System Sciences*, 12:3229–3240.

Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics*, 9(1):383 – 418.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(1):207–222.

Cisneros, D., Richards, J., Dahal, A., Lombardo, L., and Huser, R. (2024). Deep graphical regression for jointly moderate and extreme Australian wildfires. *Spatial Statistics*, 59:100811.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London, UK.

Coles, S. G., Heffernan, J. E., and Tawn, J. A. (1999). Dependence measures for extreme value analyses. *Extremes*, 2:339–365.

Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):377–392.

Coles, S. G. and Tawn, J. A. (1994). Statistical methods for multivariate extremes: An application to structural design. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):1–31.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences of the United States of America*, 117:30055–30062.

Cranmer, K., Pavez, J., and Louppe, G. (2016). Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv*, 1506.02169v2.

D'Arcy, E., Tawn, J. A., Joly, A., and Sifnioti, D. E. (2023). Accounting for seasonality in extreme sea-level estimation. *The Annals of Applied Statistics*, 17(4):3500–3525.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 52(3):393–425.

de Carvalho, M., Pereira, S., Pereira, P., and de Zea Bermudez, P. (2022). An extreme value Bayesian Lasso for the conditional left and right tails. *Journal of Agricultural, Biological and Environmental Statistics*, 27:222–239.

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771.

de Mendes, B. and Lopes, H. F. (2004). Data driven estimates for mixtures. *Computational Statistics and Data Analysis*, 47:583–598.

Diggle, P. J. and Gratton, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46:193–212.

do Nascimento, F. F., Gamerman, D., and Lopes, H. F. (2012). A semiparametric Bayesian approach to extreme value estimation. *Statistics and Computing*, 22:661–675.

Durante, F., Fernández Sánchez, J., and Sempi, C. (2013). Multivariate patchwork copulas: a unified approach with applications to partial comonotonicity. *Insurance: Mathematics and Economics*, 53(3):897–905.

Durante, F., Foschi, R., and Sarkoci, P. (2010). Distorted copulas: constructions and tail dependence. *Communications in Statistics - Theory and Methods*, 39(12):2288–2301.

Durante, F., Saminger-Platz, S., and Sarkoci, P. (2009). Rectangular patchwork for bivariate copulas and tail dependence. *Communications in Statistics - Theory and Methods*, 38(15):2515–2527.

Durrleman, V., Nikeghbali, A., and Roncalli, T. (2000). A simple tranformation of copulas. *SSRN Electronic Journal*.

Eastoe, E. F. and Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 58(1):25–45.

Eastoe, E. F. and Tawn, J. A. (2012). Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. *Biometrika*, 99(1):43–55.

Engelke, S., Opitz, T., and Wadsworth, J. (2019). Extremal dependence of random scale constructions. *Extremes*, 22:623–666.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, pages 615–629.

Finch, D. P. and Palmer, P. I. (2020). Increasing ambient surface ozone levels over the UK accompanied by fewer extreme events. *Atmospheric Environment*, 237:117627.

Frank, M. J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae*, 19:194–226.

Frigessi, A., Haug, O., and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5(2002):219–235.

Galambos, J. (1975). Order statistics of samples from multivariate distributions. *Journal of the American Statistical Association*, 70(351):674–680.

Gamet, P. and Jalbert, J. (2022). A flexible extended generalized Pareto distribution for tail estimation. *Environmetrics*, 33(6):e2744.

Gerber, F. and Nychka, D. (2021). Fast covariance parameter estimation of spatial Gaussian process models using neural networks. *Stat*, 10:e382.

Gjerloev, J. W. (2009). A global ground-based magnetometer initiative. *Eos Transactions American Geophysical Union*, 90(27):230–231.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gormley, I. C. and Frühwirth-Schnatter, S. (2019). Mixture of experts models. In *Handbook of Mixture Analysis*, pages 271–307. Chapman and Hall/CRC.

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*, 273:118975.

Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8:S85–S118.

Gräler, B. (2014). Modelling skewed spatial random fields through the spatial vine copula. *Spatial Statistics*, 10:87–102.

Grazian, C. and Fan, Y. (2020). A review of approximate Bayesian computation methods via density estimation: Inference for simulator-models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12:e1486.

Guerrero, M. B., Huser, R., and Ombao, H. (2023). Conex–Connect: Learning patterns in extremal brain connectivity from MultiChannel EEG data. *The Annals of Applied Statistics*, 17(1):178–198.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707.

Harrison, E., Drake, T., and Ots, R. (2024). finalfit: Quickly create elegant regression results tables and plots when modelling (R package).

Hastie, T., Friedman, J., and Tibshirani, R. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY, 1st edition.

Heffernan, J. E. (2000). A directory of coefficients of tail dependence. *Extremes*, 3(3):279–290.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate

extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:497–546.

Hermans, J., Begy, V., and Louppe, G. (2020). Likelihood-free MCMC with amortized approximate ratio estimators. In *Proceeding of the 37th International Conference on Machine Learning*.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Hu, C., Swallow, B., and Castro-Camilo, D. (2024). A Bayesian multivariate extreme value mixture model. *arXiv*, 2401.15703.

Hu, S. and O'Hagan, A. (2021). Copula averaging for tail dependence in insurance claims data. *arXiv*, 2103.10912.

Huang, W. K., Nychka, D. W., and Zhang, H. (2019). Estimating precipitation extremes using the log-histospline. *Environmetrics*, 30(4):e2543.

Hummel, C. (2009). Shaping tail dependencies by nesting box copulas. *arXiv*, 0906.4853.

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, 114(525):434–444.

Hüsler, J. and Reiss, R. D. (1989). Maxima of normal random vectors: Between independence and complete dependence. *Statistics and Probability Letters*, 7(4):283–286.

Inácio de Carvalho, V., de Carvalho, M., and Branscum, A. J. (2017). Nonparametric Bayesian covariate-adjusted estimation of the Youden index. *Biometrics*, 73(4):1279–1288.

Innes, M. (2018). Flux: Elegant machine learning with Julia. *Journal of Open Source Software*, 3(25):602.

Janßen, A. and Wan, P. (2020). $k$-means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211 – 1233.

Joe, H. (1996). Families of $m$-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. *Lecture Notes-Monograph Series*, 28:120–141.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall/CRC, New York, 1st edition.

Joe, H. (2014). *Dependence Modeling with Copulas*. Taylor and Francis Group, Florida.

Jonathan, P., Randell, D., Wu, Y., and Ewans, K. (2014). Return level estimation from non-stationary spatial data exhibiting multidimensional covariate effects. *Ocean Engineering*, 88:520–532.

Keef, C., Papastathopoulos, I., and Tawn, J. A. (2013a). Estimation of the conditional distribution of a multivariate variable given that one of its components is large: additional constraints for the Heffernan and Tawn model. *Journal of Multivariate Analysis*, 115:396–404.

Keef, C., Tawn, J. A., and Lamb, R. (2013b). Estimating the probability of widespread flood events. *Environmetrics*, 24(1):13–21.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Klement, E. P., Mesiar, R., and Pap, E. (2005). Transformations of copulas. *Kybernetika*, 41(4):425–434.

Krock, M., Bessac, J., Stein, M. L., and Monahan, A. H. (2022). Nonstationary seasonal model for daily mean temperature distribution bridging bulk and tails. *Weather and Climate Extremes*, 36:100438.

Krupskii, P., Huser, R., and Genton, M. G. (2018). Factor copula models for replicated spatial data. *Journal of the American Statistical Association*, 113(521):467–479.

Kyselý, J., Picek, J., and Beranová, R. (2010). Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold. *Global and Planetary Change*, 72(1-2):55–68.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer, New York, NY.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Ledford, A. W. and Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):475–499.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York, 2nd edition.

Lenzi, A., Bessac, J., Rudi, J., and Stein, M. L. (2023). Neural networks for parameter estimation in intractable models. *Computational Statistics and Data Analysis*, 185:107762.

Lenzi, A. and Rue, H. (2023). Towards black-box parameter estimation. *arXiv*, 2303.15041.

Leonelli, M. and Gamerman, D. (2020). Semiparametric bivariate modelling with flexible extremal dependence. *Statistics and Computing*, 30:221–236.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2017). Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66(1):66–82.

Liu, Y. and Tawn, J. (2014). Self-consistent estimation of conditional multivariate extreme value distributions. *Journal of Multivariate Analysis*, 127:19–35.

MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M., and Russell, G. (2011). A flexible extreme value mixture model. *Computational Statistics and Data Analysis*, 55(6):2137–2157.

Majumder, R. and Reich, B. J. (2023). A deep learning synthetic likelihood approximation of a non-stationary spatial model for extreme streamflow forecasting. *Spatial Statistics*, 55:100755.

Majumder, R., Reich, B. J., and Shaby, B. A. (2024). Modeling extremal streamflow using deep learning approaximations and a flexible spatial process. *The Annals of Applied Statistics*, 18:1519–1542.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, 18:285–296.

Mesiar, R., Jágr, V., Juráňová, M., and Komorníková, M. (2008). Univariate conditioning of copulas. *Kybernetika*, 44(6):807–816.

Mhalla, L., Opitz, T., and Chavez-Demoulin, V. (2019). Exceedance-based nonlinear regression of tail dependence. *Extremes*, 22(3):523–552.

Morillas, P. M. (2005). A method to obtain new copulas from a given one. *Metrika*, 61(2):169–184.

Murphy, C., Tawn, J. A., and Varty, Z. (2024). Automated threshold selection and associated inference uncertainty for univariate extremes. *Technometrics*, pages 1–17.

Murphy-Barltrop, C. J. and Wadsworth, J. L. (2024). Modelling non-stationarity in asymptotically independent extremes. *Computational Statistics and Data Analysis*, 199:108025.

Murphy-Barltrop, C. J. R., Majumder, R., and Richards, J. (2024). Deep learning of multivariate extremes via a geometric representation. *arXiv*, 2406.19936v2.

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York, NY.

Northrop, P. J., Attalides, N., and Jonathan, P. (2017). Cross-Validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 66(1):93–120.

Northrop, P. J. and Jonathan, P. (2011). Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics*, 22(7):799–809.

Opitz, T., Huser, R., Bakka, H., and Rue, H. (2018). INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes*, 21(3):441–462.

Papastathopoulos, I. and Tawn, J. A. (2013). Extended generalised Pareto models for tail estimation. *Journal of Statistical Planning and Inference*, 143(1):131–143.

Pasche, O. C. and Engelke, S. (2024). Neural networks for extreme quantile regression

with an application to forecasting of flood risk. *The Annals of Applied Statistics*, 18(4):2818–2839.

Pfeifer, D., Mändle, A., and Ragulina, O. (2017). New copulas based on general partitions-of-unity and their applications to risk management (part II). *Dependence Modeling*, 5(1):246–255.

Pfeifer, D., Mändle, A., Ragulina, O., and Girschig, C. (2016). New copulas based on general partitions-of-unity and their applications to risk management. *Dependence Modeling*, 4:123–140.

Pfeifer, D., Mändle, A., Ragulina, O., and Girschig, C. (2019). New copulas based on general partitions-of-unity (part III) - the continuous case. *Dependence Modeling*, 7(1):181–201.

Pfeifer, D. and Ragulina, O. (2021). Generating unfavourable VaR scenarios under Solvency II with patchwork copulas. *Dependence Modeling*, 9(1):327–346.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119 – 131.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11.

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022). The dependent Dirichlet process and related models. *Statistical Science*, 37(1):24–41.

Radev, S. T., Schmitt, M., Pratz, V., Picchini, U., Köthe, U., and Bürkner, P.-C. (2023). JANA: Jointly amortized neural approximation of complex Bayesian models.

In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 1013)*, pages 1695–1706.

Ramírez, V. P., de Carvalho, M., and Gutiérrez, L. (2024). Heavy-tailed NGG-mixture models. *Bayesian Analysis*, 1(1):1–29.

Resnick, S. (2002). Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5:303–336.

Richards, J. and Huser, R. (2022). Regression modelling of spatiotemporal extreme U.S. wildfires via partially-interpretable neural networks. *arXiv*, 2208.07581.

Richards, J., Sainsbury-Dale, M., Zammit-Mangion, A., and Huser, R. (2024). Neural Bayes estimators for censored inference with peaks-over-threshold models. *Journal of Machine Learning Research*, (to appear).

Rogers, N. C., Wild, J. A., Eastoe, E. F., Gjerloev, J. W., and Thomson, A. W. P. (2020). A global climatological model of extreme geomagnetic field fluctuations. *Journal of Space Weather and Space Climate*, 10(5):1–19.

Rohrbeck, C., Simpson, E., and Tawn, J. A. (2024). Editorial: EVA (2023) conference data challenge. *Extremes*, (to appear).

Sainsbury-Dale, M., Zammit-Mangion, A., and Huser, R. (2024a). Likelihood-free parameter estimation with neural Bayes estimators. *The American Statistician*, 78(1):1–14.

Sainsbury-Dale, M., Zammit-Mangion, A., Richards, J., and Huser, R. (2024b). Neural Bayes estimators for irregular spatial data using graph neural networks. *Journal of Computational and Graphical Statistics*, (to appear).

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical Journal*, 10(1):33–60.

Shamiri, A., Hamzah, N. A., and Pirmoradian, A. (2011). Tail dependence estimate in financial market risk management: Clayton-Gumbel copula approach. *Sains Malaysiana*, 40(8):927–935.

Shrestha, R. R., Cannon, A. J., Schnorbus, M. A., and Zwiers, F. W. (2017). Projecting future nonstationary extreme streamflow for the Fraser River, Canada. *Climatic Change*, 145:289–303.

Siburg, K. F. and Stoimenov, P. A. (2008). Gluing copulas. *Communications in Statistics - Theory and Methods*, 37(19):3124–3134.

Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532.

Sisson, S., Fan, Y., and Beaumount, M. (2018). *Handbook of Approximate Bayesian Computation*. Chapman & Hall/CRC Press.

Sklar, M. (1959). Fonctions de répartition à $n$ dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, VIII(3):229–231.

Stein, M. L. (2021). A parametric model for distributions with flexible behavior in both tails. *Environmetrics*, 32(2):e2658.

Tancredi, A., Anderson, C., and O'Hagan, A. (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes*, 9:87–106.

Tawn, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika*, 75(3):397–415.

Tencaliec, P., Favre, A. C., Naveau, P., Prieur, C., and Nicolet, G. (2020). Flexible semiparametric generalized Pareto modeling of the entire range of rainfall amount. *Environmetrics*, 31:e2582.

Vasiliades, L., Galiatsatou, P., and Loukas, A. (2015). Nonstationary frequency analysis of annual maximum rainfall using climate covariates. *Water Resources Management*, 29:339–358.

Vrac, M., Naveau, P., and Drobinski, P. (2007). Modeling pairwise dependencies in precipitation intensities. *Nonlin. Processes Geophys*, 14:789–797.

Wadsworth, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics*, 58(1):116–126.

Wadsworth, J. L. and Tawn, J. A. (2013). A new representation for multivariate tail probabilities. *Bernoulli*, 19(5B):2689–2714.

Wadsworth, J. L., Tawn, J. A., Davison, A. C., and Elton, D. M. (2017). Modelling across extremal dependence classes. *Journal of the Royal Statistical Society: Series B*, 79:149–175.

Walchessen, J., Lenzi, A., and Kuusela, M. (2024). Neural likelihood surfaces for spatial processes with computationally intensive or intractable likelihoods. *Spatial Statistics*, 62:100848.

Wixson, T. P. and Cooley, D. (2024). Neural network for asymptotic dependence/ independence classification: A series of experiments. *https://doi.org/10.21203/rs.3.rs-3994810/v1*.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, New York, 2nd edition.

Yadav, R., Huser, R., and Opitz, T. (2021). Spatial hierarchical modeling of threshold exceedances using rate mixtures. *Environmetrics*, 32:e2662.

Youngman, B. D. (2019). Generalized Additive Models for Exceedances of High Thresholds With an Application to Return Level Estimation for U.S. Wind Gusts. *Journal of the American Statistical Association*, 114(528):1865–1879.

Youngman, B. D. (2022). evgam: An R Package for generalized additive extreme value models. *Journal of Statistical Software*, 103(3):1–26.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics and Probability Letters*, 54(4):437–447.

Zaheer, M., Kottur, S., Ravanbhakhsh, S., Póczos, B., Salakhutdinov, R., and Smola, A. J. (2017). Deep Sets. *Advances in Neural Information Processing Systems*, 30.

Zammit-Mangion, A., Sainsbury-Dale, M., and Huser, R. (2025). Neural methods for amortized inference. *Annual Review of Statistics and Its Applications*, (to appear).

Zhang, L., Shaby, B. A., and Wadsworth, J. L. (2022a). Hierarchical transformed scale mixtures for flexible modeling of spatial extremes on datasets with many locations. *Journal of the American Statistical Association*, 117(539):1357–1369.

Zhang, Z., Huser, R., Opitz, T., and Wadsworth, J. L. (2022b). Modeling spatial extremes using normal mean-variance mixtures. *Extremes*, 25(2):175–197.