Semiparametric Bayesian Risk Estimation for Complex Extremes

THÈSE N^{O} 8349 (2018)

PRÉSENTÉE LE 26 AVRIL 2018

À LA FACULTÉ DES SCIENCES DE BASE
CHAIRE DE STATISTIQUE
PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Thomas LUGRIN

acceptée sur proposition du jury:

Prof. T. Mountford, président du jury
Prof. A. C. Davison, Prof. J. Tawn, directeurs de thèse
Dr Ph. Naveau, rapporteur
Dr P. Northrop, rapporteur
Prof. S. Morgenthaler, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2018

μηδὲν ἄγαν
(inscription on the temple of Apollo at Delphi)

À Maman, pour ta thèse interrompue

Acknowledgements

I am much indebted to Anthony Davison, who provided me with the support for carrying research in an excellent environment, giving the opportunity for me to participate in many international conferences and workshops. I learned a lot from his broad knowledge, patience and kindness. His rigour and expertise will continue guiding me in the future.

It is hard to express in a few words how much I owe to Jonathan Tawn, whose tireless energy and motivation steered me throughout my research. He made me feel as if I had always been at Lancaster University, encouraging me to take part in the academic activities of the STOR-i doctoral training centre, that have greatly enriched me. During my many visits to Lancaster, I was able to get ahead in my research, punctuated by very regular meetings with Jon and the inevitable pile of notes —full of ideas— that I would bring back at my desk.

I would also like to thank Paul Northrop, Philippe Naveau, Stephan Morgenthaler, and Thomas Mountford for sitting on the thesis jury and for their careful reading and detailed comments.

I acknowledge support from the Swiss National Science Foundation. Ich möchte mich besonders bei Georges Klein bedanken, denn er hat mich für meine 6-Monaten Besuch in Lancaster wegen der Universitätgebühren ganz nett und professionnell unterstüzt.

I would like to thank Kim and Rosemary, who helped me settle in at STOR-i. It would be too long here to mention all the people in Lancaster with whom I had so many fruitful discussions and spent such a great time; Jenny Wadsworth and Ioannis Papastathopoulos for the helpful discussions, and STOR-i mates Aaron & Jamie-Leigh and their inimitable accents, Christian, Ciara for guiding me through British life and habits, Dave & Aisha for their hospitality, Harjit and his endless theories, Hugo for extreme conversations and the Eurovision, and Charlotte, Jack for the muddy runs, and Sorcha, Jamie and his questions about French words, Lucy for organising outings, Matt & Helen, Paul & Zak, Rhian & Jeddy, Sam for the dry hikes, Kathryn for the Lindy-Hop. My thanks also go to Amy Cotterill for introducing me to the Lakes and the Dales.

J'aimerais également remercier mes collègues et amis de l'EPFL pour leur présence quotidienne et les discussions enrichissantes, parmi lesquels Miguel pour m'avoir donné le goût des statistiques bayésiennes non-paramétriques, Claudio pour tous les échanges, Raphaël pour la course à pieds, Léo, Hélène, Peiman, Sophie pour son rayonnement, Sebastian pour son accent vaudois. Je remercie les membres des autres chaires de statistiques, en particulier Rémy, Marie-Hélène et Yoav. Merci beaucoup à Nadia pour son travail efficace et sa disponibilité, et les échanges passionnants.

Je dois aussi énormément au professionalisme, au calme et à la patience du docteur Duccio Boscherini et de toute son équipe, pour m'avoir remis sur pieds si impeccablement et si rapidement d'une hernie discale qui a si brusquement interrompu ma thèse.

Je pense également aux amitiés forgées dans le travail et les cafés: David, Pauline et Reda, et aux amitiés partagées depuis plus longtemps: Adrien, Denis, Odile, Matthieu et Marie, pour tous les bons moments passés ensemble, les partages, les échanges d'idées et les voyages.

Do të doja të falenderoja zemrën time që më ka nxitur shumë e më ka dhënë besim. Anjeza më ka hapur shpirtin dhe më ka bërë të zbuloj një tjetër botë, plot me bujari. Faleminderit dielli im që ndriçon çdo ditë!

Je pense aussi à mes frères, avec qui j'ai très tôt partagé ce besoin d'expliquer le pourquoi et le comment, et avec certains desquels j'ai eu la chance de partager le quotidien du campus.

Merci de tout cœur à ma maman qui a fait de moi ce que je suis, qui m'a constamment soutenu et m'a donné sa confiance inconditionnelle et infrangible. Certes la théorie des nœuds m'est bien trop étrangère pour avoir pu prétendre poursuivre ta recherche interrompue, mais je suis très heureux d'avoir pu t'emboîter le pas dans un autre domaine des mathématiques.

Lausanne, le 10 avril 2018 Th. L.

Abstract

Extreme events are responsible for huge material damage and are costly in terms of their human and economic impacts. They strike all facets of modern society, such as physical infrastructure and insurance companies through environmental hazards, banking and finance through stock market crises, and the internet and communication systems through network and server overloads. It is thus of increasing importance to accurately assess the risk of extreme events in order to mitigate them. Extreme value theory is a statistical approach to extrapolation of probabilities beyond the range of the data, which provides a robust framework to learn from an often small number of recorded extreme events.

In this thesis, we consider a conditional approach to modelling extreme values that is more flexible than standard models for simultaneously extreme events. We explore the sub-asymptotic properties of this conditional approach and prove that in specific situations its finite-sample behaviour can differ significantly from its limit characterisation.

For modelling extremes in time series with short-range dependence, the standard peaks-over-threshold method relies on a pre-processing step that retains only a subset of observations exceeding a high threshold and can result in badly-biased estimates. This method focuses on the marginal distribution of the extremes and does not estimate temporal extremal dependence. We propose a new methodology to model time series extremes using Bayesian semiparametrics and allowing estimation of functionals of clusters of extremes. We apply our methodology to model river flow data in England and improve flood risk assessment by explicitly describing extremal dependence in time, using information from all exceedances of a high threshold.

We develop two new bivariate models which are based on the conditional tail approach, and use all observations having at least one extreme component in our inference procedure, thus extracting more information from the data than existing approaches. We compare the efficiency of these models in a simulation study and discuss generalisations to higher-dimensional setups.

Existing models for extremes of Markov chains generally rely on a strong assumption of asymptotic dependence at all lags and separately consider marginal and joint features. We introduce a more flexible model and show how Bayesian semiparametrics can provide

a suitable framework allowing simultaneous inference for the margins and the extremal dependence structure, yielding efficient risk estimates and a reliable assessment of uncertainty.

Key words: Asymptotic independence; Clustering; Conditional extremes; Dirichlet process mixture; Extreme value theory; Flood; Hierarchical Bayesian semiparametric inference; Markov chain Monte Carlo; Penultimate analysis; Risk assessment; River flow; Threshold-based extremal index; Time series extremes

Résumé

Les événements extrêmes causent de très importants dégâts matériels et ont un coûteux impact tant humain qu'économique. Ils touchent toutes les facettes de la société moderne, telles les infrastructures matérielles et les compagnies d'assurances par les catastrophes naturelles, le secteur bancaire et le monde de la finance par les crises boursières, ainsi que l'internet et les systèmes de communication par les surcharges des réseaux et des serveurs. Il devient donc toujours plus important d'évaluer le risque d'événements extrêmes avec précision, de manière à pouvoir les anticiper. La théorie des valeurs extrêmes est une approche statistique permettant l'extrapolation de probabilités par-delà les données observées, et fournissant un cadre robuste pour tirer parti d'un nombre souvent restreint d'événements extrêmes enregistrés.

Dans cette thèse, nous considérons la modélisation des valeurs extrêmes par une approche conditionnelle, plus flexible que les modèles standard portant sur les événements extrêmes simultanés. Nous explorons les propriétés sous-asymptotiques de cette approche conditionnelle et démontrons qu'un échantillon fini peut, dans certaines situations, se comporter très différemment de sa caractérisation asymptotique.

Pour la modélisation des extrêmes de séries chronologiques avec de la dépendance à court terme, la méthode usuelle des « pics au-dessus d'un seuil » repose sur un pré-traitement des dépassements d'un seuil élevé qui n'en retient qu'un sous-ensemble et peut conduire à des estimations particulièrement biaisées. Cette méthode se concentre sur la distribution marginale des extrêmes et ne permet pas d'estimer la dépendance temporelle extrémale. Nous développons une nouvelle méthodologie de modélisation des extrêmes de séries chronologiques utilisant une approche bayésienne semi-paramétrique, et permettant l'estimation de fonctionnelles de grappes de valeurs extrêmes. Nous appliquons notre méthodologie à la modélisation du débit de rivières en Angleterre et améliorons l'évaluation du risque d'inondation en décrivant explicitement la dépendance temporelle extrémale et en exploitant l'information de tous les dépassements d'un seuil élevé.

Nous développons deux nouveaux modèles bivariés basés sur l'approche conditionnelle pour les extrêmes et utilisant pour l'inférence toutes les observations possédant au moins une composante extrême, permettant ainsi d'extraire plus d'information des données que

les approches existantes. Nous comparons l'efficience de ces modèles par des simulations et discutons la généralisation à des dimensions supérieures.

Les modèles de chaînes de Markov pour les extrêmes reposent généralement sur une hypothèse forte de dépendance asymptotique entre tous les événements extrêmes de la chaîne et considèrent les caractéristiques marginale et conjointe de manière distincte. Nous présentons un modèle plus flexible et montrons comment une approche bayésienne semi-paramétrique peut fournir un cadre adapté à l'inférence simultanée des structures extrémales marginale et conjointe, produisant des estimations efficientes du risque et une évaluation fiable de l'incertitude.

Mots-clefs : Analyse de risque ; Analyse sous-asymptotique ; Débit de rivière ; Extrêmes conditionnels ; Extrêmes de séries chronologiques ; Indépendence asymptotique ; Indice extrémal sous-asymptotique ; Inférence semi-paramétrique bayésienne hiérarchique ; Inondation ; Mélange de processus de Dirichlet ; Méthode de Monte Carlo par chaînes de Markov ; Mise en grappes ; Théorie des valeurs extrêmes

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

1.1 Motivation

In Switzerland, floods are the most damaging natural hazards in terms of the costs incurred, as is reported for the year 2007 by Hilker *et al.* (2008). Floods have also caused more than a hundred deaths since the mid-20th century, but the number of fatalities dropped in the last few decades (Badoux *et al.*, 2016). The last three extremely damaging events which happened in Switzerland were in October 2000, when Canton Valais experienced a major flood in which 16 people died and material damage amounted to CHF 600 million; in August 2005, a widespread flood caused six deaths and cost more than CHF 3 billion; in August 2007, the River Aare and its tributaries impacted regions in the Plateau, the Jura and the Chablais, with damages of more than CHF 600 million (Spicher, 2017; Hilker *et al.*, 2009; Pfister, 2009). Figure 1.1 shows the inflation-adjusted costs due to floods, debris flows, landslides and rock falls, in Switzerland during the period 1972–2016 (Swiss Federal Institute for Forest, Snow and Landscape Research WSL, 2016); more than 90% of the costs over this period are due to floods and debris flows.

In the UK, the cost of flood damage exceeds £1 billion per year and is expected to rise significantly in the coming decades (Bennett and Hartwell-Naguib, 2014), and 13 deaths due to flooding have been reported since 2000 (Gummer and Leasom, 2016). Government spending on maintenance and extension of flood defences represents more than £600 million a year (Gummer and Leasom, 2016).

Switzerland and the UK are two examples of countries experiencing recurrent flood events, but many other countries in the world endure much more severe flood damage. In order to assign government spending adequately and to efficiently reduce the impacts of large-scale flooding events, it is thus key to be able to model and predict the likelihood of these extreme events.

Carbon dioxide and other pollutants released by human activity trap huge amounts of energy in the atmosphere. Owing to this increasing surplus of energy, we can expect environmental damage of greater magnitude and of wider extent than in the previous century. New

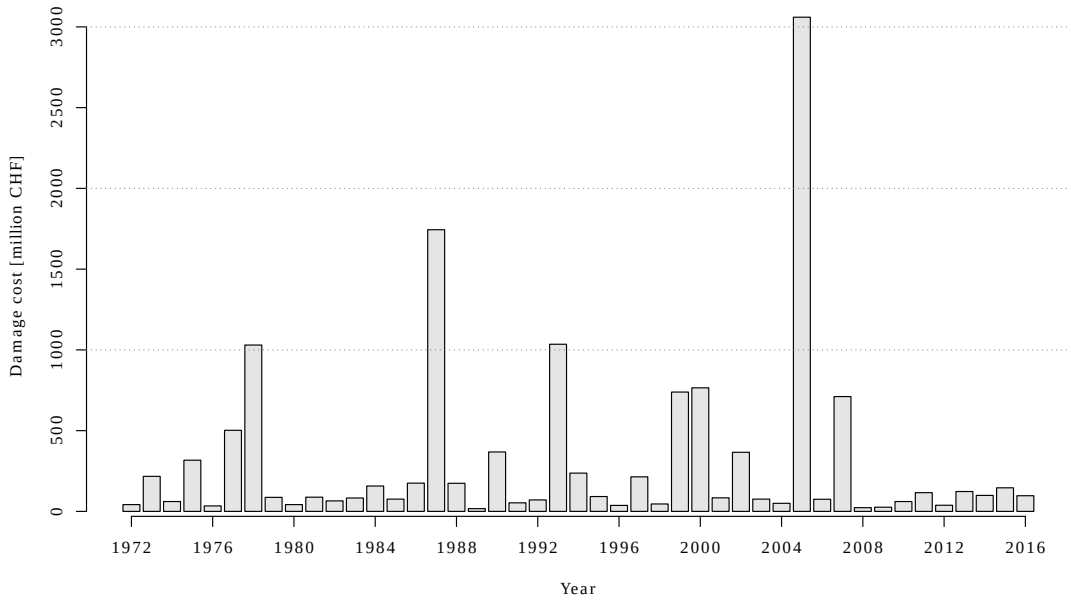

Figure 1.1 – Costs (million CHF, inflation-adjusted) of damage caused by floods, debris flows, landslides and rock falls in Switzerland from 1972 to 2016.

homes tend to be built closer to coastlines and flood areas, as safer points are already occupied by existing buildings. The combined effect of demographics and climate change means that more people and goods are at risk of natural catastrophes. Given these changes, it becomes more and more critical to measure the risk of these extreme events in order to mitigate them.

The theory of extreme values can help better understand and estimate the frequency and magnitude of extreme events and provides tools for assessing the risk of events that have yet never been measured. Statistics of extreme events gives the necessary insight for making appropriate decisions for risk mitigation.

Suppose we measure some phenomenon at regular intervals until we have n stationary observations with finite variance, e.g., a river flow recorded every day in winter over a number of years. The central limit theorem would then suggest, for n sufficiently large, that the distribution of the mean of the n observations approaches a Gaussian distribution at a $n^{-1/2}$ rate. This theorem justifies the use of the Gaussian distribution in many situations, as diverse as modelling random errors, constructing confidence intervals, and making prediction. The Gaussian distribution is so widely used that it becomes a default distribution for many practitioners. When the tail of a process is of interest, rather than its mean, using the Gaussian distribution can yield underestimation of joint tail risks, as was experienced in the banking sector in 2007 (Hartmann *et al.*, 2004). The statistical theory of extreme events provides an analogue to the central limit theorem, which establishes the limit distribution, as $n \to \infty$, of the maximum of n independent observations. For the estimation of joint tail risks, it is thus

appropriate to adopt an extreme value approach, since averages follow a very different process from that of extremes.

After the first results of the statistics of extremes were established (Fisher and Tippett, 1928; von Mises, 1936; Gnedenko, 1943), applications of the theory appear in Gumbel and Goldstein (1964), where the authors consider two data sets with different dependence properties, namely the oldest age at death for females and males in Sweden, and river discharges at two gauges situated along the Ocmulgee River in Georgia (US). The study of life expectancy is still relevant today, and has recently benefitted from the contributions of Einmahl *et al.* (2017) and Rootzén and Zholud (2017), leading to different conclusions. River flooding has attracted much attention in the extreme value literature and has been applied to high-dimensional and spatial problems (Katz *et al.*, 2002; Keef *et al.*, 2009a,b; Asadi *et al.*, 2015).

Other environmental applications of extreme value models include extreme rainfall (Coles and Tawn, 1996a,b; Süveges and Davison, 2012; Huser and Davison, 2014; Sharkey and Tawn, 2017), extreme wind speeds (Coles and Walshaw, 1994; Fawcett and Walshaw, 2006a,b; Oesting *et al.*, 2017), wind storms (Coles, 1993; Northrop *et al.*, 2017), wave height and extreme sea surge (de Haan and de Ronde, 1998; Fawcett and Walshaw, 2007), heatwaves (Reich *et al.*, 2014; Winter and Tawn, 2016) and high concentrations of air pollutants (Smith, 1989; Heffernan and Tawn, 2004; Eastoe and Tawn, 2009).

Many applications have contributed to the improvement of risk assessment in the insurance and finance industries, in particular to dependence modelling of extreme stock market losses (Poon *et al.*, 2003), hedging strategies (Hilal *et al.*, 2011), portfolio risk assessment (Hilal *et al.*, 2014), estimation of value-at-risk and expected shortfall (Chavez-Demoulin *et al.*, 2005, 2014; Cai *et al.*, 2015), and insurance losses (Embrechts *et al.*, 1997; Rohrbeck *et al.*, 2018).

1.2 Thesis outline

The thesis is structured as follows: the following two chapters review existing models, the first being focused towards statistics of extreme values, and the second dealing with a Bayesian approach to nonparametric modelling. The next three chapters cover various developments and contributions of the thesis, and the last chapter discusses potential extensions and interesting future directions of research.

In Chapter 2, we review many methods developed in the literature of extreme value modelling, and we explain how these methods tackle the problem of extrapolating probabilities from moderately extreme sets, where data have been observed, to very extreme sets, typically beyond the range of the data, and corresponding to risk levels of interest. We give a brief introduction to univariate modelling of maxima and of excesses of a high threshold, followed by extreme value modelling for time series, which often exhibit short-range dependence at high levels. We then turn our attention to models for multivariate extreme events, in particular those which can capture a dependence strength weakening as we move further

into the joint tail. The conditional model for extremes (Heffernan and Tawn, 2004) has the ability of covering many existing non-parametric and parametric models for extremes while being parsimonious. Although the inference procedure advocated by Heffernan and Tawn is very simple, its efficiency has room for improvement, and in Chapter 5 we develop a method which tackles this issue. The conditional tail model is the main concern of this thesis and is developed along different directions in Chapters 4, 5, 6 and 7.

In Chapter 3, we introduce the Dirichlet process as an approach to nonparametric modelling in the Bayesian framework, and discuss extensions of it, namely Dirichlet process mixtures. Two broad approaches exist for fitting these mixtures, both of which are illustrated with specific algorithms. The code to fit these algorithms was developed and optimised for the purpose of this thesis. The respective performances of these algorithms in terms of mixing and computational cost are compared. An illustration with real data gives an insight into the performance and flexibility of these methods. One of these algorithms is used and extended in Chapter 5 and potential applications of Dirichlet process mixtures are explored in Chapter 7.

In Chapter 4, we delve into the subasymptotic properties of the conditional model introduced in Chapter 2. Penultimate analysis was conducted by Smith (1987) and Gomes (1984, 1994) in the context of univariate extreme values, but consideration of the slow convergence of the distribution of the maxima of a Gaussian distribution already appears in Fisher and Tippett (1928). For the conditional model for extremes, the bivariate Gaussian distribution given one of its margins is large is of particular interest, as its rate of convergence to the limit conditional distribution is notably slow. The penultimate analysis of the conditional tail model is helpful in simulations for evaluation purposes when using sample sizes similar to real data sets. In this context, penultimate approximations of the parameters of the conditional tail model, instead of their respective limit values, which can lie well away from the penultimate true values, can be used as benchmarks. We shall use these penultimate approximations in Chapter 6 to help evaluate the efficiency of one of the new approaches presented there.

In Chapter 5, we present a new approach to modelling the extremes of a stationary time series. In this context, both marginal and dependence features must be described. There are standard approaches to marginal modelling, but long- and short-range dependence of extremes may both appear. In applications, an assumption of long-range independence often seems reasonable, but short-range dependence, i.e., the clustering of extremes, needs attention. The extremal index $0 < \theta \leq 1$ is a natural limiting measure of clustering, but for wide classes of dependent processes, including all stationary Gaussian processes, it cannot distinguish dependent processes from independent processes with $\theta = 1$. Eastoe and Tawn (2012) exploit methods from multivariate extremes to treat the subasymptotic extremal dependence structure of stationary time series, covering both $0 < \theta < 1$ and $\theta = 1$, through the introduction of a threshold-based extremal index. Inference for their dependence models uses an inefficient stepwise procedure that has various weaknesses and has no reliable assessment of uncertainty. We overcome these issues using a Bayesian semiparametric approach. Simulations and the

analysis of a UK daily river flow time series show that the new approach provides improved efficiency for estimating properties of functionals of clusters.

In Chapter 6, we shall introduce a new constraint for the conditional tail model. Previously, based on the initial formulation of Heffernan and Tawn (2004), Keef *et al.* (2013) derive constraints on the model parameters, which we shall present in Section 2.4.3 of Chapter 2. These constraints help increase efficiency and remove inconsistency of probabilities extrapolated from the model. Another issue remains which these constraints do not address, namely the lack of identifiability of the model parameters in specific situations. We tackle this by adding a new constraint to the model when positive or negative association can reasonably be assumed. We also develop two models for fitting joint distributions for extremes based on the conditional tail model. This gives a coherent framework for fitting multivariate extremes with at least one component being extreme, whereas the method suggested by Heffernan and Tawn (2004) uses an incorrect likelihood function. Our method adds censored data to the fit and makes simultaneous inference for the extremal marginal and dependence features.

In Chapter 7, we return to the issue of fitting time series extremes of Chapter 5, and we propose a model to fit the extremes of first order Markov chains, so that combined inference can be made on extremal marginal and joint distributions, thus giving a full account of uncertainty of the parameter estimates. We develop a semiparametric Bayesian approach to avoid the strong assumptions needed in Chapter 6, and to give an account of uncertainty on the estimated model and on probabilities derived from the model.

We conclude with a discussion of possible extensions of the results presented in the thesis, and in particular potential future work on a generalisation of the approach of Chapter 5 to the spatial context, which can be of great interest for modelling asymptotically independent data, e.g., environmental data, which generally display decreasing strength in extremal dependence between increasingly distant locations.

2 Modelling extremes

2.1 Extrapolation principle

2.1.1 Univariate setting

The study of extreme values is concerned with predicting the likelihood of potentially damaging events at levels never previously recorded. Statistical methods for estimating the probabilities of occurrence of these events help their users to learn from the history of the most harmful events in order to extend information from these events of moderate intensity towards others of higher magnitude for which information is incomplete or absent. More formally, we need to carefully choose a non-empty moderately extreme set $\mathscr{A}$ on which we can rely for extrapolation to a more extreme set $\mathscr{B}$ containing few or no data points.

In the univariate setting with $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, this extrapolation paradigm appears clearly. The founding limit theorem in extreme value theory (Fisher and Tippett, 1928; Gnedenko, 1943) states

Theorem 2.1 (Fisher–Tippett–Gnedenko)
Let $X_1, \ldots, X_n$, $n \in \mathbb{N}$, be independent and identically distributed random variables. Consider the sequence of maxima $M_n = \max\{X_1, \ldots, X_n\}$, $n \in \mathbb{N}$. If this sequence can be linearly renormalised as $(M_n - a_n)/b_n$ by sequences of locations (a_n) and scales $(b_n) > 0$ in order to converge to a non-degenerate distribution $G(x)$, then

$$G(x) = \exp\left\{-\left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right\}, \quad x \in \mathbb{R}, \tag{2.1}$$

with $c_+ = \max\{c, 0\}$, location parameter $\mu \in \mathbb{R}$, shape parameter $\xi \in \mathbb{R}$ and scale parameter $\sigma > 0$; the case with $\xi = 0$ is interpreted as the limit as $\xi \to 0$.

This powerful result suggests modelling the distribution of the maxima of observations for finite n using the single parametric family of distributions $G(x) = G(x; \mu, \sigma, \xi)$ in (2.1) as an approximation of $F^n(x)$. When the distribution F converges to G in the terms of Theorem 2.1, we say that F is in the domain of attraction of G and we denote it by $\mathscr{D}(G)$.

The extrapolation strategy from set $\mathscr{A}$ to set $\mathscr{B}$ underlying inference of extreme values and outlined at the beginning of this section is better explained through the point process representation for extremes (Hsing *et al.*, 1988). Consider the point process with points

$$P_n = \left\{ \left(\frac{i}{n+1}, \frac{X_i - a_n}{b_n} \right) : i = 1, \ldots, n \right\}. \tag{2.2}$$

From Theorem 2.1 we have

$$\Pr\left(\frac{M_n - a_n}{b_n} \leq x \right) = \{F(b_n x + a_n)\}^n \to G(x), \quad n \to \infty \tag{2.3}$$

with $G(x)$ non-degenerate, or equivalently on the log scale,

$$n\{1 - F(b_n x + a_n)\} \to -\log G(x), \quad n \to \infty. \tag{2.4}$$

We get the intensity of the point process (2.2) if we consider the number of excesses of $b_n x + a_n$, as in Leadbetter (1976). More precisely, we consider $N_n(x) = \sum_{i=1}^n \mathbb{1}(X_i > b_n x + a_n)$, whose distribution is $\text{Bin}\{n, 1 - F(b_n x + a_n)\}$. The Poisson limit and (2.4) yield

$$N_n(x) \to N(x) \sim \text{Pois}\{-\log G(x)\}.$$

We get the sizes of the excesses of $a_n x + b_n$, i.e., the sizes of the marks of (2.2), through the limit

$$\begin{aligned} \Pr(X > b_n x + a_n \mid X > a_n) &= \frac{1 - F(b_n x + a_n)}{1 - F(a_n)} \\ &= \frac{n\{1 - F(b_n x + a_n)\}}{n\{1 - F(a_n)\}} \\ &\to \frac{\log G(x)}{\log G(0)} \\ &= \left(1 + \xi \frac{x}{\sigma - \xi\mu} \right)_+^{-1/\xi}, \quad n \to \infty. \end{aligned} \tag{2.5}$$

As $n \to \infty$, $a_n \to x_F$, the upper endpoint of F, suggesting the approximation using the generalised Pareto distribution

$$\Pr(X > x + u \mid X > u) \approx (1 + \xi x / \sigma_u)_+^{-1/\xi}, \quad \text{for large } u, \tag{2.6}$$

where $\sigma_u > 0$ is a scale parameter that is a function of the threshold u.

We can now depict how extrapolation underlies inference for very extreme sets, as we can derive $\Pr(X \in \mathscr{B})$ from $\Pr(X \in \mathscr{A})$ and (2.5),

$$\Pr(X \in \mathscr{B}) = \left(1 + \xi \frac{v - u}{\sigma_u} \right)_+^{-1/\xi} \Pr(X \in \mathscr{A}),$$

with $\mathscr{A} = [u, \infty)$ and $\mathscr{B} = [v, \infty)$, $u < v$.

2.1.2 Bivariate and low-dimensional setting

Coles and Tawn (1994) consider structure variables of interest in practical design problems, reducing the dimension of the problem so that univariate methods of Section 2.1.1 can easily be applied. If $\mathscr{B}$ is an extreme set of interest and $\boldsymbol{X}$ are observed data, the structure variable $S(\boldsymbol{X})$ transforms the original multivariate problem into the much simpler problem of estimating

$$\Pr(\boldsymbol{X} \in \mathscr{B}) = \Pr\{S(\boldsymbol{X}) > v\}.$$

The authors give several examples of design settings where specific structure variables are of interest. For example, offshore platform engineers are mainly interested in the force of waves X_1 and winds X_2, and a typical structure variable in this context would be $S(\boldsymbol{X}) = S(X_1, X_2) = a_1 X_1^2 + a_2 X_2^2$; in rainfall studies for which we have gauges at sites $X_1, \ldots, X_d$, a quantity of interest is the cumulative rainfall measured over a whole region $\sum_{j=1}^{d} w_j X_j$, with $w_1, \ldots, w_d$ being weights associated with the sites. This last structure variable could also be interpreted in the financial context as a portfolio of d assets where the X_j represent the negative returns and for which we are interested in potential extreme losses. The main interest of such an approach is that it only needs a univariate fit without needing to consider the dependence between variables. On the other hand, the structure variable approach does not guarantee coherence between probabilities extrapolated from different structure variables based on the same data; it hides the connections between variables, and is unable to account for deterministic and potentially abrupt changes in the structure of S beyond the data.

Consideration of extremes of bivariate and higher-dimensional data entails defining an ordering in $\mathbb{R}^d$, $d \geq 2$. Barnett (1976) lists various orderings, a couple of which have been considered in the literature. The structure variable is one. Standard approaches in extreme value problems use componentwise maxima, which can combine either simultaneous events, e.g., recorded on the same day, or events recorded at different time points. Stephenson and Tawn (2005) raise this issue and suggest distinguishing the contributions depending on which of these two types the observed maxima correspond to.

The univariate result in Theorem 2.1 can be generalised to the d-variate setting by considering componentwise maxima of n independent replicates $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,d})$, $i = 1, \ldots, n$. The multivariate componentwise maximum is defined as

$$\boldsymbol{M}_n = \left(M_{n,1}, \ldots, M_{n,d}\right), \quad M_{n,j} = \max\left\{X_{1,j}, \ldots, X_{n,j}\right\}, \quad j = 1, \ldots, d.$$

In the following theorem, operations must be interpreted componentwise.

Definition 2.1 (Multivariate domain of attraction)
The multivariate distribution F of the d-variate independent replicates $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is in the multivariate domain of attraction of the multivariate extreme value distribution G if there exist normalising vectors of constants $\boldsymbol{a}_n$ and $\boldsymbol{b}_n > \boldsymbol{0}$ such that

$$\Pr\left(\frac{\boldsymbol{M}_n - \boldsymbol{a}_n}{\boldsymbol{b}_n} \leq \boldsymbol{x}\right) \to G(\boldsymbol{x}), \quad n \to \infty,$$

where G is non-degenerate in each margin.

Copulae (Joe, 2014) are a standard approach to modelling dependence between two random variables X_1 and X_2. Copulae are multivariate distributions with uniform marginal distributions, and any continuous multivariate distribution has its copula equivalent through an appropriate marginal change of scale (Sklar, 1959). The study of bivariate extremes embraces copulae but considers margins on scales that better reveal the structure of dependence at extreme levels. Figure 2.1 illustrates how modifying the marginal scale can display various features in the extremes, and shows that uniform margins are poor at revealing features of the extremes. In this figure and later in the text, we use X^{G}, X^{U}, X^{F} and X^{L} to refer to the random variable X on Gaussian, uniform, Fréchet and Laplace scales.

The copula of bivariate maxima using Fréchet margins,

$$C^{\mathrm{F}}(x,y) = C\left(e^{-1/x}, e^{-1/y}\right) = \exp\{-V(x,y)\}, \quad x,y>0,$$

is identified by a function $V(x,y)$ known as the exponent measure which satisfies

$$V(x,\infty) = x^{-1}, \quad V(\infty,y) = y^{-1}, \quad x,y>0,$$

and a homogeneity property of order -1, i.e., $V(tx,ty) = t^{-1}V(x,y)$, $t>0$. The exponent measure can be characterised by

$$V(x,y) = \int_0^1 \max\{\omega x^{-1}, (1-\omega)y^{-1}\}\,dH(\omega), \quad x,y>0, \tag{2.7}$$

where $H(\cdot)$ is a non-negative measure, termed the angular distribution, that satisfies the moment constraints

$$\int_0^1 \omega\,dH(\omega) = \int_0^1 (1-\omega)\,dH(\omega) = 1.$$

On Fréchet margins, the copula can be written as

$$C^{\mathrm{F}}(x,y) = \exp\{-V(x,y)\} = \exp\left\{-\left(\frac{1}{x}+\frac{1}{y}\right)A\left(\frac{x}{x+y}\right)\right\}, \quad x,y>0, \tag{2.8}$$

with $A(\omega)$, $\omega\in[0,1]$, termed Pickands' dependence function (Pickands, 1981), which is convex and satisfies $A(0) = A(1) = 1$ and $\max\{\omega, 1-\omega\} \le A(\omega) \le 1$. A central quantity derived from Pickands' function is the coefficient of extremal dependence $\theta = 2A(1/2)$, $\theta\in[1,2]$, measuring the effective number of independent variables, since

$$C^{\mathrm{F}}(x,x) = \exp\{-V(x,x)\} = \exp\{-V(1,1)/x\} = \exp\{-2A(1/2)/x\} = \{\exp(-1/x)\}^{\theta},$$

using the homogeneity of $V(\cdot,\cdot)$ and (2.8). The boundary cases $A(\omega)\equiv 1$ and $A(\omega) = \max\{\omega, 1-\omega\}$ coincide with independence and complete dependence.

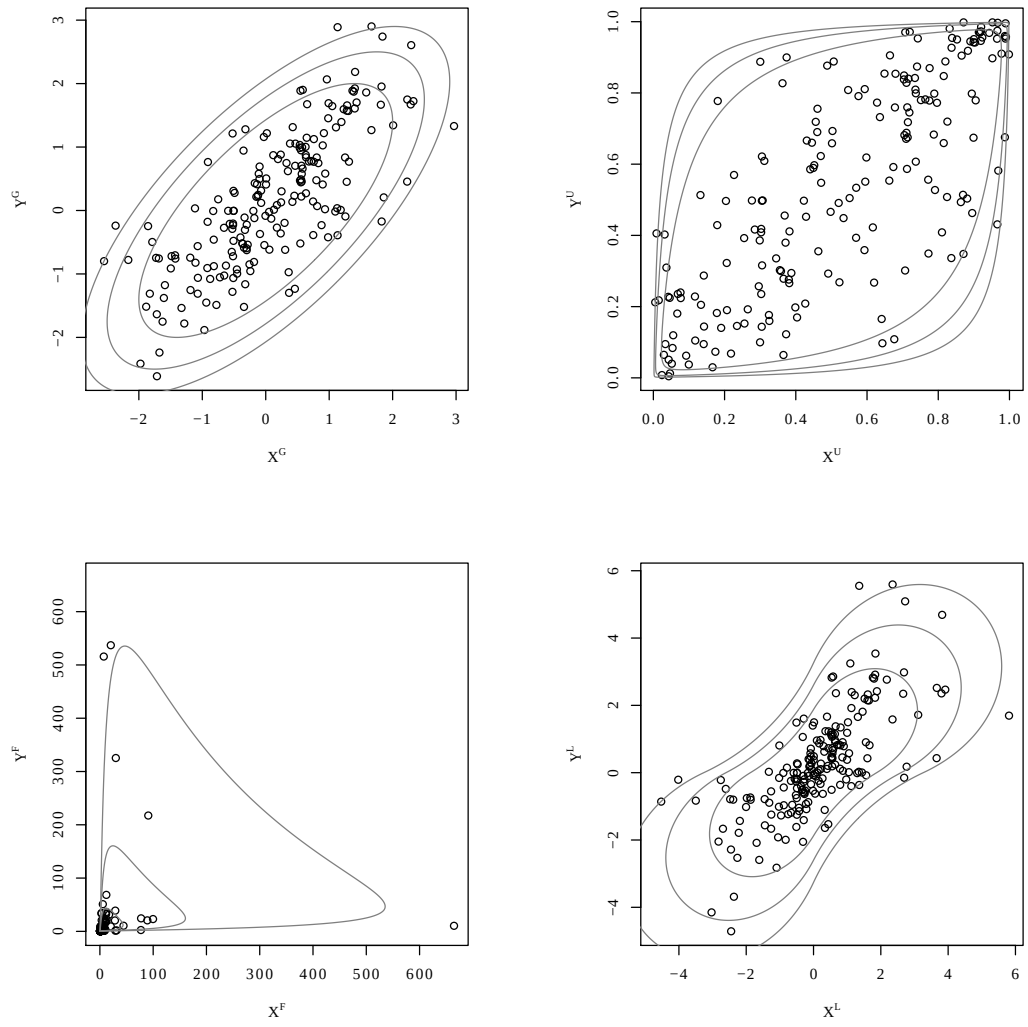

Figure 2.1 – Data from a bivariate Gaussian copula on four different marginal scales, from left to right and top to bottom: Gaussian, uniform, Fréchet and Laplace scales. Grey lines are density contours.

The three different characterisations of multivariate extreme copulae through the exponent measure V, the angular distribution H and Pickands' function A give different perspectives and each has contributed to the development of models for extremes. The literature contains a vast choice of copula-type models for componentwise maxima (Smith *et al.*, 1990), with non-differentiable models (Tiago de Oliveira, 1980), the bivariate logistic (Gumbel, 1960), its asymmetric version and the asymmetric mixed model (Tawn, 1988), the Coles and Tawn (1991) Dirichlet model, the Hüsler–Reiss model (Hüsler and Reiss, 1989) and many others. This abundance of models is explained by the fact that no closed-form distribution characterises the limiting probabilities of extreme sets in two and more dimensions.

2.2 Extremes in time series

Consider a stationary time series (X_t) with marginal distribution F, for which we want to derive probabilities of events that are more extreme than the observations recorded so far. The univariate framework of Section 2.1.1 is directly applicable to this problem if we can reasonably assume the X_t to be independent. In practice, this is rarely the case, and departures from independence have been explored in the literature.

Leadbetter (1974) showed that a weak mixing condition suffices for Theorem 2.1 to hold for the stationary sequence (X_t).

Definition 2.2 ($D(u_n)$ condition)
The stationary sequence (X_t) satisfies the $D(u_n)$ condition for a sequence (u_n), if for each n, l, and each choice of sequences $1 \le i_1 < \cdots < i_p$ and $j_1 < \cdots < j_q \le n$ with $j_1 - i_p \ge l$, we have

$$|F_{i_1,\ldots,i_p,j_1,\ldots,j_q}(u_n,\ldots,u_n) - F_{i_1,\ldots,i_p}(u_n,\ldots,u_n)F_{j_1,\ldots,j_q}(u_n,\ldots,u_n)| \le \varepsilon_{n,l},$$

with $\varepsilon_{n,l} \to 0$ as $n \to \infty$ and $l = l_n = o(n)$.

The $D(u_n)$ condition ensures that long-range dependence remains small, but does not prevent clustering of extreme values. The independence assumption in Theorem 2.1 can be replaced by the much weaker assumption that the X_t satisfy the $D(u_n)$ condition for all sequences (u_n) with $u_n = b_n x + a_n$.

Consider the series $(X_t^\star)$ of independent variables with the same marginal distribution F, and denote $M_n^\star = \max\{X_1^\star,\ldots,X_n^\star\}$. A strong link connects the limiting distribution of $M_n^\star$ with that of M_n under the $D(u_n)$ condition (Leadbetter, 1983).

Theorem 2.2 (Extremal index)
Let the $D(u_n)$ condition hold for the stationary sequence (X_t). Then there exist sequences $(a_n^\star)$, $(b_n^\star) > 0$, and a non-degenerate distribution function $G^\star$ such that

$$\Pr\left(\frac{M_n^\star - a_n^\star}{b_n^\star} \le x\right) \to G^\star(x), \quad n \to \infty,$$

if and only if there exist sequences (a_n), $(b_n) > 0$, *and a non-degenerate distribution function* G *such that*

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

In addition, we have $a_n^\star = a_n$, $b_n^\star = b_n$ *and* $G^\star(x) = \{G(x)\}^\theta$, *with* $\theta \in [0,1]$ *named the extremal index.*

As this result shows, the extremal index plays a key role when modelling time series with short-range dependence, and it will be further detailed in Section 2.3. Independent times series have $\theta = 1$, but this case appears also in time series with some dependence, under a condition on short-range dependence of (X_t) (Leadbetter, 1974; Leadbetter *et al.*, 1983, Chap. 3).

Definition 2.3 ($D'(u_n)$ condition)
The stationary sequence (X_t) *staisfies the* $D'(u_n)$ *condition for a sequence* (u_n) *if*

$$\lim_{k\to\infty} \limsup_{n\to\infty} \left\{ n \sum_{t=2}^{\lfloor n/k \rfloor} \Pr(X_1 > u_n, X_t > u_n) \right\} = 0.$$

While the $D(u_n)$ condition ensures that long-range dependence decreases at an appropriate rate, the $D'(u_n)$ condition ensures that short-range dependence remains sufficiently low, by imposing that two observations have a small probability of exceeding u_n in any block of length n.

In practice, the $D(u_n)$ and $D'(u_n)$ conditions are hard to verify, but they support the approach of making inference on time series even when weak dependence is observed at finite levels. A typical approach is maximum likelihood estimation based on the conditional distribution (2.6). Other methods for estimating the generalised Pareto distribution in this context are reviewed in Kotz and Nadarajah (2000, Chap. 1). When short-range dependence in (X_t) cannot be ignored, inference typically involves a pre-processing step that retains only observations that can be considered independent (Davison and Smith, 1990), or requires inflation of standard errors for estimates derived from all exceedances of a high threshold (Fawcett and Walshaw, 2007, 2012).

The point process theory related to clustered stationary time series (X_t) (Hsing *et al.*, 1988; Leadbetter, 1995) describes the limit process of such series as a compound Poisson process (Daley and Vere-Jones, 2003, Chap. 2), where the marks are the cluster sizes. Assuming $1 - F(u_n) \sim \tau/n$, $n \to \infty$, $\tau > 0$ and asymptotic cluster size distribution π, this compound Poisson process has intensity $\theta\tau$ and mark size distribution π. This theory also establishes that the distribution of cluster maxima is asymptotically identical to the marginal distribution of all excesses (Pickands, 1975). Several declustering schemes exist, of which the most popular are based on blocks (Leadbetter *et al.*, 1989) and on runs (Smith, 1989). The former partitions the series into n_B blocks of the same length r_B and picks the blocks with at least one exceedance

of u_n as clusters; the latter picks clusters of exceedances which are separated by at least r_{R} non-exceedances of u_n. Having defined clusters, peaks over threshold analysis is based on the maximum value in each of them. Both methods involve the choice of an arbitrary quantity r_{B} or r_{R}, as well as the threshold u_n, which can affect subsequent inferences. An automatic selection procedure is suggested by Ferro and Segers (2003) for a given u_n.

Fawcett and Walshaw (2007, 2012) showed, through extensive simulations and examples, that the pre-processing of dependent time series may yield badly-biased estimates of tail probabilities and return levels, which we shall define shortly. The authors suggest using all exceedances of a threshold and computing uncertainty by inflating standard errors with a sandwich method. Eastoe and Tawn (2012) take another approach, based on the idea that the distributions of cluster maxima and of marginal exceedances coincide only in the limit. Their model is a modified generalised Pareto distribution that better reflects the distribution of cluster maxima at subasymptotic levels, with the additional benefit of using the information contained in all exceedances of a threshold; this will be further developed in Chapter 5.

An important aspect in the field of extreme values is the communication of conclusions, for example about an estimate of $\Pr(\boldsymbol{X} \in \mathscr{B})$, which may be a very small quantity, in a way that can be grasped by common sense and can serve as a basis for decisions. Return periods are defined such that risk assessment is expressed in terms of a time span instead of tiny probabilities. For a stationary series, we say that an extreme event of size x_n has a return period of n years if the probability of experiencing an event of size exceeding x_n in a year is $1/n$. The event size x_n is called the n-year return level.

All the methods described so far in this section deal with the marginal distribution of X_t but do not consider modelling the joint distribution in time. This is of particular interest when deriving functionals of extreme events, such as the duration of extreme events, e.g., the duration of drought (Winter and Tawn, 2016), the cumulated intensity of an event, e.g., the amount of rain over a period of extreme rainfall, the rth-largest statistic of a cluster, and many others (Yun, 2000; Segers, 2003).

A natural approach to fitting the joint distribution of a series is to assume a Markov property. Smith *et al.* (1997) were the first to consider this type of modelling for extremes, with a likelihood of the form

$$\ell(x_1, \ldots, x_n; \phi_1, \phi_2) = f(x_1; \phi_1) \prod_{t=2}^{n} f(x_t \mid x_{t-1}; \phi_1, \phi_2),$$

where ϕ_1 and ϕ_2 denote the parameters of the marginal and joint distributions respectively, $f(\cdot)$ is the marginal density and $f(\cdot \mid x)$ is the conditional density given $X = x$. They use the alternative formulation

$$\frac{\prod_{t=2}^{n} f(x_t, x_{t-1}; \phi_1, \phi_2)}{\prod_{t=2}^{n-1} f(x_t; \phi_1)},$$

where $f(\cdot,\cdot)$ is the bivariate joint density, so that models of Section 2.1.2 can be used for the numerator. As we shall describe in the next section, these models can be very restrictive, and a complementary approach was developed by Bortot and Tawn (1998). This formulation was used, for example, in a study on wind speed data using a Bayesian framework (Fawcett and Walshaw, 2006b). In the same vein, Sisson and Coles (2003) use a copula model (Coles and Pauli, 2002) and Bayesian fitting procedures to give an account of the uncertainty in the estimates. Bortot and Gaetan (2014) develop a model in which the observations are independent given a latent process, along similar lines as Sang and Gelfand (2009). They present processes that have a generalised Pareto marginal distribution and that deal with temporal dependence but are restricted to cases with positive shape, i.e., $\xi > 0$ in (2.6).

For reviews of the modelling of time series extremes, see Chavez-Demoulin and Davison (2012) or Beirlant *et al.* (2004, Chap. 10).

2.3 Modelling asymptotic independence

2.3.1 Classification of limit distributions

The copula models for componentwise maxima of Section 2.1.2 yield different descriptions of the data at finite levels, but they all rely on a regular variation condition (Resnick, 1987, Chap. 5; Basrak *et al.*, 2002; Resnick, 2007, Chap. 6) on the cone $[0,\infty]^2 \setminus \{\mathbf{0}\}$, which reduces the scope of dependence structures that can be considered. Informally, in our problem of extrapolation of probabilities from $\mathscr{A}$ to $\mathscr{B}$, making this assumption when it is not appropriate would typically yield an overestimate of $\Pr(\boldsymbol{X} \in \mathscr{B})$.

In order to distinguish broad classes of extremal dependence structures between two random variables X and Y, or equivalently between X^{U} and Y^{U} on the uniform scale, we define

$$\chi = \lim_{u\to 1} \chi(u) = \lim_{u\to 1} \Pr(Y^{\mathrm{U}} > u \mid X^{\mathrm{U}} > u) = \lim_{u\to 1} \Pr(X^{\mathrm{U}} > u \mid Y^{\mathrm{U}} > u), \tag{2.9}$$

with $\chi \in [0,1]$. When $\chi > 0$, X and Y are termed asymptotically dependent; when $\chi = 0$, X and Y are said to be asymptotically independent. The models for maxima in Section 2.1.2 correspond to the former case, but the latter case is often met in practice. An example of an asymptotically independent distribution is the Gaussian bivariate distribution which, with correlation $\rho < 1$, shows dependence at any finite level, but has $\chi = 0$ (Sibuya, 1960). In the asymptotic dependence case, the measure $\chi \in (0,1]$ gives finer detail about extremal dependence (Coles *et al.*, 1999) than in the asymptotic independence case, where χ can take a single value. An equivalent definition of χ uses a copula formulation, and is the limit as $u \to 1$ of

$$\chi(u) = 2 - \frac{\log \Pr\left(X^{\mathrm{U}} < u, Y^{\mathrm{U}} < u\right)}{\log \Pr\left(X^{\mathrm{U}} < u\right)}, \quad u \in (0,1). \tag{2.10}$$

To complement the measure χ and provide more information in cases of asymptotic independence, Coles *et al.* (1999), by analogy with (2.10), define

$$\overline{\chi} = \lim_{u \to 1} \overline{\chi}(u) = \lim_{u \to 1} \frac{2 \log \Pr(X^{\mathrm{U}} > u)}{\log \Pr(X^{\mathrm{U}} > u, Y^{\mathrm{U}} > u)} - 1,$$

with $\overline{\chi} \in [-1, 1]$. When $\chi = 0$, $-1 \leq \overline{\chi} < 1$ describes the strength of extremal dependence in the asymptotic independence class, with $\overline{\chi} = 0$, $\overline{\chi} < 0$ and $\overline{\chi} > 0$ corresponding respectively to complete independence, negative association and positive association; when $\overline{\chi} = 1$, we are in the case of asymptotic dependence, and we can use $\chi > 0$ as a measure of the strength of extremal dependence in this class.

The extremal index θ, introduced in Section 2.2 in the context of time series, also indicates whether we are in a situation of asymptotic dependence, with $\theta \in (0, 1)$, or asymptotic independence, with $\theta = 1$. In terms of the clustering of extremes, a constructive definition of the extremal index is

$$\theta = \lim_{u \to 1} \Pr(X_1^{\mathrm{U}} \leq u, \ldots, X_m^{\mathrm{U}} \leq u \mid X_0^{\mathrm{U}} > u), \tag{2.11}$$

with $m \to \infty$ as $u \to 1$ appropriately (O'Brien, 1987). The reciprocal of the probability (2.11) counts the proportion of excesses with respect to the number of clusters; it can be interpreted as the asymptotic mean cluster size. In this perspective, asymptotic independence corresponds to no clusters in the limit, whereas asymptotic dependence means some level of clustering, and the limiting cluster size is θ^{-1}. Chapter 5 gives more details about the extremal index, its sub-asymptotic properties and methods of estimation.

We conclude with an illustration of a use of the extremal measures reviewed in this section. We introduce an example involving financial data, where estimation of χ and $\overline{\chi}$ is of great importance, as standard dependence diagnostics fail to describe the degree of dependence at an asymptotic level. We consider Roche and Novartis, two companies in the pharmaceutical sector and listed on the Zurich stock exchange. We compute the returns from their stocks, extracted from the daily closing prices from 1 November 2000 to 31 October 2017. The data were downloaded from `finance.yahoo.com`, and missing data are filled using closing prices from the previous day. Roche prices at the beginning of the series are incorrect and need to be multiplied by 100. Figure 2.2 shows a strong link between the prices of the two companies, and the scatterplot of the returns confirms this. This link also appears in the estimated correlation of about 0.5.

In a risk assessment perspective, we are interested in the joint behaviour of these stocks when one of them faces severe price drops, i.e., large negative returns. After transformation to the uniform scale, we compute the empirical estimates corresponding to $\chi(u)$ and $\overline{\chi}(u)$ for $u \in [0.8, 0.99]$ on the negative returns and their respective 95% block bootstrap confidence intervals. The results are shown in Figure 2.3. The estimates of $\overline{\chi}(u)$ show evidence in favour of $\overline{\chi} \in (0, 1)$, i.e., asymptotic independence with positive association. This is confirmed with the confidence intervals for $\chi(u)$, which suggest that $\chi = 0$ is reasonable .

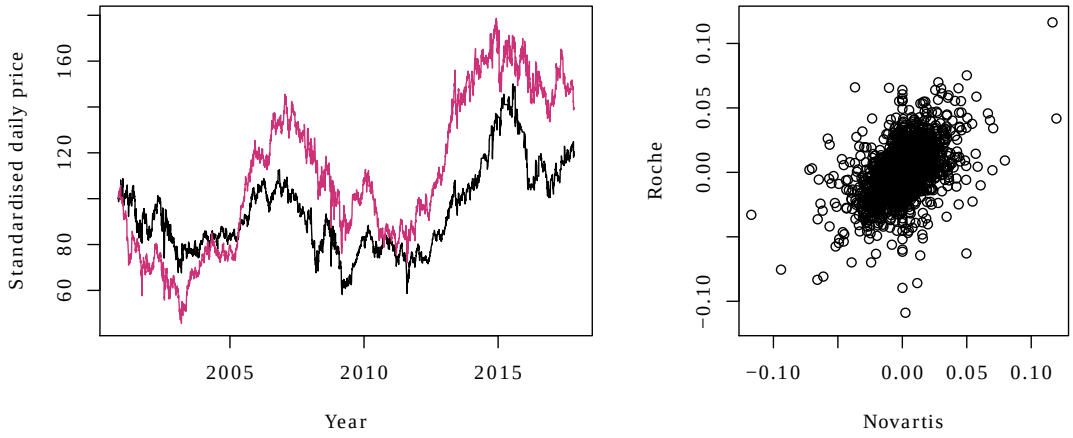

Figure 2.2 – Novartis and Roche stock prices. Left panel: daily prices of Novartis (–) and Roche (–) shares, standardised to start at 100 on the 1st of November 2000 for better interpretation. Right panel: daily returns, using the daily share prices of the two companies.

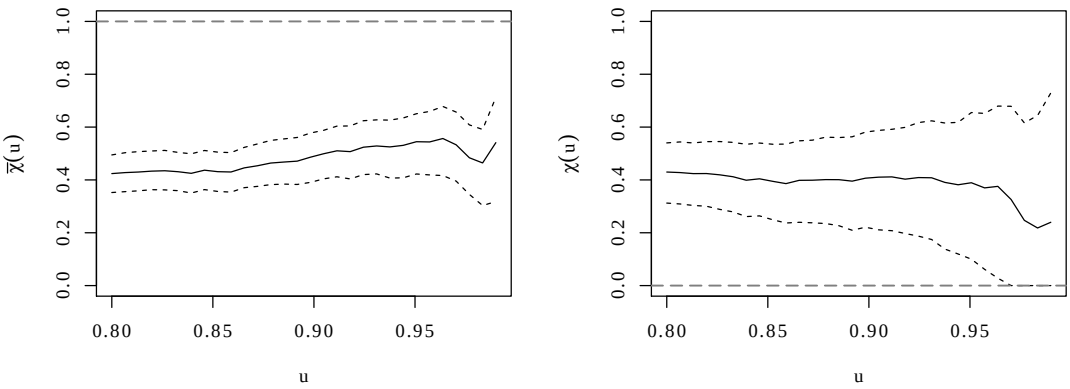

Figure 2.3 – Estimates of $\overline{\chi}(u)$ and $\chi(u)$ (solid) with 95% confidence intervals (small dashes) for Roche and Novartis negative returns. Values corresponding to asymptotic dependence ($\overline{\chi} = 1$) and asymptotic independence ($\chi = 0$) are shown as horizontal lines (dashed, grey).

2.3.2 From asymptotic dependence to asymptotic independence

Standard models for extreme values, such as those mentioned in Section 2.1.2, are based on a max-stability assumption, or equivalently on an assumption of regular variation, which entails either asymptotic dependence in the extremes or exact independence, both of which are unlikely in practice. When $\chi = 0$, these models can yield overestimation of joint risks, as they misspecify the probability that extreme events will occur simultaneously.

Ledford and Tawn (1996) focus on a class of models that connects asymptotic dependence and independence. Extrapolation from $\mathscr{A}$ to $\mathscr{B}$ under asymptotic dependence, using Pareto or Fréchet marginal distributions, with $\mathscr{A} = \{(x, y) : x > z, y > z\}$ and $\mathscr{B} = t\mathscr{A} = \{(x, y) : x > tz, y > tz\}$, has the form

$$\Pr\{(X_{\mathrm{F}}, Y_{\mathrm{F}}) \in \mathscr{B}\} = \Pr\{(X_{\mathrm{F}}, Y_{\mathrm{F}}) \in t\mathscr{A}\} \sim t^{-1}\Pr\{(X_{\mathrm{F}}, Y_{\mathrm{F}}) \in \mathscr{A}\}, \tag{2.12}$$

whereas complete independence between X^{F} and Y^{F} would yield

$$\begin{aligned}\Pr\{(X^{\mathrm{F}}, Y^{\mathrm{F}}) \in \mathscr{B}\} &= \Pr(X^{\mathrm{F}} > tz)\Pr(Y^{\mathrm{F}} > tz) \\ &\sim t^{-2}\Pr(X^{\mathrm{F}} > z)\Pr(Y^{\mathrm{F}} > z) \\ &= t^{-2}\Pr\{(X^{\mathrm{F}}, Y^{\mathrm{F}}) \in \mathscr{A}\},\end{aligned} \tag{2.13}$$

for large z. In order to bridge the gap between these two classes, Ledford and Tawn suggest modelling the joint tail as

$$\Pr(X^{\mathrm{F}} > z, Y^{\mathrm{F}} > z) \sim \mathscr{L}(z)z^{-1/\eta}, \quad z > 0, \tag{2.14}$$

with $\mathscr{L}(z)$ slowly varying at infinity and $\eta \in (0, 1]$ the coefficient of tail dependence. The variables are negatively associated when $\eta < 1/2$ and positively associated otherwise, with $\eta = 1/2$ meaning exact independence. Asymptotic dependence arises only when $\eta = 1$ and $\mathscr{L}(z) \not\to 0$, and asymptotic independence corresponds to all other cases. The coefficient of tail dependence is linked with the theory outlined in Section 2.1.2, as $\overline{\chi} = 2\eta - 1$.

The model (2.14) allows extrapolation from a moderately extreme set $\mathscr{A}$ to a more extreme set $\mathscr{B}$ only along the diagonal $x = y$, as in equations (2.12) and (2.13). An extension is described in Ledford and Tawn (1997), where directions for extrapolation consist of all rays emanating from the origin in the first quadrant and indexed by $\omega \in (0, 1)$, as in

$$\Pr\{X^{\mathrm{F}} > \omega z, Y^{\mathrm{F}} > (1 - \omega)z\} \sim \mathscr{L}(z)g(\omega)z^{-1/\eta},$$

where $g(\omega)$ describes the asymptotic ray dependence and is invariant to the rate of tail decay, which is controlled by η.

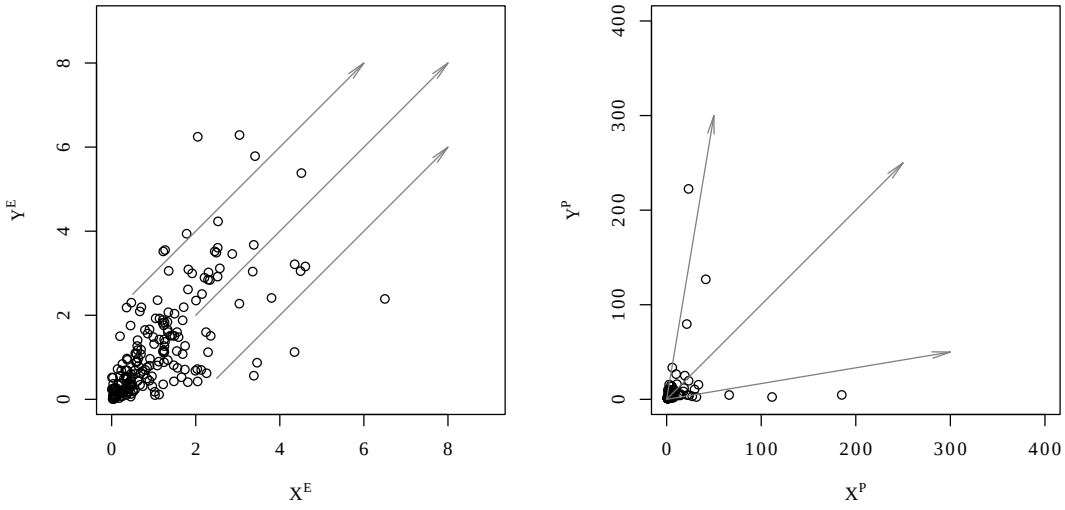

Figure 2.4 – Directions of extrapolation for probabilities of extreme sets on exponential (left) and Pareto scales (right) using the Ledford–Tawn formulation.

Consider extrapolation of $\Pr\{(X,Y)\in\mathscr{A}\}$ to $\Pr\{(X,Y)\in\mathscr{B}\}$, with $\mathscr{B}=t\mathscr{A}$, $t>0$. The Ledford–Tawn approach boils down to

$$\Pr\left\{(X^{\mathrm{F}},Y^{\mathrm{F}})\in\mathscr{B}\right\}=\Pr\left\{(X^{\mathrm{F}},Y^{\mathrm{F}})\in t\mathscr{A}\right\}\sim t^{-1/\eta}\Pr\left\{(X^{\mathrm{F}},Y^{\mathrm{F}})\in\mathscr{A}\right\},$$

which, with the factor $t^{-1/\eta}$, generalises the cases illustrated in (2.12) and (2.13). If we consider exponential margins instead of Fréchet margins, we have $\mathscr{B}=t+\mathscr{A}=\{(t+x,t+y):(x,y)\in\mathscr{A},t>0\}$ and the extrapolation has the form

$$\Pr\{(X^{\mathrm{E}},Y^{\mathrm{E}})\in\mathscr{B}\}=\Pr\{(X^{\mathrm{E}},Y^{\mathrm{E}})\in t+\mathscr{A}\}=\exp(-t/\eta)\Pr\{(X^{\mathrm{E}},Y^{\mathrm{E}})\in\mathscr{A}\}.$$

The extrapolation procedure depends on the marginal scale, as represented in Figure 2.4, with rays being parallel to the diagonal in exponential margins and emanating from the origin in Pareto margins, which correspond to Fréchet margins asymptotically. The exponential scale plot illustrates the problem arising when $\mathscr{B}$ has elements with $x\gg y$ for y large, needing $\mathscr{A}$ to lie close to the x-axis, in a region with few or no data points. In order to reduce the variance due to $\mathscr{A}$ falling in a region with little information, the extrapolation direction would need to be modified. More specifically, for $\mathscr{A}$ to be situated in a region with enough observations, we would need extrapolation rays emanating from the origin on the exponential scale.

2.3.3 Regular variation along rays in exponential margins

The multivariate regular variation condition backing standard models for extreme values arising in the context of Theorem 2.1 is restrictive, and in Section 2.3.1 we saw two complementary measures that describe two asymptotic classes of dependence and indicate when these standard models are applicable. A limitation of bivariate regular variation is that it imposes the same rate of tail decay over the whole cone $[0,\infty]^2 \setminus \{\mathbf{0}\}$. Hidden regular variation was introduced by Resnick (2002) to cope with different rates of tail decay in $(0,\infty]^2$ and on the axes, a theory that can describe asymptotic independence.

Following a different path, Wadsworth and Tawn (2013) generalise the notion of multivariate regular variation and suggest considering rays emanating from the origin in exponential margins. In Fréchet margins, this corresponds to marginal growth rates along the rays that vary depending on the ray direction. A similar approach can also be found in de Valk (2016). For all $(\beta,\gamma) \in [0,\infty]^2 \setminus \{\mathbf{0}\}$, Wadsworth and Tawn require that

$$\Pr(X^{\mathrm{P}} > x^{\beta}, Y^{\mathrm{P}} > x^{\gamma}) = \Pr(X^{\mathrm{E}} > \beta \log x, Y^{\mathrm{E}} > \gamma \log x) = \mathscr{L}(x;\beta,\gamma)x^{-\kappa(\beta,\gamma)}, \tag{2.15}$$

with $\mathscr{L}(x)$ a slowly varying function as $x \to \infty$ for all $(\beta,\gamma) \in [0,\infty]^2 \setminus \{\mathbf{0}\}$. The function $\kappa(\beta,\gamma)$ is homogeneous of order 1 (Wadsworth and Tawn, 2013, Property 1). If we define a pseudo-radial component $r = \beta + \gamma$ and a pseudo-angle component $\omega = \beta/r$, we can more simply focus on the angular dependence function $\lambda(\omega) = \kappa(\omega, 1-\omega)$, $\omega \in [0,1]$. Under non-negative association of X and Y, $\max(\omega, 1-\omega) \leq \lambda(\omega) \leq 1$, with boundary cases corresponding to asymptotic dependence and exact independence respectively. In all other cases, $\lambda(\omega)$ gives information about the degree of dependence of asymptotically independent distributions. We observe the relation between the angular dependence function $\lambda(\omega)$ and Pickands' dependence function $A(\omega)$, the first of which applies to asymptotically independent distributions, the latter being used for asymptotically dependent distributions. Both $\lambda(\omega)$ and $A(\omega)$ lie within the same boundaries, and both have a particular interpretation in extreme value theory when evaluated at $\omega = 1/2$, with $2\lambda(1/2) = 1/\eta$ the inverse coefficient of tail dependence (2.14) and $2A(1/2) = \theta$, the coefficient of extremal dependence (Smith, 1990; Coles and Tawn, 1994).

The Wadsworth–Tawn model entails assuming regular variation of order $-\lambda(\omega)$ along rays emanating from the origin, when the standard approach to modelling extremes imposes a single multivariate regular variation condition that is too rigid to take into account cases with asymptotic independence and positive association.

2.4 Conditional extremes

2.4.1 Characterising the limit

Up to this point, we have mainly discussed the standard approach to extremes, which focuses on probabilities of big events happening simultaneously and which fails to account

for asymptotically independent distributions exhibiting dependence at subasymptotic levels, an important example of which is the Gaussian distribution. This corresponds to $\chi = 0$ and $\overline{\chi} \neq 0$, a situation often met in practice. We have also mentioned alternative approaches that can capture more subtle forms of decay in tail dependence, but all methods, except the Wadsworth–Tawn approach, rely on limiting assumptions that require all variables to grow at the same rate.

Heffernan and Resnick (2007) present a conditional approach that overcomes many of the limitations of standard approaches. They give as the general framework a random vector $(X, \boldsymbol{Y}) = (X, Y_1, \ldots, Y_d)$ with X used as the conditioning variable. For ease of exposition, we take $d = 1$ and $Y_1 = Y$ in what follows. A natural assumption is that X is in the maximum domain of attraction of a Fréchet distribution, in the sense of (2.3). For the joint structure, the authors assume the existence of a Radon measure $\mu(\cdot,\cdot)$ on $[-\infty,\infty] \times (0,\infty]$ such that $\mu([-\infty, y] \times (x,\infty])$ is not a degenerate function in y for each $x > 0$, and there exist $a^{\mathrm{HR}}(\cdot)$ and $b^{\mathrm{HR}}(\cdot) > 0$ such that

$$\lim_{t\to\infty} t \times \Pr\left\{\frac{Y - a^{\mathrm{HR}}(t)}{b^{\mathrm{HR}}(t)} \leq y, \frac{X}{t} > x\right\} = \mu([-\infty, y] \times (x,\infty]), \quad y \in \mathbb{R},\ x > 0. \tag{2.16}$$

As $\Pr(X > x) \sim x^{-1}$, $x \to \infty$, in terms of the conditional probability we have

$$\begin{aligned}\Pr\left\{\frac{Y - a^{\mathrm{HR}}(t)}{b^{\mathrm{HR}}(t)} \leq y \,\middle|\, X > t\right\} &\sim t \times \Pr\left\{\frac{Y - a^{\mathrm{HR}}(t)}{b^{\mathrm{HR}}(t)} \leq y, \frac{X}{t} > 1\right\} \\ &\to \mu([-\infty, y] \times (1,\infty]), \quad t \to \infty.\end{aligned} \tag{2.17}$$

Under these mild conditions, and in particular without assuming that (X, Y) has a density, Heffernan and Resnick (2007) show that the scale function $b^{\mathrm{HR}}(\cdot)$ can be identified as regularly varying with index $\rho \in \mathbb{R}$, i.e.,

$$\lim_{t\to\infty} \frac{b^{\mathrm{HR}}(ct)}{b^{\mathrm{HR}}(t)} = c^{\rho}, \quad c > 0,$$

and the location function $a^{\mathrm{HR}}(\cdot)$ can be either identically 0, or regularly varying with index ρ if $\rho \neq 0$, up to a change of variable, and it is Π-varying (Resnick, 1987, p. 27; Bingham *et al.*, 1987, p. 158) with auxiliary function $b^{\mathrm{HR}}(\cdot)$ if $\rho = 0$. More specifically, requiring $\mu(\cdot,\cdot)$ to be a product measure, of the form

$$\mu([-\infty, y] \times (x,\infty]) = H(y)x^{-1}, \quad H(y) = \mu([-\infty, y] \times (1,\infty]), \tag{2.18}$$

is equivalent to having, for any constant $c > 0$,

$$\lim_{t\to\infty} \frac{a^{\mathrm{HR}}(ct) - a^{\mathrm{HR}}(t)}{b^{\mathrm{HR}}(t)} = 0, \qquad \lim_{t\to\infty} \frac{b^{\mathrm{HR}}(ct)}{b^{\mathrm{HR}}(t)} = 1.$$

Compared to the general framework of (2.17), here we have $\rho = 0$, i.e., $b^{\mathrm{HR}}(\cdot)$ is slowly varying.

To simplify the argument, consider $(X^{\mathrm{P}}, Y^{\mathrm{P}})$, obtained by rescaling (X, Y) onto the Pareto scale. Then assumption (2.16) becomes

$$\lim_{t\to\infty} t \times \Pr\left\{\frac{Y^{\mathrm{P}}}{b^{\mathrm{P}}(t)} \leq y, \frac{X^{\mathrm{P}}}{t} > x\right\} = \mu([1,y]\times(x,\infty]), \quad x, y > 1,$$

and $b^{\mathrm{P}}(\cdot) > 0$ is regularly varying with index $\rho \leq 1$. Following Papastathopoulos *et al.* (2017), we can consider the transformation from the Pareto to the exponential scale, from which we can derive

$$\frac{Y^{\mathrm{P}}}{b^{\mathrm{P}}(t)} \leq y \iff Y^{\mathrm{P}} \leq b^{\mathrm{P}}(t)\, y \iff Y^{\mathrm{E}} \leq \log b^{\mathrm{P}}(t) + \log y.$$

Writing $a^{\mathrm{E}}(\cdot) = \log b^{\mathrm{P}}\{\exp(\cdot)\}$, we observe that this approach reduces to a location standardisation of $Y^E = \log Y^P$, i.e., Y on the exponential scale, so that in this case (2.16) becomes

$$\lim_{t\to\infty} t \times \Pr\left\{Y^{\mathrm{E}} - a^{\mathrm{E}}\left(\log t\right) \leq y, X^{\mathrm{E}} - \log t > x\right\} = \mu([0,y]\times(x,\infty]), \quad x, y > 0,$$

or equivalently,

$$\lim_{t\to\infty} \Pr\left\{Y^{\mathrm{E}} - a^{\mathrm{E}}(t) \leq y \;\middle|\; X^{\mathrm{E}} > t\right\} = \mu([0,y]\times(0,\infty]), \quad y > 0. \tag{2.19}$$

2.4.2 Heffernan–Tawn formulation

In our extrapolation problem, we want to infer the probability of an extreme set $\mathscr{B}$ given that we are able to estimate the probability of a less extreme set $\mathscr{A}$ in a reliable fashion. A natural assumption for $\mathscr{B}$ is that all its points have at least one extreme component. Inference in such a multivariate framework with $\boldsymbol{X} = (X_1, \ldots, X_d)$ is made separately on a partition of the d-dimensional set $\mathscr{B}$ of interest. Assuming that all d marginal distributions are identically F, the partition is defined as $\mathscr{B} = \cup_{i=1}^{d} \mathscr{B}_i$, with $\mathscr{B}_i = \{\boldsymbol{x} \in \mathscr{B} : F(x_i) > F(x_j), j \neq i\}$. The inference is then split according to

$$\Pr(\boldsymbol{X} \in \mathscr{B}) = \sum_{i=1}^{d} \Pr(\boldsymbol{X} \in \mathscr{B}_i \mid X_i > u_i)\Pr(X_i > u_i), \tag{2.20}$$

with $u_i = \inf\{x_i : \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathscr{B}_i\}$.

There are two building blocks in the estimation of $\Pr(\boldsymbol{X} \in \mathscr{B})$ in (2.20): the marginal survivor probabilities $\Pr(X_i > u_i)$ can be estimated using standard threshold methods related to (2.5), details of which can be found in the literature (Coles, 2001, Chap. 3; de Haan and Ferreira, 2006, Chap. 4); the conditional probabilities in (2.20) need more careful attention, as standard extreme value approaches reduce to modelling all components as being asymptotically dependent or exactly independent.

Heffernan and Tawn (2004) present a very flexible model where the asymptotic dependence class can differ from one pair of variables to the other. Their formulation is similar to (2.17)

but assumes Gumbel margins for $\boldsymbol{X} = \boldsymbol{X}^{\mathrm{G}}$, which in practice necessitates fitting a marginal model and applying the probability integral transform twice to each element of $\boldsymbol{X}$. Another difference is that their formulation uses random norming by the conditioning variable instead of normalising by deterministic functions of the conditioning threshold u. This additional information carried by the conditioning variable X_i and passed to the norming functions $a(\cdot)$ and $b(\cdot)$ results in $\mu(\cdot,\cdot)$ being a product measure (Heffernan and Resnick, 2007), which is key for the extrapolation strategy. Assuming the same marginal distribution for all components of $\boldsymbol{X}$ gives a characterisation of the normalisation functions, i.e., that $a(\cdot)$ is regularly varying with index 1 and $b(\cdot)$ is regularly varying with index $\rho < 1$. Random norming also justifies separate inference on marginal conditional distributions, e.g., $X_j \mid X_i > u$, $X_k \mid X_i > u$, when X_j and X_k can be assumed conditionally independent given X_i, thus greatly reducing the burden of estimating a potentially high-dimensional limiting distribution $H(\cdot)$ (Papastathopoulos, 2016).

In essence, we require existence of normalising functions $a_{j|i}(\cdot)$ and $b_{j|i}(\cdot) > 0$ $(i, j = 1,\ldots,d;\, i \neq j)$, such that

$$\Pr\left\{ \frac{X_j^{\mathrm{G}} - a_{j|i}\left(X_i^{\mathrm{G}}\right)}{b_{j|i}\left(X_i^{\mathrm{G}}\right)} \leq z_{j|i},\, X_i^{\mathrm{G}} - u > x,\; j = 1,\ldots,d, j \neq i \;\middle|\; X_i^{\mathrm{G}} > u \right\} \\ \to H_{|i}(\boldsymbol{z}_{|i}) \exp(-x), \quad u \to \infty,\; i = 1,\ldots,d, \quad (2.21)$$

where the $z_{j|i}$ are the $d-1$ components of $\boldsymbol{z}_{|i}$ and the $H_{|i}(\cdot)$ are non-degenerate and have no mass at $-\infty$. The assumption (2.21) refines (2.19) by introducing a scale function $b_{j|i}(\cdot)$ satisfying $b_{j|i}(t) = o\{a_{j|i}(t)\}$ as $t \to \infty$. This scale function allows consideration of joint distributions that would be degenerate using formulations (2.16) or (2.19). For example, if we consider $(X^{\mathrm{G}}, Y^{\mathrm{G}})$ with a bivariate normal copula, correlation $\rho > 0$, and Gumbel margins, we need, after dropping the subscripts for clarity,

$$a(x) = \rho^2 x, \quad b(x) = x^{1/2},$$

for (2.21) to hold with a non-degenerate $H(\cdot)$. If (2.19) was used instead of (2.21), we would have $a^{\mathrm{E}}(x) = \rho^2 x$, but $Y^{\mathrm{E}} - a^{\mathrm{E}}$ would not be normalised and the limit (2.19) would be degenerate.

A simple class of normalising functions arises from (2.21) and is based on the study of many existing parametric models for extremes. The location function $a_{j|i}(\cdot)$ reflects the asymmetry of the upper and lower tails of the Gumbel distribution, namely,

$$\begin{aligned} a_{j|i}(x) &= \alpha_{j|i} x + \left(\gamma_{j|i} - \delta_{j|i} \log x\right) \times \mathbb{1}\left(\alpha_{j|i} = 0, \beta_{j|i} < 0\right), \\ b_{j|i}(x) &= x^{\beta_{j|i}}, \end{aligned} \quad (2.22)$$

where $\alpha_{j|i}, \beta_{j|i}, \gamma_{j|i}, \delta_{j|i}$ are parameters satisfying $\alpha_{j|i} \in [0,1]$, $\beta_{j|i} \in (-\infty, 1)$, $\gamma_{j|i} \in \mathbb{R}$, $\delta_{j|i} \in [0,1]$. We follow Heffernan and Tawn and use this notation, since for positively associated

variables we can write

$$x_j = \alpha_{j|i} x_i + x_i^{\beta_{j|i}} z_{j|i}, \quad \text{for large } x_i,$$

similar to a regression model. We shall often call $H_{|i}$ the residual distribution function. In case of negative association, the formulation for the location function $b(\cdot)$ is more involved, due to the asymmetry between the upper and lower tails of the Gumbel distribution. Interpretation of the parameters involved in (2.22) is tricky, and we postpone it to Section 2.4.3, where a different formulation allows for simpler parametric forms of $a_{j|i}(\cdot)$ and $b_{j|i}(\cdot)$.

Inference is done in three steps; first, a marginal model is used, in the form of

$$\widehat{F}_i(x) = \begin{cases} 1 - \{1 - \tilde{F}_i(u_i)\}\left(1 + \xi_i \dfrac{x - u_i}{\sigma_i}\right)_+^{-1/\xi_i}, & x > u_i, \\ \tilde{F}_i(x), & x \leq u_i, \end{cases} \tag{2.23}$$

where $\tilde{F}_i$ denotes the empirical distribution function for margin i, and the u_i are chosen large enough for the approximation (2.6) to be adequate. Using the probability integral transform twice, we get Gumbel margins by applying $-\log\{-\log \widehat{F}_i(x)\}$ to each margin of $\boldsymbol{X}$. The second step deals with dependence structure. Because the residual distribution function $H_{|i}$ is of very general form, Heffernan and Tawn introduce a working assumption that $H_{|i}$ is Gaussian with mean vector $\boldsymbol{\mu}_{|i}$ and covariance matrix $\Psi_{|i}$. If we assume that asymptotic conditional independence holds, then standard maximum likelihood estimation can be performed separately on each margin of the conditional distribution. With pairs of observations $(\boldsymbol{x}_i, \boldsymbol{x}_j) = \{(x_{1,i}, x_{1,j}), \ldots, (x_{n,i}, x_{n,j})\}$, the corresponding log-likelihood is

$$\ell\left(\alpha_{j|i}, \beta_{j|i}, \gamma_{j|i}, \delta_{j|i}, \mu_{j|i}, \psi_{j|i}; \boldsymbol{x}_i, \boldsymbol{x}_j\right) = \sum_{k=1}^{n} \log\varphi\left\{\frac{x_{k,j} - a_{j|i}(x_{k,i}) - b_{j|i}(x_{k,i})\mu_{j|i}}{b_{j|i}(x_{k,i})\psi_{j|i}}\right\} \times \mathbb{1}\left(x_{k,i} > u_i\right), \tag{2.24}$$

where $a_{j|i}(\cdot)$ and $b_{j|i}(\cdot)$ are specified in (2.22), the covariance matrix $\Psi_{|i}$ reduces to the scalar $\psi_{j|i}^2$, and φ is the standard normal density function. The parameters $\gamma_{j|i}$ and $\delta_{j|i}$ are first set to zero in (2.24). If there is evidence in favour of $\alpha_{j|i} = 0$ and $\beta_{j|i} < 0$, maximum likelihood estimation is performed for all parameters; because $\alpha_{j|i}$ is bounded below by 0, there is a non-null probability that $\widehat{\alpha}_{j|i} = 0$. In the final step, the maximum likelihood estimates $\widehat{\alpha}_{j|i}$, $\widehat{\beta}_{j|i}$, $\widehat{\gamma}_{j|i}$ and $\widehat{\delta}_{j|i}$ are used to compute the empirical distribution of the residuals, using

$$\widehat{z}_{k,j|i} = \frac{x_{k,j} - \widehat{\alpha}_{j|i} x_{k,i} - \left(\widehat{\gamma}_{j|i} - \widehat{\delta}_{j|i} \log x_{k,i}\right) \times \mathbb{1}\left(\widehat{\alpha}_{j|i} = 0, \widehat{\beta}_{j|i} < 0\right)}{x_{k,i}^{\widehat{\beta}_{j|i}}}, \quad k = 1, \ldots, n_i.$$

Extrapolation from the moderately extreme set $\mathscr{A}$ to the more extreme set $\mathscr{B}$ requires simulation from the empirical residual distribution $\widehat{H}_{|i}$, i.e., no closed-form formula such as (2.12) exists for the conditional tail model. A simple but robust method suggested by Heffernan and Tawn is to simulate data points along the lines of Algorithm 2.1 and to count the proportion

Algorithm 2.1: Simulating pairs from the conditional model.

Input: parameter estimates $\widehat{\alpha}_{j|i}, \widehat{\beta}_{j|i}$, empirical residuals $\widehat{\boldsymbol{z}}_{|i}$, threshold u_i
repeat
  Sample X_i from a Gumbel distribution conditioned on exceeding u_i
  Sample $\widehat{Z}_{j|i}$ uniformly from the elements of $\widehat{\boldsymbol{z}}_{|i}$
  Compute $X_j = \widehat{\alpha}_{j|i} X_i + X_i^{\widehat{\beta}_{j|i}} \widehat{Z}_{j|i}$
  Keep (X_i, X_j)
until *enough pairs are sampled*

of them that fall into $\mathscr{B}_i$ defined in (2.20). This gives an estimate of $\Pr(\boldsymbol{X} \in \mathscr{B}_i \mid X_i > u_i)$. The probabilities $\Pr(X_i > u_i)$ can be estimated using the marginal model (2.23).

2.4.3 Alternative formulation, extensions and additional constraints

Heffernan and Tawn (2004) assume $\boldsymbol{X}$ to be marginally on the standard Gumbel scale for the conditional model to be applicable. The asymmetry of the lower and upper tails in the Gumbel distribution yields the convoluted parametric expressions (2.22) for the normalising functions $a_{j|i}(\cdot)$ and $b_{j|i}(\cdot)$. Keef *et al.* (2013) choose to consider $\boldsymbol{X}$ on the standard Laplace scale, thus keeping the Gumbel upper tail decay and ensuring symmetry of the lower and upper tails. This yields simpler forms for the location function $a(\cdot)$, whose characterisation is valid for both tails of the conditional probability (2.21), specifically

$$\begin{aligned} a(x) &= \alpha x, \quad \alpha \in [-1, 1], \\ b(x) &= x^{\beta}, \quad \beta \in (-\infty, 1). \end{aligned} \tag{2.25}$$

The simple form of the normalising functions (2.25) covers a broad collection of standard parametric copula models, not only extreme value models such as the asymmetric logistic, but also asymptotically independent distributions arising from inverted extreme value models, and other distributions poorly described by standard theory for extreme values, e.g., the multivariate Gaussian distribution. Specifically, asymptotic dependence corresponds to similar decay rates of the conditional and the marginal distributions, i.e., $\alpha_{j|i} = 1$ and $\beta_{j|i} = 0$, but it also covers asymptotic negative dependence, when the variables are strongly negatively associated, with $\alpha_{j|i} = -1$ and $\beta_{j|i} = 0$; other values correspond to asymptotic independence. Positive association corresponds to $\alpha_{j|i} > 0$ and negative association corresponds to $\alpha_{j|i} < 0$. Examples of conditional distributions are shown in Figure 2.5, and correspond to three different classes of dependence structures under asymptotic independence that are introduced in Heffernan and Tawn (2004); X_i and X_j are positive extremal dependent if conditional quantiles of $X_j \mid X_i > u_i$ tend to ∞ when $u_i \to \infty$, i.e., $\alpha_{j|i} > 0$, they are negative extremal dependent if conditional quantiles tend to $-\infty$, i.e., $\alpha_{j|i} < 0$, and extremal near independent if conditional quantiles have a finite limit, i.e., $\alpha_{j|i} = 0$ and $\beta_{j|i} \leq 0$.

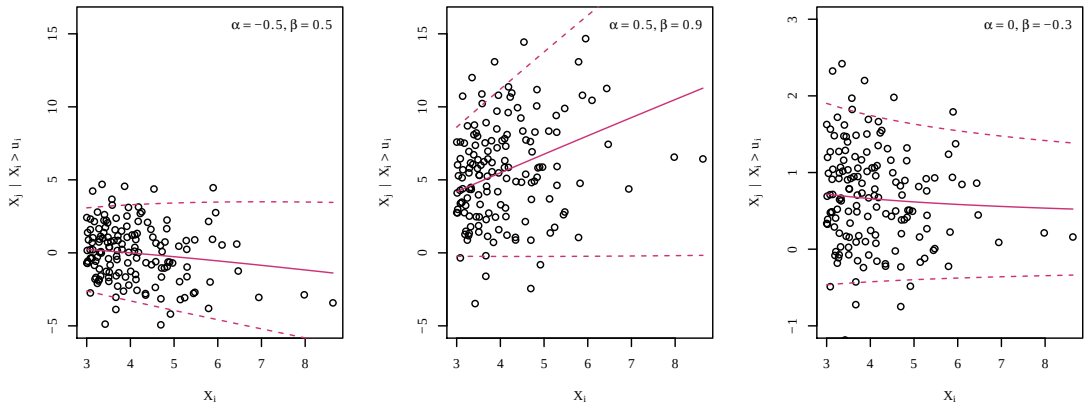

Figure 2.5 – Examples of dependence structures with asymptotic independence spanned by the conditional model. From left to right: negative association, positive association, and extremal near independence. Data are simulated using the parameters displayed in the top right corner of each panel and a Gaussian distribution with mean 1 for the residual distribution. Plain lines indicate conditional medians, dashed lines represent the 5% and 95% conditional quantiles.

The simplicity of the formulation (2.21) when using the parameterisation (2.25) has some weaknesses, such as that conditional probabilities that do not agree with marginal probabilities in general; Keef *et al.* (2013) introduce additional constraints that improve on the original model. Limiting conditional probabilities such as χ correspond to a natural ordering, with asymptotic independence, i.e., $\chi = 0$, being smaller than any type of asymptotic positive dependence, with $\chi > 0$. The measure χ defined in (2.9) is well-suited for positively associated variables, and we shall write $\chi^+ = \chi$. For negatively associated variables, the bottom right corner is of interest, leading Keef *et al.* (2013) to define

$$\chi^- = \lim_{u \to 1} \Pr\left(Y^{\mathrm{U}} \leq 1 - u \mid X^{\mathrm{U}} > u\right),$$

for which $\chi^- > 0$ defines asymptotic negative dependence, and $\chi^- = 0$ is asymptotic negative independence, from which we conclude that the same natural ordering property holds between asymptotic negative independence and asymptotic negative dependence,

The objective of additional constraints is to guarantee that these orderings hold above a sufficiently high threshold. In what follows, as a complement to Keef *et al.* (2013), we develop an argument showing how this can be done formally. Using the subasymptotic versions of χ^+ and χ^-, we have for large u,

$$\begin{aligned} 1 - \chi^+(u) &= \Pr\left(Y^{\mathrm{U}} \leq u \mid X^{\mathrm{U}} > u\right), \\ \chi^-(u) &= \Pr\left(Y^{U} \leq 1 - u \mid X^{\mathrm{U}} > u\right). \end{aligned} \tag{2.26}$$

Writing $q(\cdot) = q_u(\cdot)$ for the conditional quantile function, i.e., $q\{\Pr(Y^{\mathrm{U}} \leq p \mid X^{\mathrm{U}} = u)\} = p$, we conclude that $q\{1 - \chi^+(u)\} = u$ and $q\{\chi^-(u)\} = 1 - u$ by inverting the probabilities in (2.26).

We now write $q^+(\cdot)$, $q(\cdot)$ and $q^-(\cdot)$ for the conditional quantile functions of $Y \mid X = x$ for x large under the model of Heffernan and Tawn (2004) under asymptotic positive dependence, asymptotic independence and asymptotic negative dependence, respectively. Using the monotonicity of the quantile function, we derive, for large u,

$$\begin{aligned} q^+(1) > q^+(1-\chi^+(u)) = u = q(1-\chi^+(u)) \approx q(1), \\ q^-(0) < q^-(\chi^-(u)) = 1-u = q(\chi^-(u)) \approx q(0), \end{aligned} \tag{2.27}$$

which suggests imposing a specific ordering on the conditional quantiles, namely, $q(1) < q^+(1)$ and $q(0) > q^-(0)$.

Keef *et al.* (2013) give a more informal argument and claim that the ordering of χ^+ and χ^- under the different asymptotic limits yields a natural ordering for all levels $p \in [0,1]$ of the conditional quantiles. They argue that for any given $p \in [0,1]$, the pth conditional quantile under asymptotic independence must be larger than under asymptotic negative dependence and smaller than under asymptotic positive dependence. In terms of the conditional tail model, this translates as, for large x_i,

$$-x_i + \left(H_{|i}^-\right)^{\leftarrow}(p) \leq \alpha_{j|i} x_i + x_i^{\beta_{j|i}} H_{|i}^{\leftarrow}(p) \leq x_i + \left(H_{|i}^+\right)^{\leftarrow}(p), \quad p \in [0,1], \tag{2.28}$$

with $H_{|i}^-$ and $H_{|i}^+$ the residual distribution functions under asymptotic negative dependence and asymptotic positive dependence. By imposing the condition (2.28) for all x_i above a very high threshold $v > u_i$, Keef *et al.* (2013) derive conditions on $\alpha_{j|i}$ and $\beta_{j|i}$. They find through a range of examples that these conditions are strongest for $p \in \{0\} \cup \{1\}$ in (2.28), which matches (2.27). The threshold v must be chosen above the range of the data to give the fit more flexibility; the fit is largely insensitive to the particular choice of v, as reported by the authors and as we have also experienced in practice.

We illustrate how the constraints derived by Keef *et al.* (2013) work with an application to negative returns of Goldman Sachs and Citigroup, two major actors in the US banking sector, using the price of their stocks from 1 November 2000 to 31 October 2017. We apply the marginal model (2.23) to transform the data to the Laplace scale. We focus on the negative returns of the Goldman Sachs share given losses of the Citigroup share that exceed 3.9%, corresponding to the marginal empirical 95% quantile. A standard fit using the method described in Section 2.4.2 gives $\widehat{\alpha}_{\text{GS|Citi}} = 0.39$ and $\widehat{\beta}_{\text{GS|Citi}} = 0.72$. As shown in Figure 2.6, these estimates do not satisfy the conditions on (α, β) derived from (2.28). Another fit, where the constraints are enforced, gives $\widehat{\alpha}_{\text{GS|Citi}} = 0.46$ and $\widehat{\beta}_{\text{GS|Citi}} = 0.53$. In terms of the model parameters, the difference between the unconstrained and constrained fits is noticeable, and more importantly, conditional quantiles derived from the unconstrained fit show inconsistencies which the constrained fit eliminates. For example, a 15% loss on Citigroup gives a conditional 95% quantile of 14.1% on Goldman Sachs when the fit is unconstrained, when under asymptotic positive dependence the conditional 95% value-at-risk would be 13.5%; the constrained fit gives a consistent estimate of 13.1%.

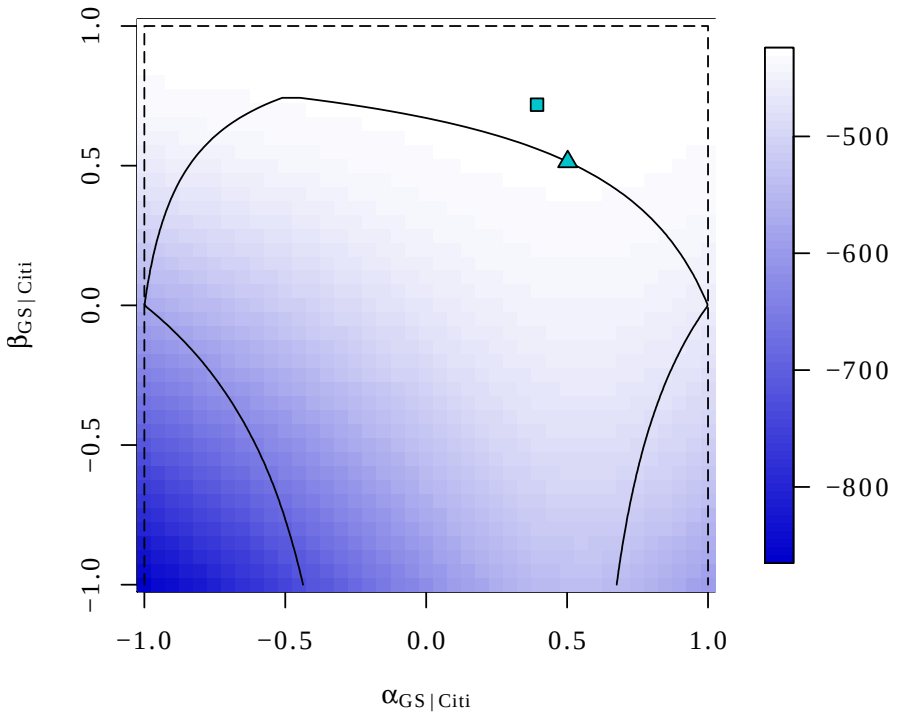

Figure 2.6 – Constraints on the parameters of the conditional model based on the negative returns of Goldman Sachs, conditioned on big negative returns of Citigroup. The solid curve surrounds the set where values of (α, β) satisfy the constraints; dashed lines correspond to the unconstrained set. Respective estimated values are shown as a triangle and a square. The profile log-likelihood levels are displayed in the background.

The parameterisation of the norming functions (2.25) allows simple inference and interpretation, and can capture a very broad class of extremal dependence structures (Heffernan and Tawn, 2004). This parameterisation is too general, however, and additional constraints are needed for sensible risk estimates. On the other hand, this parameterisation does not cover the dependence structure of some inverted max-stable processes asymptotically (Papastathopoulos and Tawn, 2016). Although this may seem restrictive from a theoretical perspective, Papastathopoulos and Tawn show that when the interest is in subasymptotic levels, the norming functions (2.25) perform well in practice.

The Heffernan–Tawn model (2.21) has been used in several applications, in particular for the dependence modelling of extremes in space and time (Keef *et al.*, 2009a,b; Winter and Tawn, 2016). Winter and Tawn (2017) extend the original model to cover kth-order Markov chains and derive the transition distribution

$$X_k \mid (X_0 = x_0, \ldots, X_{k-1} = x_{k-1}),$$

based on the joint conditional distribution of $(X_1, \ldots, X_k) \mid X_0 > u$ given by (2.21), with (X_t) a stationary time series having a Laplace marginal distribution. With this formulation, the authors are able to derive the probabilities associated with any heatwave duration and introduce a hypothesis test to facilitate the choice of the order k of the Markov chain. Extending the conditional model in a different direction, for a 1st order Markov chain, Papastathopoulos *et al.* (2017) consider finite-dimensional distributions of

$$\left\{ \frac{X_t - b_t(X_0)}{a_t(X_0)} : t = 1, 2, \ldots \right\} \,\middle|\, X_0 > u,$$

and characterise the distributions of the corresponding tail chains.

An issue raised by Heffernan and Tawn (2004) is the lack of self-consistency between probabilities derived from the separate conditional models involved in (2.20). For a pair of variables X, Y on the Laplace scale, self-consistency would require joint densities to match in regions where more than one conditional model applies, that is,

$$\frac{d}{dx}\Pr\left(X \le \alpha_{x|y} y + y^{\beta_{x|y}} x \,\middle|\, Y = y\right) e^{-y} = \frac{d}{dy}\Pr\left(Y \le \alpha_{y|x} x + x^{\beta_{y|x}} y \,\middle|\, X = x\right) e^{-x}, \tag{2.29}$$

for all $x > u_x$ and $y > u_y$, with u_x and u_y two suitably high thresholds used as an approximation to the limit (2.21). The case of asymptotic dependence, with $\alpha_{x|y} = \alpha_{y|x} = 1$ and $\beta_{x|y} = \beta_{y|x} = 0$, implies the following constraint on the residual distributions,

$$\frac{d}{dz} H_{|x}(z) = e^{-z} \frac{d}{dz} H_{|y}(-z), \quad z \in \mathbb{R}.$$

In the case of asymptotic independence, Heffernan and Tawn argue that the equality (2.29) is trivial in the limit and it is too complex to apply at finite levels.

Liu and Tawn (2014) explore self-consistency more thoroughly; they show that (2.29) cannot hold for all $x > u_x$, $y > u_y$, with $H_{|x}$ and $H_{|y}$ twice differentiable. They define a weaker version of self-consistency, termed diagonal self-consistency, for which (2.29) holds for all $x > u$, $y > u$ with $x = y$ and u a high threshold. This definition is helpful in practice, and the authors work out a parametric class of residual distributions that satisfy diagonal self-consistency and covers many asymptotically independent distributions.

Following the theoretical developments of Heffernan and Resnick (2007), self-consistency is studied by Das and Resnick (2011), who show that if convergence (2.17) holds when conditioning on either X or Y, with potentially different norming functions a and b, then (X, Y) is in the domain of attraction of a multivariate extreme value distribution. This does not help from a practical point of view, as this result does not hold when $\mu(\cdot, \cdot)$ is a product measure, of the form (2.18), as we assumed in (2.21).

2.5 Summary

In this chapter, we have reviewed how the probability of very extreme sets can be inferred from that of moderately extreme sets, in the univariate setting using maxima or exceedances of a high threshold, and in the multivariate setting, where no natural ordering exists. We have then focused on extremes in time series, where dependence in time prevents from using standard inference procedures; the peaks-over-threshold approach is one of the methods that deal with short-range dependence of extreme observations. In the two-dimensional case, we have described how asymptotic dependence can be described and measured. We have emphasised the importance of models for data that are asymptotically independent, with dependence at subasymptotic levels.

The last section was dedicated to the conditional tail approach, and we have seen how it can be used to estimate the probabilities of extreme events. In Laplace margins, this approach is parsimonious and easy to use for inference on multidimensional data. It is very flexible and covers a broad class of extremal structures of dependence, but lacks self-consistency for extreme joint probabilities extrapolated from different conditional distributions. It is also unclear to what extent the assumptions needed to make inference have an impact on the estimation of extreme probabilities and on the assessment of uncertainty for risk estimates.

We suggest a new approach to fitting the conditional tail model in Chapter 5, but before we do so, we introduce background material on the Bayesian nonparametric framework and we explore finite-sample properties of the conditional model in Chapter 4.

3 The Dirichlet process

3.1 Formal definitions

3.1.1 The Dirichlet distribution

In Bayesian modelling, the Dirichlet distribution is a conjugate prior for the parameters of a multinomial distribution. It is also a generalisation of the beta distribution to the $(d-1)$-dimensional simplex $\mathbb{S} = \{(x_1,\ldots,x_{d-1}) : x_j \geq 0,\ \sum_{j=1}^{d-1} x_j \leq 1\}$. The Dirichlet density function is

$$f(x_1,\ldots,x_{d-1} \mid \gamma_1,\ldots,\gamma_d) = \frac{\Gamma(\gamma_1+\cdots+\gamma_d)}{\Gamma(\gamma_1)\cdots\Gamma(\gamma_d)}\left(1-\sum_{j=1}^{d-1} x_j\right)^{\gamma_d-1}\prod_{j=1}^{d-1} x_j^{\gamma_j-1}, \quad (x_1,\ldots,x_{d-1}) \in \mathbb{S}, \tag{3.1}$$

with $\gamma_j > 0$, $j = 1,\ldots,d$. Ferguson (1973) uses a constructive approach that permits a more general definition. Let $Z_1,\ldots,Z_d$ be independently gamma distributed $\mathrm{Ga}(1,\gamma_j)$ variables, with shape parameters $\gamma_j \geq 0$, and $\gamma_j > 0$ for some j, $j = 1,\ldots,d$. The distribution of $(X_1,\ldots,X_d)$, with

$$X_j = \frac{Z_j}{\sum_{k=1}^{d} Z_k}, \quad j = 1,\ldots,d,$$

is Dirichlet with parameter $(\gamma_1,\ldots,\gamma_d)$, which we shall write $\mathrm{Dir}(\gamma_1,\ldots,\gamma_d)$. Any $\gamma_j = 0$ implies that $X_j \equiv 0$; if $\gamma_j > 0$ for all $j = 1,\ldots,d$, the density function of $(X_1,\ldots,X_{d-1})$ is exactly (3.1). The marginal expectation and variance are

$$\mathrm{E}(X_j) = \frac{\gamma_j}{\sum_{k=1}^{d}\gamma_k}, \qquad \mathrm{var}(X_j) = \frac{\gamma_j\left(\sum_{k=1}^{d}\gamma_k - \gamma_j\right)}{\left(\sum_{k=1}^{d}\gamma_k\right)^2\left(\sum_{k=1}^{d}\gamma_k + 1\right)}. \tag{3.2}$$

An interesting property of the Dirichlet distribution is its updating of prior beliefs after recording multinomial observations, building a useful link with the Pólya urn scheme developed in Section 3.1.3. If $(X_1,\ldots,X_d)$ have prior distribution $\mathrm{Dir}(\gamma_1,\ldots,\gamma_d)$ and observations

are such that

$$\Pr(Y = j \mid X_1, \ldots, X_d) = X_j, \quad j = 1, \ldots, d,$$

almost surely, then the posterior distribution becomes

$$(X_1, \ldots, X_d) \mid \{Y = j\} \sim \mathrm{Dir}\big(\gamma_1, \ldots, \gamma_j + 1, \ldots, \gamma_d\big). \tag{3.3}$$

3.1.2 Ferguson's definition

Given a set $\mathscr{P}$ and its associated σ-field $\mathscr{T}$, Ferguson (1973) gives the following definition of a Dirichlet process:

Definition 3.1 (Dirichlet process)
Let $\nu(\cdot)$ be a finite measure on $(\mathscr{P}, \mathscr{T})$. We say that P is a Dirichlet process on $(\mathscr{P}, \mathscr{T})$ with parameter $\nu(\cdot)$ and we write DP(ν) if, for every $k = 1, 2, \ldots$ and measurable partition $(\mathscr{C}_1, \ldots, \mathscr{C}_k)$ of $\mathscr{P}$,

$$\{P(\mathscr{C}_1), \ldots, P(\mathscr{C}_k)\} \sim Dir\{\nu(\mathscr{C}_1), \ldots, \nu(\mathscr{C}_k)\}.$$

The joint probability of any measurable sets $\mathscr{D}_1, \ldots, \mathscr{D}_l$, for any $l = 1, 2, \ldots$, can be derived from the partition with sets

$$\mathscr{C}_{k_1, \ldots, k_l} = \bigcap_{j=1}^{l} \mathscr{D}_j^{k_j},$$

where $k_j \in \{0, 1\}$ and $\mathscr{D}_j^1$ is interpreted as $\mathscr{D}_j$ and $\mathscr{D}_j^0$ as its complement $\mathscr{D}_j^c = \mathscr{X} \setminus \mathscr{D}_j$. The marginal distribution of $\{P(\mathscr{D}_1), \ldots, P(\mathscr{D}_l)\}$ is given by

$$P\left(\mathscr{D}_j\right) = \sum_{\{(k_1, \ldots, k_l) : k_j = 1\}} P\left(\mathscr{C}_{k_1, \ldots, k_l}\right).$$

A more practical expression for the finite measure $\nu(\cdot)$ is $\gamma P_0(\cdot) = \nu(\cdot)$, where $\gamma = \nu(\mathscr{P}) > 0$ is a constant termed the concentration parameter and $P_0(\cdot)$ is a probability distribution termed the baseline distribution. These terms can be understood from the expectation and variance of the Dirichlet distribution (3.2), which yield, for any set $\mathscr{C} \in \mathscr{T}$,

$$\begin{aligned} \mathrm{E}\{P(\mathscr{C})\} &= P_0(\mathscr{C}), \\ \mathrm{var}\{P(\mathscr{C})\} &= \frac{P_0(\mathscr{C})\{1 - P_0(\mathscr{C})\}}{\gamma + 1}. \end{aligned} \tag{3.4}$$

The baseline distribution $P_0(\cdot)$ can thus be interpreted as the prior belief for $P(\cdot)$, and the concentration parameter γ as the assurance we have in this prior belief; the larger the value of γ, the stronger the confidence.

In analogy with the Bayesian update of the Dirichlet distribution as a prior distribution for multinomial data in (3.3), Ferguson shows that if X is a sample from the Dirichlet process

$P(\cdot) = \gamma P_0(\cdot)$, then $P(\cdot \mid X)$ is the updated Dirichlet process $\mathrm{DP}\{\gamma P_0(\cdot) + \delta_X(\cdot)\}$, where $\delta_x(\cdot)$ is the measure on $(\mathscr{P}, \mathscr{T})$ such that $\delta_x(\mathscr{C}) = \mathbb{1}(x \in \mathscr{C})$, for any $\mathscr{C} \in \mathscr{T}$.

3.1.3 Extension of the Pólya urn scheme

Blackwell and MacQueen (1973) give another definition of the Dirichlet process based on a generalisation of Pólya urn schemes using a continuum of colours.

Definition 3.2 (Pólya sequence)
The sequence (X_n) *of random variables taking values in* $\mathscr{P}$ *is a Pólya sequence with parameter* $\nu(\cdot)$ *if for every* $\mathscr{C} \in \mathscr{T}$, $\Pr(X_1 \in \mathscr{C}) = \nu(\mathscr{C})/\nu(\mathscr{P}) = P_0(\mathscr{C})$, *and*

$$\Pr(X_{n+1} \in \mathscr{C} \mid X_1, \ldots, X_n\} = \nu_n(\mathscr{C})/\nu_n(\mathscr{P}),$$

with $\nu_n(\cdot) = \nu(\cdot) + \sum_{i=1}^{n} \delta_{X_i}(\cdot)$.

For finite $\mathscr{P}$, this definition mimics the process of drawing a ball from an urn initially containing $\nu(x)$ balls of colour x and putting the ball drawn back into the urn with an additional ball of the same colour. By extending this to the continuous setting as in Definition 3.2, Blackwell and MacQueen show that $\nu_n(\cdot)/\nu_n(\mathscr{P})$ converges with probability 1 to a discrete distribution $P(\cdot)$ and $P \sim \mathrm{DP}(\nu)$. They also show that given $P(\cdot)$, the variables $X_1, \ldots, X_n$ are independent and

$$X_1, \ldots, X_n \mid P \sim P.$$

As we shall see in Section 3.2, this is one of the building blocks of the Dirichlet process mixture model.

3.1.4 Constructive definition

Similarly to the construction of a Dirichlet distribution using gamma-distributed random variables described in Section 3.1.1, Ferguson (1973) introduces an alternative definition of the Dirichlet process through a gamma process with independent increments. A more intuitive approach is presented by Sethuraman (1994), who shows that a Dirichlet process with measure $\gamma P_0(\cdot)$ can be represented as

$$P(\cdot) = \sum_{c=1}^{\infty} w_c \delta_{X_c}(\cdot), \tag{3.5}$$

where the weights w_c are constructed using the stick-breaking process as follows,

$$\begin{aligned} w_1 &= V_1, & V_1 &\sim \mathrm{Beta}(1, \gamma), \\ w_c &= V_c \times \prod_{k=1}^{c-1} (1 - V_k), & V_c &\overset{\text{iid}}{\sim} \mathrm{Beta}(1, \gamma), \quad c = 2, 3, \ldots, \end{aligned} \tag{3.6}$$

where the V_c are mutually independent and independent of the X_c, which are independent and identically P_0-distributed. The process outlined in (3.6) corresponds to breaking a stick of

length 1 to get w_1, then breaking the remainder of the stick to get w_2, and so on. Notice that $\sum_{c=1}^{\infty} w_c = 1$, since

$$\sum_{c=1}^{N} w_c = V_1 + \sum_{c=2}^{N} V_c \prod_{k=1}^{c-1}(1-V_k) = 1 - \prod_{c=1}^{N}(1-V_c) \xrightarrow{\Pr} 1, \quad N \to \infty,$$

where the convergence holds with $\mathscr{P}$-probability 1, and the last equality is obtained using a simple recursion argument.

The very simple representation offered by the stick-breaking process is widely used in the conditional approach to fitting Dirichlet process mixtures, as we shall see in Section 3.3.2.

3.2 The Dirichlet process mixture

Using different approaches, Blackwell (1973) and Ferguson (1973) show that the Dirichlet process introduced in Sections 3.1.2 and 3.1.3 almost surely has a discrete measure, and this property appears explicitly in Sethuraman's representation (3.5). Although the Dirichlet process is a very flexible Bayesian approach to nonparametric problems, its discreteness limits the range of applications in practice.

Antoniak (1974) introduces the Dirichlet process mixture model, which builds on the attractive properties of the Dirichlet process and enriches the class of problems to which it can be applied. This model features a mixture of a countably infinite number of distributions, where the Dirichlet process is used as a mixing distribution for the parameters of the mixture distributions. Given observations $X_1, \ldots, X_n$, the model is specified as

$$\begin{aligned} X_i \mid \theta_i &\overset{\text{ind}}{\sim} F(\theta_i), && i = 1, \ldots, n, \\ \theta_i \mid P &\overset{\text{iid}}{\sim} P, && i = 1, \ldots, n \\ P &\sim \text{DP}(\gamma P_0), \end{aligned} \tag{3.7}$$

where $F(\theta)$ belongs to a family of distributions indexed by a parameter θ, and conditional independence is assumed to hold for the X_i given the θ_i and for the θ_i given P.

The introduction of an auxiliary variable which plays the role of an indicator variable simplifies (3.7) and helps understand the clustering nature of the Dirichlet process mixture model (Müller *et al.*, 1996; MacEachern, 1994), namely taking the limit as $N \to \infty$ of

$$\begin{aligned} X_i \mid c_i, (\theta_1, \ldots, \theta_N) &\overset{\text{ind}}{\sim} F(\theta_{c_i}), && i = 1, \ldots, n, \\ c_i \mid (w_1, \ldots, w_N) &\overset{\text{ind}}{\sim} \text{Mult}(w_1, \ldots, w_N), && i = 1, \ldots, n, \\ \theta_c &\overset{\text{iid}}{\sim} P_0, && c = 1, \ldots, N, \\ (w_1, \ldots, w_N) &\sim \text{Dir}(\gamma/N, \ldots, \gamma/N), \end{aligned} \tag{3.8}$$

where $\text{Mult}(w_1,\ldots,w_N)$ denotes the multinomial distribution, with $\sum_{c=1}^{N} w_c = 1$, and the c_i are the indicator variables, so that if $c_i = c_j$, X_i and X_j share the same parameter θ and thus belong to the same component in the mixture; the weights $w_1,\ldots,w_N$ have a symmetric Dirichlet prior distribution with concentration parameter γ/N approaching 0 as $N \to \infty$.

Theoretical results establish the flexibility and the suitability of the Dirichlet process mixture model in estimating a broad class of densities. Using various metrics on densities, and under mild conditions, Ghosal *et al.* (1999) and Barron *et al.* (1999) develop consistency results that guarantee that the posterior density of the Dirichlet process mixture can arbitrarily closely approximate any density lying in the support of the Dirichlet process mixture prior. Ghosal *et al.* (2000) derive convergence rates for infinite-dimensional Bayesian models, and in particular for Dirichlet process mixture models. Good references for Dirichlet process mixture models and Bayesian nonparametrics in general are Dey *et al.* (1998), Gosh and Ramamoorthi (2003), Hjort *et al.* (2010), and Ghosal and van der Vaart (2017).

3.3 Algorithms

3.3.1 Marginal approach

The first approach to fitting a Dirichlet process mixture model is based on the generalised Pólya urn scheme of Blackwell and MacQueen (1973), which gives the joint distribution of the θ_i in (3.7) as the product of the conditional distributions

$$\theta_i \mid \theta_1,\ldots,\theta_{i-1} \sim \frac{\gamma}{\gamma+i-1} P_0 + \frac{1}{\gamma+i-1} \sum_{j=1}^{i-1} \delta_{\theta_j}.$$

This yields the following conditional prior probabilities for the indicator variables c_i introduced in the model (3.8), integrating over the weights $w_1,\ldots,w_N$ and taking the limit as $N \to \infty$,

$$\Pr(c_i = c \mid c_1,\ldots,c_{i-1}) \to \frac{n_{-i,c}}{n-1+\gamma}, \quad c = c_j,\ j = 1,\ldots,i-1, \tag{3.9}$$

$$\Pr\big(c_i \neq c_j,\ j \neq i,\ j = 1,\ldots,n \mid \boldsymbol{c}_{-i}\big) \to \frac{\gamma}{n-1+\gamma}, \tag{3.10}$$

where $n_{-i,c} = \sum_{j=1}^{i-1} \mathbb{1}(c_j = c)$ corresponds to the number of observations among $X_1,\ldots,X_{i-1}$ assigned to component c. Given a state where all n observations are allocated to several components, a new observation X_{n+1} has thus more probability to join a component with many observations assigned to it than a component with few observations assigned to it, but there is a non-null probability $\gamma/(n-1+\gamma)$ that X_{n+1} will create a new singleton component. This is another perspective of the role of the precision parameter γ, in addition to its importance in the variance (3.4).

A lot of research has been done to develop efficient algorithms to sample from the posterior distribution of the Dirichlet process mixture, including MacEachern and Müller (1998) and Escobar and West (1998). Neal (2000) reviews these techniques and develops more advanced methods with the objective of improving the mixing of the Markov chain Monte Carlo procedure.

We now briefly present here Algorithms 5 and 8 of Neal (2000), more details being available in Appendix A. We begin with the description of a fundamental sampling procedure in Bayesian statistics. In the general case when the priors are non-conjugate, the Metropolis–Hastings (1970) algorithm provides a rejection procedure from which posterior samples are drawn. Given a current state θ, a proposal density $q(\cdot \mid \cdot)$ and a posterior density $\pi(\cdot \mid x)$ conditionally on observing x, the procedure consists of the following steps: first, sample a candidate $\theta^\star$ from $q(\cdot \mid \theta)$; second, compute the acceptance ratio

$$a\left(\theta^\star, \theta\right) = \min\left\{1, \frac{q(\theta \mid \theta^\star)\pi(\theta^\star \mid x)}{q(\theta^\star \mid \theta)\pi(\theta \mid x)}\right\}; \tag{3.11}$$

finally, assign $\theta^\star$ to the new state with probability $a(\theta^\star, \theta)$, otherwise assign θ to the new state. In practice, $\pi(\cdot \mid x)$ is often unknown, and the product of the prior density $\pi(\theta)$ and the density $f(x \mid \theta)$ is used instead, as the norming constant cancels in the acceptance ratio (3.11).

In the context of the Dirichlet process mixture, we can simplify the acceptance probability of the indicator variable by using the limiting conditional prior probabilities (3.9) and (3.10) as the proposal distribution. Instead of using a standard proposal distribution $q(\cdot \mid c_i)$ to draw a new candidate $c_i^\star$ given the current state c_i, the proposal distribution is of the form $q(\cdot \mid \boldsymbol{c}_{-i})$, considering the ith observation as being the last in (3.9) and (3.10), giving

$$\begin{aligned} a\left(c_i^\star, c_i\right) &= \min\left\{1, \frac{q(c_i \mid \boldsymbol{c}_{-i}) f(x_i \mid \theta_{c_i^\star})\pi(c_i^\star \mid \boldsymbol{c}_{-i})}{q(c_i^\star \mid \boldsymbol{c}_{-i}) f(x_i \mid \theta_{c_i})\pi(c_i \mid \boldsymbol{c}_{-i})}\right\} \\ &= \min\left\{1, \frac{f(x_i \mid \theta_{c_i^\star})}{f(x_i \mid \theta_{c_i})}\right\}, \qquad i = 1, \ldots, n, \end{aligned} \tag{3.12}$$

as the prior and the proposal distributions cancel. This acceptance probability can be used to sweep through the n observations and update the indicator variables according to the Metropolis–Hastings scheme. Theoretically, we would not need to update the θ_i, as newly-created components naturally introduce new values of θ, but much better mixing of the Markov chain is achieved by implementing an update of the mixture parameters, along the lines of Algorithm A.1 in Appendix A.

In order to improve mixing further, Neal (2000) introduces auxiliary variables in the form of multiple empty component candidates in Algorithm 8. For some positive integer $m \geq 1$, m new components are drawn each time before a new candidate $c_i^\star$ is sampled, increasing the chances of effectively creating a new component. Since the ordering of the observations is arbitrary and the c_i are exchangeable, we can assume that for a fixed c_i, the $c_j \neq c_i$ take values

in $\{1,\dots,k_{-i}\}$ and have conditional prior probabilities

$$\Pr(c_i = c \mid \boldsymbol{c}_{-i}) = \frac{n_{-i,c}}{n-1+\gamma}, \quad i = 1,\dots,n,\ 1 \le c \le k_{-i}, \tag{3.13}$$

as we can use the probability (3.9) and consider c_i as being the last of $c_1,\dots,c_n$, since these are exchangeable. The m new components are indexed by $k_{-i}+1,\dots,k_{-i}+m$ and have the conditional prior probability (3.10), which is split between the m new components so that

$$\Pr(c_i = c \mid \boldsymbol{c}_{-i}) = \frac{\gamma/m}{n-1+\gamma}, \quad k_{-i} < c \le k_{-i}+m. \tag{3.14}$$

From the conditional prior probabilities (3.13) and (3.14), the posterior probability given observation x_i follows from the Bayes rule, namely

$$\begin{aligned}
&\Pr\left(c_i = c \mid \boldsymbol{c}_{-i}, x_i, \theta_1,\dots,\theta_{k_{-i}+m}\right) \\
&\qquad \propto \Pr\left(x_i \,\middle|\, c_i = c, \boldsymbol{c}_{-i}, \theta_1,\dots,\theta_{k_{-i}+m}\right) \times \Pr\left(c_i = c \,\middle|\, \boldsymbol{c}_{-i}, \theta_1,\dots,\theta_{k_{-i}+m}\right) \\
&\qquad = \Pr\left(x_i \mid \theta_c\right) \times \Pr\left(c_i = c \mid \boldsymbol{c}_{-i}\right),
\end{aligned}$$

which is specified as

$$\Pr\left(c_i = c \,\middle|\, \boldsymbol{c}_{-i}, x_i, \theta_1,\dots,\theta_{k_{-i}+m}\right) = \begin{cases} \kappa \dfrac{n_{-i,c}}{n-1+\gamma} f\left(x_i \mid \theta_c\right), & 1 \le c \le k_{-i} \\ \kappa \dfrac{\gamma/m}{n-1+\gamma} f\left(x_i \mid \theta_c\right), & k_{-i} < c \le k_{-i}+m, \end{cases} \tag{3.15}$$

where $\kappa > 0$ is a normalising constant ensuring that the probabilities sum to 1. This yields a Gibbs sampler, or equivalently a Metropolis–Hastings algorithm where the acceptance probability is constantly 1, and the choice of m can be viewed as a trade-off between computational cost and mixing. Algorithm A.2 in Appendix A provides more details.

3.3.2 Conditional approach

Another approach uses an approximation of the Dirichlet process closely related to (3.8) for finite N. Ishwaran and Zarepour (2000) and Ishwaran and James (2003) consider Sethuraman's representation (3.5) and approximate the series by a finite sum, so that

$$P(\cdot) = \sum_{c=1}^{N} w_c \delta_{X_c}(\cdot).$$

In practice, the value of N should not be interpreted as an upper bound for the number of components in the mixture, and should be set to a value much larger than the a priori expected number of components. The precise value of N generally does not impact on the inference, as long as it is large enough; in particular, Ishwaran and James (2001, 2002) show that the joint posterior distribution of the c_i rapidly converges to its limit as $N \to \infty$. According to these results, it is advisable to choose $N > 50$ for 1,000 observations and $\gamma = 3$. More details are

given in Section 5.6. If we believe that the number of components should be small, or large, then the concentration parameter γ should reflect this prior belief, as the larger γ, the more likely it is to have many components in the mixture, and vice versa.

The truncation of (3.5) is key in this approach, and conjugate priors can be used for sampling the finite number of parameters in the model. Specifically, a generalised Dirichlet distribution (Connor and Mosimann, 1969) corresponds to the prior distribution for the weights, for which the truncation implies $w_N = 1 - \prod_{c=1}^{N-1}(1 - V_c)$, in contrast with (3.6). Because the parameter space is finite, $(\theta_1, \ldots, \theta_N)$ are sampled in one block from the joint posterior; this may seem as a waste of computational time, since potentially most of the N components will remain empty during the sampling process, but the cost is bearable, as empty components only imply sampling from the prior distribution.

Extensions and improvements have been suggested, e.g., Papaspiliopoulos and Roberts (2008) develop a retrospective sampling scheme that does not need truncation of Sethuraman's representation (3.5), Walker (2007) introduces a latent variable that also does not involve approximation of (3.5) and Papaspiliopoulos (2008) brings these two approaches together in a better-mixing and computationally less expensive algorithm.

3.4 Example: 3-year bond yields

We now illustrate the three algorithms presented in Section 3.3 with data from the sovereign bond market. We consider 3-year bond yields across 53 countries and their respective rating published by Moody's. More details about the data are available in Appendix C. Our goal is to examine whether the Dirichlet process can capture any structure in the yields corresponding to Moody's ratings. We look at cross-sectional data at a given time point, since our objective is to illustrate the different Dirichlet process algorithms rather than give a complete study of the evolution of yield spreads and ratings across time.

Figure 3.1 shows different perspectives of the data; the default histogram produced by R (R Core Team, 2017) hides its clustered nature, and a thinner bin width gives more insight. We do not expect the rating of a country to be directly linked with its bond yield, as the rating does not cover all aspects that are important in explaining the yield, such as the lending rates of the central bank. Another reason partially explaining the difficulty in linking rating and yield is the time lag needed by the agency to update its rating. Mexico, for example, borrows money on a 3-year horizon at around 7% and is rated A. However, its exact rating is A3 with a negative perspective, meaning that it is likely to be attributed a Baa rating soon.

We run Algorithms 5 and 8 of Neal (2000), corresponding to the marginal approach to fitting a Dirichlet process mixture, and the blocked Gibbs algorithm of Ishwaran and James (2002), following the conditional approach. I implemented the code for all three algorithms, the last of which is available in my R package `tsxtreme`.

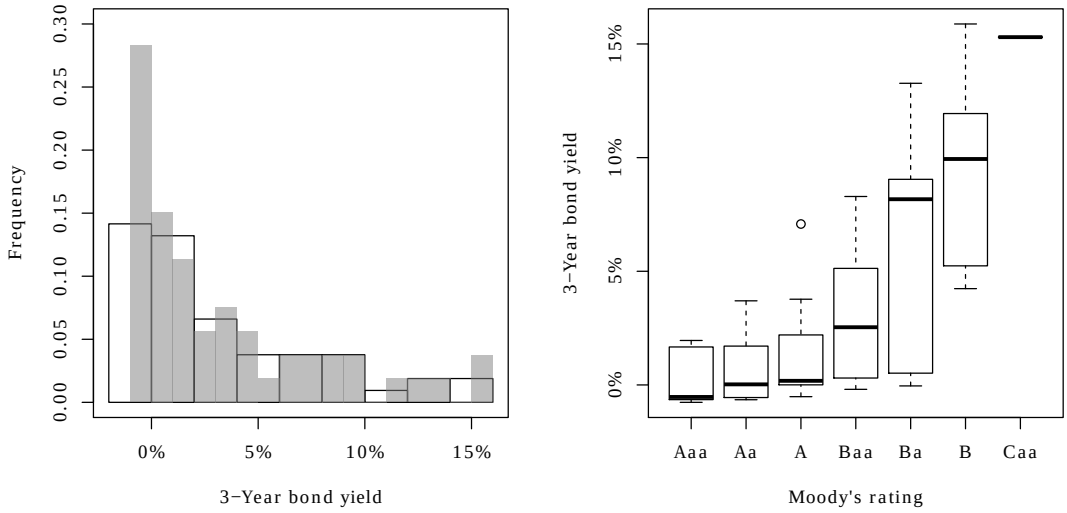

Figure 3.1 – Bond yields for 53 countries for a 3-year horizon. Left panel: histograms of the yields with bin width 2 (white) and 1 (grey). Right panel: boxplots of the yields stratified by rating; Ukraine is the only country with a Caa rating and is represented with a single line.

We use the same vague but well-defined priors for the three algorithms, in particular a gamma distribution with scale and shape equal to 2 for the hyperprior of the concentration parameter γ, thus having prior expectation equal to 4, and prior variance equal to 8. The update procedure for the concentration parameter in Algorithm 5 and Algorithm 8 is detailed in Appendix B. For Algorithm 5 of Neal, we update the c_i only once per sweep, corresponding to $R = 1$ in Algorithm A.1, thus not trying to reduce autocorrelation between consecutive samples. For Algorithm 8 of Neal, we use $m = 1$, meaning that we propose only one new component when $c_i = c_j$ for some j, or none if c_i belongs to a singleton. We compute 20,000 iterations after a burn-in period of 1,000 iterations.

The computing times for each algorithm were in the following proportions, compared to the conditional approach: Algorithm 5 takes about twice the time and Algorithm 8 takes more than eight times longer. Additional computing time comes with some benefits, namely a reduction in the autocorrelation of successive samples. A particular quantity of interest is the autocorrelation time (Robert and Casella, 2004, Chap. 12), which is defined for any function h evaluated at posterior samples $\theta^{(1)}, \ldots$, specifically

$$\tau(h) = 1 + 2\sum_{t=2}^{\infty} \operatorname{corr}\left\{h\left(\theta^{(1)}\right), h\left(\theta^{(t)}\right)\right\},$$

which can be estimated by truncation of the infinite sum in practice. A more explicit measure of the information carried by n consecutive posterior samples $\theta^{(1)}, \ldots, \theta^{(n)}$ from a Markov chain, compared to independent posterior samples, is the effective sample size (Kish, 1965,

Chap. 8; Kass *et al.*, 1998)

$$\mathrm{ESS}(n) = \frac{n}{\tau(h)}.$$

Another definition is (Gong and Flegal, 2016)

$$\mathrm{ESS}(n) = n\frac{\lambda^2}{\sigma^2},$$

where σ^2 stands for the variance of the mean of the dependent samples $h(\theta^{(1)}),\ldots,h(\theta^{(n)})$, typically estimated with batch means (Jones *et al.*, 2006; Geyer, 2011), and λ^2 is the posterior variance of $h(\theta^{(1)}),\ldots,h(\theta^{(n)})$, for which a standard estimator can be used.

Because of the lack of identifiability of the components in the Dirichlet process mixture, the Markov chains for the means and variances cannot be considered for computing an estimate of the effective sample size. The precision parameter is a better candidate for this purpose. We use the R routine of Flegal *et al.* (2017) to compute effective sample sizes on the three algorithms mentioned above, plus variants of Algorithms 5 and 8. These variants consist in setting $R = 5$ for Algorithm 5, i.e., we compute a nested Markov chain of length 5 for the c_i and only store the last sample; for Algorithm 8, we set $m = 5$, i.e., we create 4 or 5 new candidate components at each iteration, depending on whether c_i or not corresponds to a singleton. Both variants improve the mixing of the output compared to the base cases where $R = 1$ or $m = 1$, either by repeated updates of the component assignments of which 1 in 5 is saved, or by proposing more than one new component at each update. We standardise the effective sample sizes and computing times and plot them in Figure 3.2 for comparison. It appears that the blocked Gibbs algorithm used for the conditional approach is best in terms of computing time; the algorithm can overcome relatively poor mixing by computing longer chains. For a given computational time available, with the current implementation, the conditional approach seems to be the best choice, along with Algorithm 5 with $R = 1$.

Going back to our sovereign bond data, we plot the distribution of the distributions fitted to the 3-year yield, for which the three algorithms give similar outputs. An example with the conditional approach is shown in Figure 3.3, where sample densities are compared to a histogram of the data and shows how bumps in the tail are well-captured by the algorithm. The distribution of the yields with a pointwise 90% coverage interval is also constructed from 1,000 posterior distribution samples and shows a good fit of the empirical distribution.

A feature of interest of the Dirichlet process mixture is the distribution of the number of components containing at least one observation (right panel of Figure 3.4). The conditional approach and Algorithm 8 give similar results and higher probabilities for larger numbers of non-empty components than Algorithm 5. We do not expect the algorithms to capture the structure in terms of Moody's ratings, but we observe that the respective posterior medians (8,7,6) are close to the 7 grades in the sample.

The Bayesian approach to fitting a mixture of distributions provides a broad range of statistics that can be calculated from posterior traces. Figure 3.4 shows (left panel) an example

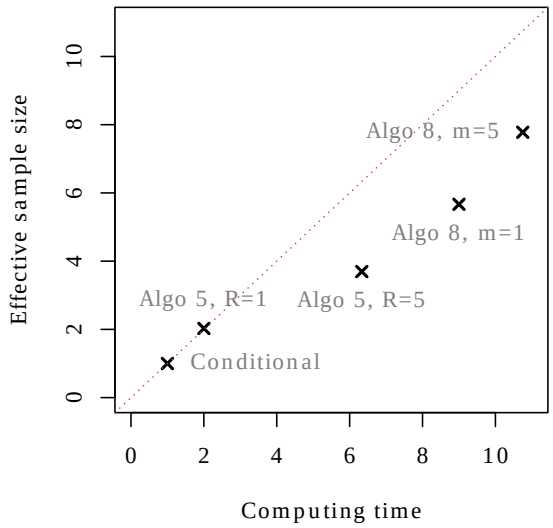

Figure 3.2 – Efficiency of algorithms measured as units of time versus units of effective sample size. The conditional approach ('Conditional') is set to one unit of time and one unit of effective sample size; the two algorithms ('Algo 5' and 'Algo 8') using the marginal approach with different settings are benchmarked against the conditional approach using these standardised units. The dotted line represents effective sample sizes that are reachable with the blocked Gibbs sampler of the conditional approach; any cross below this line represents an algorithm that produces a smaller number of effective samples by unit of time than the conditional approach.

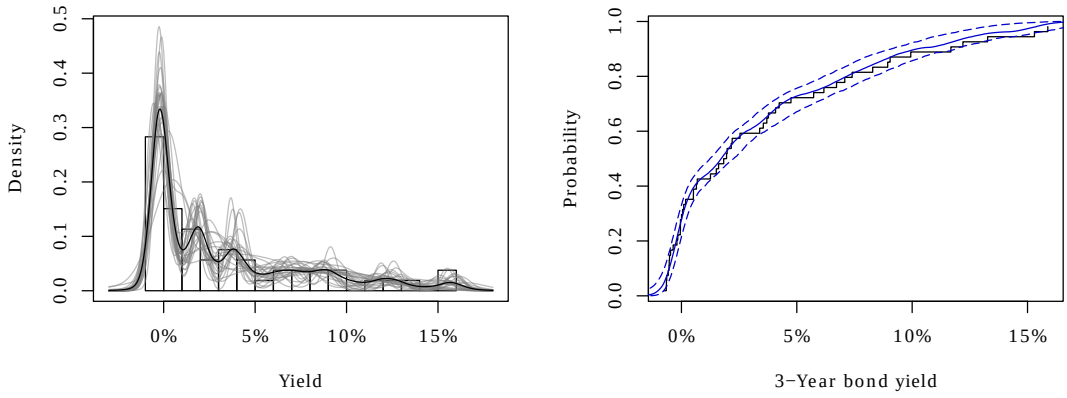

Figure 3.3 – Conditional approach to fitting the Dirichlet process. Left panel: 25 sample densities and their pointwise mean density superimposed on a histogram of the bond yield data. Right panel: pointwise median (solid, blue), 5% and 95% quantiles (dashed, blue) with empirical distribution (black).

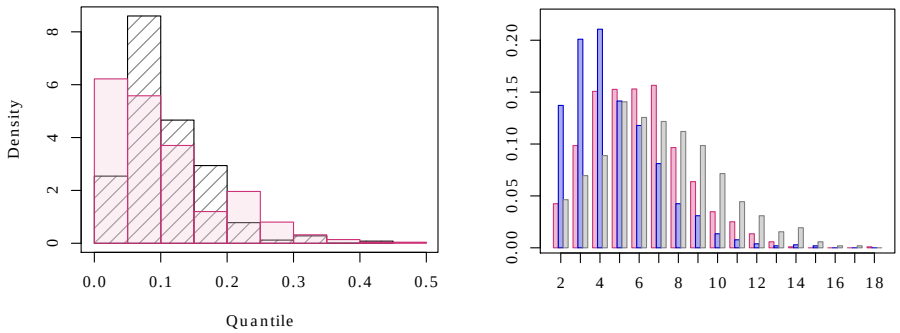

Figure 3.4 – Posterior features of the Dirichlet process. Left panel: posterior distribution of the quantile (percentage) for the Swiss yield in its component given that there are seven non-empty components in the mixture; the output from the conditional approach (hatched bins) is compared with that of Algorithm 8 (light red). Right panel: posterior number of non-empty components in the mixture using the same colour code, with Algorithm 5 superimposed (blue).

of this richness, when we are interested in how Switzerland, rated Aaa by Moody's, positions itself among countries with similar yields. In this figure, we condition on the mixture having seven non-empty components and we compute the posterior quantile of Switzerland's yield in the component it belongs to. Algorithms 5 and 8 give similar results, and only the latter is displayed, along with the conditional approach. Since there are seven countries rated Aaa and Switzerland is the least generous borrower among them, an empirical estimate would be 12.5%, while all three algorithms seem to indicate the range 5% to 10% as more likely, which we can explain since many countries rated Aa and A enjoy similar 3-year yields as countries rated Aaa, thus having a non-negligible probability to belong to the same mixture component as Switzerland.

3.5 Summary

In this chapter, we reviewed several definitions and properties of the Dirichlet process. Because Dirichlet processes are not appropriate to fitting continuous distributions, we have presented an extension in the form of Dirichlet process mixtures, which are very flexible and powerful tools for fitting density functions with unknown shapes, including multimodal densities.

We then presented two different kinds of approaches to make inference on Dirichlet process mixtures. The marginal approach uses a type of Pólya urn scheme to assign an observation to a component in the mixture, while the conditional approach approximates the Dirichlet process so as to deal with a finite sample space, thus allowing successive conditional updates of the model parameters. An illustration with real data highlighted the performances of algorithms stemming from the marginal and conditional approaches to fitting Dirichlet

process mixtures, and showed how the output from these algorithms can be used to answer a very broad class of questions.

If we can take advantage of this Bayesian nonparametric framework in the context of the conditional tail model introduced in Section 2.4.2 of Chapter 2, then we may be able to derive an efficient methodology for modelling extremes. This will be the topic of Chapter 5, but we shall first look into the penultimate properties of the conditional tail model, which the next chapter is about.

4 Penultimate analysis of the conditional tail model

4.1 Related research

4.1.1 Univariate case

The founding theorem in the theory of extreme values that we introduced as Theorem 2.1 in Chapter 2 characterises the extreme value distribution $G(x)$ which arises as the limit of the distribution of suitably normalised independent maxima $F^n(a_n x + b_n)$. This asymptotic distribution $G(x)$ is useful in practice, i.e., when a finite amount of data is available, and usually, the tail of $F^n(x)$ is identified to its limit distribution and inference is conducted using $G\{(x-b_n)/a_n\}$, with a_n and b_n two parameters to be estimated.

Fisher and Tippett (1928) raised the question of the accuracy of this approximation in the Gaussian case, i.e., $F(x) = \Phi(x)$, for which they prove that the limit distribution $G(x)$ is Gumbel. They show experimentally that $\Phi^n(a_n x + b_n)$ would be better approximated at finite levels by a distribution belonging to the Weibull class, corresponding in our notation to a generalised extreme value distribution with shape parameter $\xi < 0$. The default approximation of $F^n(a_n x + b_n)$ by $G(x)$ in extreme value applications is thus of great concern when the convergence of the former towards the latter is particularly slow.

The study of rates of convergence of $|F^n(a_n x + b_n) - G_\xi(x)|$ towards 0, where we emphasise that $G(x) = G_\xi(x)$ is parametrised by ξ, has a long history. The first attempts at characterising this rate of convergence for any value of ξ date to Gomes (1984, 1994) and an unpublished report by Smith (1987). Smith considers convergence in Hellinger distance, which implies uniform convergence of the distribution functions, and for this he assumes existence of the density $f(x) = F'(x)$ and its derivative $f'(x)$. If $F(\cdot)$ has support on $[x_F, x^F]$, we can write

$$-\log F(x) = \exp\left\{-\int_{x_F}^{x} \frac{dt}{\tilde{h}(t)}\right\}, \quad x \in [x_F, x^F],$$

where $\tilde{h}(x) = -F(x)\log F(x)/f(x)$; it is convenient to use the equivalent characterisation for x large, namely

$$\overline{F}(x) = \exp\left\{-\int_{x_F}^{x} \frac{dt}{h(t)}\right\}, \quad x \to x^F,$$

where $h(x) = \{1-F(x)\}/f(x)$ is the hazard function, and $\overline{F}(x) = 1-F(x)$ is the survival distribution function.

Assume

$$\lim_{x\to x^F} h'(x) = \xi, \quad \xi \in \mathbb{R},$$

which is one of the von Mises (1936) conditions and is necessary for $F(\cdot)$ to be in the domain of attraction of an extreme value distribution. Pickands (1986) shows that this condition is necessary and sufficient for convergence of F^n in Theorem 2.1, and related convergences for f and f'. For $X \sim F$, Smith elegantly shows that

$$\Pr(X > u + xh(u) \mid X > u) = \frac{1-F\{u+xh(u)\}}{1-F(u)} = \left\{1+xh'(y)\right\}^{-1/h'(y)}, \tag{4.1}$$

for some fixed $y \in [u, u+sh(u)]$. Substituting $b_n = u$, $a_n = h(b_n)$ and $1-F(b_n) = 1/n$ in (4.1), this yields

$$-\log F^n(a_n x + b_n) \sim n\{1-F(a_n x+b_n)\} \sim (1+x\xi_n)^{-1/\xi_n}, \quad n\to\infty,$$

or equivalently

$$F^n(a_n x + b_n) \sim \exp\left\{-(1+x\xi_n)^{-1/\xi_n}\right\}, \quad n\to\infty,$$

where Smith defines $\xi_n = h(b_n)$, effectively choosing $y = u$ in (4.1). The details of this development are explained in Appendix D.

Using the above expressions for a_n, b_n and ξ_n, Gomes and Pestana (1987) and Gomes (1994) show that

$$F^n(a_n x + b_n) - G_\xi(x) = O(\xi_n - \xi), \quad n\to\infty, \tag{4.2}$$

and give the structure of the remainder term on the right-hand side for a broad class of distribution functions $F(\cdot)$, including the Gaussian distribution (Anderson, 1971). For this class, Gomes (1994) also gives

$$F^n(a_n x + b_n) - G_{\xi_n}(x) = O\{(\xi_n - \xi)^2\}, \tag{4.3}$$

thus showing that the subasymptotic approximation for the shape parameter greatly improves the rate of convergence.

In order to illustrate the properties stated above in equations (4.2) and (4.3), consider the standard Gaussian distribution with $F(\cdot) = \Phi(\cdot)$ and $f(\cdot) = \varphi(\cdot)$. Mills' ratio can be used to derive the approximation to the survival distribution function $\overline{F}(x) \approx \varphi(x)\left\{1/x - 1/x^3\right\}$, from which we get the approximate hazard function $h(x) \approx 1/x - 1/x^3$ and its derivative

$h'(x) \approx -1/x^2 + 3/x^4$. We also know (Leadbetter *et al.*, 1983, p. 14) that $b_n = \sqrt{2\log n} + o(1)$ in the Gaussian case, so that $\xi_n = h'(b_n) = -1/(2\log n) + O\{1/(\log n)^2\}$. The convergence rate (4.2) is $O(1/\log n)$ using the limit shape parameter, and it improves to $O\{1/(\log n)^2\}$ when using ξ_n instead of ξ, as in (4.3). In this development, we observe that the penultimate shape parameter is negative for any finite n, which is consistent with the analysis by Fisher and Tippett (1928) of a Weibull-type distribution yielding better approximations at finite levels.

4.2 Bivariate case

4.2.1 Componentwise maxima

Marshall and Olkin (1983) derive the general form of the norming sequences $(a_n) > 0$ and (b_n) in the bivariate context, extending the univariate results of Gnedenko (1943). They also give examples of bivariate distributions and show whether they belong to a domain of attraction.

Only a few studies have focused on the penultimate properties of bivariate maxima. Among these are Bofinger and Bofinger (1965), who derive the correlation of componentwise maxima in samples with $n = 2, \ldots, 50$ replicates from a bivariate Gaussian distribution. This study gives the penultimate form of this correlation, which is more subtle than the limit correlation, which we know from Sibuya (1960) is always zero. In a subsequent article, Bofinger (1970) extends this analysis to non-Gaussian bivariate distributions, in particular the bivariate Gamma and Morgenstern (1956) distributions. This analysis sheds light on the form of the penultimate correlation for small n, in contrast with the asymptotic correlation being always zero, e.g., for the Morgenstern distribution, whatever the dependence at finite levels.

In a different context, Ledford and Tawn (1996) use penultimate properties of $\Pr(X^{\mathrm{F}} > z, Y^{\mathrm{F}} > z)$ on Fréchet margins to derive their model, which we introduced in Section 2.3.2. A better understanding of the subasymptotic tail behaviour yields a model that smoothly connects perfect dependence and complete independence.

4.2.2 Conditional extremes

The conditional model for extremes covers a broad class of extremal dependence structures, and in particular most of the parametric models of Section 4.2.1. It is also interesting in that it can capture positive association for random variables that are asymptotically independent, in contrast with standard theory for multivariate extremes in which asymptotic independence can only arise under complete independence.

In our analysis, we focus on the bivariate case, as it yields clearer notation, and extension to the general multivariate case is straightforward when conditional independence holds in the limit. We consider marginally Laplace-distributed (X, Y) and look at pairs for which X exceeds a high threshold u.

The conditional model for extreme values (Heffernan and Tawn, 2004) introduced in Section 2.4.2 is based on the assumption that there exist norming functions $a(\cdot)$ and $b(\cdot) > 0$ such that

$$\lim_{x\to\infty} \Pr\left\{ \frac{Y - a(X)}{b(X)} \leq z \,\middle|\, X = x \right\} = H(z), \tag{4.4}$$

where the distribution $H(\cdot)$ is non-degenerate and has no mass at ∞.

Recall from Section 2.4.2 that Heffernan and Resnick (2007) consider an equivalent setup with $a^{\mathrm{HR}}(\cdot)$ and $b^{\mathrm{HR}}(\cdot) > 0$ that are deterministic functions of the modelling threshold u, that is,

$$\lim_{u\to\infty} \Pr\left\{ \frac{\tilde{Y} - a^{\mathrm{HR}}(u)}{b^{\mathrm{HR}}(u)} \leq z \,\middle|\, X^{\mathrm{F}} > u \right\} = \mu([-\infty, z] \times (1, \infty]) =: H^{\star}(z),$$

where $H^{\star}(\cdot)$ is a non-degenerate distribution function and the marginal distribution of $\tilde{Y}$ is unknown. We do not consider this formulation here, as in practice we have information about the exact values of the observations X^{F}, and this additional information guarantees factorisation of the measure $\mu(\cdot,\cdot)$, which is key for extrapolation of probability of extreme sets from the model (4.4).

By considering a broad class of parametric models, Heffernan and Tawn derive parametric forms for $a(\cdot)$ and $b(\cdot)$ that yield a parsimonious model and cover a broad range of extremal dependence structures not described by models arising from the standard theory for multivariate extremes. In our penultimate analysis, we interpret these parametric forms of $a(\cdot)$ and $b(\cdot)$ as the first order behaviour of the norming functions, and we write

$$\begin{aligned} a(x) &\sim \alpha x := a_0(x), \quad \alpha \in [-1, 1], \\ b(x) &\sim x^{\beta} := b_0(x), \quad \beta \in (-\infty, 1). \end{aligned} \tag{4.5}$$

Our goal is to characterise the behaviour of the remainder terms defined as

$$\begin{aligned} a(x) - a_0(x) &\sim r_a(x), \\ b(x) - b_0(x) &\sim r_b(x). \end{aligned}$$

We can now consider the second-order normalisation for $a(\cdot)$ and $b(\cdot)$, with

$$\begin{aligned} a_1(x) &= a_0(x) + r_a(x), \\ b_1(x) &= b_0(x) + r_b(x). \end{aligned} \tag{4.6}$$

With these penultimate forms, we are able to refine the normalisation of Y in (4.4), yielding the subasymptotic conditional distribution

$$\Pr\left\{ \frac{Y - a_1(x)}{b_1(x)} \leq z \,\middle|\, X = x \right\} = H_x(z), \quad x > u, \tag{4.7}$$

with $H_x(z) \to H(z)$ as $x \to \infty$.

Heffernan and Tawn (2004) give the rate of convergence of the conditional distribution for data arising from various copula models in terms of the order of convergence towards zero, as $x \to \infty$, of

$$\Pr\left\{ \frac{Y - a_0(X)}{b_0(X)} \leq z \;\middle|\; X = x \right\} - H(z), \tag{4.8}$$

with (X, Y) on the Gumbel scale, and transform the marginal scale in order to have $\Pr(X > x) = n^{-1}$, so that rates are invariant to the specific choice of marginal distribution.

We compute these rates on the Laplace scale and also use a marginal transform in order to have $\Pr(X > x) = n^{-1}$. We consider how much we can improve on these rates when using the penultimate norming, by studying the rate of convergence to zero of

$$\Pr\left\{ \frac{Y - a_1(X)}{b_1(X)} \leq z \;\middle|\; X = x \right\} - H(z). \tag{4.9}$$

We also want to quantify the subasymptotic remainder, using

$$\sup_{x>u} \left| \Pr\left\{ \frac{Y - a_1(x)}{b_1(x)} \leq z \;\middle|\; X = x \right\} - H_x(z) \right|, \tag{4.10}$$

along the lines of Gomes (1994) in the univariate context.

The subasymptotic norming functions $a_1(\cdot)$ and $b_1(\cdot)$ are of particular interest when conducting simulation studies, as the rate of convergence of (4.8) towards zero can be slow, so that estimates of $a_0(\cdot)$ and $b_0(\cdot)$ can misleadingly suggest a poor fit.

In the next sections, we consider three examples of copula structures. We first consider the Gaussian copula for which the convergence of (4.8) to zero was reported by Heffernan and Tawn to be the slowest in the examples they considered, namely $O\{\log(\log n)/(\log n)^{1/2}\}$; it is thus of great interest to derive the subasymptotic remainder size $r_a(x)$ and $r_b(x)$. Second, we consider another example with asymptotic independence; the inverted logistic dependence structure has a faster convergence rate than the Gaussian copula, and we shall see that it is harder to determine a subasymptotic behaviour. Heffernan and Tawn reported the rate to be $O(1/\log n)$ in this case. Third, the logistic dependence structure represents the case of asymptotic dependence, and this is also a situation where convergence is known to be fast, with $O(1/n)$.

4.3 Gaussian distribution

We now turn our attention to the first data structure, the bivariate Gaussian copula, and show that in this case, similarly to what was first observed by Fisher and Tippett (1928) in the univariate context, penultimate approximations of the normalising parameters differ significantly from their limit form.

Theorem 4.1
Let (V,W) have a bivariate standard normal distribution with correlation parameter $\rho \neq 0$ and let $(X,Y) = (V^L, W^L)$ be its marginal transform to the Laplace scale, with

$$X = \begin{cases} -\log 2\{1-\Phi(V)\}, & V > 0, \\ \log 2\Phi(V), & V \leq 0, \end{cases}$$

and similarly for Y as a function of W. Then, the ultimate and penultimate normings (4.5) *and* (4.6) *for $Y \mid X = x$, with x large, are*

$$\begin{aligned} a_0(x) &= \rho^2 x, & b_0(x) &= x^{1/2}, \\ a_1(x) &= \rho^2 x + \frac{(1-\rho^2)}{2}\log(x), & b_1(x) &= x^{1/2-1/(4x)}. \end{aligned}$$

The limit distribution $H(z)$ in (4.4) *is a centred Gaussian with variance $2\rho^2(1-\rho^2)$, and the penultimate distribution* (4.7) *is*

$$H_x(z) \sim \mathcal{N}\left\{0, 2\rho^2(1-\rho^2)\left(1+\frac{\log x}{2\sqrt{2\rho^2 x}}\right)^2\right\}.$$

If we write $n^{-1} = \Pr(X > u)$, the rates of convergence to the limit distribution are as follows: $O\{\log(\log n)/\sqrt{\log n}\}$ using the ultimate norming in (4.8)*, which is not improved using the penultimate norming in* (4.9)*. The subasymptotic remainder* (4.10) *behaves like $O(1/\sqrt{\log n})$.*

Proof The details of the proof for the asymptotic quantities $a_0(\cdot)$, $b_0(\cdot)$ and $H(z)$ can be found in Heffernan and Tawn (2004). We follow a similar path to derive the penultimate approximations and use Mill's ratio to get a tail approximation for v and x on the normal and Laplace scales, respectively. Since the Laplace distribution is symmetric, we focus on the right tail:

$$\begin{aligned} x &\sim -\log\left\{2\frac{\varphi(v)}{v}\right\} = -\log 2 + \log v + \frac{1}{2}\log(2\pi) + \frac{1}{2}v^2, \qquad (4.11) \\ v &\sim \sqrt{2x}, \end{aligned}$$

for large x and v. In order to get a second-order approximation for v, we define a small quantity $\varepsilon > 0$ and set $v \sim \sqrt{2x}(1+\varepsilon)$; plugging this back in (4.11), we get

$$\begin{aligned} x &\sim -\log 2 + \log\left\{\sqrt{2x}(1+\varepsilon)\right\} + \frac{1}{2}\log(2\pi) + \frac{1}{2}\left\{\sqrt{2x}(1+\varepsilon)\right\}^2 \\ &\sim -\log 2 + \frac{1}{2}\log\left(\sqrt{2x}\right) + \frac{1}{2}\log(2\pi) + x + 2x\varepsilon \qquad (4.12) \\ \implies v &= \sqrt{2x} + \frac{2\log 2 - \log(2x) - \log(2\pi)}{2\sqrt{2x}} + O\left\{\frac{(\log x)^2}{x^{3/2}}\right\}. \qquad (4.13) \end{aligned}$$

We want the conditional distribution of $W \mid V = v$ to be well-behaved in its upper tail when $v \to \infty$. We have $\Pr(W - \rho V \le z \mid V = v) = \Phi(z/\sqrt{1-\rho^2})$. On the Laplace scale, W is transformed to

$$Y = a_1(x) + b_1(x)Z, \tag{4.14}$$

with $a_1(x)$ and $b_1(x) > 0$ norming functions to be determined, and Z a random variable with a fixed distribution non-degenerate at $+\infty$. We derive these penultimate norming functions by writing $a_1(x) = a_0(x)(1+\varepsilon) = \rho^2 x(1+\varepsilon)$ in (4.14), and using (4.13). We get the second-order approximation by first expanding

$$\begin{aligned} W &\sim \sqrt{2\rho^2 x(1+\varepsilon) + 2b(x)Z} - \frac{\log\pi + \log\{\rho^2 x(1+\varepsilon) + b(x)Z\}}{2\sqrt{2\rho^2 x(1+\varepsilon) + 2b(x)Z}} \\ &\sim \rho\sqrt{2x(1+\varepsilon)}\left\{1 + \frac{b(x)Z}{2\rho^2 x(1+\varepsilon)}\right\} \\ &\quad - \left[\log\pi + \log\left\{\rho^2 x(1+\varepsilon)\right\} + \frac{b(x)Z}{\rho^2 x(1+\varepsilon)}\right] \frac{1}{2\sqrt{2\rho^2 x(1+\varepsilon)}}\left\{1 - \frac{b(x)Z}{2\rho^2 x(1+\varepsilon)}\right\}. \end{aligned} \tag{4.15}$$

We can use this expression for W in $W - \rho V \mid V = v$, in which we keep only the terms that are not functions of Z, as they are disconnected from $a_1(\cdot)$, and we get

$$\rho\sqrt{2x(1+\varepsilon)} - \frac{\log\left\{\rho^2 x(1+\varepsilon)\right\}}{2\sqrt{2\rho^2 x(1+\varepsilon)}} - \rho\sqrt{2x} + \rho\frac{\log x}{2\sqrt{2x}} + O\left(x^{-1/2}\right).$$

Expanding further, we arrive at

$$\rho\sqrt{2x}\left(1 + \frac{\varepsilon}{2}\right) - \frac{\log(\rho^2 x) + \varepsilon}{2\sqrt{2\rho^2 x}}\left(1 - \frac{\varepsilon}{2}\right) - \rho\sqrt{2x} + \rho\frac{\log x}{2\sqrt{2x}} + O\left(x^{-1/2}\right),$$

and cancellation of the leading term yields

$$\varepsilon = \frac{(1-\rho^2)\log x}{2\rho^2 x}, \tag{4.16}$$

or equivalently $a_1(x) = \rho^2 x + \frac{(1-\rho^2)}{2}\log x$.

The penultimate scale function $b_1(\cdot)$ stems from the Z-terms in (4.15), namely

$$\frac{b(x)Z}{\rho\sqrt{2x(1+\varepsilon)}} + \frac{\log\left\{\rho^2 x(1+\varepsilon)\right\} b(x)Z}{2\left\{2\rho^2 x(1+\varepsilon)\right\}^{3/2}} - \frac{b(x)Z}{\left\{2\rho^2 x(1+\varepsilon)\right\}^{3/2}}, \tag{4.17}$$

which we expand as follows,

$$\begin{aligned} b(x) &\sim \rho\sqrt{2x(1+\varepsilon)}\left[1+\frac{\log\{x(1+\varepsilon)\}}{4\rho^2 x(1+\varepsilon)}-\frac{1}{2\rho^2 x(1+\varepsilon)}\right]^{-1} \\ &\sim \rho\sqrt{2x}\left(1+\frac{\varepsilon}{2}\right)\left\{1-\frac{\log x}{4\rho^2 x}(1-\varepsilon)\right\} \\ &\sim \rho\sqrt{2x}-\frac{\log x}{2\rho\sqrt{2x}}+\rho\sqrt{2x}\frac{\varepsilon}{2}. \end{aligned}$$

Substituting ε into (4.16) gives

$$b(x) \sim \rho\sqrt{2x}-\frac{\log x}{2\rho\sqrt{2x}}+\frac{(1-\rho^2)\log x}{2\rho\sqrt{2x}} \sim x^{1/2-1/(4x)} = b_1(x).$$

We now compute the penultimate distribution $H_x(z)$ by substituting the expression for ε in (4.16), writing $A(x)+B(x)Z \sim H_x(z)$, with

$$\rho\sqrt{2x}\frac{(1-\rho^2)\log x}{4\rho^2 x}-\frac{\log(\rho^2 x)}{2\sqrt{2\rho^2 x}}-\frac{(1-\rho^2)\log x}{4\rho^2 x\sqrt{2\rho^2 x}}+\frac{(1-\rho^2)\log(\rho^2 x)\log x}{4\sqrt{2\rho^2 x}(2\rho^2 x)}+\rho\frac{\log x}{2\sqrt{2x}},$$

which equals

$$-\frac{\log(\rho)}{\sqrt{2\rho^2 x}}-\frac{(1-\rho^2)\log(x)}{2(2\rho^2 x)^{3/2}}+\frac{(1-\rho^2)(\log x)^2}{4(2\rho^2 x)^{3/2}}+\frac{(1-\rho^2)\log(\rho)}{2(2\rho^2 x)^{3/2}} \sim -\frac{\log(\rho)}{\sqrt{2\rho^2 x}} = A(x), \qquad (4.18)$$

and

$$\frac{b(x)Z}{\sqrt{2\rho^2 x}}\left\{(1-\varepsilon/2)+\frac{\log(\rho^2 x)+\varepsilon}{2\sqrt{2\rho^2 x}}(1-\varepsilon)-\frac{1-3\varepsilon/2}{2\rho^2 x}\right\},$$

which equals

$$\begin{aligned} \frac{b(x)Z}{\sqrt{2\rho^2 x}}&\left\{1-\frac{\varepsilon}{2}+\frac{\log x}{2\sqrt{2\rho^2 x}}(1-\varepsilon)+\frac{\varepsilon}{2\sqrt{2\rho^2 x}}+\frac{\log\rho}{\sqrt{2\rho^2 x}}(1-\varepsilon)-\frac{1}{2\rho^2 x}+O(\varepsilon^2)\right\} \\ &= Z+Z\frac{\log x}{2\sqrt{2\rho^2 x}}+Z\frac{\log\rho}{\sqrt{2\rho^2 x}}+Z\times O\left(\frac{\log x}{x}\right). \qquad (4.19) \end{aligned}$$

Taking the leading and penultimate terms in (4.18) and (4.19), which corresponds to setting $B(x) = 1+(8\rho^2 x)^{-1/2}\log x$, we find that $H_x(z) = H_x^{(1)}(z)$ is a centred normal distribution with variance

$$2\rho^2(1-\rho^2)\left(1+\frac{\log x}{2\sqrt{2\rho^2 x}}\right)^2,$$

i.e., it has larger variance than the asymptotic $H(z)$ found in Heffernan and Tawn (2004). If we consider the antepenultimate terms in (4.18) and (4.19), we get $H_x(z) = H_x^{(2)}(z)$ as

$$\mathcal{N}\left\{-\frac{\log(\rho)}{\sqrt{2\rho^2 x}}, 2\rho^2(1-\rho^2)\left(1+\frac{\log x + 2\log\rho}{2\sqrt{2\rho^2 x}}\right)^2\right\}.$$

We now give the order of convergence of (4.10) with the penultimate approximations $a_1(\cdot)$ and $b_1(\cdot)$ that we derived. After a marginal transform in order to get $n^{-1} = \Pr(X > u)$, we get the rate of convergence for $H_x^{(1)}(z)$, namely $O(1/\sqrt{\log n})$, which improves on the version of $H(z)$ with first-order approximation for the normalising functions, for which the rate of convergence is of order $\log\log n/\sqrt{\log n}$. Taking $H_x^{(2)}(z)$ improves even more on $H(z)$, as its rate of convergence in (4.10) is of order $O(\log\log n/\log n)$. ■

The penultimate norming $a_1(\cdot)$, $b_1(\cdot)$ can be used to assess the goodness-of-fit at a finite level. By replacing x by the threshold u in (4.16), we derive a second-order approximation for $\alpha = a_1(x)/x$ of the form

$$\alpha_1 = \rho^2 + \frac{(1-\rho^2)\log u}{2u}. \tag{4.20}$$

Similarly, we derive a second-order approximation for $\beta = \log\{b_1(x)\}/\log(x)$,

$$\beta_1 = \frac{1}{2} + \frac{\log\{1-\log u/(4u)\}}{\log u} \approx \frac{1}{2} - \frac{1}{4u}. \tag{4.21}$$

Convergence of the second-order approximations for α_1 and β_1 towards their respective limits is illustrated in Figure 4.1 with correlation $\rho = 0.5$, and for values of x corresponding to the 97.5% up to the 99.998% Laplace quantile. It appears that convergence is very slow and it makes sense to consider second-order approximations when measuring the goodness-of-fit of finite-sample estimates. In order to give an idea of the amount of data needed to reach such quantiles, we change the scale of the abscissa to the return period scale, using

$$\frac{1}{1-F(x^{\mathrm{L}})} \times \frac{1}{n_Y},$$

with $F(\cdot)$ the Laplace distribution function, x^{L} any quantile on the Laplace scale and $n_Y = 365.25$ the number of observations per year. In Figure 4.1, we observe that even with the equivalent of 120 years of daily data, the location and scale parameters differ significantly from their asymptotic values.

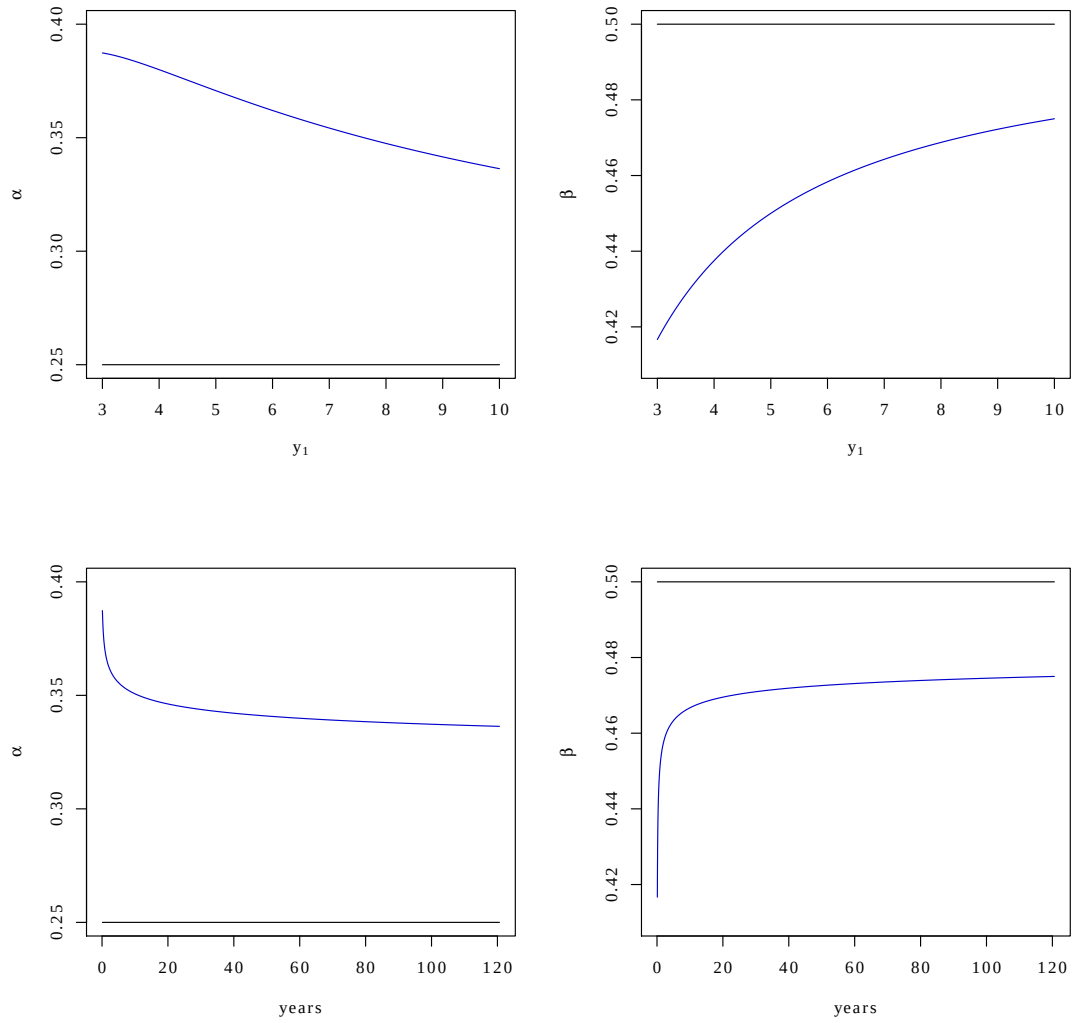

Figure 4.1 – Comparison of first- (black) and second-order (blue) approximations to the Heffernan–Tawn parameters α and β for a Gaussian copula with covariance parameter $\rho = 0.5$. Top line: abscissa on Laplace scale; bottom line: abscissa on the return period scale, assuming daily observations.

4.4 Inverted logistic distribution

In this section, we consider the bivariate random vector (X, Y) with inverted logistic distribution and Laplace margins. Its joint survival distribution function is

$$\Pr(X > x, Y > y) = \exp\left[-V\left\{\frac{-1}{\log\left(\frac{1}{2}e^{-x}\right)}, \frac{-1}{\log\left(\frac{1}{2}e^{-y}\right)}\right\}\right], \quad x, y > 0,$$

where $V(z,w) = (z^{-1/\gamma} + w^{-1/\gamma})^{\gamma}$, $0 < \gamma \leq 1$, is the exponent measure function of the logistic distribution.

Theorem 4.2
Let (X, Y) have a bivariate inverted logistic distribution with dependence parameter $0 < \gamma \leq 1$ and Laplace margins. Then the ultimate and penultimate normings (4.5) and (4.6) for $Y \mid X = x$, with x large, are

$$a_0(x) \equiv 0, \qquad b_0(x) = x^{1-\gamma},$$
$$a_1(x) \equiv -\log 2, \quad b_1(x) = x^{1-\gamma},$$

so no penultimate form for $b_1(\cdot)$ exists in the standard location and scale norming formulation.

The limit distribution $H(z)$ in (4.4) is Weibull, specifically $H(z) = \exp(-\gamma z^{1/\gamma})$, and the penultimate distribution $H_x(\cdot)$ in (4.7) is such that

$$-\log \overline{H}_x(z) = \gamma z^{1/\gamma} + \begin{cases} \dfrac{(1-\gamma)(1-\log 2)}{x} z^{1/\gamma} - \dfrac{\gamma(1-\gamma)}{2x} z^{2/\gamma}, & 0 < \gamma < 2/3, \\ \dfrac{1-\log 2}{3x} z^{3/2} - \dfrac{1}{9x} z^3 - \dfrac{(\log 2)^2}{8x}\left(4 - \dfrac{13\log 2}{3}\right) z^{-3/2}, & \gamma = 2/3, \\ -\dfrac{(\log 2)^2}{6\gamma^2}(1-\gamma)\left\{6\gamma + (1-8\gamma)\log 2\right\} x^{3\gamma-3} z^{1/\gamma-3}, & 2/3 < \gamma < 1. \end{cases} \tag{4.22}$$

When $0 < \gamma < 2/3$, $H_x(\cdot)$ has finite support

$$\left[0, \left\{\frac{x}{1-\gamma} + \frac{1-\log 2}{\gamma}\right\}^{\gamma}\right] \longrightarrow \mathbb{R}_+, \quad x \to \infty,$$

and $H_x(z^H) \sim 1 - \exp\{-\gamma x/(2-2\gamma)\}$ as $x \to \infty$. When $\gamma = 2/3$, $H_x(\cdot)$ has approximate finite support

$$\left[(12 - 13\log 2)^{1/3}\left(\frac{\log 2}{4}\right)^{2/3} x^{-1/3}, 9^{1/3} x^{2/3}\right] \to \mathbb{R}_+, \quad x \to \infty,$$

and $H_x(z^H) \sim 1 - \exp(-x)$ as $x \to \infty$. When $2/3 < \gamma < 1$, $H_x(\cdot)$ has approximate finite support

$$\left[\frac{(\log 2)^{2/3}}{\gamma}\left(\frac{1-\gamma}{6}\right)^{1/3}\left\{6\gamma + (1-8\gamma)\log 2\right\}^{1/3} x^{\gamma-1}, +\infty\right) \to \mathbb{R}_+, \quad x \to \infty.$$

If we write $n^{-1} = \Pr(X > u)$, the rates of convergence to the limit distribution are as follows: $O\{(\log n)^{-1}\}$ using the ultimate norming in (4.8), $O\{(\log n)^{\gamma-1}\}$ using the penultimate norming

in (4.9)*; the subasymptotic remainder* (4.10) *behaves like*

$$\begin{aligned} &O\{(\log n)^{\alpha-2}\}, \quad \alpha\in(0,1/2], \quad O\{(\log n)^{3\alpha-3}\}, \quad \alpha\in(1/2,2/3),\\ &O\{(\log n)^{-4/3}\}, \quad \alpha=2/3, \qquad\quad O\{(\log n)^{-1}\}, \quad \alpha\in(2/3,1). \end{aligned} \tag{4.23}$$

Proof We start by computing the conditional survival distribution of $Y \mid X = x$ for large x and deriving a tail approximation to it. We have, for non-negative x and y,

$$\begin{aligned} &\Pr(Y>y\mid X=x)\\ &\qquad = 2e^x \times \frac{\partial \Pr(X>x,\, Y>y)}{\partial x}\\ &\qquad = -2\exp\left\{x - V\left(\frac{1}{x+\log 2}, \frac{1}{y+\log 2}\right)\right\} V_1\left(\frac{1}{x+\log 2}, \frac{1}{y+\log 2}\right)(x+\log 2)^{-2}, \end{aligned}$$

where the partial derivative of the exponent measure is

$$V_1(x,y) = -\left(x^{-1/\gamma}+y^{-1/\gamma}\right)^{\gamma-1} x^{-1/\gamma-1}.$$

To ease the following developments, we examine the log-survival conditional probability

$$\begin{aligned} &\log\Pr\left(Y>y\mid X=x\right)\\ &\qquad = \log 2 + x - \left[\left\{x(1+x^{-1}\log 2)\right\}^{1/\gamma} + \left\{y(1+y^{-1}\log 2)\right\}^{1/\gamma}\right]^{\gamma}\\ &\qquad\quad + (\gamma-1)\log\left[\left\{x(1+x^{-1}\log 2)\right\}^{1/\gamma} + \left\{y(1+y^{-1}\log 2)\right\}^{1/\gamma}\right]\\ &\qquad\quad + \frac{1-\gamma}{\gamma}\log\left\{x(1+x^{-1}\log 2)\right\}\\ &\qquad \approx \log 2 + x - x\left[1 + \frac{\log 2}{\gamma x} + \frac{1-\gamma}{2\gamma^2}\frac{(\log 2)^2}{x^2} + \left(\frac{y}{x}\right)^{1/\gamma}\left\{1+\frac{\log 2}{\gamma y}+\frac{1-\gamma}{2\gamma^2}\frac{(\log 2)^2}{y^2}\right\}\right]^{\gamma}\\ &\qquad\quad + (\gamma-1)\left[\frac{1-\gamma}{2\gamma^2}\frac{(\log 2)^2}{x^2} + \left(\frac{y}{x}\right)^{1/\gamma}\left\{1+\frac{\log 2}{\gamma y}+\frac{1-\gamma}{2\gamma^2}\frac{(\log 2)^2}{y^2}\right\}\right], \end{aligned}$$

for large x and y. We can expand further, using the fact that x and y are positively associated and asymptotically independent, so large values of x occur with large values of y with large ratio x/y, so the log conditional probability can be approximated as follows,

$$\begin{aligned} -x\Bigg[\frac{1-\gamma}{2\gamma}\frac{(\log 2)^2}{x^2} + \left(\frac{y}{x}\right)^{1/\gamma}\left\{\gamma + \frac{\log 2}{y} + \frac{1-\gamma}{2\gamma}\frac{(\log 2)^2}{y^2}\right\} + \frac{\gamma(\gamma-1)}{2}\left\{\frac{\log 2}{\gamma x} + \left(\frac{y}{x}\right)^{1/\gamma}\left(1+\frac{\log 2}{\gamma y}\right)\right\}^2\Bigg]\\ - \frac{(1-\gamma)^2}{2\gamma}\frac{(\log 2)^2}{x^2} - (1-\gamma)\left(\frac{y}{x}\right)^{1/\gamma}\left\{1+\frac{\log 2}{\gamma y}+\frac{1-\gamma}{2\gamma^2}\frac{(\log 2)^2}{y^2}\right\}. \end{aligned}$$

For $Y = a(x) + b(x)Z$, the first order behaviour is cancelled by choosing $a_0(x) \equiv 0$ and $b_0(x) = x^{1-\gamma}$ (Heffernan and Tawn, 2004). We find $a_1(\cdot)$ by setting $a_1(x) = \varepsilon$, with $\varepsilon = \varepsilon(x) =$

$o(x^{1-\gamma})$, namely with $Y = \varepsilon + x^{1-\gamma}Z$,

$$
\begin{aligned}
&\log\Pr(Y > y \mid X = x) \\
&\quad\approx -\frac{1-\gamma}{2\gamma}(\log 2)^2 x^{-1} - x^{1-1/\gamma}\left(\varepsilon + x^{1-\gamma}z\right)^{1/\gamma} \\
&\qquad\qquad \times\left\{\gamma + (\log 2)\left(\varepsilon + x^{1-\gamma}z\right)^{-1} + \frac{1-\gamma}{2\gamma}(\log 2)^2\left(\varepsilon + x^{1-\gamma}z\right)^{-2}\right\} \\
&\qquad - \frac{(\gamma-1)(\log 2)^2}{2\gamma x} - \frac{\gamma(\gamma-1)}{2}x^{1-2/\gamma}\left(\varepsilon + x^{1-\gamma}z\right)^{1/\gamma}\left\{1 + \frac{\log 2}{\gamma}\left(\varepsilon + x^{1-\gamma}z\right)^{-1}\right\} \\
&\qquad + (1-\gamma)(\log 2)x^{-1/\gamma}\left(\varepsilon + x^{1-\gamma}z\right)^{1/\gamma}\left\{1 + \frac{\log 2}{\gamma}\left(\varepsilon + x^{1-\gamma}z\right)^{-1}\right\} - \frac{(1-\gamma)^2}{2\gamma}(\log 2)^2 x^{-2} \\
&\qquad - (1-\gamma)x^{-1/\gamma}\left(\varepsilon + x^{1-\gamma}z\right)^{1/\gamma}\left\{1 + \frac{\log 2}{\gamma}\left(\varepsilon + x^{1-\gamma}z\right)^{-1} + \frac{1-\gamma}{2\gamma^2}(\log 2)^2\left(\varepsilon + x^{1-\gamma}z\right)^{-2}\right\} \\
&\quad= -z^{1/\gamma}\left\{1 + \frac{\varepsilon}{\gamma}x^{\gamma-1}z^{-1} + \frac{1-\gamma}{2\gamma^2}\varepsilon^2 x^{2\gamma-2}z^{-2} + \frac{(1-\gamma)(1-2\gamma)}{6\gamma^3}\varepsilon^3 x^{3\gamma-3}z^{-3} + O\left(x^{4\gamma-4}\right)\right\} \\
&\qquad \times\left\{\gamma + (\log 2)\left(x^{\gamma-1}z^{-1} - \varepsilon x^{2\gamma-2}z^{-2} + \varepsilon^2 x^{3\gamma-3}z^{-3}\right)\right. \\
&\qquad\qquad \left. + \frac{1-\gamma}{2\gamma}(\log 2)^2\left(x^{2\gamma-2}z^{-2} - 2\varepsilon x^{3\gamma-3}z^{-3}\right) + O\left(x^{4\gamma-4}\right)\right\} \\
&\qquad + \frac{\gamma(1-\gamma)}{2}z^{2/\gamma}x^{-1} + (1-\gamma)(\log 2)x^{-1}z^{1/\gamma} - (1-\gamma)x^{-1}z^{1/\gamma} + O\left(x^{\gamma-2}\right).
\end{aligned}
$$

Expanding this expression and rearranging the terms yields

$$
\begin{aligned}
&-\gamma z^{1/\gamma} - (\log 2 + \varepsilon)x^{\gamma-1}z^{1/\gamma-1} + \left\{(1-\gamma)(\log 2 - 1)z^{1/\gamma} + \frac{\gamma(1-\gamma)}{2}z^{2/\gamma}\right\}x^{-1} \\
&+ \left\{(\log 2)\varepsilon z^{1/\gamma-2} - \frac{\varepsilon\log 2}{\gamma}z^{1/\gamma-2} - \frac{1-\gamma}{2\gamma}\varepsilon^2 z^{1/\gamma-2} - \frac{1-\gamma}{2\gamma}(\log 2)^2 z^{1/\gamma-2}\right\}x^{2\gamma-2} \\
&- \left\{\varepsilon^2 - \frac{1-\gamma}{\gamma}(\log 2)^2\varepsilon - \frac{\varepsilon^2}{\gamma} + \frac{1-\gamma}{2\gamma^2}(\log 2)^2\varepsilon + \frac{1-\gamma}{2\gamma^2}(\log 2)\varepsilon^2 + \frac{(1-\gamma)(1-2\gamma)}{6\gamma^2}\varepsilon^3\right\}z^{1/\gamma-3}x^{3\gamma-3} \\
&+ O\left(x^{\max\{\gamma-2,4\gamma-4\}}\right).
\end{aligned}
\tag{4.24}
$$

We obtain $\varepsilon = -\log 2$, or equivalently $a_1(x) \equiv -\log 2$, cancelling the leading term in x in (4.24). Higher-order terms imply different powers of Z, hence making any further approximation of the norming functions not feasible because of the linearity in Z stemming from the location-scale norming.

Plugging $a_1(x)$ into (4.24) results in

$$
\begin{aligned}
&-\gamma z^{1/\gamma} + (1-\gamma)\left\{\log 2 - 1 + \frac{\gamma}{2}z^{1/\gamma}\right\}z^{1/\gamma}x^{-1} \\
&\qquad + \frac{(1-\gamma)(\log 2)^2}{6\gamma^2}\left\{6\gamma + (1-8\gamma)\log 2\right\}z^{1/\gamma-3}x^{3\gamma-3} + O\left(x^{\max\{\gamma-2,4\gamma-4\}}\right).
\end{aligned}
$$

The various expressions for $H_x(\cdot)$ depending on the value of γ directly follow.

We now derive the support of $H_x(\cdot)$ when $0 < \gamma < 2/3$. A necessary condition for $H_x(\cdot)$ to be well-defined is that the density $h_x(\cdot) = H'_x(\cdot)$ is non-negative, that is

$$\begin{aligned}
&\overline{H}_x(z)\left\{z^{1/\gamma-1} + \frac{(1-\gamma)(1-\log 2)}{\gamma x} z^{1/\gamma-1} - \frac{1-\gamma}{x} z^{2/\gamma-1}\right\} \geq 0 \\
\Longrightarrow\ & \frac{1-\gamma}{x} z^{2/\gamma} \leq z^{1/\gamma}\left\{1 + \frac{(1-\gamma)(1-\log 2)}{\gamma x}\right\} \\
\Longrightarrow\ & z \leq \left(\frac{x}{1-\gamma} + \frac{1-\log 2}{\gamma}\right)^{\gamma},
\end{aligned}$$

and the upper bound is also the upper endpoint z^H; the value of $H_x(z^H)$ follows directly. The lower endpoint is attained when the exponent in (4.22) vanishes, in other terms when

$$z^{1/\gamma}\left(2\frac{1-\log 2}{\gamma} + \frac{2x}{1-\gamma} - z^{1/\gamma}\right) = 0,$$

for which the root of interest is $z = 0$, which concludes this part of the proof.

The case $\gamma = 2/3$ is treated similarly, as we require the derivative of $H_x(z)$ to be non-negative,

$$\begin{aligned}
&\frac{1}{3x} z^2 - z^{1/2} - \frac{3}{16x}(\log 2)^2\left(4 - \frac{13}{3}\log 2\right) z^{-5/2} \leq 0 \\
\Longrightarrow\ & w^3 - (3x+a)\, w^2 - c \leq 0, \qquad\qquad z > 0, x > 0, \qquad (4.25)
\end{aligned}$$

with $w = z^{3/2}$, $a = 3(1-\log 2)/2$ and $c = 3(\log 2)^2(12 - 13\log 2)/16$. When $x \to \infty$, we have $w \to \infty$, and a leading term is given by

$$w - (3x+a) \leq 0 \Longrightarrow w \leq 3x + a.$$

In order to ensure that we are not missing a term in this approximation, we consider (4.25) with $w = 3x + a + \delta$, $0 < \delta = \delta(x) = O(x)$, as follows,

$$(3x+a+\delta)^3 - (3x+a)(3x+a+\delta)^2 - c = \delta(3x+a)^2 + 2\delta^2(3x+a) + \delta^3 - c \leq 0, \quad x > 0.$$

For this inequality to hold, we need at least $9x^2\delta \leq c$, $x > 0$, i.e., $\delta = O(x^{-2})$. We conclude that $z^H = \{3x + 3(1-\log 2)/2\}^{2/3}$ is an approximation of the upper endpoint when $\gamma = 2/3$, with

$$\overline{H}_x\left(z^H\right) = \exp\left\{-\frac{1}{x}\left(x + \frac{1-\log 2}{2}\right)^2 + O\left(x^{-2}\right)\right\} \sim \exp(-x) \to 0, \quad x \to \infty.$$

The lower endpoint is computed by finding an approximation to the root of interest of the exponent in (4.22), or equivalently with $w = z^{3/2}$,

$$w^3 - \left\{3(1-\log 2) + 6x\right\} w^2 + \frac{3}{8}(\log 2)^2(12 - 13\log 2) = 0,$$

for which we know $w \to 0$ when $x \to \infty$, leading to the approximation

$$(3x+a)w^2 + c = 0 \implies w = \left(\frac{c}{3x+a}\right)^{1/2} \approx \sqrt{c}\left\{\frac{1}{\sqrt{3x}} - \frac{a}{2(3x)^{3/2}}\right\}, \quad w > 0, x > 0. \tag{4.26}$$

Consider $w = \sqrt{c/(6x+a)} + \delta$, $0 < \delta = \delta(x) = O(1/\sqrt{x})$, in order to confirm that (4.26) is a sensible approximation as follows,

$$\left(\frac{c}{3x+a}\right)^{3/2} + 3\frac{c\delta}{3x+a} + 3\left(\frac{c}{3x+a}\right)^{1/2}\delta^2 + \delta^3 - (3x+a)\left\{2\delta\left(\frac{c}{3x+a}\right)^{1/2} + \delta^2\right\} = 0,$$

and expanding the expressions in brackets yields

$$c^{3/2}(3x)^{-3/2} + c\delta x^{-1} + \sqrt{3c}\delta^2 x^{-1/2}\left(1 - \frac{a}{6x}\right) + \delta^3 \\ - 2\sqrt{3c}\delta x^{1/2}\left(1 + \frac{a}{6x} - \frac{a^2}{72x^2}\right) - \delta^2 a - 6\delta^2 x + O\left(x^{-5/2}\right) = 0.$$

From this we observe that we require $\delta = o(x^{-1/2})$ for the equality to hold as $x \to \infty$, so we can simplify further and get

$$\delta^2(3x+a) + \delta\left(2\sqrt{3c}x^{1/2} + \frac{a\sqrt{3c}}{3}x^{-1/2}\right) - c^{3/2}(3x)^{-3/2} = 0,$$

which we solve in δ. We get an approximate square root discriminant

$$x^{1/2}\left(2\sqrt{3c} + \frac{a\sqrt{3c}}{3}x^{-1} + \frac{c}{3}x^{-3/2}\right),$$

from which we compute the approximate root of interest

$$\delta = \frac{c}{6}x^{-1}(6x+2a)^{-1} \sim \frac{c}{36}x^{-2}.$$

This ends the proof for the support of $H_x(z)$ when $\gamma = 2/3$.

In the case when $\gamma \in (2/3, 1)$, we require

$$\frac{\partial \overline{H}_x(z)}{\partial z} = \overline{H}_x(z)\left[-z^{1/\gamma-1} + \left(\frac{1}{\gamma} - 3\right)\frac{(\log 2)^2}{6\gamma^2}(1-\gamma)\left\{6\gamma + (1-8\gamma)\log 2\right\}x^{3\gamma-3}z^{1/\gamma-4}\right] \leq 0,$$

which is true for all $z > 0$ and $x > 0$. The density is well-defined and we can verify that the upper endpoint of $H_x(z)$ is $+\infty$. We work out the lower endpoint by considering the exponent

in (4.22), with

$$\gamma z^{1/\gamma} - \frac{(\log 2)^2}{6\gamma^2}(1-\gamma)\left\{6\gamma + (1-8\gamma)\log 2\right\} x^{3\gamma-3} z^{1/\gamma-3} \\ = \gamma z^{1/\gamma}\left[1 - \frac{(\log 2)^2}{6\gamma^3}(1-\gamma)\left\{6\gamma + (1-8\gamma)\log 2\right\} x^{3\gamma-3} z^{-3}\right],$$

which vanishes when $z = 0$, and

$$z = \frac{(\log 2)^{2/3}}{\gamma}\left[\frac{1-\gamma}{6}\left\{6\gamma + (1-8\gamma)\log 2\right\}\right]^{1/3} x^{\gamma-1}, \quad x > 0. \tag{4.27}$$

The root of interest is (4.27), giving the desired result.

We now give the convergence rate of (4.10) using the penultimate approximation $a_1(\cdot)$. The convergence rate is linked with the value of the dependence parameter γ and can be found from (4.24). The powers of x of interest appearing in (4.24) are -1, $3\gamma-3$, $\gamma-2$, $4\gamma-4$, depending on the precise value of γ. For $\gamma \in (0, 1/2]$, convergence is the fastest, as subtraction of $H_x(z)$ in this case removes terms in x^{-1}, so we conclude that (4.10) has a leading term in $x^{\gamma-2}$. Similarly we conclude that the order of convergence for $\gamma \in (1/2, 2/3)$ is $x^{3\gamma-3}$. For $\gamma = 2/3$, the $\gamma-2$ and $4\gamma-4$ powers coincide and give $x^{-4/3}$ as the leading term. When $\gamma \in (2/3, 1)$, convergence is slowest with x^{-1} as the leading term. ■

4.5 Logistic distribution

Consider (X, Y) having a bivariate logistic distribution with Laplace margins. Its joint distribution for (x, y), $x, y > 0$, is

$$\Pr(X \le x,\, Y \le y) = \exp\left[-V\left\{\frac{-1}{\log\left(1-\frac{1}{2}e^{-x}\right)}, \frac{-1}{\log\left(1-\frac{1}{2}e^{-y}\right)}\right\}\right],$$

with

$$V(z, w) = \left(z^{-1/\gamma} + w^{-1/\gamma}\right)^{\gamma}, \quad \gamma \in (0, 1].$$

Theorem 4.3
Let (X, Y) have a bivariate inverted logistic distribution with dependence parameter $0 < \gamma \le 1$ and Laplace margins. Then, the ultimate normings (4.5) *for $Y \mid X = x$, with x large, are*

$$a_0(x) = x, \quad b_0(x) = 1,$$

and no penultimate normings (4.6) *exist.*

Proof We now focus on the conditional probability

$$\begin{aligned}\Pr(Y \le y \mid X = x) = 2\exp\left[x - V\left\{\frac{-1}{\log\left(1-\frac{1}{2}e^{-x}\right)}, \frac{-1}{\log\left(1-\frac{1}{2}e^{-y}\right)}\right\}\right] \\ \times\left[-V_1\left\{\frac{-1}{\log\left(1-\frac{1}{2}e^{-x}\right)}, \frac{-1}{\log\left(1-\frac{1}{2}e^{-y}\right)}\right\}\right]\frac{d}{dx}\left\{-\frac{1}{\log\left(1-\frac{1}{2}e^{-x}\right)}\right\}.\end{aligned} \tag{4.28}$$

We can approximate the last term in (4.28), for large x,

$$\begin{aligned}\frac{d}{dx}\left\{-\frac{1}{\log\left(1-\frac{1}{2}e^{-x}\right)}\right\} &= \frac{1}{\left\{\log\left(1-\frac{1}{2}e^{-x}\right)\right\}^2}\frac{1}{\left(1-\frac{1}{2}e^{-x}\right)}\frac{1}{2}e^{-x} \\ &\approx \left(\frac{1}{2}e^{-x}+\frac{1}{8}e^{-2x}\right)^{-2}\left(\frac{1}{2}e^{-x}+\frac{1}{4}e^{-2x}\right) \\ &\approx 2e^{x}-\frac{1}{8}e^{-x}.\end{aligned} \tag{4.29}$$

The partial derivative of $V(\cdot,\cdot)$ in (4.28) can be approximated as

$$\begin{aligned}&V_1\left\{-\frac{1}{\log\left(1-\frac{1}{2}e^{-y}\right)}\right\} \\ &\approx -\left\{\left(\frac{1}{2}e^{-x}+\frac{1}{8}e^{-2x}\right)^{1/\gamma}+\left(\frac{1}{2}e^{-y}+\frac{1}{8}e^{-2y}\right)^{1/\gamma}\right\}^{\gamma-1}\left(\frac{1}{2}e^{-x}+\frac{1}{8}e^{-2y}\right)^{1/\gamma+1} \\ &\approx -\frac{1}{4}\left\{e^{-x/\gamma}+\frac{1}{4\gamma}e^{-x(1+1/\gamma)}+\frac{1-\gamma}{32\gamma^2}e^{-x(2+1/\gamma)}+e^{-y/\gamma}+\frac{1}{4\gamma}e^{-y(1+1/\gamma)}+\frac{1-\gamma}{32\gamma^2}e^{-y(2+1/\gamma)}\right\}^{\gamma-1} \\ &\quad\times e^{-x(1+1/\gamma)}\left(1+\frac{1}{4}e^{-x}\right)^{1+1/\gamma},\end{aligned} \tag{4.30}$$

for large x. From (4.29) and (4.30), the conditional probability (4.28) on the log scale is

$$\begin{aligned}&\log\{\Pr(Y \le y \mid X = x)\} \\ &\quad= \log 2 + x - \left\{\left(\frac{1}{2}e^{-x}\right)^{1/\gamma}\left(1+\frac{1}{4\gamma}e^{-x}\right)+\left(\frac{1}{2}e^{-y}\right)^{1/\gamma}\left(1+\frac{1}{4\gamma}e^{-y}\right)\right\}^{\gamma} \\ &\qquad-\log 4-(1-\gamma)\log\left\{e^{-x/\gamma}+\frac{1}{4\gamma}e^{-x(1+1/\gamma)}+e^{-y/\gamma}+\frac{1}{4\gamma}e^{-y(1+1/\gamma)}\right\} \\ &\qquad-\frac{1+\gamma}{\gamma}x+\frac{1+\gamma}{4\gamma}e^{-x}+\log 2+x+O\left(e^{-2x}\right),\end{aligned} \tag{4.31}$$

where the constant terms cancel with each other. Imposing $Y = a_0(x) + b_0(x)Z$, with $a_0(x) = x$ and $b_0(x) = 1$ in (4.31) removes the linear terms in x.

We now try and remove the next-order terms by using $a_1(x) = (1+\varepsilon)x$, $\varepsilon = \varepsilon(x) = o(1)$, and $b_1(x) = 1+\delta$, $\delta = \delta(x) = o(1)$. Starting from (4.31), we get

$$
\begin{aligned}
&-\frac{1}{2}e^{-x}\left\{1+\frac{1}{4\gamma}e^{-x}+e^{-\{\varepsilon x+(1+\delta)Z\}/\gamma}\left(1+\frac{1}{4\gamma}e^{-\{(1+\varepsilon)x+(1+\delta)Z\}}\right)\right\}^{\gamma}\\
&-(1-\gamma)\log\left\{1+\frac{1}{4\gamma}e^{-x}+e^{-\{\varepsilon x+(1+\delta)Z\}/\gamma}\left(1+\frac{1}{4\gamma}e^{-\{(1+\varepsilon)x+(1+\delta)Z\}}\right)\right\}+\frac{1+\gamma}{4\gamma}e^{-x}+O\left(e^{-2x}\right)\\
=&-\frac{1}{2}e^{-x}\left(1+e^{-\{\varepsilon x+(1+\delta)Z\}/\gamma}\right)^{\gamma}\left\{1+\frac{e^{-x+\{\varepsilon x+(1+\delta)Z\}/\gamma}}{4\left(e^{\{\varepsilon x+(1+\delta)Z\}/\gamma}+1\right)}+\frac{e^{-\{(1+\varepsilon)x+(1+\delta)Z\}}}{4\left(e^{\{\varepsilon x+(1+\delta)Z\}/\gamma}+1\right)}\right\}\\
&-(1-\gamma)\left\{\log\left(1+e^{-\{\varepsilon x+(1+\delta)Z\}/\gamma}\right)+\frac{e^{-x+\{\varepsilon x+(1+\delta)Z\}/\gamma}}{4\gamma\left(e^{\{\varepsilon x+(1+\delta)Z\}/\gamma}+1\right)}+\frac{e^{-\{(1+\varepsilon)x+(1+\delta)Z\}}}{4\gamma\left(e^{\{\varepsilon x+(1+\delta)Z\}/\gamma}+1\right)}\right\}\\
&+\frac{1+\gamma}{4\gamma}e^{-x}+O\left(e^{-2x}\right),
\end{aligned}
\tag{4.32}
$$

We deduce from the leading order term $-(1-\gamma)\log(1+\exp[-\{\varepsilon x+(1+\delta)Z\}/\gamma])$ that ε can only be 0 as it cannot help remove any other term and comply with $\varepsilon = o(1)$. For δ, there is no terms involving both Z and x in (4.32), that would need to be cancelled, thus $\delta = 0$. We conclude that it is impossible to find a penultimate norming of the distribution of $\{Y - a_0(x)\}/b_0(x) \mid X = x$ for the logistic dependence structure. ■

4.6 Summary

In this chapter, we first considered, in the univariate case, the problem of the speed of convergence of $F^n(a_n x + b_n)$ towards $G(x)$ in Theorem 2.1. Penultimate analyses in the 1990's showed that using a penultimate version ξ_n of the shape parameter ξ in $G(\cdot)$ can provide a much better approximation to $F^n(\cdot)$, such as when $F(\cdot)$ is the Gaussian distribution, in which case the convergence of $F^n(\cdot)$ towards $G(\cdot)$ is especially slow.

We reviewed penultimate analyses in the bivariate case, but the results are not as general as in the univariate case. Focusing on a different approach to bivariate extremes, we then considered the conditional tail approach and various measures that can be used to describe penultimate properties. We investigated three parametric bivariate models, illustrating the cases of asymptotic independence and asymptotic dependence. The bivariate Gaussian distribution given one of its margins is large shows interesting subasymptotic features of the conditional tail model; the conditional inverted logistic distribution has a faster convergence towards its limit, but we saw that $H_x(\cdot)$ can have finite support depending on the precise value of the dependence parameter γ; we were not able to extract a penultimate behaviour for the bivariate logistic distribution, which has a fast rate of convergence towards its limit.

In the next chapter, we propose a Bayesian methodology for making inference on the conditional tail model, naturally summarising uncertainty of functions of the model parameters. We show how this new methodology can be used to derive efficient estimates of cluster functionals in the context of time series extremes.

5 Bayesian uncertainty management in temporal dependence of extremes

5.1 Foreword

This chapter was published in the journal *Extremes* under the same title (Lugrin *et al.*, 2016). The body of the text is reproduced here, with some minor changes, and Section 5.6.4 and Appendix F are additions to the paper. The code used to perform the simulation studies and the real data analysis was developed in R (R Core Team, 2017) and C and is available under a public licence in the `tsxtreme` package on the Comprehensive R Archive Network (CRAN) repository.

5.2 Introduction

Extreme value theory provides an asymptotically justified framework for the statistical modelling of rare events. In the univariate case with independent variables there is a broadly-used framework involving modelling exceedances of a high threshold by a generalised Pareto distribution (Coles, 2001). For extremes of stationary univariate time series, standard procedures use marginal extreme value modelling but consideration of the dependence structure between the variables is essential when assessing risk due to clusters of extremes. Leadbetter *et al.* (1983) and Hsing *et al.* (1988) described the roles of long- and short-range dependence on extremes of stationary time series. Typically an assumption of independence at long range is reasonable, with Ledford and Tawn (2003) giving diagnostic methods for testing this. In practice independent clusters are often identified using the runs method (Smith and Weissman, 1994), which deems successive exceedances to be in separate clusters if they are separated by at least m consecutive non-exceedances of the threshold; Ferro and Segers (2003) provide automatic methods for the selection of m.

Short-range dependence has the most important practical implications, since it leads to local temporal clustering of extreme values. Leadbetter (1983) and O'Brien (1987) provide different asymptotic characterisations of the clustering though the extremal index $0 < \theta \leq 1$, the former with θ^{-1} being the limiting mean cluster size. The case $\theta = 1$ corresponds to there

being no more clustering than if the series were independent, and decreasing θ corresponds to increased clustering.

Although the extremal index is a natural limiting measure for such clustering there is a broad class of dependent processes with $\theta = 1$, including all stationary Gaussian processes. Thus the extremal index cannot distinguish the clustering properties of this class of dependent processes from those of white noise. Furthermore, many other functionals of clusters of extremes may be of interest (Segers, 2003), so for practical modelling of the clusters of extreme values above a high threshold, the restriction to $\theta < 1$ is a major weakness.

Smith *et al.* (1997), Ledford and Tawn (2003), Eastoe and Tawn (2012) and Winter and Tawn (2016) draw on methodology for multivariate extremes to provide models for univariate clustering, and thereby enable the properties of a wide range of cluster functionals to be estimated. The focus of this paper is the improvement of inference techniques for the most general model for these cases, namely the semiparametric conditional extremes model of Heffernan and Tawn (2004). Our ideas can be applied to any cluster functional, but we focus here primarily on the threshold-based extremal index introduced by Ledford and Tawn (2003).

Given the strong connections between multivariate extreme value and clustering modelling, here and in Section 5.4 we present the developments of the model in parallel for the two situations. Examples of applications for multivariate cases include assessing the risk of joint occurrence of extreme river flows or sea-levels at different locations (Keef *et al.*, 2009b; Asadi *et al.*, 2015), the concurrent high levels of different pollutants at the same location (Heffernan and Tawn, 2004), and simultaneous stock market crashes (Poon *et al.*, 2003). For the time series case, applications include assessing heatwave risks (Reich *et al.*, 2014; Winter and Tawn, 2016), modelling of extreme rainfall events (Süveges and Davison, 2012) and wind gusts (Fawcett and Walshaw, 2006a).

For the stationary time series (X_t), Ledford and Tawn (2003) define the threshold-based extremal index

$$\theta(x, m) = \Pr\left(X_1 \leq x, \ldots, X_m \leq x \mid X_0 > x\right), \tag{5.1}$$

where x is large, which is the key measure of short-range clustering of extreme values, with $1/\theta(x, m)$ being the mean cluster size when clusters of exceedances of the threshold x are defined via the runs method of Smith and Weissman (1994) with run length m. Furthermore $\theta(x, m)$ converges to the extremal index θ as $x \to x^F$ and $m \to \infty$ appropriately (O'Brien, 1987; Kratz and Rootzén, 1997). Many studies have focused on estimating the limit θ (Ferro and Segers, 2003; Süveges, 2007; Robert, 2013), but in applications, all these consider a finite level u as an approximation to the limit x^F. This is equivalent to assuming $\theta(x, m)$ to be constant above u, which is generally not the case in applications (see Figure 5.1). Additionally Eastoe and Tawn (2012) find that $\theta(x, m)$ is fundamental to modelling the distributions of both cluster maxima, i.e., peaks over threshold, and block maxima, e.g., annual maxima. See Section 5.5 for more on the relevance of $\theta(x, m)$ for time series extremes.

When considering asymptotically motivated models for the joint distribution of $\mathbf{X}_{-0} = (X_1, \ldots, X_m)$ given that $X_0 > x$ for the estimation of $\theta(x, m)$, it is helpful to have a simple characterisation of extremal dependence. The standard pairwise measure of extremal dependence for (X_0, X_j) is

$$\chi_j = \lim_{x \to x^F} \Pr\left(X_j > x \mid X_0 > x\right), \quad j = 1, \ldots, m, \tag{5.2}$$

with the cases $\chi_j > 0$ and $\chi_j = 0$ respectively termed asymptotic dependence and asymptotic independence at lag j. A plot of χ_j against j has been termed the extremogram (Davis and Mikosch, 2009), by analogy with the correlogram of time series analysis. When $\chi_j = 0$ for all $j \geq 1$, the extremogram fails to distinguish between different levels of asymptotic independence, but the rate of convergence to zero of $\Pr(X_j > x \mid X_0 > x)$ determines the key characteristics of the tail of the joint distribution (Ledford and Tawn, 1996). Ledford and Tawn (2003) propose using such a measure at each time lag j when the variables are asymptotically independent. An alternative is to combine both approaches by studying a threshold-based version of χ_j,

$$\chi_j(x) = \Pr\left(X_j > x \mid X_0 > x\right), \quad j = 1, \ldots, m, \tag{5.3}$$

for a range of large values of x.

In Section 5.3, we review classical multivariate extreme models, which all entail $\chi_j(x) = \chi_j$, $x > u$ for some high threshold u, and often even $\chi_j > 0$. Instead we consider the conditional formulation of Heffernan and Tawn (2004) that has been subsequently studied more theoretically by Heffernan and Resnick (2007), Das and Resnick (2011), and Mitra and Resnick (2013). This class of models covers $\chi_j \geq 0$ and $\chi_j(x)$ changing with large x ($j = 1, \ldots, m$) through modelling dependence within the asymptotic independence class. This model gives estimates of $\theta(x, m)$ that can be constant or vary with x, $x > u$. This additional flexibility comes at a price: inference is required for up to $2m$ parameters, and for an arbitrary m-dimensional distribution G.

The asymptotic arguments for the Heffernan–Tawn model are given in Section 5.4, for an $(m+1)$-dimensional variable with Laplace marginal distributions. Suppose that the monotone increasing transformation $T = (T_0, \ldots, T_m)$ transforms $\mathbf{X} = (X_0, \ldots, X_m)$ to $\mathbf{Y} = (Y_0, \ldots, Y_m)$, with $Y_i = T_i(X_i)$ ($i = 0, \ldots, m$), so that $\mathbf{Y}$ has Laplace marginal distributions. In applications the Heffernan–Tawn model corresponds to a multivariate regression with

$$\mathbf{Y}_{-0} \mid \{Y_0 = y\} = \boldsymbol{\alpha} y + y^{\boldsymbol{\beta}} \boldsymbol{Z} = \boldsymbol{\alpha} y + \boldsymbol{\mu} y^{\boldsymbol{\beta}} + \boldsymbol{\psi} y^{\boldsymbol{\beta}} \mathbf{Z}^* \tag{5.4}$$

where here, and subsequently, the arithmetic is to be understood componentwise, with $\boldsymbol{\alpha} \in [-1, 1]^m$, $\boldsymbol{\beta} \in [-\infty, 1]^m$, $\boldsymbol{\mu} \in \mathbb{R}^m$, $\boldsymbol{\psi} \in \mathbb{R}_+^m$, and $\mathbf{Z}$ an m-dimensional random variable with $\mathbf{Z} \sim G$, with G corresponding to H in the rest of the thesis; $\mathbf{Z}^*$ has zero mean and unit variance for all marginal variables, with $\mathbf{Z}^* = (\boldsymbol{Z} - \boldsymbol{\mu})/\boldsymbol{\psi}$. We require that (5.4) holds for all $y > u$, where u is a high threshold on the Laplace marginal scale. The parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\psi})$ determine the

conditional mean and variance through

$$\mathrm{E}(\mathbf{Y}_{-0} \mid Y_0 = y) = \boldsymbol{\alpha} y + \boldsymbol{\mu} y^{\boldsymbol{\beta}}, \quad \mathrm{var}(\mathbf{Y}_{-0} \mid Y_0 = y) = \boldsymbol{\psi}^2 y^{2\boldsymbol{\beta}}.$$

Thus $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\psi})$ can be estimated by multivariate regression. The complication for inference is that the error distribution G is in general unknown and arbitrary, apart from the first two moment properties mentioned above. One exception to this is when $\boldsymbol{\alpha} = \mathbf{1}$ and $\boldsymbol{\beta} = \mathbf{0}$, in which case the Heffernan–Tawn model reduces to known asymptotically-dependent models with G directly related to the angular distribution H, as detailed in Section 5.4.

Heffernan and Tawn (2004) and Eastoe and Tawn (2012) used a stepwise inference procedure, estimating $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\psi})$ under a working assumption that $\mathbf{Z}^*$ are independent normal variables. After obtaining parameter estimates, they estimated G nonparametrically using the empirical joint distribution of the observed standardised multivariate residuals, i.e., values of $\mathbf{Z}$ for $Y_0 > u$. There are weaknesses in this approach, which loses efficiency in the estimation of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\psi})$ and G and in subsequent inferences due to the generally incorrect working assumption of normality. Moreover, as noted by Peng and Qi (2004), the empirical estimation of G leads to poor estimation of the upper tail of the conditional distribution of $\mathbf{Y}_{-0} \mid \{Y_0 = y\}$, so it would be preferable to have a better, yet general, estimator of G. Furthermore, the uncertainty of the parameter estimation is unaccounted-for in the estimation of G and of cluster functionals such as $\theta(x, m)$.

Cheng *et al.* (2014) proposed a Bayesian approach to estimating the Heffernan–Tawn model in a single stage, but their estimation procedure involves changing the structure of the model and adding a noise term in (5.4), thereby allowing the likelihood term to be split appropriately. They also need strong prior information extracted from the stepwise inference method in order to get valid estimates for the model parameters, so this procedure does not really tackle the loss of efficiency of the stepwise estimation procedure.

We propose to overcome these weaknesses by using Bayesian semiparametric inference to estimate the model parameters and the distribution G, simultaneously performing the entire fitting procedure for the dependence model. This gives a new model for G, namely, a mixture of Gaussian distributions, which provides estimates of the conditional distribution of $\mathbf{Y}_{-0} \mid Y_0 = y$ beyond the range of the current estimator and which in theory provides an arbitrary good approximation for G (Marron and Wand, 1992). The Bayesian approach also provides a coherent framework for fitting a parsimonious parametric model; joint estimation of the model parameters enables the imposition of structure between them. For example, in multivariate problems the context may suggest that different components of $\boldsymbol{\alpha}$ may be identical. In the context of time series extremes, for first-order Markov models, it can be shown that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ involve at most two unknown parameters (Papastathopoulos *et al.*, 2017). Furthermore, when the $\mathbf{X}_{-0}$ are known to be asymptotically dependent on X_0, this method provides a new approach to modelling.

We show the practical importance of the new approach by applying it to the daily mean flow time series of the River Ray at Grendon Underwood in north-west London, with observations from October 1962 to December 2008. We consider only flows from October to March, as this period typically contains the largest flows and forms an approximately stationary series. An empirical estimate of $\theta(x,m)$, with $m = 4$, is shown in Figure 5.1 with bootstrap-derived 95% confidence intervals. A major weakness with this estimate is that it cannot be evaluated beyond the range of the data, so a model is needed to evaluate $\theta(x,m)$ for larger x. We select our modelling threshold u to be the empirical 98% marginal quantile of the data. Using the methods in Eastoe and Tawn (2012) we have an estimate of $\theta(x,4)$ for all $x > u$ using the stepwise estimation method. As seen in Figure 5.1, this estimate converges to 1 as $x \to x^F$ but we have no reliable method for deriving confidence intervals. Figure 5.1 also shows posterior median estimates and 95% credibility intervals obtained using our Bayesian semiparametric method. These show broad agreement with both of the other estimates within the range of the data, but with tighter uncertainty intervals and statistically significant differences in extrapolation of $\theta(x,4)$, indicating that the new method has the potential to offer marked improvement for estimating $\theta(x,m)$ and other cluster functionals.

The chapter is structured as follows. We first briefly present the standard approaches to multivariate extremes in Section 5.3. We introduce the conditional model of interest in a multivariate framework in Section 5.4, followed by a section about modelling of dependent time series. Section 5.6 explains the Bayesian semiparametric inference procedure, which is used in Section 5.7 to illustrate the efficiency gains of this new inference method on simulated data. In Section 5.8 we fit our model to the River Ray flow data and show its ability to estimate functionals of time series clusters other than the threshold-based index.

5.3 Multivariate setup and classical models

Both multivariate and time series extremes involve estimating the probability of events that may never yet have been observed. Suppose that $\mathbf{X} = (X_0,\dots,X_m)$ is an $(m+1)$-dimensional variable with joint distribution function $F_{\mathbf{X}}$ and marginal distribution functions $F_0,\dots,F_m$. We need to estimate the probability $\Pr(\mathbf{X} \in A)$, where $A \subset \mathbb{R}^{m+1}$ is an extreme set, i.e., a set such that for all $\mathbf{x} \in A$, at least one component of $\mathbf{x} = (x_0,\dots,x_m)$ is extreme. To do this we must model $F_{\mathbf{X}}(\mathbf{x})$ for all $\mathbf{x} \in B$, where B is an extreme set that contains A. Let A_i be the subset of A for which component i is largest on a quantile scale, i.e.,

$$A_i = A \cap \left\{\mathbf{x} \in \mathbb{R}^{m+1} : F_i(x_i) > F_j(x_j),\, j \in \{0,\dots,m\}\backslash\{i\}\right\}, \quad i = 0,\dots,m,$$

and let $v_i = \inf\{x_i : \mathbf{x} \in A_i\}$, so that we can write

$$\Pr(\mathbf{X} \in A) = \sum_{i=0}^{m} \Pr(\mathbf{X} \in A_i \mid X_i > v_i)\Pr(X_i > v_i). \tag{5.5}$$

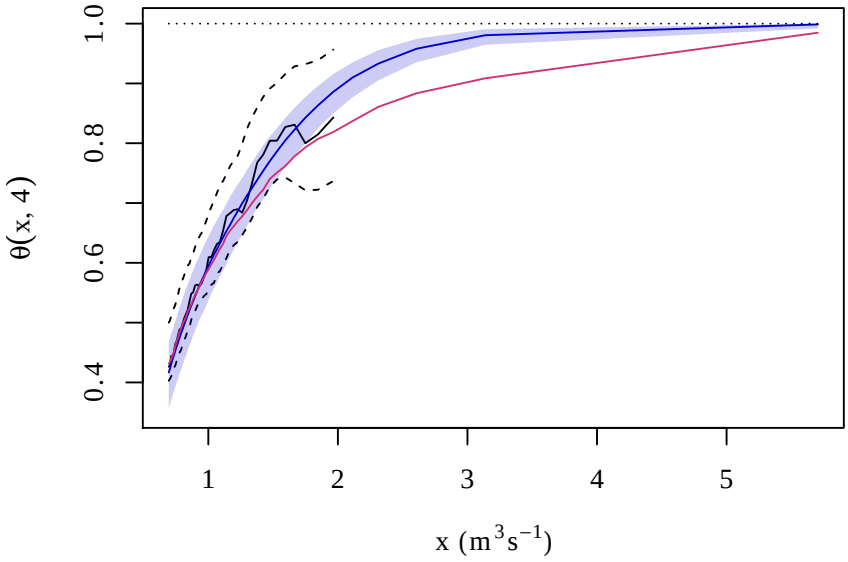

Figure 5.1 – Comparison of the empirical, stepwise and Bayesian semiparametric estimates of $\theta(x,4)$ described in Sections 5.4 and 5.6 respectively. Empirical estimate (black) with bootstrap-derived 95% confidence interval (dashed lines), stepwise estimate (light red) and Bayesian semiparametric estimate (blue) with its 95% pointwise coverage interval (shaded), estimated on winter flows (m^3s^{-1}) of the River Ray from 1962 to 2008.

Thus estimates of marginal and dependence features are required for estimating the conditional probabilities in the sum and marginal distributions determine the second terms of the products in the sum.

Although our approach applies to any form of set A, to focus our arguments we restrict ourselves to identical marginal distributions F with upper endpoint x^F and we set

$$A = A_0 = \{X_0 > x, X_1 \leq x, \ldots, X_m \leq x\},$$

so $v_0 = x$ and we estimate only the conditional probability term in (5.5), i.e., $\theta(x,m)$, as defined in (5.1).

Early approaches to modelling the conditional distribution appearing in (5.1) assumed that $\boldsymbol{X}$ lies in the domain of attraction of a multivariate extreme value distribution (Coles and Tawn, 1994; de Haan and de Ronde, 1998) and applied these asymptotic models above a high threshold. Unlike in the univariate case, there is no finite parametrisation of the dependence structure; it can only be restricted to functions of a distribution H on the m unit simplex S_m with $\int_{S_m} w_i \, dH(\mathbf{w}) = (m+1)^{-1}$ $(i = 0, \ldots, m)$, where $\mathbf{w} = (w_0, \ldots, w_m)$. Both parametric and non-parametric inference for this class of models has been proposed. Numerous parametric models are available (Kotz and Nadarajah, 2000, Ch. 3; Cooley *et al.*, 2010; Ballani and Schlather, 2011). Nonparametric estimation is also widely studied, mostly based on empirical estimators

(de Haan and de Ronde, 1998; Hall and Tajvidi, 2000; Einmahl *et al.*, 2001; Einmahl and Segers, 2009).

A major weakness of these early methods is that for these models either $\chi_j > 0$ or (X_0, X_j) are independent for all $j = 1, \ldots, m$, whereas there are distributions, such as the multivariate Gaussian, with $\chi_j = 0$ but (X_0, X_j) dependent. If $\chi_j > 0$ for any $j = 1, \ldots, m$, these models give estimates of $\theta(x, m) \to c_m$ as $x \to x^F$, where $c_m < 1$. This class of models is not flexible enough to cover distributions that are dependent at finite levels, but asymptotically independent for all pairs of variables.

5.4 Threshold-based model for conditional probabilities

5.4.1 Heffernan–Tawn model

In order to provide a model characterising conditional probabilities, such as those which describe the clustering behaviour, we need a model for multivariate extreme values, and in particular we require a model for the joint distribution of $X_1, \ldots, X_m$ given that $X_0 > x$. In this section we suppose that the marginal distributions of $\boldsymbol{X}$ are not identical, and that $X_0 \sim F_0$ is in the domain of attraction of a generalised Pareto distribution, i.e., there exists a function $\sigma_u > 0$ such that as $u \to x^{F_0}$, $(X_0 - u)/\sigma_u$, conditional on $X_0 > u$, converges to a generalised Pareto variable with unit scale parameter and shape parameter ξ.

The joint distribution is modelled via the marginal distributions F_j and a copula for the dependence structure. To study the conditional behaviour of extremes, the copula formulation is most transparent when expressed with Laplace marginals. Let T_j denote the transformation of the marginal distribution of X_j to the Laplace scale, i.e.,

$$T_j(X_j) = \begin{cases} \log\{2F_j(X_j)\}, & X_j < F_j^{-1}(1/2), \\ -\log[2\{1 - F_j(X_j)\}], & X_j > F_j^{-1}(1/2); \end{cases} \quad j = 0, \ldots, m,$$

the specification of the F_j is discussed in Section 5.4.2. Assume there exist m-dimensional functions $\boldsymbol{a}(x) = \{a_1(x), \ldots, a_m(x)\}$ and $\boldsymbol{b}(x) = \{b_1(x), \ldots, b_m(x)\} > \boldsymbol{0}$ for which

$$\Pr\left[\frac{T_j(X_j) - a_j\{T_0(X_0)\}}{b_j\{T_0(X_0)\}} \leq z_j,\ j = 1, \ldots, m \,\middle|\, X_0 > u\right] \to G(\boldsymbol{z}), \quad u \to x^{F_0}, \tag{5.6}$$

where all marginal distributions of G are non-degenerate and $\boldsymbol{z} = (z_1, \ldots, z_m) \in \mathbb{R}^m$. Hereafter we write the standardised X_j, or residual, as

$$Z_j = \frac{T_j(X_j) - a_j\{T_0(X_0)\}}{b_j\{T_0(X_0)\}}, \quad X_0 > u,\ j = 1, \ldots, m.$$

Under assumption (5.6), the rescaled conditioning variable $(X_0 - u)/\sigma_u$ is asymptotically conditionally independent of the residual $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$ given $X_0 > u$, as $u \to x^{F_0}$. That is,

$$\begin{aligned}\Pr\{\boldsymbol{Z} \le \boldsymbol{z},\ (X_0 - u)/\sigma_u > x \mid X_0 > u\} \\ &= \Pr\{\boldsymbol{Z} \le \boldsymbol{z} \mid (X_0 - u)/\sigma_u > x\} \Pr\{(X_0 - u)/\sigma_u > x \mid X_0 > u\} \\ &\to G(\boldsymbol{z})\,\overline{K}(x), \quad u \to x^{F_0}, \end{aligned} \tag{5.7}$$

where $\overline{K}$ is the generalised Pareto distribution survivor function (5.12) with scale and shape parameters $(1, \xi)$ and G is the limit distribution of the residuals.

Equation (5.6) can be illustrated through the particular case when $\boldsymbol{X}$ is a centred multivariate Gaussian distribution with correlation matrix elements ρ_{ij}, $i, j = 0, \ldots, m$, $i \neq j$. In this case we can derive $a_j(x) = \text{sign}(\rho_{0j})\rho_{0j}^2 x$ and $b_j(x) = x^{1/2}$, and $G(\boldsymbol{z})$ is a centred multivariate Gaussian distribution with variances $\rho_{0j}^2(1 - \rho_{0j}^2)$ and correlation matrix elements

$$\rho'_{ij} = \frac{\rho_{ij} - \rho_{0i}\rho_{0j}}{\sqrt{\left(1 - \rho_{0i}^2\right)\left(1 - \rho_{0j}^2\right)}}, \quad i \neq j.$$

Heffernan and Tawn (2004) and Keef *et al.* (2013) showed that under broad conditions, the component functions of $\boldsymbol{a}(x)$ and $\boldsymbol{b}(x)$ can be modelled by

$$a_j(x) = \alpha_j x, \quad b_j(x) = x^{\beta_j}, \quad -1 \le \alpha_j \le 1, \quad -\infty < \beta_j \le 1, \quad j = 1, \ldots, m.$$

In terms of the dependence structure, α_j and β_j reflect the flexibility of the model. It turns out that (X_0, X_j) are asymptotically dependent only if $\alpha_j = 1$, $\beta_j = 0$, and then

$$\chi_j = \lim_{x \to x^{F_0}} \Pr\left\{T_j(X_j) > x \mid T_0(X_0) > x\right\} = \int_0^\infty \overline{G}_j(-z)e^{-z}dz;$$

with $\overline{G}_j$ the jth marginal survivor function of G; if $-1 < \alpha_j < 1$, then (X_0, X_j) are asymptotically independent, with positive extremal dependence if $\alpha_j > 0$, negative extremal dependence if $\alpha_j < 0$, and with extremal near-independence if $\alpha_j = 0$ and $\beta_j = 0$.

We set $\beta_j \ge 0$, as when $\beta_j < 0$ all the conditional quantiles for X_j converge to the same value as X_0 increases, which is unlikely in most environmental contexts. If the conditioning threshold u is high enough that the conditional probability on the left of (5.6) is close to its limit, then the Heffernan–Tawn model can be stated as

$$T_j(X_j) = \alpha_j T_0(x) + \{T_0(x)\}^{\beta_j} Z_j, \quad X_0 = x > u, \quad j = 1, \ldots, m, \tag{5.8}$$

where $(Z_1, \ldots, Z_m) \sim G$ is independent of X_0, and G can be any distribution with non-degenerate margins.

5.4.2 Existing inference procedure

We now outline the approach to inference suggested by Heffernan and Tawn (2004). Consider a vector $(X_0, \ldots, X_m)$ whose marginal distributions $F_0, \ldots, F_m$ each lie in the domain of attraction of a generalised Pareto distribution. We estimate them using the semiparametric estimator of Coles and Tawn (1994),

$$\widehat{F}_j(x) = \begin{cases} \widetilde{F}_j(x), & x < u, \\ 1 - \left\{1 - \widetilde{F}_j(u)\right\} \left(1 + \widehat{\xi}_j \dfrac{x-u}{\widehat{\sigma}_{u,j}}\right)_+^{-1/\widehat{\xi}_j}, & x \geq u, \end{cases} \tag{5.9}$$

where $\widetilde{F}_j$ is the empirical marginal distribution function of X_j. Here $\widehat{\sigma}_{u,j}$ and $\widehat{\xi}_j$ are maximum likelihood estimates based on all exceedances of u, ignoring any dependence; their variances can be evaluated by a sandwich procedure (Fawcett and Walshaw, 2007) or by a block bootstrap. The margins F_j are transformed to the Laplace scale through the transformation $\widehat{T}_j\left(X_j\right)$.

Estimation of the probability of any extreme set of interest involves inference for model (5.8) with the estimators of parameters of the dependence model assumed independent of the parameter estimators of the marginal distribution. This assumption has been found not to be restrictive in other copula inference contexts (Genest *et al.*, 1995). Heffernan and Tawn (2004) proposed a stepwise inference procedure for estimating the extremal dependence structure, based on the working assumption that the residual variables $Z_1, \ldots, Z_m$ are independent and Gaussian with means $\mu_1, \ldots, \mu_m$ and variances $\psi_1^2, \ldots, \psi_m^2$. This assumption allows likelihood inference based on the assumed marginal densities,

$$T_j\left(X_j\right) \mid \{T_0(X_0) = x\} \sim \mathcal{N}\left(\alpha_j x + x^{\beta_j}\mu_j,\ x^{2\beta_j}\psi_j^2\right), \quad x > u,\ j = 1, \ldots, m.$$

The first step of their procedure consists of a likelihood maximisation performed separately for each j, giving estimates of α_j, β_j and the nuisance parameters μ_j and ψ_j. Additional constraints, arising from results of Keef *et al.* (2013), lead to the likelihood function being zero for certain combinations of parameters (α_j, β_j). Thus the maximisation is over a subset of $[-1,1] \times [0,1]$ for these two parameters. These constraints ensure that the conditional quantiles of $T_j(X_j) \mid T_0(x)$ are ordered in a decreasing sequence for all large x under fitted models corresponding to positive asymptotic dependence, asymptotic independence and negative asymptotic dependence respectively. For details of these constraints see Keef *et al.* (2013), who show that imposing these additional constraints improves inference of the conditional extremes model; Section 5.6.4 explains how these constraints are implemented in our Bayesian framework. Given model (5.8) and the estimates $\widehat{\alpha}_j$ and $\widehat{\beta}_j$, the second step of the estimation procedure involves multivariate residuals $\boldsymbol{Z}$ for each data point, using the relation

$$\widehat{Z}_j = \frac{\widehat{T}_j\left(X_j\right) - \widehat{\alpha}_j \widehat{T}_0\left(X_0\right)}{\left\{\widehat{T}_0\left(X_0\right)\right\}^{\widehat{\beta}_j}}, \quad j = 1, \ldots, m, \quad X_0 > u,$$

and hence constructing the joint empirical distribution function $\widehat{G}(\boldsymbol{z})$.

An estimator for $\Pr(A \mid X_0 > x)$ for any extreme set A is obtained as follows: sample R independent replicates $X_0^{(1)},\ldots,X_0^{(R)}$ of X_0 conditional on $X_0 > x$ from a generalised Pareto distribution with threshold x; independently sample $\boldsymbol{Z}^{(1)},\ldots,\boldsymbol{Z}^{(R)}$ from the joint empirical distribution function $\widehat{G}$; compute

$$\boldsymbol{X}_{-0}^{(r)} = \widehat{\boldsymbol{T}}^{-1}\left[\widehat{\boldsymbol{\alpha}}\,\widehat{T}_0(X_0^{(r)}) + \left\{\widehat{T}_0(X_0^{(r)})\right\}^{\widehat{\boldsymbol{\beta}}}\boldsymbol{Z}^{(r)}\right], \quad r = 1,\ldots,R,$$

where vector arithmetic is to be understood componentwise and $\widehat{\boldsymbol{T}}^{-1} = (T_1^{-1},\ldots,T_m^{-1})$ is a componentwise back-transformation to the original scale; then the estimator for $\Pr(A \mid X_0 > x)$ is

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{1}\left\{\left(X_0^{(r)},\boldsymbol{X}_{-0}^{(r)}\right)\in A\right\},$$

where $\mathbb{1}$ is the indicator function.

In the rest of the paper, we are interested in estimating the conditional probability $\theta(x,m)$, corresponding to $A = \{X_0 > x, X_1 < x,\ldots,X_m < x\}$, for which the Heffernan–Tawn model provides a characterisation. Under the assumption that the limit (5.6) approximately holds for some subasymptotic u, Eastoe and Tawn (2012) obtain

$$\theta(x,m) = \int_x^\infty G\{\boldsymbol{z}(x,y)\}k_x(y)\delta y, \quad x > u, \tag{5.10}$$

where $k_x(y)$ is the generalised Pareto density for threshold x, with scale parameter $1+\xi(x-u)$ and shape parameter ξ, and $\boldsymbol{z}(x,y)$ is an m-dimensional vector with elements

$$z_j\left(x,y\right) = \frac{T_j\left(x\right) - \alpha_j T_0\left(y\right)}{\left\{T_0\left(y\right)\right\}^{\beta_j}}, \quad j = 1,\ldots,m. \tag{5.11}$$

A Monte Carlo approximation to the integral (5.10) gives the estimator

$$\widehat{\theta}(x,m) = \frac{1}{R}\sum_{r=1}^{R}\widehat{G}\left\{\boldsymbol{z}\left(x,X_0^{(r)}\right)\right\},$$

where $\boldsymbol{z}(x,X_0^{(r)})$ is given by expression (5.11) with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ replaced by estimates. Monte Carlo variability can be reduced by using the same pseudo-random sequence when generating samples for different values of x. Eastoe and Tawn (2012) use a bootstrap method to get confidence bounds for $\widehat{\theta}(x,m)$, but de Carvalho and Ramos (2012) found this to be unreliable.

Four main weaknesses of this inference procedure justify developing a more comprehensive approach. First, the working assumption needed for the likelihood maximisation is that the residuals have independent Gaussian distributions, and it is hard to quantify how this affects inference. Second, ignoring the variability of $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$ estimated in the first step leads to underestimation of the uncertainty in the estimate for the residual distribution G and

hence also for $\theta(x,m)$. Third, the empirical estimation of G restricts estimates of extremal conditional probabilities, as simulated $\boldsymbol{Z}$ values provide no extrapolation over observed values of $\boldsymbol{Z}$. Fourth, the inability to impose natural constraints on $(\alpha_1,\ldots,\alpha_m)$ and $\left(\beta_1,\ldots,\beta_m\right)$ leads to inefficiency.

5.5 Modelling dependence in time

Consider a stationary time series $\{X_t\}$ satisfying appropriate long-range dependence properties and with marginal distribution F. The threshold-based extremal index $\theta(x,m)$ summarises the key extremal dependence in time series. In the block-maxima context, the distribution of the block maximum $M_n = \max\{X_1,\ldots,X_n\}$ at a level x is approximately $\{F(x)\}^{n\theta(x,m)}$ for large x, n and m (O'Brien, 1987; Kratz and Rootzén, 1997). The associated independent series $\{X_t^*\}$, having the same marginal distribution as $\{X_t\}$ but independent observations, has $M_n^* = \max\{X_1^*,\ldots,X_n^*\}$ with distribution function $\{F(x)\}^n$. So $\Pr(M_n < x) \approx \{\Pr(M_n^* < x)\}^{\theta(x,m)}$, with $\theta(x,m)$ accounting for the dependence.

The most popular approach to dealing with short-range dependent in such series is the peaks over threshold (POT) approach formalised by Davison and Smith (1990). This approach consists of selecting a high threshold u, identifying independent clusters of exceedances of u, picking the maximum Y of each cluster, and then fitting to these cluster maxima the generalised Pareto distribution

$$\Pr(Y < x \mid Y > u) = 1 - \left(1 + \xi\frac{x-u}{\sigma_u}\right)_+^{-1/\xi}, \quad x > u. \tag{5.12}$$

The limiting results are used as an approximation for data at subasymptotic levels with limit distribution (5.12) taken as exact above a selected value of u.

Alternatives to the POT approach include modelling the series of all exceedances, for example using a Markov chain (Smith *et al.*, 1997; Winter and Tawn, 2016), but they depend heavily on the validity of the underlying modelling assumptions and so may be inappropriate.

Eastoe and Tawn (2012) consider the threshold-based extremal index as part of a model for the distribution of cluster maxima. Specifically they show that, for a given high threshold u, the cluster maxima, defined by the runs method with run-length m, have approximate distribution function

$$1 - \frac{\theta(x,m)}{\theta(u,m)}\left(1 + \xi\frac{x-u}{\sigma_u}\right)_+^{-1/\xi}, \quad x > u, \tag{5.13}$$

where the parameters ξ and $\sigma_u > 0$ determine the marginal distribution of the original series, and $x_+ = \max(x,0)$. Eastoe and Tawn (2012) show how using the information in $\theta(x,m)$ can improve over the POT approach. Distribution (5.13) reduces to the generalised Pareto model asymptotically as $u \to x^F$, and more generally when $\theta(x,m) = \theta(u,m)$ for all $x > u$. When

estimates of $\theta(x,m)$ vary appreciably above u, this equality condition for $\theta(x,m)$ provides a diagnostic for situations where the POT method is inappropriate.

In our approach, when a Markov property can reasonably be assumed for a time series, the α_j and β_j have a structure that we want to exploit. Papastathopoulos *et al.* (2017) and Kulik and Soulier (2015) characterise the form of $a_j(x)$ and $b_j(x)$ under very weak assumptions. If the conditions needed for the Heffernan–Tawn simplification — $a_1(x) = \alpha_1 x$ and $b_1(x) = x^{\beta_1}$ — hold, then for positively associated first order Markov processes, Papastathopoulos *et al.* (2017) show that either $(\alpha_j, \beta_j) = (1,0)$, or $(0, \beta^j)$ or (α^j, β) for some $\alpha \in [0,1)$ and $\beta = [0,1)$. The first case corresponds to asymptotic dependence at all time lags, and the other two to different forms of decaying dependence under asymptotic independence. If $\{X_t\}$ follows an asymptotically dependent Markov process, then no parameters need be estimated, rather than $2m$. If the process is well-approximated by an asymptotically independent Markov process then either of the last two cases applies, and the number of parameters in the parametric component of the model reduces from $2m$ to 1 or 2. In the case of a Gaussian AR(1) process (X_t) with standard Gaussian margins

$$X_{t+1} = \rho X_t + \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}\left(0, 1-\rho^2\right),\ \rho \in (-1,1),$$

Heffernan and Tawn (2004) get normalising parameters $\alpha_j = \text{sign}(\rho)\rho^{2j}$ and $\beta_j = 1/2,\ j = 1,\ldots,m$; the distribution $G(\boldsymbol{z})$ is a centred multivariate Gaussian with variances $\rho^{2j}(1-\rho^{2j})$ and correlation matrix elements

$$\rho'_{ij} = \frac{\text{sign}(\rho^{i+j})\rho^{j-i}\sqrt{1-\rho^{2i}}}{\sqrt{1-\rho^{2j}}}, \quad i < j.$$

See Papastathopoulos *et al.* (2017) for many more examples of first order Markov processes and their resulting forms for α_j, β_j and G.

5.6 Bayesian Semiparametrics

5.6.1 Overview

Since the m-dimensional residual distribution G in the Heffernan–Tawn model (5.8) is unknown, the approach described in Section 5.4.2 uses the joint empirical distribution function, which cannot model the tails of the conditional distribution of X_j in (5.8). Our proposed Bayesian approach instead takes G to be a mixture of a potentially infinite number of multivariate Gaussian distributions through the use of a Dirichlet process. This approach can model any G and capture its tails, and has the major benefit of allowing joint estimation of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and G.

Below we introduce Dirichlet processes and describe an approach to approximate Monte Carlo sampling from them. We then describe Bayesian semiparametric inference and the spec-

ification of prior distributions, and discuss implementation issues. Throughout we assume that we have n observations from the distribution of $(X_1,\ldots,X_m) \mid X_0 > u$, or equivalently from $\boldsymbol{Z} = (Z_1,\ldots,Z_m)$ if $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ were known.

5.6.2 Dirichlet process mixtures for the residual distribution

Consider a bivariate problem, $m = 1$; with known α and β. If we are to estimate the distribution function of $X_1 \mid X_0 > u$ this is equivalent to estimating the distribution function G of the univariate random variable $Z_1 \sim G$. If G is estimated nonparametrically, its prior must be a distribution over a space of distributions. In this context, a widely used prior is the Dirichlet process (Ferguson, 1973) mixture. A simple model structure takes G to be a mixture of an unknown number of distributions Q_k having parameters $\boldsymbol{\lambda}_k$, $k = 1,2,\ldots$, so that the Dirichlet process boils down to a distribution on the space of mixture distributions P for $\{\boldsymbol{\lambda}_k\}$. If $\boldsymbol{\lambda}_k \mid P \sim P$, then the distribution of P is the Dirichlet process $\mathrm{DP}\left(\gamma P_0\right)$, where P_0 is the centre distribution and $\gamma > 0$ the concentration parameter (Hjort *et al.*, 2010).

The definition of a Dirichlet process states that for any $p = 1,2,\ldots$, and any finite measurable partition $\{B_1,\ldots,B_p\}$ of the space of the $\boldsymbol{\lambda}_k$,

$$\left\{P(B_1),\ldots,P(B_p)\right\} \sim \mathrm{Dirichlet}\{\gamma P_0(B_1),\ldots,\gamma P_0(B_p)\}.$$

The interpretation of the Dirichlet process parameters stems from the properties

$$\mathrm{E}\{P(B_i)\} = P_0(B_i), \quad \mathrm{var}\{P(B_i)\} = \frac{P_0(B_i)\{1-P_0(B_i)\}}{\gamma+1}, \quad i = 1,\ldots,p,$$

so the $\mathrm{DP}(\gamma P_0)$ prior is closer to its mean P_0 and less variable for large values of γ. A constructive characterisation of the Dirichlet process is the stick-breaking representation (Sethuraman, 1994)

$$P(\cdot) = \sum_{k=1}^{\infty} w_k \delta_{\boldsymbol{\lambda}_k}(\cdot), \tag{5.14}$$

where $\delta_{\boldsymbol{\lambda}}$ denotes a distribution concentrated on $\boldsymbol{\lambda}$, and $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ are independent, P_0-distributed, and independent of the random weights $w_k \geq 0$, which satisfy $\sum_{k=1}^{\infty} w_k = 1$. The stick-breaking process takes its name from the computation of the weights: define $V_1, V_2, \ldots$ as the breaks independently sampled from a $\mathrm{Beta}(1,\gamma)$ distribution. The weights are then

$$w_1 = V_1, \quad w_k = V_k \prod_{i=1}^{k-1}(1-V_i), \quad k = 2,3,\ldots. \tag{5.15}$$

Ishwaran and Zarepour (2000) use formulation (5.14) to express the Dirichlet process in terms of the random variables w_k and $\boldsymbol{\lambda}_k$. They also introduce index variables $c_1,\ldots,c_n$ that describe the components of the mixture in which the observations $z_1,\ldots,z_n$ lie, giving

a stick-breaking representation in terms of the index variables c_i rather than the random variables $\boldsymbol{\lambda}_k$.

A key step in deriving a posterior distribution is the truncation of the sum in the stick-breaking representation, i.e., replacing the infinite sum in (5.14) by a sum up to N. This is achieved by imposing $V_N = 1$ in (5.15). The accuracy of the stick-breaking approximation improves exponentially fast in terms of the $\mathscr{L}_1$ error (Ishwaran and James, 2001), as

$$\|M_N - M_\infty\|_1 \leq 4\left[1 - \mathrm{E}\left\{\left(\sum_{k=1}^{N-1} w_k\right)^n\right\}\right] \approx 4n\exp\left(-\frac{N-1}{\gamma}\right),$$

where M_N is the marginal density

$$\int \prod_{i=1}^{n}\left\{\sum_{k=1}^{N} w_k Q_k\left(dZ_i \mid \boldsymbol{\lambda}_i\right)\right\}\mathrm{DP}\left(dP_N\right).$$

For example, the error is smaller than 10^{-29} when truncating the stick-breaking sum representation at $N = 150$ as in our real data analysis (Section 5.8) for which $\gamma \leq 2$ and $n = 154$.

Taking into account the transformations discussed above, a simple model for G involving the Dirichlet process prior is

$$\begin{aligned} Z \mid c, \Lambda &\overset{\text{ind}}{\sim} Q_c = Q(\cdot; \boldsymbol{\lambda}_c), \\ c \mid P_N &\overset{\text{iid}}{\sim} P_N = \sum_{k=1}^{N} w_k \delta_k(\cdot), \\ (\Lambda, \boldsymbol{w}) &\sim \pi_\Lambda(\cdot)\pi_{\boldsymbol{w}}(\cdot), \end{aligned} \tag{5.16}$$

where Λ is the matrix with rows $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N$, and $\boldsymbol{w} = (w_1, \ldots, w_N)$, with $\sum_{k=1}^{N} w_k = 1$ for some suitably large N. To lighten the notation we write $Z_1 \mid \boldsymbol{\lambda}_c$ instead of $Z_1 \mid c, \Lambda$ in what follows. Taking the Q_k $(k = 1, \ldots, N)$ to be normal distributions with means $\mu_{1,k}$ and variances $\psi^2_{1,k}$ leads to Λ being a $2 \times N$ matrix, with rows $\boldsymbol{\lambda}_k = (\mu_{1,k}, \psi^2_{1,k})$. Model (5.16) is made more flexible by adding a hyperprior for the concentration parameter γ.

5.6.3 Multivariate semiparametric setting

We now specify the features of our algorithm, finally yielding model (5.18). We must add a further element to (5.16): covariates, i.e., the parametric part of the Heffernan–Tawn model (5.8), to recognise that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown. This is achieved using a covariate-dependent Dirichlet process, and it can be formulated in terms of the truncated stick-breaking representation as

$$P_{|x}(\cdot) = \sum_{k=1}^{N} w_k \delta_{\boldsymbol{\lambda}_k(x)}(\cdot), \tag{5.17}$$

so that a single output of the stick-breaking procedure gives rise to a whole family of distributions indexed by x. Our data are the n observations from m-dimensional variables $\boldsymbol{X}_{-0}$, given X_0 is large. We assume that, conditional on $T(X_0) > u$, $T(\boldsymbol{X}_{-0})$ has a mixture of multivariate normal distributions, $\sum_{k=1}^{\infty} w_k \mathcal{N}_m$, where the mean vector $M_k(x)$ and the covariance matrix $\Psi_k(x)$ of the kth normal component depend on the value x of $T(X_0)$. For parsimony, the variance matrix $\Psi_k(x)$ is taken to be diagonal with diagonal elements $\{\Psi_{1,k}(x), \ldots, \Psi_{m,k}(x)\}$, as the mixture structure is considered flexible enough to capture the dependence between the elements of $\boldsymbol{X}_{-0}$.

As we use the truncated version of the stick-breaking representation (5.14), the conditional distribution for the weights w_i is a generalised Dirichlet distribution (Connor and Mosimann, 1969), written as GDirichlet. This gives the final form of our semiparametric model:

$$\begin{gathered} T(X_j) \mid \left\{T(X_0) = x,\ \alpha_j,\ \beta_j,\ M_{j,c}(x),\ \Psi_{j,c}(x)\right\} \overset{\text{ind}}{\sim} \mathcal{N}\left\{M_{j,c}(x), \Psi_{j,c}(x)\right\},\ j = 1, \ldots, m,\ x > u, \\ M_{j,c}(x) = \alpha_j x + \mu_{j,c} x^{\beta_j}, \quad \Psi_{j,c}(x) = x^{2\beta_j} \psi_{j,c}^2, \\ c \mid \boldsymbol{w} \sim \sum_{k=1}^{N} w_k \delta_k(\cdot), \end{gathered} \tag{5.18}$$

where the prior for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}_k, \boldsymbol{\psi}_k)$, with $\boldsymbol{\mu}_k = (\mu_{1,k}, \ldots, \mu_{m,k})$ and $\boldsymbol{\psi}_k = (\psi_{1,k}, \ldots, \psi_{m,k})$, takes the form

$$\begin{aligned} \boldsymbol{w} \mid \gamma &\sim \text{GDirichlet}\left(1, \gamma, \ldots, 1, \gamma\right), \\ \gamma &\sim \text{Gamma}\left(\eta_1, \eta_2\right), \\ \alpha_j &\overset{\text{iid}}{\sim} \mathcal{U}(0,1), \quad \beta_j \overset{\text{iid}}{\sim} \mathcal{U}(0,1), \quad j = 1, \ldots, m, \\ \mu_{j,k} &\overset{\text{ind}}{\sim} \mathcal{N}\left(0, \psi_{(\mu)}^2\right), \quad \psi_{j,k}^2 \overset{\text{iid}}{\sim} \text{Inv-Gamma}\left(\nu_{1,j}, \nu_{2,j}\right), \quad j = 1, \ldots, m,\ k = 1, \ldots, N, \end{aligned}$$

with positive hyper-parameters $(\eta_1, \eta_2, \psi_{(\mu)}^2, \nu_1, \nu_2)$. The Keef *et al.* (2013) conditions mentioned in Section 5.4.2 are built into the likelihood terms for α_j and β_j, so the Metropolis–Hastings scheme systematically rejects proposals outside the support of the posterior.

5.6.4 Implementation of existing constraints in the Bayesian framework

In this section, we give more details about how the constraints of Keef *et al.* (2013) are dealt with in the Bayesian framework. We state the form of the constraints, which hold if Conditions A and B are both satisfied in Definition 5.1, for conditional quantiles evaluated at $p = 0$ and $p = 1$, as we have shown in Section 2.4.3 of Chapter 2. We recall here that we define the conditional quantile functions for x large

$$q_j^-(p) = -x + z_j^-(p), \quad q_j(p) = \alpha_j x + x^{\beta_j} z_j(p), \quad q_j^+(p) = x + z_j^+(p),$$

under asymptotic negative dependence ($\chi_j^- > 0$), asymptotic independence ($\chi_j^- = \chi_j^+ = 0$) and asymptotic positive dependence ($\chi_j^+ > 0$) respectively, with

$$\chi_j^+ = \lim_{x \to x^F} \Pr\left(X_j > x \mid X_0 > x\right), \quad \chi_j^- = \lim_{x \to x^F} \Pr\left(X_j \leq x \mid X_0 > x\right), \quad j = 1, \ldots, m,$$

and

$$G_j^-\left\{z_j^-(p)\right\} = G_j\left\{z_j(p)\right\} = G_j^+\left\{z_j^+(p)\right\} = p,$$

where $G_j^-(\cdot)$ and $G_j^+(\cdot)$ correspond to the residual distributions of the normalised limit of $X_j \mid X_0 = x$ for x large under asymptotic negative and positive dependence respectively.

Definition 5.1 (Keef *et al.* (2013))
The constraints of Keef et al. *are satisfied if Conditions A and B both hold, with $v > u$ and $p \in \{0\} \cup \{1\}$.*

Condition A *Either*

$$\alpha_j \leq \min\left\{1, 1 - \beta_j z_j(p) v^{\beta - 1}, 1 - v^{\beta_j - 1} z_j(p) + v^{-1} z_j^+(p)\right\},$$

or

$$1 - \beta_j z_j(p) v^{\beta_j - 1} < \alpha_j \leq 1 \quad \textit{and} \quad \left(1 - \beta_j^{-1}\right)\left\{\beta_j z_j(p)\right\}^{1/(1-\beta_j)} (1-\alpha_j)^{-\beta_j/(1-\beta_j)} + z_j^+(p) > 0.$$

Condition B *Either*

$$-\alpha_j \leq \min\left\{1, 1 + \beta_j v^{\beta_j - 1} z_j(p), 1 + v^{\beta_j - 1} z_j(p) - v^{-1} z_j^-(p)\right\},$$

or

$$1 + \beta_j v^{\beta_j - 1} z_j(p) < -\alpha_j \leq 1 \quad \textit{and} \quad \left(1 - \beta_j^{-1}\right)\left\{-\beta_j z_j(p)\right\}^{1/(1-\beta_j)} (1+\alpha_j)^{-\beta_j/(1-\beta_j)} - z_j^-(p) > 0.$$

From a Bayesian point of view these constraints do not represent prior knowledge, as they depend on the data, which are used to compute $z_j(0)$ and $z_j(1)$ from the maximum and minimum value of the residuals $z_j = \{T_j(x_j) - \alpha_j T_j(x_0)\} / \{T_j(x_0)\}^{\beta_j}$. The constraints of Definition 5.1 correspond to constraints on the support of the likelihood function for α_j and β_j, so that any candidate $\alpha_j^\star$ or $\beta_j^\star$ not satisfying the constraints can be rejected before computing the acceptance ratio in a Metropolis–Hastings scheme, thus saving computational time. Because the constraints cannot be expressed in closed form, it would require a lot of effort to shape a proposal distribution which would sample only candidates satisfying the constraints, and we do not explore this further here. We suggest an approach to more efficiently sample from the proposal distribution of (α_j, β_j) in Appendix F, using a state-dependent proposal distribution.

5.6.5 Implementation issues

The semiparametric multivariate Bayesian model (5.18) has the added benefit of allowing us to structure the parametric component of the model. Assuming $\boldsymbol{X}$ to be a first order Markov process yields different structures discussed in Section 5.5. For example, the different forms of decaying dependence in the class of asymptotic independence can be modelled by setting the priors to be $\alpha \sim \mathcal{U}(0,1)$ and $\beta \sim \mathcal{U}(0,1)$, independently. The appropriate structure of model (5.18) can be determined using standard diagnostics, and if adopted in the modelling will lead to substantially improved efficiency. Imposing continuous priors on α and β induces a restriction to the class of asymptotically independent series, but both parameters can be arbitrary close to the boundaries of their support, ensuring that the behaviour of $\theta(x,m)$ and the extremal structure of dependence of the series are not affected at the high levels of interest. A reversible jump procedure (Green, 1995) could be added to the current algorithm in order to enable α and β to have prior masses on the support boundaries to ensure positive posterior probability of asymptotic dependence; see Coles and Pauli (2002) for an example of this type of construction.

The shape and scale for the prior variances of the components $\psi^2_{j,k}$ are taken to be $\nu_1 = \nu_2 = 2$ to make the model prefer numerous components with smaller variances to a few dispersed components. The posterior distribution for γ depends on the logarithm of the last weight in the truncation (cf. Appendix E) and can be numerically unstable, so a vague gamma prior truncated at small values is needed to ensure convergence. Conjugacy of the prior densities allows analytical calculation of the posterior distributions for all parameters in model (5.18) except $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, for which a Metropolis–Hastings step is needed. We use a regional adaptive scheme in Roberts and Rosenthal (2009) to avoid the choice of specific proposal variances. The posterior densities are mainly derived from Ishwaran and James (2002), and are given in Appendix E.

As noticed by Porteous *et al.* (2006), the Gibbs sampler used for model (5.18) leads to a clustering bias, because the weights do not satisfy the weak ordering $\mathrm{E}(w_1) \geq \cdots \geq \mathrm{E}(w_{N-1})$. Papaspiliopoulos and Roberts (2008) suggested two different label switching moves to improve the mixing of the algorithm. Components cannot be simply swapped, as this would change the joint distribution of the weights. Label switching is not to be understood in its exact sense within this framework: if a switch between two mixture components is proposed and accepted, then only their means and variances are swapped and the index variables c_i of the data points belonging to these components are renumbered accordingly. We use this approach and adapt it to our semiparametric framework.

The results presented in Sections 5.7 and 5.8 are promising, but two aspects would benefit from improvement. Bayesian semiparametric inference provides a valuable approach to uncertainty in the Heffernan–Tawn model, but the procedure is not fully Bayesian, since the marginal distribution is fitted using maximum likelihood estimation. With further work we could include the fit for the marginal distribution within the fit for the dependence structure,

but we would have to account for the temporal dependence between observations in order not to introduce bias. The second possibility for improvement pertains to the sampling of α_j and β_j in (5.8): the special cases corresponding to the boundaries of their support should correspond to Dirac masses, so reversible jump Markov chain Monte Carlo sampling (Green, 1995) could be used.

5.7 Simulation study

5.7.1 Bivariate data

We start by showing how the Bayesian semiparametric approach to inference can improve over the stepwise approach in a bivariate setting. The working assumption of Gaussianity for the residual variable Z is key to the stepwise process, and if it fails badly then the stepwise approach may perform poorly relative to the Bayesian semiparametric approach. To illustrate this we take Z to have a bimodal density, either a mixture of Gaussian densities, or a mixture of Laplace densities. As the former is a special case of the structure of the mixture components in the dependent Dirichlet process, we may expect a clear improvement in that case, but it is less clear what to expect in the latter case.

We generated data (X, Y) directly from the Heffernan–Tawn model with parameters (α, β) subject to $X > u$, for large $u > 0$, as follows:

1. Simulate X as $u + \text{Exp}(1)$;
2. Independently simulate Z from the required mixture model;
3. Let $Y = \alpha X + X^{\beta} Z$.

We selected the mixture for the bimodal distribution of Z such that the simulated (X, Y) data are split into two clear clusters for large X (see left panel of Figure 5.2 for an example).

We simulated 1000 data sets each with 400 points, roughly twice the number of exceedances available in our river flow application, and fitted the conditional model using the stepwise and the Bayesian semiparametric approaches. We compare the methods through the relative efficiency, measured as the ratio of the root mean squared error (RMSE) for the Bayesian approach to the RMSE of the stepwise approach. The estimators we consider in order to compute the efficiency are the mode, the mean, and the median of the posterior distribution of α and β for the Bayesian approach and the maximum likelihood estimators of α and β for the stepwise approach.

The benefits of the Bayesian semiparametric approach are clearly found, with similar relative efficiencies whether Z is simulated with a Gaussian or Laplace mixture. The relative efficiency is broadly 0.6 for α and in the range $0.5 - 0.65$ for β depending on which of the three summary measures of the posterior distribution is chosen. The posterior number of

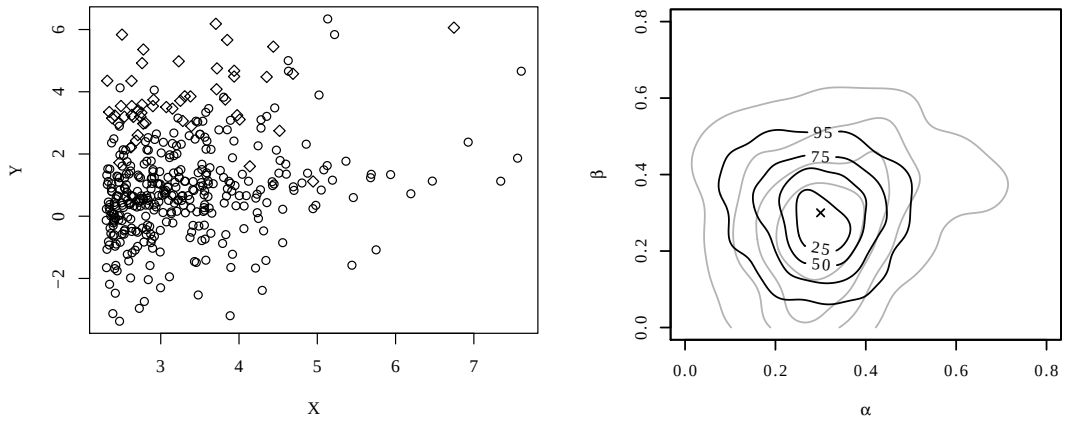

Figure 5.2 – Results for the simulation study based on the Heffernan–Tawn model with a bimodal residual distribution with two Laplace components. Left panel: simulated data, with the two components shown as squares and circles. Right panel: kernel density estimate based on 1000 estimates of (α, β) from the stepwise method (grey) and from the Bayesian semiparametric method (black) using the posterior medians as the summary statistic. The same contour levels are used for both density estimates. The true value is shown as a cross.

components in the mixture is concentrated around 2 and 3, so the Bayesian semiparametric approach seems to capture the distribution of Z well. Figure 5.2 shows the joint sampling distribution of the estimators of (α, β) based on the two inference methods. The contours are similar, but suggest that the Bayesian approach estimates the true (α, β) more precisely.

Of key importance is the practical implication of this improvement, which is more naturally measured in terms of improved performance for estimating the threshold-based extremal index. Specifically we estimate $\theta(x,1) = \Pr(Y < x \mid X > x)$, which requires accurate estimation of the distribution of Z as well as of α and β. The relative efficiency for $\theta(x,1)$ is computed, where the true value for $\theta(x,1)$ is obtained from a huge simulation from the true model. The relative efficiency varies over x, with values of 0.95 and 0.90 for the 99% and 99.9% quantiles, suggesting slight improvements within the range of the data. The relative efficiency reduces to 0.69 at the 99.99% quantile, suggesting that the real benefits in the Bayesian semiparametric approach arise when we extrapolate.

5.7.2 AR(1) process

We now compare the performances of the empirical, the stepwise and the Bayesian semiparametric inference procedures in estimating the threshold-based extremal index of a stationary time series. The data are generated from a first-order Markov process with Gaussian copula and exponential margins. This is equivalent to having a standard Gaussian AR(1) process and using the probability integral transform to obtain exponential marginal distributions. In Gaussian margins this process has lag τ autocorrelation $\rho_\tau = \rho^\tau$, where we consider the set of $\{-0.75, -0.5, \ldots, 0.5, 0.75\}$ for the true value of ρ. For each of these values of ρ, the process

is asymptotically independent, with extremal index $\theta = 1$, but it exhibits dependence at any subasymptotic threshold when $\rho \neq 0$. The true value for $\theta(x, m)$ is evaluated by computing the ratio of multivariate normal integrals

$$\theta(x, m) = \Pr(X_0 > x, X_1 < x, \ldots, X_m < x)/\Pr(X_0 > x). \tag{5.19}$$

using the methods of Genz and Bretz (2009) and Genz *et al.* (2014). The use of exponential margins ensures that the GPD marginal model is exact for all thresholds, so any bias in the estimation of $\theta(x, m)$ can be attributed to inference for the dependence structure. A similar approach was taken by Eastoe and Tawn (2012).

The three methods are applied to 1000 data sets of length 8000, approximately the length of the winter flow data studied in Section 5.8. This procedure is repeated for each value of ρ in the range $\{-0.75, \ldots, 0.75\}$. The empirical method simply estimates each of the probabilities in expression (5.19) empirically. Often called the runs estimate (Smith and Weissman, 1994), this is not defined beyond the largest value of the sample, whereas the other two methods do not suffer this weakness. In each case the marginal threshold u for the modelling and inference is fixed at the 95% empirical quantile of each series. Unlike the stepwise procedure, the Bayesian semiparametric approach enables us to constrain $\{(\alpha_j, \beta_j) : j = 1, \ldots, m\}$, and this allows us to exploit our knowledge of the Markovian structure of the process to impose the constraints on the α_j and β_j discussed in Section 5.5, thus reducing the number of parameters from $2m$ to 1 or 2.

We estimate $\theta(x, m)$ for a range of high quantiles x and for declustering run-lengths $m = 1$ and 4. Table 5.1 shows the ratios of RMSEs of the posterior median of $\theta(x, m)$ from the Bayesian semiparametric approach and the empirical and the stepwise estimators in the particular case when $\rho = 0.5$. The Bayesian semiparametric estimator is always superior to the empirical estimator, with the advantage improving as x increases. For the stepwise approach the results are similar to those in Section 5.7.1: the two estimators are similar at low levels but the Bayesian semiparametric estimator performs better at higher levels. Figure 5.3 summarises the results for all values of ρ, showing a major improvement of our method over the stepwise approach for negative autocorrelation and short run-length, with increased gain at higher levels. The particularly good performance of our estimator for negative values of ρ stems from the constraint on $\{(\alpha_j, \beta_j) : j = 1, \ldots, m\}$, which dramatically improves identifiability of the parameters in the conditional tail model; more details on this topic are provided in Section 6.6.5. In order to assess the effectiveness of imposing the Markovian structure in the Bayesian semiparametric approach, we also fitted the 1000 simulated time series with unconstrained $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the case $\rho = 0.5$. The efficiency of the unconstrained approach only declines relative to the constrained approach at high quantiles. For example the 57% in the bottom right of Table 5.1 increases to 75%.

We expect the Bayesian approach to gain accuracy in terms of frequentist coverage of $\theta(x, m)$, as it fits the data in one stage and thus provides a better measure of uncertainty. To

	m=1		m=4	
Level	Empirical	Stepwise	Empirical	Stepwise
98%	88	101	88	100
99%	68	92	71	98
99.99%	–	59	–	57

Table 5.1 – Ratios (%) of RMSEs computed with estimates of $\theta(x,m)$; the numerator of these efficiencies is always the RMSE estimate derived from the posterior median in the Bayesian semiparametric approach, and the denominator is either the RMSE corresponding to the runs estimate (Empirical) or to the stepwise estimate (Stepwise). Empty cells correspond to high levels of x for which estimates of $\theta(x,m)$ cannot be evaluated.

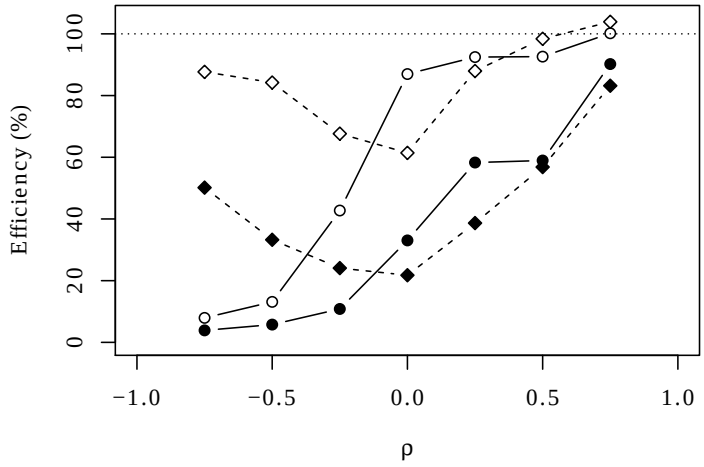

Figure 5.3 – Ratios (%) of RMSEs computed with estimates of $\theta(x,m)$, the numerator being the RMSE derived from the posterior median in our approach and the denominator being the RMSE corresponding to the stepwise estimate. The cases $m = 1$ (circles) and $m = 4$ (diamonds) are illustrated for x at the 99% (empty symbols) and 99.99% (filled symbols) levels.

assess this we considered bootstrap confidence intervals for the stepwise method and credible intervals for the Bayesian method, both of the type $[L_\alpha, \infty)$. Here L_α is the αth quantile of the distribution of the estimator, considering bootstrap estimates for the former and posterior samples for the latter. Using the same 1000 simulated data sets as earlier in this section, we computed the proportion of times that the true value of $\theta(x,m)$ would fall in these confidence or credible intervals, for a range of α-confidence levels from 5% to 95%, different run-lengths m, and several levels x for $\theta(x,m)$. The coverage performance is summarised in Figure 5.4, which shows the difference between the calculated and the nominal coverage α. Zero coverage error means perfect uncertainty assessment; positive and negative errors mean one-sided over- and under-coverage respectively. The stepwise approach over-estimates coverage for both levels of x and all α. At relatively low x-levels of $\theta(x,m)$, the gain in coverage accuracy

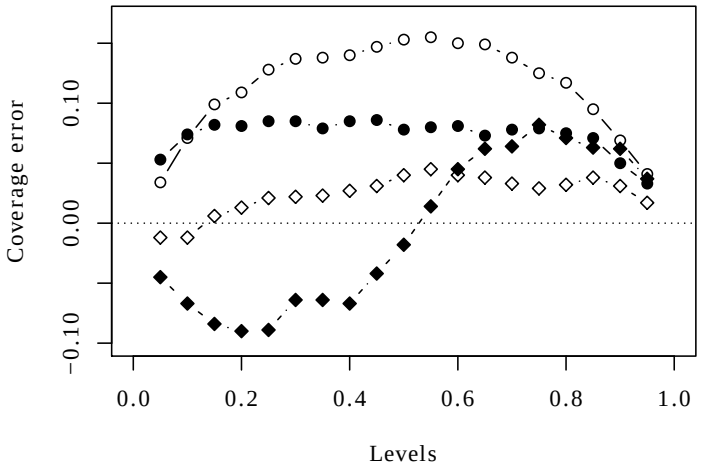

Figure 5.4 – Coverage error for $\theta(x,m)$ computed for confidence levels $0.05, 0.1, \ldots, 0.95$, for $x = 98\%$ (empty symbols) and $x = 99.999\%$ (filled symbols), with $m = 1$, for the stepwise approach (circles) and the Bayesian semiparametric approach (diamonds).

by the Bayesian approach is remarkable in particular at mid-coverage levels, but it shows no marked improvement for larger x.

5.8 Data analysis

River flooding can badly damage properties and have huge insurance costs. Large-scale floods in the UK since the year 2000 have caused insurance losses of £5 billion, and more than £400 million is spent each year on flood defences. Modelling the dependence of extreme water-levels is key to accurate prediction of flood risk.

Our application uses daily flows of the River Ray at Grendon Underwood, north-west of London, for the 47 winters from 1962 to 2008. We assume stationarity of the series over the winter months. River flows in this catchment are typically short-range dependent: after heavy rainfall the flow can reach high values before decreasing gradually as the river returns to its baseflow regime. We thus expect the flow to be dependent at extreme levels and at small lags, so a small run length m is required. For illustration we take $m = 1, 7$, with the former being the more appropriate.

Standard graphical methods (Coles, 2001) were used to choose the 95% empirical quantile as the marginal threshold. A sensitivity analysis on a range of thresholds gave results similar to those below. The Heffernan–Tawn model was then fitted to the data transformed to Laplace margins, with u as the 98% empirical quantile and $m = 1, 7$. A higher threshold was selected for the dependence modelling to ensure the independence of X_0 and $\boldsymbol{Z}$ in approximation (5.8).

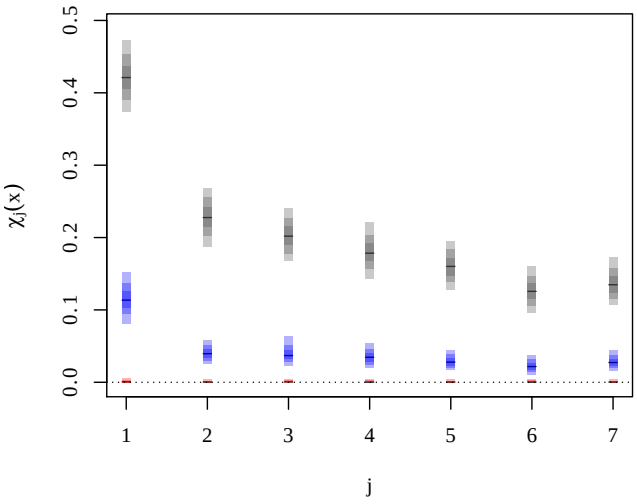

Figure 5.5 – Posterior summaries for $\chi_j(x)$, $j = 1,\ldots,7$, with x the 95% (grey), 99% (blue), and 99.99% (red) marginal quantiles of the River Ray data. The different shades of colour indicate the 95%, 80%, and 50% highest density regions of the posterior densities. The black segments indicate the posterior medians.

We first investigate the asymptotic structure of the data at different lags. We use $\chi_j(x) = \Pr(X_j > x \mid X_0 > x)$, $j = 1,\ldots,m$, whose limit χ_j (cf. Section 5.2) either measures the degree of association within the asymptotic dependence class when $\chi_j > 0$ or indicates asymptotic independence when $\chi_j = 0$. A Monte Carlo integration similar to that used for estimating the posterior distribution of $\theta(x,m)$ is applied to get the posterior distribution of $\chi_j(x)$ for selected values of x and $j = 1,\ldots,7$ depicted in Figure 5.5, where the posterior densities are summarised using highest density regions (Hyndman, 1996). Convergence of $\chi_j(x)$ to 0 at all lags is supported by the model. As expected we observe a monotone decay in extremal dependence over time lag. The flexibility of the conditional model is well illustrated here, as the procedure establishes positive dependence at any finite level but anticipates asymptotic independence of successive daily flows.

From the estimates of $\chi_j(x)$ we expect $\theta = 1$ but $\theta(x,m) < 1$ for $x < x^F$. We computed estimates of $\theta(x,m)$ for several values of x based on the posterior distribution fitted with the Bayesian semiparametric approach and compared them with the stepwise approach and the empirical estimates. We give block bootstrap confidence intervals for the stepwise and empirical estimates, with a block length ensuring that winters are not split between blocks. Figure 5.1 shows estimates of $\theta(x,4)$ obtained with the three methods, with $m = 4$ to show an intermediate estimate, as $\theta(x,7) \leq \theta(x,4) \leq \theta(x,1)$. The three methods broadly agree at historical levels, with wider confidence intervals for the runs estimate. For higher x, the stepwise estimate predicts stronger dependence than does the Bayesian estimate.

	$m=1$			$m=7$		
Level	Empirical	Stepwise	Bayesian	Empirical	Stepwise	Bayesian
95%	$62_{(58,68)}$	$62_{(61,70)}$	$62_{(57,66)}$	$35_{(33,41)}$	$36_{(33,42)}$	$33_{(27,38)}$
99%	$90_{(83,95)}$	$87_{(83,91)}$	$88_{(83,91)}$	$80_{(70,88)}$	$69_{(66,79)}$	$76_{(71,80)}$
99.9%	–	$96_{(93,99)}$	$97_{(94,99)}$	–	$90_{(86,96)}$	$98_{(96,99)}$
99.99%	–	$99_{(96,100)}$	$100_{(98,100)}$	–	$98_{(93,100)}$	$100_{(99,100)}$

Table 5.2 – Estimation of the threshold-based extremal index $\theta(x,m)$ (%) for four different levels of x and $m = 1,7$ on the Ray River winter flow data, with 95% confidence intervals (CI) given as subscripts. Empirical: runs estimator (block bootstrap CI); Stepwise: Heffernan–Tawn method (block bootstrap CI); Bayesian: posterior median from the Bayesian semiparametric approach (quantiles of the posterior distribution).

Table 5.2 shows that the three methods agree closely for $m = 1$, with the posterior distribution giving slightly tighter credible intervals than the two other methods. For $m = 7$, the Bayesian approach seems to improve a little on the stepwise estimates when compared to the empirical estimates at low levels, partly because of the Markov constraints on the α_j and β_j, which also reduce the uncertainty. In terms of convergence of $\theta(x,m)$ to the extremal index θ, we observe that at very high levels and for both values of m, the estimates of $\theta(x,m)$ tend to the same values, which indicates coherence in the approach. This also illustrates the lesser concern of the choice of the run-length when we are interested in tail probabilities, typically when estimating cluster maxima distributions.

The Bayesian semiparametric approach appears to offer a more coherent basis for the extrapolation to the required levels and uncertainty quantification for design purposes than does the stepwise method. We assessed the performance of Bayesian semiparametric estimates of $\theta(x,m)$, but other conditional probabilities could also be estimated using this approach.

5.9 Summary

In this chapter, we reviewed the finite-sample clustering properties in stationary time series, as measured by the subasymptotic extremal index $\theta(x,m)$ and by $\chi_j(x)$, defined in (5.1) and (5.2) respectively. We observed that the subasymptotic versions of θ and χ_j can converge very slowly to their respective limits, leading classical methods to overestimate clustering of extremes when extrapolating beyond the range of the data.

We then devised a Dirichlet process mixture appropriate for our approach, introducing a semiparametric model which allows estimation of the conditional tail model in a single stage, thus improving efficiency. We improved the mixing of successive posterior samples

by introducing label switching moves in the algorithm. Another benefit of the Bayesian semiparametric approach is to provide posterior distributions for the model parameters and for the conditional distribution G, thus naturally giving estimates of uncertainty for functions of the model parameters. Our new approach brings additional structure for fitting a parsimonious model, e.g., in time series, when a Markov property is a reasonable assumption.

We showed the multiple benefits of the Bayesian approach on simulated data, namely the efficiency improved compared to the stepwise inference procedure of Eastoe and Tawn (2012) and good coverage properties. We concluded with an illustration on river flow data and we showed, with the example of $\chi_j(x)$, how the Bayesian approach can naturally provide estimates of cluster functionals other than $\theta(x, m)$.

In the next chapter, we shall begin with the description of a new constraint for the conditional tail model, followed by the development of a joint likelihood for bivariate observations based on the conditional tail model, and completing the thesis with a contribution to fitting the extremes of first order Markov chains.

6 New improvements to the conditional tail model

6.1 Joint likelihood for bivariate extremes

Consider a bivariate random vector (X, Y) with marginal and joint distributions functions $F_X(\cdot)$ and $F_Y(\cdot)$, and $F_{X,Y}(\cdot,\cdot)$, respectively. A standard modelling approach, supported by the representation theorem of Sklar (1959), is to consider the marginal and joint features separately. In this setting, the dependence structure is often represented using a copula, as we saw in Section 2.1.2, which suggests fitting $F_X(\cdot)$ and $F_Y(\cdot)$ first, and then using the probability integral transform in order to convert the marginal distributions to the uniform scale, using $X^{\mathrm{U}} = \widehat{F}_X(X)$ and $Y^{\mathrm{U}} = \widehat{F}_Y(Y)$, with $\widehat{F}_X(\cdot)$ and $\widehat{F}_Y(\cdot)$ estimates of $F_X(\cdot)$ and $F_Y(\cdot)$ respectively.

A similar methodology exists for modelling bivariate componentwise maxima, with marginal distributions often transformed from generalised extreme value to a particular subclass of distributions, such as Fréchet or Gumbel, instead of the uniform distribution used in the standard copula approach. The joint extremal structure is fitted with models that are justified asymptotically. Many parametric and non-parametric models have been derived from Pickands' function $A(\cdot)$, from the spectral measure $V(\cdot,\cdot)$ or from the angular distribution $H(\cdot)$ (Section 2.1.2). When bivariate excesses of a threshold are considered, models such as the generalised Pareto distribution are used to transform the marginal distributions to a specific distribution, e.g., Fréchet, exponential, or Pareto. The key aspect of all these procedures is that they transform the problem of modelling (X, Y) with general margins to one where the margins are on the same scale, so that joint features are easier to model.

Standard inference for threshold methods for bivariate extremes involves generalised Pareto marginal distributions above a high threshold. In this approach, the margins are also considered as fixed when fitting the dependence structure. Uncertainty of the parameters of the joint structure is generally reported assuming known marginal distributions, which can yield standard errors that are too small. The work of Genest *et al.* (1995) suggests that a single-step procedure, i.e., fitting the marginal and joint features simultaneously, can yield moderate efficiency gains in the case of Clayton's copula (Clayton, 1978).

Max-stable bivariate distributions, stemming from the limit of normalised componentwise maxima, display complete independence or asymptotic dependence, meaning that the maxima of X and Y occur simultaneously with a positive probability (Section 2.3.1). Wadsworth and Tawn (2012) introduce the class of inverted max-stable distributions, whose copula's upper joint tail corresponds to the lower joint tail of the corresponding max-stable distribution: $(X^{\mathrm{E}}, Y^{\mathrm{E}})$, with $X^{\mathrm{E}} = 1/X^{\mathrm{F}}$ and $Y^{\mathrm{E}} = 1/Y^{\mathrm{F}}$, has an inverted max-stable distribution with exponential margins if $(X^{\mathrm{F}}, Y^{\mathrm{F}})$ has a max-stable distribution with Fréchet margins. Inverted max-stable distributions display asymptotic independence in their upper tails, with a dependence strength that depends on the magnitude of the events considered.

The conditional approach to extremes is not restricted to events that are simultaneously extreme in X and Y, and covers a broad class of max-stable and inverted max-stable distributions (Heffernan and Tawn, 2004; Papastathopoulos and Tawn, 2016). It is based on the assumption that there exist normalising functions $a(\cdot)$ and $b(\cdot) > 0$ such that, for $(X^{\mathrm{L}}, Y^{\mathrm{L}})$ in Laplace margins,

$$\lim_{u\to\infty} \Pr\left\{ \frac{Y^{\mathrm{L}} - a\left(X^{\mathrm{L}}\right)}{b\left(X^{\mathrm{L}}\right)} \leq z,\, X^{\mathrm{L}} - u > x \,\middle|\, X > u \right\} = H(z)\exp(-x), \quad x > 0, \tag{6.1}$$

with $H(\cdot)$ a non-degenerate distribution with no mass at infinity. Recall from Section 2.4.2, that the conditional tail model arising from (6.1) uses $a(x) = \alpha x$ and $b(x) = x^{\beta}$, so that the model takes the form

$$Y^{\mathrm{L}} \mid \{X^{\mathrm{L}} = x\} = \alpha x + x^{\beta} Z, \quad X^{\mathrm{L}} > u, \tag{6.2}$$

for some large threshold u, with $\alpha \in [-1, 1]$, $\beta \in (-\infty, 1]$ and $Z \sim H$ independent of X^{L}. Inference for the conditional tail model is easy to implement even in high-dimensional settings, but in the approach suggested by Heffernan and Tawn (2004) it involves multiple steps, where estimates computed in one step are considered as fixed in the next step, thus losing efficiency. In the first step, the marginal distribution is transformed to the Laplace scale using a rank transform or the semiparametric model (5.9) of Coles and Tawn (1994). In the second step, the norming parameters α and β are estimated by maximising a likelihood function, temporarily assuming the conditional tail distribution $H(\cdot)$ to be Gaussian. In the final step, $H(\cdot)$ is estimated by the empirical distribution function of the fitted residuals $\widehat{Z}_i = (Y_i^{\mathrm{L}} - \widehat{\alpha} X_i^{\mathrm{L}})/(X_i^{\mathrm{L}})^{\widehat{\beta}}$, $X_i^{\mathrm{L}} > u$, and $(X_i^{\mathrm{L}}, Y_i^{\mathrm{L}})$ $(i = 1, \ldots, n)$ represent independent replicates of $(X^{\mathrm{L}}, Y^{\mathrm{L}})$.

In Chapter 5, we developed a methodology which yields more efficient inference for the conditional tail model, but which still relies on a first step where the marginal distributions are fitted separately. This approach does not permit uncertainty of the marginal fit to be accounted for when fitting the dependence structure, thus potentially impacting the assessment of uncertainty for probability estimates of extreme sets, leading to a loss in efficiency.

Another important aspect is that this approach only considers one conditional distribution, i.e., $Y \mid X$ with X large, without looking at $X \mid Y$ with Y large at the same time. In the approach of Heffernan and Tawn (2004), the two conditional models are fitted as if they were

independent, since the likelihood function was taken as

$$\prod_{(x_i,y_i):x_i>u} f_{Y|X}(y_i \mid x_i) \times \prod_{(x_i,y_i):y_i>u} f_{X|Y}(x_i \mid y_i), \tag{6.3}$$

for some large threshold u, observations (x_i, y_i) $(i = 1,\ldots,n)$ transformed to the Laplace scale, and where $f_{Y|X}$ is the conditional density based on the model (6.2) with $H(\cdot)$ being Gaussian, and similarly for $f_{Y|X}$ with X and Y flipped. Another incorrect aspect of the joint density (6.3) is its double-counting of observations (x_i, y_i) having $\min(x_i, y_i) > u$, thus over-weighting jointly extreme events. In Section 6.2, we shall give two different approaches to a correct likelihood formulation.

The lack of self-consistency in the joint tail between each conditional distribution was recognised by Heffernan and Tawn (2004), since the formulation of the conditional model does not guarantee equality of estimates of, e.g., $\Pr(X^{\mathrm{L}} > v, Y^{\mathrm{L}} > v)$, for any $v > u$, extrapolated from $X \mid Y > u$ and from $Y \mid X > u$, for large u. Heffernan and Tawn suggest imposing weak pairwise exchangeability, i.e., $\alpha_{|x} = \alpha_{|y}$ and $\beta_{|x} = \beta_{|y}$, or strong pairwise exchangeability, i.e., weak pairwise exchangeability and $H_{|x}(\cdot) \equiv H_{|y}(\cdot)$, where the subscripts refer to the conditioning variables used in the corresponding models. When X and Y are not exchangeable, these assumptions lead to biased estimates. Liu and Tawn (2014) explore self-consistency and consider the densities

$$h_{|x}(z) = \frac{d}{dz} H_{|x}(z), \quad h_{|y}(z) = \frac{d}{dz} H_{|y}(z).$$

Self-consistency of the conditional models $Y \mid X = x$, with x large, and $X \mid Y = y$, with y large, requires

$$f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y),$$

for all x and y large, which implies

$$h_{|x}\left(\frac{y - \alpha_{|x} x}{x^{\beta_{|x}}}\right) e^{-x} = h_{|y}\left(\frac{x - \alpha_{|y} y}{y^{\beta_{|y}}}\right) e^{-y}, \tag{6.4}$$

for all x and y large. Liu and Tawn show that (6.4) cannot be satisfied for all points in the set $\{(x, y) : x > u, y > u\}$, assuming the densities $h_{|x}(\cdot)$ and $h_{|y}(\cdot)$ to be differentiable. They show that a weaker version of self-consistency holds, termed diagonal self-consistency, in which equality of the densities is imposed only on $\{(x, y) : x = y > u\}$, if and only if

$$h_{|x}(z) = \frac{1 - \alpha_{|x}}{z} h_0\left\{\left(\frac{1 - \alpha_{|x}}{z}\right)^{1/(\beta_{|x}-1)}\right\}, \quad z > (1 - \alpha_{|x}) u^{1-\beta_{|x}},$$
$$h_{|y}(z) = \frac{1 - \alpha_{|y}}{z} h_0\left\{\left(\frac{1 - \alpha_{|y}}{z}\right)^{1/(\beta_{|y}-1)}\right\}, \quad z > (1 - \alpha_{|y}) u^{1-\beta_{|y}},$$

with $h_0(\cdot) \colon (u, \infty) \to [0, \infty)$. Liu and Tawn (2014) propose the parametric family

$$h_0(z) = \omega_1 z^{\omega_2} e^{-\omega_3 z}, \quad \omega_i > 0,\ i = 1, 2, 3,$$

as a means to enforce diagonal self-consistency while covering several known asymptotically independent parametric families of distributions. This approach is interesting in that it can improve self-consistency across several conditional tail models applied to the same set of multivariate observations, but it is restricted by a stepwise fitting procedure not giving a full account of uncertainty for the model parameters, and by the incorrect likelihood based on (6.3).

In this chapter, we introduce a new methodology to jointly fit the two conditional tail models for $Y \mid X$ and $X \mid Y$, while setting a proper density for observations for which both conditional models are valid. This methodology allows the sharing of information between the two conditional models, thus improving the consistency of estimates of joint tail probabilities, while carrying the benefits of the conditional tail approach, and in particular its flexibility and its ability to fit asymptotically independent distributions. We censor non-extreme observations in order to add information to the fit and to capture extremal dependence better. In a second part, we also aim at reducing inconsistency of the conditional tail model, following a different route than Liu and Tawn. We tackle the issue of multi-stage inference and the loss of efficiency that it implies, and present a method to fit the marginal distributions of X and Y and their joint extremal dependence structure in a single stage. This offers a reliable and complete assessment of uncertainty for the model parameters and for risk estimates extrapolated from the fit. We also introduce new constraints for the conditional model under the hypothesis of positive or negative association, thus reinforcing the robustness of the fit in cases of weak identifiability of the model parameters.

6.2 New methodology for modelling extremes in one component

6.2.1 General framework

In order to simplify the following developments, we first assume strong exchangeability of the parameters and of the residual distributions for the conditional tail models corresponding to $X \mid Y$ and $Y \mid X$, i.e., we set $\alpha = \alpha_{|x} = \alpha_{|y}$, $\beta = \beta_{|x} = \beta_{|y}$, and $H(\cdot) \equiv H_{|x}(\cdot) \equiv H_{|y}(\cdot)$. We leave extensions where weak or no exchangeability is assumed to Section 6.3.3. We define several regions used in the subsequent developments, namely

$$\begin{aligned} R_{01} &= \left\{(x,y) \in \mathbb{R}^2 : x \leq u, y > u\right\}, \quad R_{11} = \left\{(x,y) \in \mathbb{R}^2 : x > u, y > u\right\}, \\ R_{00} &= \left\{(x,y) \in \mathbb{R}^2 : x \leq u, y \leq u\right\}, \quad R_{10} = \left\{(x,y) \in \mathbb{R}^2 : x > u, y \leq u\right\}, \\ R_{|x} &= \left\{(x,y) \in \mathbb{R}^2 : x > u, y \leq x\right\}, \quad R_{|y} = \left\{(x,y) \in \mathbb{R}^2 : y > u, x < y\right\}, \end{aligned} \tag{6.5}$$

and we show an illustration of them in Figure 6.1.

A joint fit of models $M_{|x}$, in which $Y \mid X = x$, $X > u$, corresponding to region $R_{10} \cup R_{11}$, and $M_{|y}$, in which $X \mid Y = y$, $Y > u$, corresponding to region $R_{01} \cup R_{11}$, requires careful attention for data lying in R_{11}, where both X and Y are large, as these data influence the fit of both $M_{|x}$ and $M_{|y}$. In their inference procedure, Heffernan and Tawn (2004) consider $M_{|x}$ and $M_{|y}$ as

Figure 6.1 – Regions where likelihood contributions differ. Left panel: setup used when contributions from $Y \mid X$ and $X \mid Y$ are averaged; right panel: setup used when contributions are split along the diagonal where $x = y$. The grey areas represent regions where the observations are censored.

separate models, yielding a log-likelihood of the form

$$\ell^{\mathrm{HT}}\left(\boldsymbol{\theta}_{|x}^{\mathrm{HT}}, \boldsymbol{\theta}_{|y}^{\mathrm{HT}}; x_i, y_i \, i = 1, \ldots, n\right) = \sum_{i=1}^{n} \mathbb{1}(x_i > u) \log f_{Y|X}\left(y_i \mid x_i; \boldsymbol{\theta}_{|x}^{\mathrm{HT}}\right) + \mathbb{1}(y_i > u) \log f_{X|Y}\left(x_i \mid y_i; \boldsymbol{\theta}_{|y}^{\mathrm{HT}}\right), \quad (6.6)$$

where $\boldsymbol{\theta}_{|x}^{\mathrm{HT}} = (\alpha_{|x}, \beta_{|x}, \mu_{|x}, \psi_{|x})$ and $\boldsymbol{\theta}_{|y}^{\mathrm{HT}} = (\alpha_{|y}, \beta_{|y}, \mu_{|y}, \psi_{|y})$ are the vectors of parameters for $M_{|x}$ and $M_{|y}$ respectively. The log-likelihood formulation (6.6) entails using the observations in R_{11} twice; non-extreme observations in R_{00} are not considered.

Our approach deals more carefully with the observations lying in R_{11}, where both $M_{|x}$ and $M_{|y}$ are valid, and as a consequence also incorporates information about the number of data points in R_{00}. Given (X, Y) with Laplace marginal distributions, the general form of our model is based on the joint distribution

$$F(x, y) = \begin{cases} F_{X,Y}(x, y), & \max(x, y) > u, \\ \widetilde{F}_{X,Y}(x, y) \dfrac{F_{X,Y}(u, u)}{\widetilde{F}_{X,Y}(u, u)}, & \max(x, y) \leq u, \end{cases} \quad (6.7)$$

where u is a large threshold on the Laplace scale, $\widetilde{F}_{X,Y}(\cdot, \cdot)$ is a nonparametric estimate of the joint distribution of (X, Y), and $F_{X,Y}(\cdot, \cdot)$ is a semiparametric estimate of $F(\cdot, \cdot)$ which we shall detail later. The term $F_{X,Y}(u, u)/\widetilde{F}_{X,Y}(u, u)$ ensures continuity of the joint distribution at (u, u), which is key in our approach, although discontinuity may occur on the half-lines (x, u), $x \leq u$ and (u, y), $y \leq u$. From (6.7), we derive the joint density, specifically

$$f(x, y) = \begin{cases} f_{X,Y}(x, y), & \max(x, y) > u, \\ \widetilde{f}_{X,Y}(x, y) \dfrac{F_{X,Y}(u, u)}{\widetilde{F}_{X,Y}(u, u)}, & \max(x, y) \leq u, \end{cases} \quad (6.8)$$

so that the parametric contribution of an observation in R_{00} to the likelihood function reduces to $F_{X,Y}(u, u)$, thus the information for data points in R_{00} is censored and is limited to the fact that these points lie below the bivariate threshold (u, u). Formulation (6.8) yields a bivariate

log-likelihood for observations (x_i, y_i), $i = 1, \ldots, n$, assuming Laplace margins, specifically

$$\ell\left(\boldsymbol{\theta}_{x,y}; x_i, y_i, i = 1, \ldots, n\right) = \sum_{i=1}^{n} \left[\mathbb{1}\left\{(x_i, y_i) \notin R_{00}\right\} \log f_{X,Y}\left(x_i, y_i; \boldsymbol{\theta}_{x,y}\right) \right] + n_{00} \log F_{X,Y}\left(u, u; \boldsymbol{\theta}_{x,y}\right),$$

where $\boldsymbol{\theta}_{x,y}$ is the vector of parameters for the joint distribution; $n_{00} = \sum_{i=1}^{n} \mathbb{1}\{(x_i, y_i) \in R_{00}\}$ is the number of censored observations and $f_{X,Y}(\cdot,\cdot)$ and $F_{X,Y}(\cdot,\cdot)$ are the joint density and distribution functions of (X, Y).

We now model the joint density function $f_{X,Y}(\cdot,\cdot)$ in the L-shaped region $R_{10} \cup R_{11} \cup R_{01}$. We suggest a first approach where the likelihood contributions from the two conditional models $M_{|x}$ and $M_{|y}$ are averaged out in R_{11}, $M_{|x}$ is used in R_{10} and $M_{|y}$ is used in R_{01}. This setup corresponds to the left panel of Figure 6.1. An alternative setup is to split R_{11} along the diagonal where $x = y$, i.e., to consider contributions from $M_{|x}$ and $M_{|y}$ only when $x > y$ and $y < x$ respectively. This setup corresponds to the regions depicted in the right panel of Figure 6.1. This setup may seem to resemble the extrapolation procedure of Heffernan and Tawn (2004) outlined in Section 2.4.2, where probabilities in $R_{|x}$ and $R_{|y}$ are estimated separately to form an estimate of a joint tail probability in R_{11}. In effect, after having fitted the incorrect log-likelihood (6.6), Heffernan and Tawn post-process probabilities estimated from $M_{|x}$ and $M_{|y}$ to compute extrapolations. In contrast, our proposal is to use the correct likelihood from the start. We take a fundamentally different approach to that of Heffernan and Tawn, as we consider splitting R_{11} in order to construct a model and likelihood function that are coherent in R_{11}.

In the following developments, we shall use the terms *average* and *split* to refer to the two models arising from the two different setups pictured in Figure 6.1.

6.2.2 Likelihood contributions averaged in R_{11}

In the average likelihood setup, we need to define the different likelihood contributions in the four regions R_{11}, R_{00}, R_{10} and R_{01}. Note that the threshold u in Figure 6.1 is chosen large enough such that $u > 0$. We first consider pairs of observations $(x, y) \in R_{11}$, where we construct a coherent likelihood which we derive as follows,

$$\begin{aligned} F_{X,Y}(x, y) &= \Pr\left(X \leq x, Y \leq y\right) \\ &= \frac{1}{2}\Pr\left(Y \leq y \mid X \leq x\right)\Pr\left(X \leq x\right) + \frac{1}{2}\Pr\left(X \leq x \mid Y \leq y\right)\Pr\left(Y \leq y\right) \\ &= \frac{1}{2}\int_{-\infty}^{x} \Pr\left(Y \leq y \,\middle|\, X = s\right)\frac{1}{2}e^{-|s|}ds + \frac{1}{2}\int_{-\infty}^{y} \Pr\left(X \leq x \mid Y = t\right)\frac{1}{2}e^{-|t|}dt. \end{aligned} \tag{6.9}$$

By differentiating both sides of (6.9), we get

$$f_{X,Y}(x, y) = \frac{1}{4}\left\{f_{Y|X}(y \mid x)e^{-x} + f_{X|Y}(x \mid y)e^{-y}\right\}. \tag{6.10}$$

Since both x and y are extreme in R_{11}, we can model the conditional densities $f_{Y|X}(\cdot \mid \cdot)$ and $f_{X|Y}(\cdot \mid \cdot)$ using the conditional tail approach. We shall describe the detailed inference procedure in Section 6.3.

The likelihood contribution of any of the censored observations $(x_i, y_i) \in R_{00}$ is

$$F_{X,Y}(u,u) = 1 - \overline{F}_X(u) - \overline{F}_Y(u) + \overline{F}_{X,Y}(u,u) = 1 - e^{-u} + \overline{F}_{X,Y}(u,u), \tag{6.11}$$

where the last term corresponds to $\Pr\{(X,Y) \in R_{11}\}$ and can be derived from

$$\begin{aligned} \overline{F}_{X,Y}(u,u) &= \Pr(X > u, Y > u) \\ &= \frac{1}{2}\int_u^\infty \Pr(Y > u \mid X = x)\frac{1}{2}e^{-x}dx + \frac{1}{2}\int_u^\infty \Pr(X > u \mid Y = y)\frac{1}{2}e^{-y}dy. \end{aligned} \tag{6.12}$$

Similarly to the modelling of the likelihood contribution (6.10), the conditional probabilities in (6.12) can be modelled using $M_{|x}$ and $M_{|y}$ respectively.

The likelihood contributions of observations (x_i, y_i) in R_{10} and R_{01}, i.e., when exactly one of x_i and y_i exceeds the threshold u, can be directly modelled using the conditional tail approach, using the formulation

$$f_{X,Y}(x,y) = \begin{cases} f_{Y|X}(y \mid x)\frac{1}{2}e^{-x}, & (x,y) \in R_{10}, \\ f_{X|Y}(x \mid y)\frac{1}{2}e^{-y}, & (x,y) \in R_{01}, \end{cases} \tag{6.13}$$

so that the conditioning variable in the conditional density terms corresponds to an threshold excess in both situations.

6.2.3 Likelihood contributions split in R_{11}

In this alternative setup, we need to define likelihood contributions in the three regions R_{00}, $R_{|x}$ and $R_{|y}$. We first consider $(x,y) \in R_{00}$, for which we write (6.11) as

$$F_{X,Y}(u,u) = 1 - \Pr\{\max(X,Y) > u\} = 1 - \Pr\big\{(X,Y) \in R_{|x} \cup R_{|y}\big\}, \tag{6.14}$$

where the probability of (X,Y) being in the L-shaped region $R_{|x} \cup R_{|y}$ is

$$\begin{aligned} \Pr\big\{(X,Y) \in R_{|x} \cup R_{|y}\big\} &= \Pr\big\{(X,Y) \in R_{|x}\big\} + \Pr\big\{(X,Y) \in R_{|y}\big\} \\ &= \int_{R_{|x}} f_{Y|X}(y \mid x)\frac{1}{2}e^{-x}dydx + \int_{R_{|y}} f_{X|Y}(x \mid y)\frac{1}{2}e^{-y}dxdy \\ &= \int_u^\infty \Pr(Y \le x \mid X = x)\frac{1}{2}e^{-x}dx + \int_u^\infty \Pr(X \le y \mid Y = y)\frac{1}{2}e^{-y}dy. \end{aligned} \tag{6.15}$$

Compared to the survival distribution (6.12) for which integrals are averaged over R_{11}, expression (6.15) builds on separate calculations of probabilities in the region where at least X or Y is extreme.

The likelihood contributions of observations in $R_{|x}$ and $R_{|y}$ in the split likelihood setup are identical to those in R_{10} and R_{01} for the average likelihood setup, but differ in $R_{|x} \setminus R_{10}$ and $R_{|y} \setminus R_{01}$. These contributions are

$$f_{X,Y}(x,y) = \begin{cases} f_{Y|X}(y \mid x)\frac{1}{2}e^{-x}, & (x,y) \in R_{|x}, \\ f_{X|Y}(x \mid y)\frac{1}{2}e^{-y}, & (x,y) \in R_{|y}. \end{cases} \tag{6.16}$$

Although we do not consider a conditional likelihood here, we now give the density of observations (x,y) conditional on $(x,y) \in R_{|x}$, namely

$$f_{X,Y|X>Y}(x,y) = \frac{f_{Y|X}(y \mid x)\frac{1}{2}e^{-x}}{\int_{R_{|x}} f_{Y|X}(t \mid s)\frac{1}{2}e^{-s}dtds} = \frac{f_{Y|X}(y \mid x)e^{-x}}{\int_u^\infty \Pr(Y < s \mid X = s)e^{-s}ds}, \tag{6.17}$$

and similarly, the density of (x,y) conditional on $(x,y) \in R_{|y}$ is

$$f_{X,Y|X<Y}(x,y) = \frac{f_{X|Y}(x \mid y)\frac{1}{2}e^{-y}}{\int_{R_{|y}} f_{X|Y}(s \mid t)\frac{1}{2}e^{-t}dsdt} = \frac{f_{X|Y}(x \mid y)e^{-y}}{\int_u^\infty \Pr(X < s \mid Y = s)e^{-s}ds}, \tag{6.18}$$

thus conditioning implies dividing the likelihood contributions (6.16) by the terms in the denominator of (6.17) and (6.18).

6.2.4 Full methodology including the margins

We now present the methodology for fitting the marginal and joint distributions simultaneously, without assuming particular common marginal distributions for (X,Y) in the first place. We need the mappings $T_X(x)\colon x \mapsto x^{\mathrm{L}}$ and $T_Y(y)\colon y \mapsto y^{\mathrm{L}}$ to transform x and y from the original marginal scales of (X,Y) to the Laplace scale, with

$$\begin{aligned} T_X(x) &= \begin{cases} \log\{2F_X(x)\}, & F_X(x) \le 1/2, \\ -\log[2\{1-F_X(x)\}], & F_X(x) > 1/2, \end{cases} \\ T_Y(y) &= \begin{cases} \log\{2F_Y(y)\}, & F_Y(y) \le 1/2, \\ -\log[2\{1-F_Y(y)\}], & F_Y(y) > 1/2. \end{cases} \end{aligned} \tag{6.19}$$

We use the semiparametric model of Coles and Tawn (1994) for $F_X(\cdot)$ and $F_Y(\cdot)$ for given high thresholds u_x and u_y respectively, assuming X and Y are in the domain of attraction of the

generalised Pareto distribution, namely

$$F_X(x) = \begin{cases} \widetilde{F}_X(x), & x \le u_x, \\ 1 - \{1 - \widetilde{F}_X(u_x)\} \left(1 + \xi_x \dfrac{x - u_x}{\sigma_{u,x}}\right)_+^{-1/\xi_x}, & x > u_x, \end{cases}$$
$$F_Y(y) = \begin{cases} \widetilde{F}_Y(y), & y \le u_y, \\ 1 - \{1 - \widetilde{F}_Y(u_y)\} \left(1 + \xi_y \dfrac{y - u_y}{\sigma_{u,y}}\right)_+^{-1/\xi_y}, & y > u_y, \end{cases} \tag{6.20}$$

with ξ_x, ξ_y the shape parameters, $\sigma_{u,x}, \sigma_{u,y} > 0$ the scale parameters of the generalised Pareto distribution (GPD) and $\widetilde{F}_X(\cdot), \widetilde{F}_Y(\cdot)$ the empirical distribution functions associated with X and Y respectively.

We can work out a joint likelihood function for the marginal parameters $\boldsymbol{\theta}_x = (\sigma_{u,x}, \xi_x)$ and $\boldsymbol{\theta}_y = (\sigma_{u,y}, \xi_y)$ and for the vector of dependence parameters $\boldsymbol{\theta}_{x,y}$ which can be used for a joint fit, detailed in Section 6.3. This joint likelihood can be derived using either the average or split likelihood approach joined with (6.19) and (6.20), and it has the form

$$\begin{aligned} &\ell\left(\boldsymbol{\theta}_x, \boldsymbol{\theta}_y, \boldsymbol{\theta}_{x,y}; x_i, y_i,\, i = 1, \ldots, n\right) \\ &\qquad = \ell\left(\boldsymbol{\theta}_x, \boldsymbol{\theta}_y, \boldsymbol{\theta}_{x,y}; x_i^{\mathrm{L}}, y_i^{\mathrm{L}},\, i = 1, \ldots, n\right) + \sum_{i=1}^{n} \left\{\log J_X\left(x_i; \boldsymbol{\theta}_x\right) + \log J_Y\left(y_i; \boldsymbol{\theta}_y\right)\right\}, \end{aligned}$$

where $x_i^{\mathrm{L}} = T_X(x_i; \boldsymbol{\theta}_x)$, $y_i^{\mathrm{L}} = T_Y(y_i; \boldsymbol{\theta}_y)$, and $J_X(\cdot)$ and $J_Y(\cdot)$ are the Jacobians of the transformations $T_X(\cdot)$ and $T_Y(\cdot)$. Keeping only terms depending on the likelihood parameters $\boldsymbol{\theta}_x, \boldsymbol{\theta}_y, \boldsymbol{\theta}_{x,y}$, we get

$$\begin{aligned} \ell\left(\boldsymbol{\theta}_x, \boldsymbol{\theta}_y, \boldsymbol{\theta}_{x,y}; x_i, y_i,\, i = 1, \ldots, n\right) \propto \sum_{i=1}^{n} \Big[&\mathbb{1}\{(x_i, y_i) \notin R_{00}\} \log f_{X,Y}^{\mathrm{L}}\left(x_i^{\mathrm{L}}, y_i^{\mathrm{L}}; \boldsymbol{\theta}_x, \boldsymbol{\theta}_y, \boldsymbol{\theta}_{x,y}\right) \\ &+ \mathbb{1}(x_i > u_x) \log J_X\left(x_i; \boldsymbol{\theta}_x\right) + \mathbb{1}(y_i > u_y) \log J_Y\left(y_i; \boldsymbol{\theta}_y\right) \Big] \\ &+ n_{00} \log F_{X,Y}^{\mathrm{L}}\left(u^{\mathrm{L}}, u^{\mathrm{L}}; \boldsymbol{\theta}_{x,y}\right), \end{aligned} \tag{6.21}$$

where $u^L = T_X(u_x) = T_Y(u_y)$, assuming for simplicity that the thresholds u_x and u_y correspond to the same quantile, and $F_{X,Y}^{\mathrm{L}}(\cdot,\cdot)$ and $f_{X,Y}^{\mathrm{L}}(\cdot,\cdot)$ represent the joint distribution and density functions of $(X^{\mathrm{L}}, Y^{\mathrm{L}})$. Assuming $u_x, u_y > 0$, we have

$$\begin{aligned} J_X(x) = \frac{\partial T_X(x)}{\partial x} &= \frac{1}{1 - F_X(x)} \frac{1 - \widetilde{F}_X(u_x)}{\sigma_{u,x}} \left(1 + \xi_x \frac{x - u_x}{\sigma_{u,x}}\right)_+^{-1/\xi_x - 1} \\ &= \frac{1}{\sigma_{u,x}} \left(1 + \xi_x \frac{x - u_x}{\sigma_{u,x}}\right)_+^{-1}, \qquad x > u_x, \\ J_Y(y) = \frac{\partial T_Y(y)}{\partial y} &= \frac{1}{\sigma_{u,y}} \left(1 + \xi_y \frac{y - u_y}{\sigma_{u,y}}\right)_+^{-1}, \qquad y > u_y, \end{aligned} \tag{6.22}$$

and $J_X(x)$ and $J_Y(y)$ are constant in the GPD parameters for $x \leq u$ and $y \leq u$ respectively, thus when maximising the log-likelihood (6.21) we only need to compute expressions in (6.22) on the subset of observations in the L-shaped region $R_{10} \cup R_{11} \cup R_{01} = R_{|x} \cup R_{|y}$.

In the log-likelihood (6.21), the exact forms of the joint density $f^{\mathrm{L}}_{X,Y}(\cdot,\cdot)$ and of the joint distribution $F^{\mathrm{L}}_{X,Y}(\cdot,\cdot)$ depend on whether the average or split setup is used, implying a different treatment of observations in R_{11}, which are extreme in both X and Y. We develop the inference procedures for these two approaches in more details in the next section.

6.3 Inference

6.3.1 Strong exchangeability and Gaussian residuals

Here we consider a simplified setup of the model developed in Section 6.2. We assume strong exchangeability of (X,Y), and normality of $H(\cdot)$ in (6.1), with mean μ and variance ψ^2, so that $\alpha_{|x} = \alpha_{|y} = \alpha$, $\beta_{|x} = \beta_{|y} = \beta$, $H_{|x}(\cdot) \equiv H_{|y}(\cdot) \equiv H(\cdot)$. Using the notation of Section 6.2, the vector of parameters for the joint structure is $\boldsymbol{\theta}_{X,Y} = (\alpha, \beta, \mu, \psi^2)$. Assuming a Gaussian conditional distribution for the residuals, we get simple expressions for the different likelihood contributions for the average likelihood of Section 6.2.2; contribution (6.10) becomes

$$f_{X,Y}(x,y) = \frac{1}{4}\left\{ \frac{e^{-x}}{\psi x^{\beta}} \varphi\left(\frac{y - \alpha x - \mu x^{\beta}}{\psi x^{\beta}} \right) + \frac{e^{-y}}{\psi y^{\beta}} \varphi\left(\frac{x - \alpha y - \mu y^{\beta}}{\psi y^{\beta}} \right) \right\}, \quad (x,y) \in R_{11},$$

where $\phi(\cdot)$ is the standard normal density function; the survival distribution in contribution (6.11), corresponding to observations in R^c_{00}, becomes

$$\begin{aligned} \overline{F}_{X,Y}(u,u) &= \frac{1}{2}\int_u^{\infty} \overline{\Phi}\left(\frac{u - \alpha x - \mu x^{\beta}}{\psi x^{\beta}} \right) \frac{1}{2} e^{-x} dx + \frac{1}{2}\int_u^{\infty} \overline{\Phi}\left(\frac{u - \alpha y - \mu y^{\beta}}{\psi y^{\beta}} \right) \frac{1}{2} e^{-y} dy \\ &= \frac{1}{2}\int_u^{\infty} \overline{\Phi}\left(\frac{u - \alpha s - \mu s^{\beta}}{\psi s^{\beta}} \right) e^{-s} ds, \end{aligned} \tag{6.23}$$

thanks to the strong exchangeability of X and Y, where $\overline{\Phi}(\cdot) = 1 - \Phi(\cdot)$ is the survival function of the standard normal distribution; for the contributions in R_{10} and R_{01} discussed generally in (6.13), we have

$$f_{X,Y}(x,y) = \begin{cases} \frac{e^{-x}}{2\psi x^{\beta}} \varphi\left(\frac{y - \alpha x - \mu x^{\beta}}{\psi x^{\beta}} \right), & (x,y) \in R_{10}, \\ \frac{e^{-y}}{2\psi y^{\beta}} \varphi\left(\frac{x - \alpha y - \mu y^{\beta}}{\psi y^{\beta}} \right), & (x,y) \in R_{01}. \end{cases} \tag{6.24}$$

Similarly, for the split likelihood described in Section 6.2.3, contribution (6.14) for observations in R_{00} needs the computation of the probability (6.15) in R_{00}^c, which is

$$\begin{aligned}\Pr\{(X,Y)\in R_{|x}\cup R_{|y}\} &= \int_u^\infty \Phi\left\{\frac{(1-\alpha)x-\mu x^\beta}{\psi x^\beta}\right\}\frac{1}{2}e^{-x}dx + \int_u^\infty \Phi\left\{\frac{(1-\alpha)y-\mu y^\beta}{\psi y^\beta}\right\}\frac{1}{2}e^{-y}dy \\ &= \int_u^\infty \Phi\left\{\frac{(1-\alpha)s-\mu s^\beta}{\psi s^\beta}\right\}e^{-s}ds, \end{aligned} \tag{6.25}$$

thanks to the strong exchangeability of X and Y; contributions for observations in $R_{|x}$ and $R_{|y}$ are identical in form to those in (6.24) for R_{10} and R_{01} respectively, except over the extended sets $R_{|x}\supset R_{10}$ and $R_{|y}\supset R_{01}$, namely

$$f_{X,Y}(x,y)=\begin{cases}\frac{e^{-x}}{2\psi x^\beta}\varphi\left(\frac{y-\alpha x-\mu x^\beta}{\psi x^\beta}\right), & (x,y)\in R_{|x},\\ \frac{e^{-y}}{2\psi y^\beta}\varphi\left(\frac{x-\alpha y-\mu y^\beta}{\psi y^\beta}\right), & (x,y)\in R_{|y}.\end{cases}$$

The expressions for the likelihood contributions presented in this section offer a way to estimate joint tail probabilities using the flexibility of the conditional approach of Heffernan and Tawn (2004) while correctly characterising the joint tail of X and Y. The integrals (6.23) and (6.25) have no closed form even if $H(\cdot)$ is assumed to be Gaussian. Many statistical packages and computer libraries for general-purpose languages offer routines that compute $\Phi(x)=\int_{-\infty}^x \varphi(s)ds$ very efficiently, and also routines that compute univariate integrals using quadratic approximations, such as the QUADPACK library in FORTRAN, which has been ported to R (Piessens *et al.*, 1983). Even if such routines give accurate and fast approximations and are very helpful in practice, we give results that help understand the behaviour of this type of integral as u approaches infinity.

We start with a simplified setup corresponding to independence of X and Y and standard Gaussian residuals, i.e., $\alpha=\beta=\mu=0$ and $\psi=1$.

Theorem 6.1
The following asymptotic expansion holds as $u\to\infty$,

$$\int_u^\infty \Phi(x)e^{-x}dx = e^{-u} - e^{1/2}\frac{\varphi(u+1)}{u^2} + o\left\{\frac{\varphi(u+1)}{u^2}\right\}. \tag{6.26}$$

Proof An approximation of the left-hand side of (6.26) for large u can be derived using Mill's ratio, i.e., $1-\Phi(x)\sim\varphi(x)/x$ as $x\to\infty$, giving

$$\begin{aligned}\int_u^\infty e^{-x}\left\{1-\frac{\varphi(x)}{x}\right\}dx &= e^{-u} - \int_u^\infty \frac{1}{x\sqrt{2\pi}}e^{-(x^2+2x)/2}dx \\ &= e^{-u} - e^{1/2}\int_u^\infty \frac{1}{x\sqrt{2\pi}}e^{-(x+1)^2/2}dx \\ &= e^{-u} - e^{1/2}\int_{u+1}^\infty \frac{\varphi(y)}{y-1}dy,\end{aligned}$$

where $y = x + 1$. Integration by parts yields

$$\int_{u+1}^{\infty} \frac{\varphi(y)}{y-1} dy = \left\{ \left[-\frac{\varphi(y)}{y(y-1)} \right]_{u+1}^{\infty} - \int_{u+1}^{\infty} \frac{(2y-1)\varphi(y)}{y^2(y-1)^2} dy \right\} = \frac{\varphi(u+1)}{u(u+1)} + o\left\{ \frac{\varphi(u+1)}{u(u+1)} \right\},$$

hence the result. ■

In the case of the average likelihood and assuming strong exchangeability of X and Y, the probability mass in R_{00} can be derived from Theorem 6.1 and is approximately

$$\Pr\{(X,Y) \in R_{00}\} \approx 1 - e^{-u} + e^{1/2} \frac{\varphi(u+1)}{u^2} = 1 - e^{-u} \left(1 - \frac{e^{-u^2/2}}{\sqrt{2\pi}\, u^2} \right).$$

Theorem 6.2
If (X,Y) have Laplace marginal distributions and are asymptotically independent ($\alpha < 1$) and strongly exchangeable in the sense of the conditional tail approach, i.e.,

$$Y \mid \{X = x\} = \alpha x + x^{\beta} Z_x, \quad x > u, \qquad X \mid \{Y = y\} = \alpha y + y^{\beta} Z_y, \quad y > u,$$

for large u and Gaussian residuals $Z_x, Z_y \sim \mathcal{N}(\mu, \psi^2)$, then, with

$$\mu(x) = \alpha x + \mu x^{\beta}, \quad \psi(x) = \psi x^{\beta},$$

we have as $u \to \infty$,

$$e^{-u} - \int_u^{\infty} \Phi\left\{ \frac{x - \mu(x)}{\psi(x)} \right\} e^{-x} dx \sim \frac{1}{\sqrt{2\pi}} \frac{\psi^3}{(1-\alpha)^3(1-\beta)} u^{3\beta-2} \exp\left[-\frac{u^2 - 2u\{\mu(u) + \psi(u)^2\} + \mu(u)^2}{2\psi(u)^2} \right],$$

provided $\beta < 1/2$.

Proof The proof is detailed in Appendix G. ■

Theorem 6.2 assumes that (X,Y) are asymptotically independent. For (X,Y) on the Laplace scale and asymptotically dependent, we have $\alpha = 1$ and $\beta = 0$, thus

$$\begin{aligned} \int_u^{\infty} \Phi\left\{ \frac{x - \mu(x)}{\psi(x)} \right\} e^{-x} dx &= \int_u^{\infty} \Phi\left(-\mu/\psi\right) e^{-x} dx \\ &= \Phi\left(-\mu/\psi\right) e^{-u}. \end{aligned}$$

A numerical approach to estimating the type of integrals of Theorem 6.2 was suggested by Hugo Winter in his Ph.D. thesis (Winter, 2015, Chap. 2), and we discuss the approximations implied by his approach and improve it in Appendix H, where we present other approaches to numerical integration which can prove useful in higher dimensions.

6.3.2 Estimation of conditional quantiles

In this section, we detail the procedure to estimate conditional quantiles using the split approach. In this section, we assume that (X, Y) have Laplace margins to simplify notation. We show how to sample replicates from the conditional distribution $Y \mid X = x$ with x large; any conditional quantile given $X = x$ can then be derived using the empirical distribution of these replicates.

For a given value x, we can compute the probability that $y \sim Y \mid X = x$ belongs to $R_{|x}$, namely $\Phi\{(x - \alpha x - \mu x^{\beta})/(\psi x^{\beta})\}$. With probability p, we sample in $R_{|x}$ using a $\mathcal{N}(\alpha x + \mu x^{\beta}, \psi^2 x^{2\beta})$ truncated on $(-\infty, x]$. With probability $1-p$, we sample in $R_{|y}$ by sampling $L^{(r)}$ from a Laplace distribution truncated on $[x, \infty)$ and by solving

$$\int_{-\infty}^{y} f_{Y|X}(t \mid x)\, dt = L^{(r)} \tag{6.27}$$

in y, using

$$\begin{aligned} 1 - \int_{y}^{\infty} f_{Y|X}(t \mid x)\, dt &= 1 - \int_{y}^{\infty} f_{X|Y}(x \mid t) f(t)/f(x)\, dt \\ &= 1 - \int_{y}^{\infty} \Phi\left(\frac{x - \alpha t - \mu t^{\beta}}{\psi t^{\beta}}\right) e^{x-t}\, dt, \end{aligned} \tag{6.28}$$

where $f(\cdot)$ denotes the Laplace density function. Since (6.28) is only applicable for $y > x$, we reject $L^{(r)}$ if $y < x$ solves (6.27) and we sample $L^{(r)}$ again. The detailed procedure is described in Algorithm 6.1, where we use $F_{\mathrm{L}}(\cdot)$ to denote the Laplace distribution function.

6.3.3 Generalisations and extensions

In the developments of Section 6.3.1, we assumed strong exchangeability of (X, Y) for clarity, but relaxation of this assumption and consideration of weak exchangeability or no exchangeability of X and Y only imply that some simplifications are not possible. Under weak exchangeability and without exchangeability, we have respectively

$$\boldsymbol{\theta}_{x,y} = \begin{cases} (\alpha, \beta, \mu_{|x}, \mu_{|y}, \psi^2_{|x}, \psi^2_{|y}), \\ (\alpha_{|x}, \alpha_{|y}, \beta_{|x}, \beta_{|y}, \mu_{|x}, \mu_{|y}, \psi^2_{|x}, \psi^2_{|y}). \end{cases}$$

Without exchangeability, (6.23) becomes

$$\overline{F}_{X,Y}(u,u) = \frac{1}{2}\int_{u}^{\infty} \overline{\Phi}\left(\frac{u - \alpha_{|x} x - \mu_{|x} x^{\beta_{|x}}}{\psi_{|x} x^{\beta_{|x}}}\right)\frac{1}{2}e^{-x}dx + \frac{1}{2}\int_{u}^{\infty} \overline{\Phi}\left(\frac{u - \alpha_{|y} y - \mu_{|y} y^{\beta_{|y}}}{\psi_{|y} y^{\beta_{|y}}}\right)\frac{1}{2}e^{-y}dy,$$

with $\mu_{|x} = \mu_{|y}$ and $\psi_{|x} = \psi_{|y}$ under weak exchangeability. Similar conclusions apply to the formulation of $\Pr\{(X, Y) \in R_{|x} \cup R_{|y}\}$ in (6.23), and conditional densities (6.24).

Algorithm 6.1: Sampling from the conditional distribution given $X = x$ with x extreme.

Input: $\alpha, \beta, \mu, \psi, x, \tilde{n}$
Set $\mu(x) \leftarrow\!\shortmid \alpha x + \mu x^{\beta}$
Set $\psi(x) \leftarrow\!\shortmid \psi x^{\beta}$
`# Differentiate regions` $R_{|x}$ `and` $R_{|y}$
Set $p \leftarrow\!\shortmid \Phi\left(\dfrac{x-\mu(x)}{\psi(x)}\right)$
for $r = 1$ **to** $\tilde{n}$ **do**
  Sample B from Bernoulli(p)
  if $B = 1$ **then** `# sample in` $R_{|x}$
    Sample $U^{(r)}$ from $U(0, p)$
    Set $Y^{(r)} \leftarrow\!\shortmid \Phi^{\leftarrow}\left(U^{(r)}\right)$
  else `# sample in` $R_{|y}$
    repeat
      Sample $U^{(r)}$ from $U(p, 1)$
      Set $L^{(r)} \leftarrow\!\shortmid F_L^{\leftarrow}\left(U^{(r)}\right)$
      Find y such that (6.27) holds
    until $y > x$
    Set $Y^{(r)} \leftarrow\!\shortmid y$
  end
end

Earlier in this section, we assumed the distribution of the residuals to be Gaussian, i.e., $H_{|x}(z) = \Phi\{(z-\mu_{|x})/\psi_{|x}\}$, $H_{|y}(z) = \Phi\{(z-\mu_{|y})/\psi_{|y}\}$, but this is a restrictive assumption that can yield badly-biased estimates. In practice, ideas developed in Chapters 3 and 5 can be useful for removing restrictions on the specific form of $H_{|x}(\cdot)$ and $H_{|y}(\cdot)$, using the flexibility of Dirichlet process mixtures not only as a way to alleviate prior assumptions, but also as a means for quantifying uncertainty for the model fit and for extrapolation of risk measures through Bayesian posterior analysis. We develop this approach in the context of extremes of Markov chains in Chapter 7.

Previous developments have focused on (X, Y) in the bivariate setup, but both the average and the split likelihood functions of Sections 6.2.2 and 6.2.3 can be extended to higher-dimensional setups, provided a sufficiently flexible model such as a Dirichlet process mixture is used for the residual distributions $H_{|1}(\cdot), \ldots, H_{|d}(\cdot)$ corresponding to $(d-1)$-dimensional distributions conditioned on $X_1, \ldots, X_d$ respectively. The 2-dimensional regions defined in Figure 6.1 can be extended to this d-dimensional setup, and we give an example with $d = 3$ in

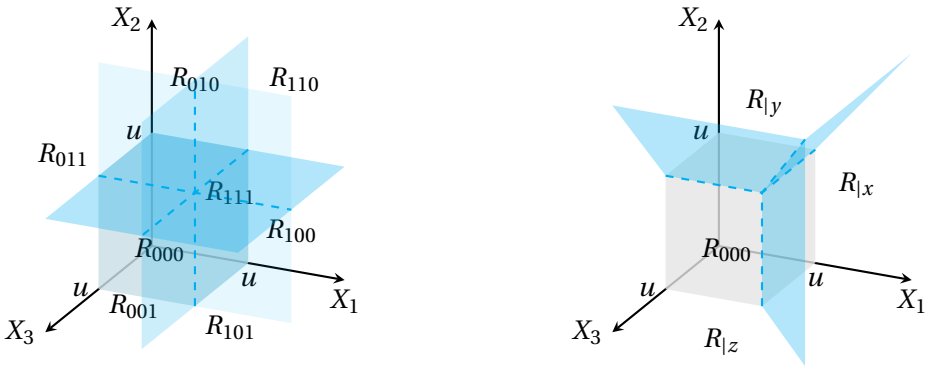

Figure 6.2 – Regions where likelihood contributions differ, in a 3-dimensional setup. Left panel: setup used when contributions from $(X_2, X_3) \mid X_1$, $(X_1, X_3) \mid X_2$ and $(X_1, X_2) \mid X_3$ are averaged; right panel: setup used when contributions are split along half-planes defined by $x_1 = x_2$ with $x_3 \leq x_1$, $x_2 = x_3$ with $x_1 \leq x_2$ and $x_1 = x_3$ with $x_2 \leq x_3$. The grey volumes represent regions where the observations are censored.

Figure 6.2 corresponding to the regions

$$R_{000} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : \max(x_1, x_2, x_3) \leq u\right\}, \qquad R_{001} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : \max(x_1, x_2) \leq u,\, x_3 > u\right\},$$
$$R_{010} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : \max(x_1, x_3) \leq u,\, x_2 > u\right\}, R_{100} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1 > u, \max(x_2, x_3) \leq u\right\},$$
$$R_{011} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1 \leq u, \min(x_2, x_3) > u\right\}, R_{101} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : \min(x_1, x_3) > u,\, x_2 \leq u\right\},$$
$$R_{110} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : \min(x_1, x_2) > u,\, x_3 \leq u\right\}, R_{111} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : \min(x_1, x_2, x_3) > u\right\},$$

for the average approach, and

$$R_{|1} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1 > \max(u, x_2, x_3)\right\}, \quad R_{|2} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 > \max(u, x_1, x_3)\right\},$$
$$R_{|3} = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 > \max(u, x_1, x_2)\right\},$$

for the split approach. As it appears in this list of regions and in the figure, the number of regions to consider in the average approach increases like 2^d as d increases, whereas in the split approach, the number of regions to consider is linear in d. This does not mean that the split approach is more parsimonious than the average approach, but the increased complexity of the likelihood function of the latter not only impacts inference, but also extrapolation and prediction.

6.4 New constraints for the conditional tail model

6.4.1 Existing fitting procedure

In the conditional tail approach, the parametric model (6.2) arising from the limit (6.1) has $(\alpha, \beta) \in [-1, 1] \times (-\infty, 1]$, which Keef *et al.* (2013) have shown to be too general, i.e., not all combinations of (α, β) yield consistent joint probabilities. In Section 2.4.3, we detailed the

informal approach of Keef *et al.* for deriving new constraints that tackle this issue. We recall here the main aspects leading to these constraints, for which Keef *et al.* introduce, on the uniform scale,

$$\chi^+ = \lim_{u\to 1} \Pr\left(Y^{\mathrm{U}} > u \mid X^{\mathrm{U}} > u\right), \qquad \chi^- = \lim_{u\to 1} \Pr\left(Y^{\mathrm{U}} \leq 1-u \mid X^{\mathrm{U}} > u\right),$$

which define asymptotic positive and negative dependence when $\chi^+ > 0$, $\chi^- > 0$ respectively, and asymptotic positive and negative independence when $\chi^+ = 0$, $\chi^- = 0$ respectively. The orderings of each of these two measures χ^+ and χ^- under different asymptotic regimes imply an ordering on the conditional quantiles of $Y \mid \{X = x\}$. In Section 2.4.3, we have formally shown that these orderings are satisfied when imposing constraints only for some specific conditional quantiles of $Y \mid \{X = x\}$ for large x.

As Figure 2.6 shows, the implementation of the constraints when fitting the conditional tail model can cut off large portions of $[-1,1] \times (-\infty, 1]$, thus reducing the space in which (α, β) live and improving the consistency of estimates of joint tail probabilities. Based on the conditional model (6.2), the conditional mean of $Y \mid \{X = x\}$ is

$$\mathrm{E}(Y \mid X = x) = \alpha x + \mu x^{\beta}, \quad \mathrm{E}(Z) = \mu. \tag{6.29}$$

It appears from this conditional mean that when β approaches 1, i.e., $\mathrm{E}(Y \mid X = x) \approx \alpha x + \mu x$, μ and α become unidentifiable, which can greatly hamper inference and derived risk estimates. By shrinking the space of (α, β), the constraints of Keef *et al.*, help reduce this identifiability issue in some cases, but in practice they are not as restrictive as in Figure 2.6. This lack of identifiability can yield parameter estimates for which the model fits the data well but gives bad extrapolations of probabilities at moderately extreme levels.

6.4.2 New constraints under positive and negative association

In many situations, it is reasonable, after collecting observations or as a prior belief, to assume positive or negative association of bivariate data (Sibuya, 1960; Lehmann, 1966). Standard extreme value theory induces models that only cover the former case, but the conditional approach to extremes provides enough flexibility to cover both. In particular, the sign of the α parameter in the model (6.2) reflects the type of association present in the data. In practice, because of the identifiability issue described in Section 6.4.1, the estimate for α can be negative and the mean of the residual μ overestimated even in the presence of positive association.

We suggest new constraints to cope with this identifiability issue and force the conditional tail model to reflect the structure of the data. The positive, respectively negative, association property implies that the conditional mean (6.29) has positive, respectively negative, slope in

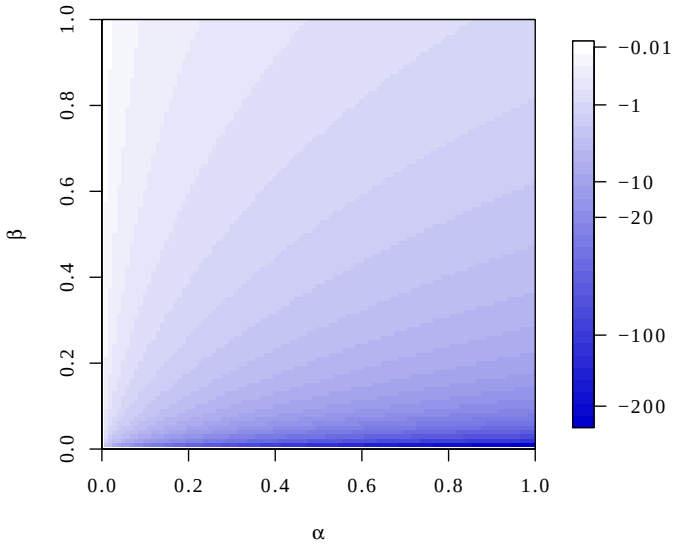

Figure 6.3 – Minimum values of μ satisfying the conditional mean constraint as a function of α and β under positive association, for a threshold u equal to the 95% Laplace quantile. Negative values of α are not shown as they do not satisfy the constraint.

x, for all large x. In order for this to hold, we require under positive association that

$$\alpha + \beta x^{\beta-1}\mu \geq 0, \quad x > u, \tag{6.30}$$

for some large threshold u, and similarly with the inequality sign reversed under negative association. As in Chapter 5, we restrict our attention to $\beta \geq 0$, as negative values of this parameter are unlikely in most environmental contexts, since it implies convergence of all conditional quantiles of $Y \mid \{X = x\}$ to a single value as $x \to \infty$. With $\beta \in [0,1]$, we derive the new constraints, equivalent to (6.30) but easier to verify,

$$\alpha + \beta u^{\beta-1}\mu \geq 0, \quad \alpha \geq 0,$$

by taking the limits of (6.30) as $x \to u$ and $x \to \infty$ respectively; negative association implies similar constraints with the inequality signs reversed. Figure 6.3 illustrates the case of positive association by showing the lower bound implied on μ for given values of α and β for a threshold u set at the 95% Laplace quantile. The constraint takes effect mainly when $\beta \approx 1$, e.g., $\beta = 0.8$ in this case yields $\mu > -1.5$, and smaller values of α imply larger values of μ.

6.5 Improving self-consistency of the conditional tail model

Liu and Tawn (2014) tackled the lack of self-consistency of the conditional tail models $M_{|x}$ and $M_{|y}$ by using a parametric family of distributions to model $H_{|x}(\cdot)$ and $H_{|y}(\cdot)$ such that their

corresponding densities $h_{|x}(\cdot)$ and $h_{|y}(\cdot)$ coincide along the upper diagonal $\{(x,y) \in \mathbb{R}^2 : x = y > u\}$, with u a high threshold. We aim to impose a less restrictive structure on $H_{|x}(\cdot)$ and $H_{|y}(\cdot)$. The main idea behind the following proposals is to penalise differences in probability estimates in R_{11}, where both $M_{|x}$ and $M_{|y}$ are valid.

A first approach uses a penalisation based on the Hellinger distance over R_{11} between the joint density functions derived from $M_{|x}$ and $M_{|y}$, specifically

$$\begin{aligned}\frac{1}{2}\int_u^\infty\int_u^\infty \left\{\sqrt{f(y)f_{X|Y}(x \mid y)} - \sqrt{f(x)f_{Y|X}(y \mid x)}\right\}^2 dxdy \\ = 1 - \int_u^\infty\int_u^\infty \sqrt{f(x)f(y)}\sqrt{f_{X|Y}(x \mid y)f_{Y|X}(y \mid x)}dxdy \\ = 1 - 8\int_u^\infty\int_u^\infty g(x)g(y)\sqrt{f_{X|Y}(x \mid y)f_{Y|X}(y \mid x)}dxdy,\end{aligned} \tag{6.31}$$

where $f(x) = \frac{1}{2}e^{-|x|}$ and $g(x) = \frac{1}{4}e^{-|x|/2}$ are the Laplace density functions with parameter 1 and 1/2 respectively. A similar approach is used by Wu and Hooker (2013) and Hooker and Vidyashankar (2014) in a different context, where they combine density functions based on a parametric family and on a nonparametric estimator to derive a robust density estimator. In a Bayesian context, Shemyakin (2014) defines prior functions using a Hellinger distance. An approximation of (6.31) can be calculated by Monte Carlo integration or using quadratic approximations. The Hellinger distance is bounded above by 1, so we suggest the penalisation

$$\lambda\left\{\int_u^\infty\int_u^\infty g(x)g(y)\sqrt{f_{X|Y}(x \mid y)f_{Y|X}(y \mid x)}dxdy\right\}^{-1}, \quad \lambda > 0, \tag{6.32}$$

in order to penalise density functions that are more distant more strongly. According to results in Liu and Tawn (2014), there exist no smooth conditional density functions for which the penalty term (6.32) would vanish, so we must relax the equality condition for the joint density functions so that it holds on a subset of R_{11}.

We propose to use a penalisation based on a discrete version of the Hellinger distance,

$$\sum_{k=1}^{K}\left\{\Pr(Y > v_k \mid X > v_k)^{1/2} - \Pr(X > v_k \mid Y > v_k)^{1/2}\right\}^2, \quad v_K > \cdots > v_1 > u, \tag{6.33}$$

with u the conditional threshold used in $M_{|x}$ and $M_{|y}$. Penalisation (6.33) can be incorporated in the likelihood function of the original procedure of Heffernan and Tawn (2004). It corresponds to weighting the difference between the kth joint survival probability by $2e^{v_k}$, thus more encouraging self-consistency of $M_{|x}$ and $M_{|y}$ in the tail. The use of the penalty (6.33) is delicate, as small joint survival probabilities are favoured, thus artificially reducing the risk of simultaneous extremes extrapolated from the fit.

An interesting penalisation that would cope with this issue is

$$\sum_{k=1}^{K} \left| \log \frac{\Pr(Y > v_k \mid X > v_k)}{\Pr(X > v_k \mid Y > v_k)} \right|, \quad v_K > \cdots > v_1 > u,$$

and we leave further investigations for future work.

6.6 Simulation study

6.6.1 Simulation processes

We consider four different simulation processes featuring asymptotic independence and asymptotic dependence. In each cases, the series of bivariate random variables are independent and identically distributed. We shall simulate pairs of observations from the conditional tail model as a benchmark. We also consider observations simulated from a Gaussian bivariate distribution, which is a classic example of a distribution not covered by standard extreme value models, and also features a Gaussian residual distribution $H(\cdot)$ in the conditional tail model. The last two simulation processes are the inverted logistic distribution, for which we know the assumption of Gaussian residuals is incorrect, and the logistic distribution, which is asymptotically dependent.

We now describe how to simulate bivariate replicates that are consistent with the conditional tail model. The simulation procedure requires careful attention in order to preserve both marginal and joint distributions. To sample data points in the L-shaped region $R_{|x} \cup R_{|y}$, we use a rejection procedure which we describe in Algorithm 6.2, yielding approximately exponential marginal distributions for a large enough threshold u. We sample $\tilde{n}$ data points that are extreme in one of their components using this procedure. For data points in R_{00}, we only need to sample the number of data points falling in this region, since censoring will be used in the fitting procedure. We derive the number of censored data points n_{00} using a negative binomial distribution with probability of success $\Pr\{(X, Y) \in R_{|x} \cup R_{|y}\}$ and number of successes $\tilde{n}$, i.e., the number of observations sampled in $R_{|x} \cup R_{|y}$. The probability of success can be computed using (6.25).

In the other three cases, for the Gaussian, inverted logistic and logistic bivariate distributions, we also set the number of extreme observations in $R_{|x} \cup R_{|y}$ to a fixed number $\tilde{n}$ so that the information available is equivalent in all cases considered.

We developed the code used in the following simulation studies in R (R Core Team, 2017). In particular, we compute parameter estimates with an optimiser using several different initial values for the parameters in order to improve convergence to the global maxima of the likelihood functions used in the simulation studies.

Algorithm 6.2: Sampling with rejection from multiple conditional models.

Input: $\alpha, \beta, \mu, \psi^2, u, \tilde{n}$
repeat
  Sample K from Bernoulli(1/2)
  Sample X from Exp(1)
  Set $X \leftarrow\!\shortmid X + u$
  Sample Z from $\mathcal{N}(\mu, \psi^2)$
  Set $Y \leftarrow\!\shortmid \alpha X + X^{\beta} Z$
  if $X > Y$ **then**
    if $K = 0$ **then**
      Keep (X, Y)
    else if $K = 1$ **then**
      Swap $X \leftrightarrow Y$
      Keep (X, Y)
    else
      Reject the pair (X, Y)
    end
  else
    Reject the pair (X, Y)
  end
until $\tilde{n}$ *pairs are sampled*

6.6.2 Fixed margins

In this first setup, we consider replicates of the pair $(X^{\mathrm{L}}, Y^{\mathrm{L}})$ in Laplace margins for all four simulation processes, so that marginal features are known and we can focus on the joint features. This setup also permits comparison with the original method of Heffernan and Tawn (2004), which assumes fixed margins. These comparisons must be taken with caution, as inference and extrapolation in our approaches are performed under the simplifying but restrictive assumption of normality of the residuals. As described in Section 6.6.1, the margins are approximately on the correct scale when using Algorithm 6.2; for the three other cases, namely Gaussian, inverted logistic and logistic bivariate distributions, we transform the margins using the probability integral transform to obtain exponential margins above u, and we set u to the 98% Laplace quantile.

In each of the four cases, we sample 1,000 data sets with $\tilde{n} = 1{,}000$ uncensored data points. We tried different combinations of the parameters, with all 54 combinations of $\alpha = -0.6, 0, 0.4$, $\beta = 0, 0.3, 0.9$, $\mu = -1, 0, 0.5$ and $\psi^2 = 1, 3$ for the conditional model, $\rho = -0.8, -0.3, 0, 0.3, 0.8$ for the correlation of the Gaussian distribution, and $\gamma = 0.2, 0.5, 0.8$ for the inverted logistic and logistic dependence parameter.

Table 6.1 gives the bias and relative efficiency of $\widehat{\alpha}$ and $\widehat{\beta}$ for representative cases, namely $(\alpha, \beta, \mu, \psi^2) = (0.4, 0.3, 0.5, 1)$, $\rho = 0.3$, and $\gamma = 0.5$, where the relative efficiency is the ratio of

		variance ×1000			bias ×1000			rel. efficiency	
		avg	spl	std	avg	spl	std	avg	spl
conditional	α	5.3	7.4	7.3	22.1	3.3	3.9	0.9	1.0
	β	16.1	19.3	13.6	−57.8	0.7	−54.1	1.1	1.1
Gaussian	α	12.8	6.1	18.9	14.3	−57.0	−95.1	0.7	0.6
	β	21.0	19.0	28.5	−10.5	−71.1	−250.7	0.5	0.5
inverted logistic	α	17.6	17.6	5.7	341.1	200.7	170.1	2.0	1.3
	β	13.9	15.1	21.8	82.2	94.6	−138.9	0.7	0.8
logistic	α	2.5	2.0	1.6	−9.2	−11.7	−22.7	1.1	1.0
	β	19.2	16.6	12.1	241.7	191.6	155.1	1.5	1.2

Table 6.1 – Bias ×1000, variance ×1000 and relative efficiency for $\widehat{\alpha}$ and $\widehat{\beta}$, with the average (avg) and split (spl) approach compared to the standard (std) approach of Heffernan and Tawn (2004). For the relative efficiency, values smaller than 1 indicate a better performance of one of our approaches. From top to bottom: conditional tail model with rejection, with $(\alpha, \beta, \mu, \psi^2) = (0.4, 0.3, 0.5, 1)$; Gaussian bivariate distribution with Laplace margins and correlation $\rho = 0.3$; inverted logistic bivariate distribution with Laplace margins and dependence parameter $\gamma = 0.5$; logistic bivariate distribution with Laplace margins and dependence parameter $\gamma = 0.5$.

root mean squared errors (RMSE) derived from one of our approaches in the numerator, and the approach of Heffernan and Tawn (2004) in the denominator. Values smaller than 1 indicate a better performance of one of our approaches, but values larger than 1 may only indicate a bias due to assuming $H_{|x}(\cdot) \equiv H_{|y}(\cdot)$ to be Gaussian, which can be more restrictive in our joint models than in the setup of Heffernan and Tawn in which (6.3) wrongly specifies the joint density. The Bayesian setup described in Chapter 7 could be used to remove the Gaussian assumption in our approaches. We use the penultimate approximations for α and β developed in Chapter 4 in the Gaussian case, as convergence of (6.1) is particularly slow in this case. The fits producing Table 6.1 had no constraints implemented, i.e., $(\alpha, \beta) \in [-1, 1] \times [0, 1)$.

As expected, the model performs poorly on the data simulated from the inverted logistic and the logistic distributions, as assuming the residual distributions $H_{|x}(\cdot) \equiv H_{|y}(\cdot)$ to be Gaussian is a misspecification which causes badly-biased estimates of α and β. In their inference procedure, Heffernan and Tawn (2004) also assume normality of the residual distributions to construct a likelihood function, but then compute the empirical residuals of the form $\widehat{z} = (y - \widehat{\alpha} x)/x^{\widehat{\beta}}$, thus $\widehat{H}_{|x}(\cdot)$ and $\widehat{H}_{|y}(\cdot)$ are not Gaussian.

For data simulated with Algorithm 6.2, the split approach has less bias than the average approach, as the rejection sampling of the algorithm mimics the structure of the split likelihood. In the Gaussian case, the average approach appears to perform similarly to the split approach, but for data with more correlation, the split approach performs better, i.e., it

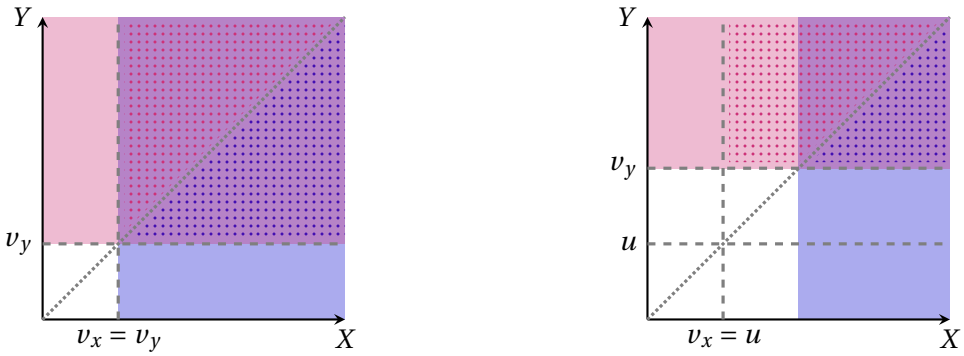

Figure 6.4 – Diagram of the method of proportions for computing joint tail probabilities of the type $\Pr(X > v_x, Y > v_y)$. Left panel: $v_x = v_y > u$; right panel: $v_x = u \ll v_y$. The dotted areas correspond to the probability to be estimated, and the models $M_{|x}$ and $M_{|y}$ are used to sample points in the regions shaded in blue and red respectively.

better captures the behaviour of jointly extreme events, corresponding to data points in R_{11}. Assessing the performance of our new methods based on α and β solely is insufficient, due to their inter-dependence, so it is more appropriate to consider tail probabilities.

We assess the performance of our approaches on two types of tail probabilities $\Pr(X > v_x, Y > v_y) = p$, with fixed p and either $v_x = v_y$ or $v_x = u \ll v_y$. Here, we explain how we compute these joint probabilities using the approach of Heffernan and Tawn (2004), and we give details about how to tackle this problem using our approaches in Section 6.6.3. The procedure of Heffernan and Tawn considers the joint probability $\Pr(X > v_x, Y > v_y)$ as the sum

$$\Pr\left(X > Y, Y > v_y \,\middle|\, X > v_x\right) \Pr\left(X > v_x\right) + \Pr\left(Y > X, X > v_x \,\middle|\, Y > v_y\right) \Pr\left(Y > v_y\right), \tag{6.34}$$

where the conditional probabilities are estimated by computing the empirical estimates based on simulated data, and the marginal probabilities are Laplace. For example, the first probability in (6.34) is estimated by first independently sampling R replicates of $X \mid X > v_x$ from an exponential distribution and R replicates of Z from the empirical distribution $\widehat{H}_{|x}(\cdot)$, then computing the corresponding Y replicates using the relation (6.2), in order to the proportion of the R points (X, Y), $X > v_x$, in the region $\{(x, y) \in \mathbb{R}^2 : x > y, y > v_y\}$. The proportions used in this procedure are illustrated in Figure 6.4, where the number of replicates of (X, Y) in the dotted regions are divided by R, the number of replicates in the shaded regions.

For data sampled with Algorithm 6.2, the relative efficiency for estimating $\Pr(X > v_x, Y > v_y)$ is 81% when $v_x = v_y$ and 85% when $v_x = u \ll v_y$ for the split approach, and 107% and 162% respectively for the average approach. The better performance of the split approach is due to its model matching exactly the simulation process. For data sampled from the other three processes, our methods show a relatively poor performance in estimating $\Pr(X > v_x, Y > v_y)$ compared to the method of Heffernan and Tawn (2004), but this is expected as, for extrapolation, $H(\cdot)$ in our approaches is a Gaussian distribution function, but in their method

it is an empirical estimate, thus being more flexible. If we use a Gaussian distribution for extrapolation from the original method of Heffernan and Tawn, relative efficiency is below 50% in all cases except for the Gaussian distribution when $v_x \ll v_y$ for which it is below 85%, and the split approach beats the average approach in all cases except for the logistic when $v_x = v_y$. As emphasised in Section 6.1, the likelihood used by Heffernan and Tawn is incorrect and their approach does the marginal and joint fit in several steps. In the next section, we perform a simulation study where we use the average and split likelihoods to simultaneously fit the marginal and dependence features of processes with different asymptotic behaviours.

6.6.3 Joint fit

In this section, we use the same four simulation setups as in Section 6.6.2, but we transform the marginal distribution above u to GPD with scale $\sigma_u = 1$ and shape $\xi = 0.1$. We fit the likelihood (6.21), covering both average and split forms, to each of the 1,000 data sets in each case, and we report the bias and relative RMSE of the marginal parameters in Table 6.2. For comparison, we fit a univariate GPD to all exceedances of u, falsely assuming independence of X and Y, thus treating the pairs $(x_1, y_1), \ldots, (x_{1000}, y_{1000})$ as a set of independent observations $z_1, \ldots, z_{2000}$. Table 6.2 gives the bias and relative RMSE of the scale and shape parameters estimated in each of the four cases.

For data simulated with Algorithm 6.2, the marginal distribution is only approximately GPD above u, so the results are harder to interpret. In general, the split approach beats the average approach, with reduced bias and relative RMSE, and it has the same efficiency as the GPD marginal fit for the first two cases displayed in the table, for which assuming normality of $H(\cdot)$ is reasonable. The misspecification of the dependence model for the inverted logistic and logistic distributions has an impact on the marginal fit, and biases in the joint and marginal structures compensate each other.

To supplement the comparison between the methods, and since we are interested in estimating probabilities in the joint tail, we compare joint probabilities $\Pr(X > v_x, Y > v_y)$ in two cases, namely $v_x = v_y > u$ and $v_x = u \ll v_y$ (Table 6.3). In these two cases, we ensure that the joint probability is equal to a fixed $p \in (0,1)$ so that comparisons between all different simulation setups are easier. In both cases, we compute the theoretical v_y, under the condition that the joint survival probability equals p, and we set either $v_x = v_y$ or $v_x = u$ depending on which case we are considering. In what follows, we add a superscript L to denote quantities on the Laplace scale. For the split likelihood, the condition is

$$\Pr\left(X^{\mathrm{L}} > v_x^{\mathrm{L}}, Y^{\mathrm{L}} > v_y^{\mathrm{L}}, X^{\mathrm{L}} > Y^{\mathrm{L}}\right) + \Pr\left(X^{\mathrm{L}} > v_x^{\mathrm{L}}, Y^{\mathrm{L}} > v_y^{\mathrm{L}}, Y^{\mathrm{L}} > X^{\mathrm{L}}\right) = p. \tag{6.35}$$

We deal with the two probabilities on the left-hand side of (6.35) separately. We have

$$\Pr\left(X^{\mathrm{L}} > v_x^{\mathrm{L}}, Y^{\mathrm{L}} > v_y^{\mathrm{L}}, X^{\mathrm{L}} > Y^{\mathrm{L}}\right) = \Pr\left(X^{\mathrm{L}} > v_x^{\mathrm{L}}, X^{\mathrm{L}} > Y^{\mathrm{L}}\right) - \Pr\left(X^{\mathrm{L}} > v_x^{\mathrm{L}}, Y^{\mathrm{L}} \leq v_y^{\mathrm{L}}, X^{\mathrm{L}} > Y^{\mathrm{L}}\right),$$

		bias ×100		variance ×100		RMSE	
		scale	shape	scale	shape	scale	shape
conditional tail	average	1.49	3.20	0.20	0.13	1.50	15.09
	split	0.29	−0.28	0.20	0.11	1.40	10.39
	marginal	1.18	−0.90	0.21	0.11	1.49	11.00
Gaussian	average	4.24	0.19	0.24	0.13	2.06	11.29
	split	2.91	−0.59	0.23	0.12	1.79	11.17
	marginal	0.18	−0.19	0.22	0.11	1.48	10.73
inverted logistic	average	6.56	1.38	0.23	0.13	2.58	12.37
	split	3.68	−0.62	0.22	0.12	1.87	11.03
	marginal	0.10	−0.18	0.21	0.11	1.44	10.57
logistic	average	−5.18	−5.22	0.15	0.09	2.04	19.12
	split	−4.14	−2.90	0.15	0.08	1.80	12.88
	marginal	0.10	−0.21	0.22	0.14	1.47	11.66

Table 6.2 – Bias ×100, variance ×100 and relative RMSE of the GPD scale and shape parameters. From top to bottom: conditional tail model with rejection, with $(\alpha, \beta, \mu, \psi^2) = (0.4, 0.3, 0.5, 1)$; Gaussian distribution with Laplace margins and correlation $\rho = 0.3$; inverted logistic and logistic distributions with Laplace margins and dependence parameter $\gamma = 0.5$. Each setup has a line for the average and split approach, and for a marginal fit of the exceedances of u.

		bias $\times 10^4$		variance $\times 10^8$		RMSE	
		$v_x = v_y$	$v_x \ll v_y$	$v_x = v_y$	$v_x \ll v_y$	$v_x = v_y$	$v_x \ll v_y$
conditional tail	avg	1.08	1.03	1.27	0.71	4.93	4.20
	spl	−0.07	−0.04	1.08	0.42	3.29	2.04
Gaussian	avg	−4.7	−4.6	1.3	1.3	15.3	14.9
	spl	−3.5	−3.0	1.9	2.0	11.9	10.5
inverted logistic	avg	−1.9	−2.3	1.3	0.8	7.1	7.7
	spl	−1.2	0.4	2.1	1.5	5.9	4.1
logistic	avg	−0.7	1.5	0.1	0.2	2.5	5.0
	spl	−1.9	−0.6	0.3	0.3	6.2	2.6

Table 6.3 – Bias $\times 10^4$ and relative RMSE of joint probabilities of the type $\Pr(X > v_x, Y > v_y)$ when simultaneously fitting the marginal and joint distributions, with data simulated from the conditional tail model with parameters $(\alpha, \beta, \mu, \psi) = (0.4, 0.3, 0.5, 1)$, the Gaussian copula ($\rho = 0.3$), the inverted logistic distribution ($\gamma = 0.5$), and the logistic distribution ($\gamma = 0.5$). In all cases the true probability is 0.001.

which can be expressed as

$$\begin{aligned}
&\int_{v_x^{\mathrm{L}}}^{\infty} \left\{ \Pr\left(Y^{\mathrm{L}} < X^{\mathrm{L}} \mid X^{\mathrm{L}} = x\right) \frac{1}{2} e^{-x} - \Pr\left(Y^{\mathrm{L}} \le v_y^{\mathrm{L}}, Y^{\mathrm{L}} < X^{\mathrm{L}} \,\middle|\, X^{\mathrm{L}} = x\right) \frac{1}{2} e^{-x} \right\} dx \\
&\qquad = \int_{v_y^{\mathrm{L}}}^{\infty} \left\{ \Pr\left(Y^{\mathrm{L}} < X^{\mathrm{L}} \mid X^{\mathrm{L}} = x\right) - \Pr\left(Y^{\mathrm{L}} \le v_y^{\mathrm{L}}, Y^{\mathrm{L}} < X^{\mathrm{L}} \,\middle|\, X^{\mathrm{L}} = x\right) \right\} \frac{1}{2} e^{-x} dx \\
&\qquad + \int_{v_x^{\mathrm{L}}}^{v_y^{\mathrm{L}}} \left\{ \Pr\left(Y^{\mathrm{L}} < X^{\mathrm{L}} \mid X^{\mathrm{L}} = x\right) - \Pr\left(Y^{\mathrm{L}} \le v_y^{\mathrm{L}}, Y^{\mathrm{L}} < X^{\mathrm{L}} \,\middle|\, X^{\mathrm{L}} = x\right) \right\} \frac{1}{2} e^{-x} dx.
\end{aligned} \tag{6.36}$$

In the last integral of (6.36), we have

$$\Pr\left(Y^{\mathrm{L}} \le v_y^{\mathrm{L}}, Y^{\mathrm{L}} < X^{\mathrm{L}} \,\middle|\, X^{\mathrm{L}} = x\right) = \Pr\left(Y^{\mathrm{L}} < X^{\mathrm{L}} \mid X^{\mathrm{L}} = x\right), \quad x \le v_y^{\mathrm{L}}.$$

We conclude that the last integral in (6.36) vanishes, and we are left with

$$\int_{v_y^{\mathrm{L}}}^{\infty} \left\{ \Pr\left(Y^{\mathrm{L}} < x \mid X^{\mathrm{L}} = x\right) - \Pr\left(Y^{\mathrm{L}} \le v_y^{\mathrm{L}} \,\middle|\, X^{\mathrm{L}} = x\right) \right\} \frac{1}{2} e^{-x} dx. \tag{6.37}$$

For the second term in the left-hand side of (6.35), we follow the same path to get

$$\begin{aligned}
&\Pr\left(X^{\mathrm{L}} > v_x^{\mathrm{L}},\, Y^{\mathrm{L}} > v_y^{\mathrm{L}},\, Y^{\mathrm{L}} > X^{\mathrm{L}}\right) \\
&\qquad = \Pr\left(Y^{\mathrm{L}} > v_y^{\mathrm{L}}, Y^{\mathrm{L}} > X^{\mathrm{L}}\right) - \Pr\left(X^{\mathrm{L}} \le v_x^{\mathrm{L}}, Y^{\mathrm{L}} > v_y^{\mathrm{L}}, Y^{\mathrm{L}} > X^{\mathrm{L}}\right) \\
&\qquad = \int_{v_y^{\mathrm{L}}}^{\infty} \left\{ \Pr\left(X^{\mathrm{L}} < y \mid Y^{\mathrm{L}} = y\right) - \Pr\left(X^{\mathrm{L}} \le v_x^{\mathrm{L}} \mid Y^{\mathrm{L}} = y\right) \right\} \frac{1}{2} e^{-y} dy.
\end{aligned} \tag{6.38}$$

In terms of the conditional tail model with Gaussian residuals, (6.35) becomes, using (6.37) and (6.38) and assuming strong exchangeability,

$$\int_{v_y^{\mathrm{L}}}^{\infty}\left\{2\Phi\left(\frac{x-\alpha x-\mu x^{\beta}}{\sigma x^{\beta}}\right)-\Phi\left(\frac{v_y^{\mathrm{L}}-\alpha x-\mu x^{\beta}}{\sigma x^{\beta}}\right)-\Phi\left(\frac{u^{\mathrm{L}}-\alpha x-\mu x^{\beta}}{\sigma x^{\beta}}\right)\right\}\frac{1}{2}e^{-x}dx=p,$$

for the case $v_x = u \ll v_y$, and reduces to

$$\int_{v^{\mathrm{L}}}^{\infty}\left\{\Phi\left(\frac{x-\alpha x-\mu x^{\beta}}{\sigma x^{\beta}}\right)-\Phi\left(\frac{v^{\mathrm{L}}-\alpha x-\mu x^{\beta}}{\sigma x^{\beta}}\right)\right\}e^{-x}dx=p,$$

when $v_x = v_y = v > u$.

For the likelihood with averaged contributions in R_{11}, we have

$$\begin{aligned}&\Pr\left(X^{\mathrm{L}}>v_x^{\mathrm{L}},Y^{\mathrm{L}}>v_y^{\mathrm{L}}\right)\\&\qquad=\frac{1}{2}\int_{v_x^{\mathrm{L}}}^{\infty}\Pr\left(Y^{\mathrm{L}}>v_y^{\mathrm{L}}\,\middle|\,X^{\mathrm{L}}=x\right)\frac{1}{2}e^{-x}dx+\frac{1}{2}\int_{v_y^{\mathrm{L}}}^{\infty}\Pr\left(X^{\mathrm{L}}>v_x^{\mathrm{L}}\mid Y^{\mathrm{L}}=y\right)\frac{1}{2}e^{-y}dy,\end{aligned}$$

which in terms of the conditional tail model with Gaussian residuals takes the form

$$\frac{1}{2}\int_{v_y^{\mathrm{L}}}^{\infty}\overline{\Phi}\left(\frac{v_x^{\mathrm{L}}-\alpha x-\mu x^{\beta}}{\sigma^{\beta}}\right)\frac{1}{2}e^{-x}dx+\frac{1}{2}\int_{v_x^{\mathrm{L}}}^{\infty}\overline{\Phi}\left(\frac{v_y^{\mathrm{L}}-\alpha x-\mu x^{\beta}}{\sigma^{\beta}}\right)\frac{1}{2}e^{-x}dx.$$

These integrals are used with the estimates $\widehat{\alpha},\widehat{\beta},\widehat{\mu},\widehat{\sigma}$ of the parameters of the conditional model in order to derive estimates $\widehat{p}$ of p for the four different dependence structures used for the simulations. We set $p = 0.001$ and derive the bias and relative RMSE for the average and split approaches and the four simulation setups. The results are shown in Table 6.3. Except for the logistic case, the split approach performs much better than the average approach, especially in the situation where $v_x \ll v_y$.

6.6.4 Conditional quantile constraints

We now investigate the improvements of imposing the constraints of Keef *et al.* (2013) to the joint fit of Section 6.6.3. The first two rows of Figure 6.5 show the effect of imposing the constraints for the estimates $(\widehat{\alpha},\widehat{\beta})$ estimated on Gaussian data with correlation $\rho = 0.3, 0.8$. For the higher correlation, the gain appears to be significant, as it helps remove the overestimated values of α, which may be an effect of the weak identifiability of α and μ. This improvement in estimating α and β is not seen in the estimation of the joint probabilities $\Pr(X > v_x, Y > v_y) = p$, $v_x = v_y$ and $v_x = u \ll v_y$, as shown by the small gains in efficiency in Table 6.4. In the table, we show the relative efficiency of estimates of the joint probability with and without the constraints of Keef *et al.*, with values smaller than 100 indicating a better performance of the fits incorporating the constraints.

	Gaussian (ρ)				inv. logistic (γ)			logistic (γ)		
	-0.8	-0.3	0.3	0.8	0.2	0.5	0.8	0.2	0.5	0.8
$v_x = v_y$	103	99	100	78	99	99	100	73	100	100
$v_x = u \ll v_y$	100	100	100	101	96	97	100	82	98	100

Table 6.4 – Efficiency gain (%) on (α, β) from using the constraints of Keef *et al.* (2013). From left to right: data simulated from the Gaussian copula ($\rho = -0.8, -0.3, 0.3, 0.8$), the inverted logistic distribution ($\gamma = 0.2, 0.5, 0.8$), and the logistic distribution ($\gamma = 0.2, 0.5, 0.8$). Values smaller than 100 indicate a performance improved by imposing the constraints, using the split likelihood approach.

6.6.5 Conditional mean constraint

We now explore how efficiency is improved by adding the conditional mean constraint. In the Gaussian case, for which we know that the asymptotic value for β is 0.5 for $\rho \neq 0$, we observe that for any $\rho = -0.8, -0.3, 0.3, 0.8$, estimates of α and β are unchanged when adding the new constraint to the fit. When identifiability issues are expected, specifically when β is close to 1, efficiency can be gained by adding the conditional mean constraint. We consider data simulated from the conditional model using Algorithm 6.2 with parameters $(\alpha, \beta, \mu, \psi^2) = (0, 0.9, 0.5, 1)$, and we investigate efficiency gains using the average and split forms of the likelihood (6.21) for estimating the joint probability $\Pr(X > v_x, Y > v_y) = p$. We define relative efficiency by the ratio of RMSE of 1,000 joint tail probability estimates, using the fit under the constraint at the numerator and the fit without the constraint at the denominator. Values lower than 1 indicate the improvement of the fit due to imposing the new constraint. We fix the joint tail probability to $p = 0.001$, and consider $v_x = v_y$ in one case and $v_x = u \ll v_y$ in the other case, and we get efficiency gains of 82% and 93% respectively.

We illustrate the Gaussian case in Figure 6.5, with $\rho = 0.3, 0.8$, thus data show positive association, justifying the use of the mean constraint. We compare the estimates of (α, β) with and without the mean constraint, so we have four different setups. Inference for all four includes the marginal features, partly explaining the variation in $(\widehat{\alpha}, \widehat{\beta})$. Without the constraint, the estimates $\widehat{\alpha}$ tend to be more erratic when $\widehat{\beta} \approx 1$, and are more concentrated for lower values of $\widehat{\beta}$. Adding the constraint to the fit removes those inconsistent estimates. Figure 6.5 also shows the penultimate and ultimate values of (α, β), which are known in the case of bivariate Gaussian data (4). When $\rho = 0.3$, the estimates tend to lie at least as close to the penultimate value as they do to the ultimate value, illustrating the slow convergence of the Gaussian distribution in the conditional limit (6.1).

The conditional mean constraint also restricts the space in which α, β and μ live, so that optimisation of the likelihood function is easier. As we explained at the end of Section 6.6.1, the optimisation needs to be run with several different initial parameter values to reach a global optimum, and adding the new constraints permits fewer runs of the optimiser, hence a gain in computational time. In a multivariate problem, when maximising the likelihood

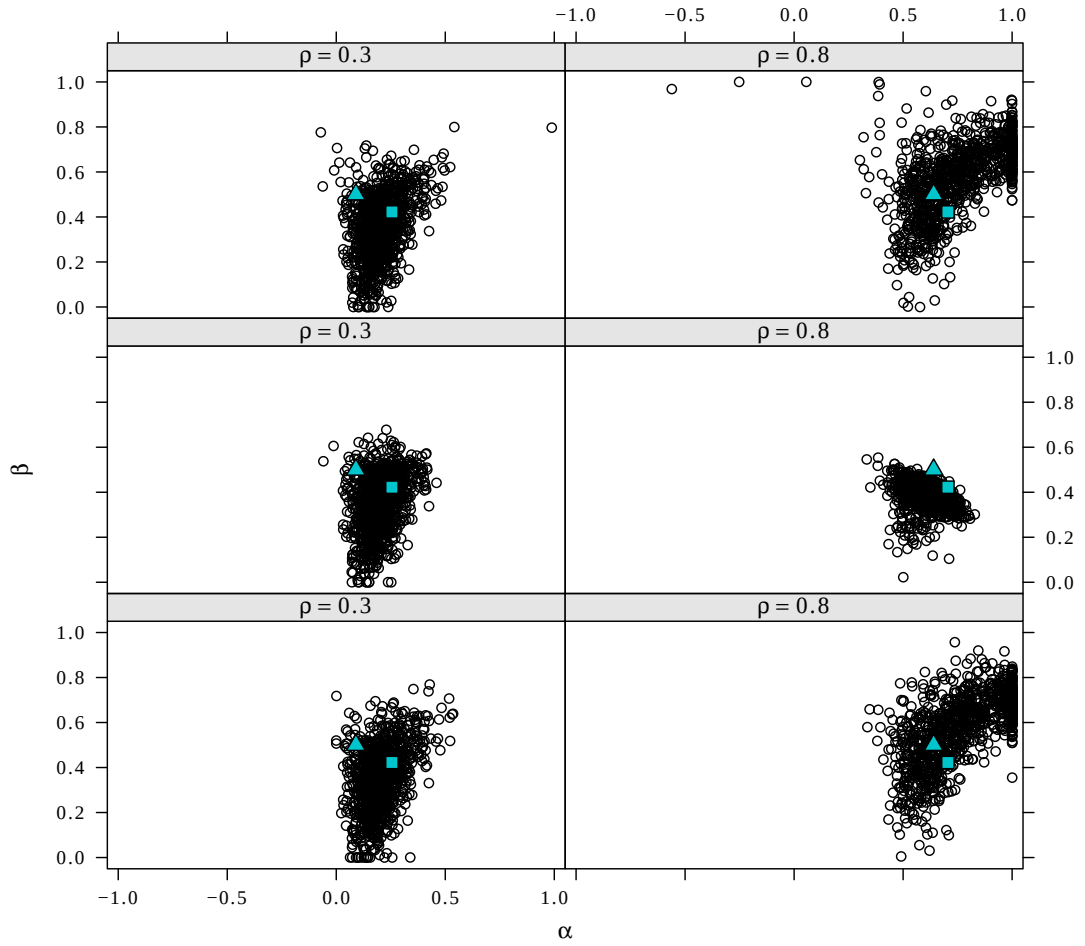

Figure 6.5 – Joint fit of the marginal and dependence features for bivariate Gaussian simulated data. Top row: 1,000 estimates of (α, β) without any constraints; middle row: 1,000 estimates of (α, β) with the conditional quantile constraints; bottom row: 1,000 estimates of (α, β) with the conditional mean constraint assuming positive association. The 1,000 data sets are simulated with correlation $\rho = 0.3$ in the left column, and $\rho = 0.8$ in the right column. The blue symbols represent the ultimate (■) and penultimate (▲) values of (α, β).

over a high-dimensional parameter space, imposing this additional constraint could improve convergence of the optimiser and could be more beneficial than in the bivariate case studied in this chapter.

6.7 Summary

In this chapter, we reviewed the weaknesses of the conditional tail model and the fitting procedure introduced by Heffernan and Tawn (2004), with two main weaknesses being the lack of self-consistency of estimates of joint tail probabilities from different conditional distributions, and the multi-stage estimation procedure, which makes it hard to measure the uncertainty of the model parameters and of probabilities extrapolated from the model. We mentioned the work of Liu and Tawn (2014), which tackled the self-consistency, and the approach developed in Chapter 5, which tackles the multi-stage estimation, but both approaches adopt the incorrect likelihood function of Heffernan and Tawn (2004).

We then introduced two new approaches which combine the flexibility of the conditional tail model with the ability of fitting the extremal joint distribution of a bivariate pair (X, Y) with Laplace margins, assuming a Gaussian residual distribution $H(\cdot)$ and strong exchangeability of (X, Y). These approaches permit information to be shared between the two conditional distributions $Y \mid X$ and $X \mid Y$, and the introduction of censored non-extreme observations help better capture the extremal dependence structure. We generalised this setting to arbitrary margins for (X, Y) and we described the likelihood function for fitting the marginal and joint features simultaneously.

We presented how this bivariate methodology can be extended to model d-dimensional data, with $d > 2$, and we observed that the split likelihood approach scales better than the average likelihood approach as d grows, where the former splits $\mathbb{R}^d$ in $d+1$ regions, whereas the latter needs 2^d regions, thus the split approach is in general preferable to the average approach. Another aspect that makes the split approach more appealing than the average approach is its simple procedure to estimate conditional quantiles. We also presented how the model can be generalised to cover weak exchangeability or no exchangeability of (X, Y). We suggested a more general approach to avoid the strong assumption that $H(\cdot)$ is Gaussian, and we shall give more details about this in the next chapter.

We introduced new constraints that tackle the issue of parameter identifiability when $\beta \approx 1$, provided it is reasonable to assume either positive or negative association of (X, Y). These new constraints also help the optimiser by shrinking the parameter space.

We compared the performance of the different approaches presented in this chapter with the original approach of Heffernan and Tawn (2004), and we observed that assuming Gaussian residuals generally is too strong and can yield badly-biased estimates. The split approach performs better than the average approach in many cases, and extends more easily to high-dimensional setups. We saw that the constraints of Keef *et al.* (2013) can improve the

parameter estimates, but has little effect on probability estimates derived from the model; the new constraints help specifically when identifiability issues are expected, without impacting fits where this is not the case.

In the next chapter, we shall present how we can take advantage of this new approach to fitting bivariate extreme values and use it in the context of time series with short-range dependence, introducing a Bayesian framework which enables relaxation of the assumption of Gaussian residuals used throughout this chapter.

7 Modelling extremes of Markov chains

7.1 Background

7.1.1 Existing approaches

The first approach to fitting exceedances of a threshold for time series was a marginal fit, where a pre-processing step selects exceedances that can be considered independent, thus enabling inference using a maximum likelihood method (Davison and Smith, 1990). Smith *et al.* (1997) deal with the problem of estimating the joint structure of the extremes in a time series by assuming a first order Markov property. This allows for using bivariate parametric models, namely bilogistic (Joe *et al.*, 1992), negative bilogistic (Coles and Tawn, 1994), Dirichlet and asymmetric logistic (Tawn, 1988), for modelling the transition distribution. Other contributions to modelling extremes of Markov chains of arbitrary order exist (Yun, 2000; Fawcett and Walshaw, 2006b; Ribatet *et al.*, 2009), but they all rely on an assumption of asymptotic dependence, e.g., for stationary Markov chains of order 1 and marginal distribution $F(\cdot)$ with upper endpoint x^F,

$$\lim_{x \to x^F} \Pr(X_{t+1} > x \mid X_t > x) > 0.$$

As noted by Winter and Tawn (2017), assuming that a first order Markov chain has asymptotic dependence between X_t and X_{t+1} implies asymptotic dependence between X_t and X_{t+h} ($h = 1, \ldots$). This can be restrictive in applications, where asymptotic independence is expected between observations that are distant in time. Reich *et al.* (2014) present a Bayesian framework in which extremal dependence in time is modelled using a hidden structure with a sliding window modelling neighbouring dependence, but in their approach asymptotic dependence also holds between X_t and X_{t+h} ($h = 1, \ldots$).

Bortot and Tawn (1998) use the model of Ledford and Tawn (1997) for the transition distribution, thus being able to fit processes from the asymptotic independence class, but they rely on an assumption that requires consecutive observations to be simultaneously large.

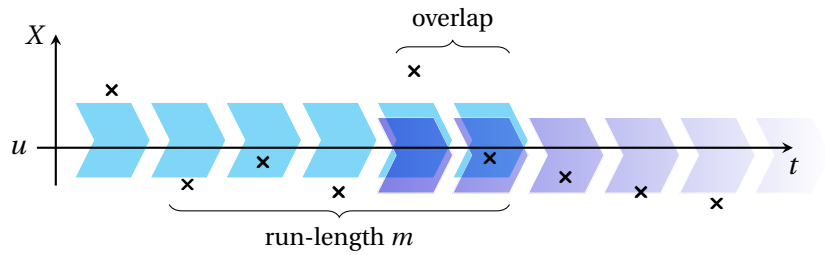

Figure 7.1 – Multiple use of data points occurring when fitting dependence structures of the type $X_1, \ldots, X_m \mid X_0 = x$ for x extreme in time series, using the standard conditional tail model approach.

Their method separately fits the marginal distribution and the extremal dependence structure, and does not provide a measure of uncertainty for the derived cluster functionals.

7.1.2 Modelling time series extremes with the conditional tail approach

Winter and Tawn (2017) use a heuristic approach based on the conditional tail model to fit the transition distribution of kth-order Markov chains. Multiple steps are used to fit the marginal distribution, the parametric part and the non-parametric part of the conditional tail distribution. As a consequence, they need a computationally intensive algorithm to estimate cluster functionals, and they base their block bootstrap confidence intervals on very few replicates, namely 20 in their application (Winter, 2015, Chap. 4). Their inference procedure considers only observations in a neighbourhood of an exceedance of a high threshold, whereas the method developed in the following takes advantage of all observations through a censoring approach.

In Chapter 5, we developed a new methodology to estimate cluster functionals with uncertainty, using Bayesian semiparametrics as a coherent and more efficient approach to fitting the conditional tail model. Although this method provides a consistent approach to modelling the dependence structure of $X_1, \ldots, X_m \mid X_0 = x$ for x extreme, it relies on an incorrect likelihood function which can make multiple use of the same data points, as sketched in Figure 7.1. Another weakness in this approach is that the marginal distribution is fitted separately, thus potentially giving overly optimistic uncertainty measures of cluster functionals estimated from the posterior distributions.

7.2 Bayesian semiparametrics for modelling first-order Markov chains

7.2.1 General

We consider observations x_t $(t = 1, \ldots, n+1)$ of a stationary first order Markov chain, so that we have pairs of consecutive observations (x_t, x_{t+1}) $(t = 1, \ldots, n)$. We use the split approach of

Chapter 6, so we define the regions

$$R_{|t} = \left\{(x_t, x_{t+1}) \in \mathbb{R}^2 : x_t > \max(u, x_{t+1})\right\}, \qquad R_{|t+1} = \left\{(x_t, x_{t+1}) \in \mathbb{R}^2 : x_{t+1} > \max(u, x_t)\right\},$$
$$R_{00} = \left\{(x_t, x_{t+1}) \in \mathbb{R}^2 : \max(x_t, x_{t+1}) \leq u\right\}.$$

Since all pairs in R_{00} are censored, we can re-number the observations in $R_{|t} \cup R_{|t+1}$ such that the first $n_>$ pairs have at least one extreme component, i.e., $(x_t, x_{t+1}) \notin R_{00}$ $(t = 1, \ldots, n_>)$ and n_{00} censored pairs, with $n_{00} + n_> = n$.

We use the generalised Pareto distribution above a high threshold u with shape and scale parameters ξ and $\sigma_u > 0$ to model the marginal tail of (x_t). To transform (x_t) to the Laplace scale, we use the semiparametric model (6.20) (Coles and Tawn, 1994).

We consider a semiparametric Bayesian approach to model the residual distributions $H_f(\cdot)$ and $H_b(\cdot)$ arising from the conditional limits of the forward and backward chains, namely

$$\begin{aligned} \lim_{u\to\infty} \Pr\left\{ \frac{X^{\mathrm{L}}_{t+1} - \alpha_f X^{\mathrm{L}}_t}{\left(X^{\mathrm{L}}_t\right)^{\beta_f}} \leq z \,\middle|\, X^{\mathrm{L}}_t \geq u \right\} = H_f(z), \\ \lim_{u\to\infty} \Pr\left\{ \frac{X^{\mathrm{L}}_t - \alpha_b X^{\mathrm{L}}_{t+1}}{\left(X^{\mathrm{L}}_{t+1}\right)^{\beta_b}} \leq z \,\middle|\, X^{\mathrm{L}}_{t+1} \geq u \right\} = H_b(z), \end{aligned} \tag{7.1}$$

where the superscript L denotes a quantity transformed to the Laplace scale. Previously, in the approaches of Winter and Tawn (2017) and Chapter 6, $H_f(\cdot)$ and $H_b(\cdot)$ were modelled as Gaussian distributions. Here, we use a mixture of a potentially infinite number of Gaussian components to model $H_f(\cdot)$ and $H_b(\cdot)$ in (7.1). As we saw in Chapter 3, we can define auxiliary index variables that link observations with components in the mixture. In the split likelihood setup, each (x_t, x_{t+1}) $(t = 1, \ldots, n_>)$ is considered only once, thus we attach an index variable c_t to each pair $(x_t, x_{t+1}) \notin R_{00}$, indicating which component in the mixture it belongs to.

In Chapter 5, we used a Gibbs sampler based on an approximation of the Dirichlet process, and posterior sampling was possible thanks to the closed-form of the likelihood function that we were considering. Since we shall base our proposal on the split model of the previous chapter, which features analytically intractable integrals, we cannot derive a similar Gibbs sampler here. In Appendix I, we discuss how we can use approximate posterior distributions as state-dependent proposal distributions in a Metropolis–Hastings scheme, however it appears to mix poorly and we do not explore this alley further. In the following, we suggest using a different approach to fitting Dirichlet processes, which has the advantage of not requiring any approximation to the process.

7.2.2 Pólya urn scheme

Algorithm A.2 of Neal (2000) is more appealing than the approach of Ishwaran and Zarepour (2000) in the context of this chapter, as it will appear in the following developments. We

recall from Chapter 3 that, for the update of the c_t, the algorithm not only proposes candidate components among those to which observations are already attached, but also newly created components, according to a fixed parameter $m \geq 1$. Iterating through each observation (x_t, x_{t+1}), we write k_{-t} for the number of different components to which observations (x_s, x_{s+1}) $(s = 1, \ldots, n_>,\ s \neq t)$ are attached, and we re-number these components $1, \ldots, k_{-t}$. If $c_t \notin \{1, \ldots, k_{-t}\}$, the algorithm creates $m-1$ new components and retains the component c_t as an additional candidate, so that there is a non-null probability for c_t to remain unchanged; otherwise, if $c_t = c$ for some $c \in \{1, \ldots, k_{-t}\}$, it creates m new components, so that the number of candidate components in any iteration is always $k_{-t} + m$.

According to the generalised Pólya urn scheme of Section 3.1.3, the prior probabilities associated with existing components are

$$\Pr(c_t = c \mid \boldsymbol{c}_{-t}, \gamma) = \frac{n_{-t,c}}{n_> - 1 + \gamma}, \quad t = 1, \ldots, n_>,\ 1 \leq c \leq k_{-t}, \tag{7.2}$$

where $n_{-t,c} = \sum_{s=1, s\neq t}^{n_>} \mathbb{1}(c_s = c)$ is the number of observations assigned to component c, without counting (x_t, x_{t+1}), and those associated with newly created components are

$$\Pr(c_i = c \mid \boldsymbol{c}_{-i}, \gamma) = \frac{\gamma / m}{n_> - 1 + \gamma}, \quad k_{-i} < c \leq k_{-i} + m. \tag{7.3}$$

From the conditional prior probabilities (7.2) and (7.3), the posterior probability of an index variable c_t given observation $(x_t, x_{t+1}) \in R_{|t}$ is

$$\begin{aligned} &\Pr\left(c_t = c \mid \boldsymbol{c}_{-t},\, x_t, x_{t+1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{-t}+m}, \gamma\right) \\ &\qquad \propto \Pr\left(x_{t+1} \,\middle|\, x_t,\, c_t = c, \boldsymbol{c}_{-t}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{-t}+m}, \gamma\right) \times \Pr\left(c_t = c \,\middle|\, \boldsymbol{c}_{-t}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{-t}+m}, \gamma\right) \\ &\qquad = \Pr\left(x_{t+1} \,\middle|\, x_t,\, c_t = c, \boldsymbol{c}_{-t}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{-t}+m}, \gamma\right) \times \Pr\left(c_t = c \,\middle|\, \boldsymbol{c}_{-t}, \gamma\right), \end{aligned} \tag{7.4}$$

where the index variables and the parameters $\boldsymbol{\theta}_k$ in the conditioning provide information about the size of each component, since $n_k = \sum_{t=1}^{n} \mathbb{1}(c_t = k)$ is the number of pairs of observations attached to component k.

We now need to detail the likelihood function used to build our model for extremes of Markov chains, so that we can give the exact form of the first conditional probability in the right-hand side of (7.4).

7.2.3 Likelihood function

From the two approaches presented in Chapter 6, the split approach (Section 6.2.3) seems to be the more promising, because of the better efficiency of its tail probability estimates and its ability to generalise nicely to higher dimensions. We use x_t^{L} to denote the transformation of x_t to the Laplace scale, and Θ represents the set of all $\boldsymbol{\theta}_c = (\mu_c, \psi_c^2)$ in a given iteration. The

general form of the likelihood where (x_t) has unknown marginal distribution is

$$\begin{aligned} L(\Theta,\xi,\sigma_u;x_1,\ldots,x_{n+1}) &= L\left(\Theta,\xi,\sigma_u;x_1^{\mathrm{L}},\ldots,x_{n+1}^{\mathrm{L}}\right)\prod_{t=1}^{n+1} J(x_t;\xi,\sigma_u) \\ &\propto \frac{\prod_{t=1}^{n} f_{t,t+1}\left(x_t^{\mathrm{L}},x_{t+1}^{\mathrm{L}};\Theta,\xi,\sigma_u\right)}{\prod_{t=2}^{n}\frac{1}{2}e^{-|x_t^{\mathrm{L}}|}}\prod_{t=1}^{n+1} J(x_t;\xi,\sigma_u)\,\mathbb{1}(x_t>u), \end{aligned} \tag{7.5}$$

for some joint density function $f_{t,t+1}(\cdot,\cdot)$, a high threshold $u>0$ on the marginal scale of (x_t), transformed to u^{L} on the Laplace scale, and

$$J(x;\xi,\sigma_u) = \frac{1}{\sigma_u}\left(1+\xi\frac{x-u}{\sigma_u}\right)_+^{-1}, \qquad x>u,$$

corresponding to the Jacobian of the marginal transformation of x to x^{L}; see Chapter 6. For the split likelihood approach, (7.5) becomes

$$\begin{aligned} e^{-|x_1^{\mathrm{L}}|}\left\{F_{t,t+1}\left(u^L,\,u^L;\Theta\right)\right\}^{n_{00}}\prod_{t=1}^{n_>} f_{t+1|t}\left(x_{t+1}^{\mathrm{L}}\mid x_t^{\mathrm{L}};\Theta,\xi,\sigma_u\right)\prod_{t=1}^{n} J(x_t;\xi,\sigma_u) \\ \propto e^{-|x_1^{\mathrm{L}}|}\left\{F_{t,t+1}\left(u^L,u^L;\Theta\right)\right\}^{n_{00}}\prod_{t:(x_t,x_{t+1})\in R_{|t}} f_{t+1|t}\left(x_{t+1}^{\mathrm{L}}\mid x_t^{\mathrm{L}};\Theta,\xi,\sigma_u\right) \\ \times\prod_{t:(x_t,x_{t+1})\in R_{|t+1}} f_{t|t+1}\left(x_t^{\mathrm{L}}\mid x_{t+1}^{\mathrm{L}};\Theta,\xi,\sigma_u\right)e^{|x_t^{\mathrm{L}}|-|x_{t+1}^{\mathrm{L}}|}\prod_{t=1}^{n} J(x_t;\xi,\sigma_u), \end{aligned} \tag{7.6}$$

using the Bayes rule. If $x_1<u$, the first term can be removed from (7.6) since the marginal transform $x_t\mapsto x_t^{\mathrm{L}}$ is nonparametric below u.

Removing the parameterisation for clarity, the likelihood contributions for observations $(x_t,x_{t+1})\in R_{|t}$ in a given iteration are

$$f_{t+1|t}\left(x_{t+1}^{\mathrm{L}}\mid x_t^{\mathrm{L}}\right) = \sum_{k=1}^{k_{-t}+m}\frac{w_{f,k}}{\left(x_t^{\mathrm{L}}\right)^{\beta_f}\psi_{f,k}}\,\varphi\left\{\frac{x_{t+1}^{\mathrm{L}}-\alpha_f x_t^{\mathrm{L}}-\left(x_t^{\mathrm{L}}\right)^{\beta_f}\mu_{f,k}}{\left(x_t^{\mathrm{L}}\right)^{\beta_f}\psi_{f,k}}\right\}, \tag{7.7}$$

where the subscript f stands for *forward* and denotes the parameters of the conditional tail model where $X_{t+1}\mid X_t=x$, $x>u$. The weights correspond to the proportion of observations in $R_{|t}$ attached to component k, with

$$w_{f,k} = \frac{\sum_{t=1}^{n_>}\mathbb{1}(c_t=k)\times\mathbb{1}\left\{(x_t,x_{t+1})\in R_{|t}\right\}}{\sum_{t=1}^{n_>}\mathbb{1}\left\{(x_t,x_{t+1})\in R_{|t}\right\}}, \quad k=1,\ldots,k_{-t}+m.$$

Similarly, likelihood contributions for observations $(x_t,x_{t+1})\in R_{|t+1}$ in a given iteration are

$$\begin{aligned} f_{t+1|t}\left(x_{t+1}^{\mathrm{L}}\mid x_t^{\mathrm{L}}\right) &= f_{t|t+1}\left(x_t^{\mathrm{L}}\mid x_{t+1}^{\mathrm{L}}\right)e^{x_t^{\mathrm{L}}-x_{t+1}^{\mathrm{L}}} \\ &= \sum_{k=1}^{k_{-t}+m}\frac{w_{b,k}}{\left(x_{t+1}^{\mathrm{L}}\right)^{\beta_b}\psi_{b,k}}\,\varphi\left\{\frac{x_t^{\mathrm{L}}-\alpha_b x_{t+1}^{\mathrm{L}}-\left(x_{t+1}^{\mathrm{L}}\right)^{\beta_b}\mu_{b,k}}{\left(x_{t+1}^{\mathrm{L}}\right)^{\beta_b}\psi_{b,k}}\right\}, \end{aligned} \tag{7.8}$$

where b stands for *backward* and denotes the parameters of the conditional tail model where $X_t \mid X_{t+1} = x$, $x > u$, and

$$w_{b,k} = \frac{\sum_{t=1}^{n_>} \mathbb{1}(c_t = k) \times \mathbb{1}\{(x_t, x_{t+1}) \in R_{|t+1}\}}{\sum_{t=1}^{n_>} \mathbb{1}\{(x_t, x_{t+1}) \in R_{|t+1}\}}, \quad k = 1, \ldots, k_{-t} + m.$$

For censored observations $(x_t, x_{t+1}) \in R_{00}$, we have

$$F_{t,t+1}(u^{\mathrm{L}}, u^{\mathrm{L}}) = 1 - \Pr\{\max(X_t^{\mathrm{L}}, X_{t+1}^{\mathrm{L}}) > u^{\mathrm{L}}\} = 1 - \Pr\left\{(X_t^{\mathrm{L}}, X_{t+1}^{\mathrm{L}}) \in R_{|t}^{\mathrm{L}} \cup R_{|t+1}^{\mathrm{L}}\right\}, \tag{7.9}$$

where the probability equals

$$\int_{R_{|t}^{\mathrm{L}}} \sum_{k=1}^{k_{-t}+m} \frac{w_{f,k}}{x^{\beta_f}\psi_{f,k}} \varphi\left(\frac{y - \alpha_f x - x^{\beta_f}\mu_{f,k}}{x^{\beta_f}\psi_{f,k}}\right) \frac{1}{2} e^{-x} dy dx + \int_{R_{|t+1}^{\mathrm{L}}} \sum_{k=1}^{k_{-t}+m} \frac{w_{b,k}}{y^{\beta_b}\psi_{b,k}} \varphi\left(\frac{x - \alpha_b y - y^{\beta_b}\mu_{b,k}}{y^{\beta_b}\psi_{b,k}}\right) \frac{1}{2} e^{-y} dx dy, \tag{7.10}$$

where $R_{|t}^{\mathrm{L}}$ and $R_{|t+1}^{\mathrm{L}}$ correspond to $R_{|t}$ and $R_{|t+1}$ on the Laplace scale. Expression (7.10) equals

$$\frac{1}{2} \sum_{k=1}^{k_{-t}+m} w_{f,k} \int_{u^{\mathrm{L}}}^{\infty} e^{-x} \Phi\left(\frac{x - \alpha_f x - x^{\beta_f}\mu_{f,k}}{x^{\beta_f}\psi_{f,k}}\right) dx + \frac{1}{2} \sum_{k=1}^{k_{-t}+m} w_{b,k} \int_{u^{\mathrm{L}}}^{\infty} e^{-y} \Phi\left(\frac{y - \alpha_b y - y^{\beta_b}\mu_{b,k}}{y^{\beta_b}\psi_{b,k}}\right) dy. \tag{7.11}$$

If we assume strong exchangeability of the forward and backward chains, i.e., $\alpha_f = \alpha_b$, $\beta_f = \beta_b$, and $\{(\mu_{f,k}, \psi_{f,k}^2) : k = 1, \ldots\} = \{(\mu_{b,k}, \psi_{b,k}^2) : k = 1, \ldots\}$, (7.11) becomes

$$\sum_{k=1}^{k_{-t}+m} w_k \int_{u^{\mathrm{L}}}^{\infty} e^{-x} \Phi\left(\frac{x - \alpha x - x^{\beta}\mu_k}{x^{\beta}\psi_k}\right) dx,$$

with $w_k = \sum_{t=1}^{n_>} \mathbb{1}(c_t = k)/n_>$.

7.2.4 Posterior assignment probabilities

With the three different likelihood contributions in $R_{|t}$, $R_{|t+1}$ and R_{00} based on (7.7), (7.8) and (7.11) respectively, we can compute the posterior probabilities (7.4) for the index variables c_t $(t = 1, \ldots, n_>)$. In the following, we assume strong exchangeability for clarity, but assuming

no exchangeability is a straightforward generalisation. We have

$$\begin{aligned}
&\Pr\left(c_t = c \mid \boldsymbol{c}_{-t}, x_t^{\mathrm{L}}, x_{t+1}^{\mathrm{L}}, \alpha, \beta, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{-t}+m}\right) \\
&\quad = \widetilde{\kappa} \begin{cases}
\dfrac{n_{-t,c}}{n-1+\gamma} \begin{cases} f_{t+1|t}\left(x_{t+1}^{\mathrm{L}} \mid x_t^{\mathrm{L}}, \alpha, \beta, \boldsymbol{\theta}_c\right), & c \le k_{-t}, (x_t, x_{t+1}) \in R_{|t}, \\ f_{t|t+1}\left(x_t^{\mathrm{L}} \mid x_{t+1}^{\mathrm{L}}, \alpha, \beta, \boldsymbol{\theta}_c\right) e^{|x_t^{\mathrm{L}}| - |x^{t+1}|}, & c \le k_{-t}, (x_t, x_{t+1}) \in R_{|t+1}, \end{cases} \\
\dfrac{\gamma/m}{n-1+\gamma} \begin{cases} f_{t+1|t}\left(x_{t+1}^{\mathrm{L}} \mid x_t^{\mathrm{L}}, \alpha, \beta, \boldsymbol{\theta}_c\right), & c > k_{-t}, (x_t, x_{t+1}) \in R_{|t}, \\ f_{t|t+1}\left(x_t^{\mathrm{L}} \mid x_{t+1}^{\mathrm{L}}, \alpha, \beta, \boldsymbol{\theta}_c\right) e^{|x_t^{\mathrm{L}}| - |x^{t+1}|}, & c > k_{-t}, (x_t, x_{t+1}) \in R_{|t+1}, \end{cases}
\end{cases}
\end{aligned} \tag{7.12}$$

where $\widetilde{\kappa} = \widetilde{\kappa}(u^{\mathrm{L}}, \boldsymbol{c}_{-t}, c_t = c, \alpha, \beta, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{-t}+m})$, with

$$\widetilde{\kappa}\left(u^{\mathrm{L}}, \boldsymbol{c}_{-t}, c_t = c, \alpha, \beta, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{-t}+m}\right) = \kappa \times F_{t,t+1}\left(u^{\mathrm{L}}, u^{\mathrm{L}} \mid \boldsymbol{c}_{-t}, c_t = c, \alpha, \beta, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{-t}+m}\right)$$

with $\kappa > 0$ a normalising constant ensuring that the probabilities in (7.12) sum to 1.

For updates of $\boldsymbol{\theta}_c$, α and β, we use a Metropolis–Hastings scheme, since the likelihood function (7.6), and in particular (7.9), has no closed form.

7.3 Summary and future work

In this chapter, we reviewed the weaknesses of current methods for modelling extremes of Markov chains, which until recently were using models from the standard extreme value theory which entail assuming asymptotic dependence at all lags of the series. The method of Winter and Tawn (2017) does not rely on this assumption, but uses a heuristic approach to construct the transition probability and a multi-step inference procedure for which computationally intensive bootstrap methods are needed.

We developed a model for extremes of first-order Markov chains based on our new methodology of Chapter 6 and showed how Bayesian semiparametrics can be used to efficiently sample from the posterior distribution of the model. This is ongoing work, and preliminary results are promising.

Generalisations to modelling extremes of higher-order Markov chains is possible, and we could use the formulation of Section 6.3.3 to build a model for kth-order Markov chains with a likelihood function of the form

$$\begin{aligned}
L(\Theta, \xi, \sigma_u; x_1, \ldots, x_{n+k}) &= L\left(\Theta, \xi, \sigma_u; x_1^{\mathrm{L}}, \ldots, x_{n+k}^{\mathrm{L}}\right) \prod_{t=1}^{n+1} J(x_t; \xi, \sigma_u) \\
&\propto \frac{\prod_{t=1}^{n} f_{t:t+k}\left(x_t^{\mathrm{L}}, \ldots, x_{t+k}^{\mathrm{L}}; \Theta, \xi, \sigma_u\right)}{\prod_{t=2}^{n} f_{t+1:t+k}\left(x_{t+1}^{\mathrm{L}}, \ldots, x_{t+k}^{\mathrm{L}}; \Theta, \xi, \sigma_u\right)} \prod_{t=1}^{n+1} J(x_t; \xi, \sigma_u)\, \mathbb{1}(x_t > u),
\end{aligned} \tag{7.13}$$

where $f_{s:t}(x_s, \ldots, x_t)$ $(s < t)$ denotes the joint density of $X_s, \ldots, X_t$. Writing (7.13) as

$$\frac{1}{2} e^{-|x_1^{\mathrm{L}}|} \frac{\prod_{t=1}^{n} f_{t+1:t+k|t}\left(x_{t+1}^{\mathrm{L}}, \ldots, x_{t+k}^{\mathrm{L}} \mid x_t^{\mathrm{L}}; \Theta, \xi, \sigma_u\right)}{\prod_{t=2}^{n} f_{t+2:t+k|t+1}\left(x_{t+2}^{\mathrm{L}}, \ldots, x_{t+k}^{\mathrm{L}} \mid x_{t+1}^{\mathrm{L}}; \Theta, \xi, \sigma_u\right)} \prod_{t=1}^{n+1} J\left(x_t; \xi, \sigma_u\right) \mathbb{1}(x_t > u),$$

we can then use the split approach of Chapter 6 to model the conditional densities $f_{s:t|s-1}$ $(s < t)$ of $X_s, \ldots, X_t \mid X_{s-1}$.

8 Discussion and extensions

This thesis has two main facets. The first contributes to research in extreme value theory by investigating and developing a conditional tail approach, and the second explores Bayesian nonparametric inference in order to provide algorithms that are well-suited for our new extreme value methodology. Chapter 2 introduces and reviews the first facet, supplemented by Chapter 4, which presents a detailed analysis of the conditional tail approach with new findings on its subasymptotic characteristics, and by Chapter 6, which develops the conditional tail model in many new directions. Chapter 3 gives an overview of the second facet. The two facets meet in Chapters 5 and 7 to provide new approaches for modelling time series extremes.

In Chapter 1, we motivated our research by presenting how extreme events strike all aspects of modern society. We emphasised how important it is to be able to evaluate the risk of catastrophic events by estimating their frequency and magnitude.

In Chapter 2, we reviewed many models that are suited for the class of asymptotic independent processes, which are often encountered in environmental data and incorrectly specified as completely independent or asymptotically dependent by standard extreme value models. It would be interesting to investigate the performance of the models covering asymptotic independence relative to the new models of Chapter 6 with Bayesian semiparametric inference in an extensive and systematic simulation study. For the models covering both asymptotic independence and asymptotic dependence, it would be interesting to evaluate their ability to distinguish the two classes in simulated data.

In Chapter 3, we defined the Dirichlet process and described how it can be used in Bayesian nonparametric inference to estimate continuous densities using a Dirichlet process mixture. Two different approaches for inference exist, namely the marginal approach reviewed by Neal (2000), and the conditional approach introduced by Ishwaran and Zarepour (2000), which approximates the Dirichlet process by truncating the infinite sum of the stick-breaking representation. The method developed by Papaspiliopoulos and Roberts (2008) avoids approximation of the conditional approach to fitting Dirichlet process mixtures through the use of a retrospective sampling algorithm. We did not consider this algorithm, and it would be

interesting to examine its performance relative to the algorithm of Ishwaran and Zarepour (2000) so as to measure the impact of the truncation on simulated and real data sets. Extensions of the Dirichlet process exist, e.g., the Beta two-parameter and the Pitman and Yor (1997) processes, and would be flexible candidates to consider for the methodology of Chapter 5. We did not investigate reversible-jump algorithms (Green, 1995; Richardson and Green, 1997), as they require the delicate construction of transition probabilities across parameter spaces of different dimensions, but this could be a potential path for future research.

In Chapter 4, we conducted an analysis that sheds light on the penultimate properties of the conditional approach for extremes, which has formerly been carried out only in the univariate and, to some extent, the bivariate theory of extremes. Extreme value models are based on approximations of asymptotic results, and can require a significant amount of data for this approximation to be reasonable for deriving accurate risk measures. In the univariate setup, asymptotic models may poorly describe finite-sample behaviour, it is important to better understand the subasymptotic behaviour of extreme value models in general. We analysed the subasymptotic properties of the conditional tail model for a few parametric copulae with various asymptotic properties; it would be interesting to investigate the many other examples scrutinised by Heffernan and Tawn (2004), who give their limit forms, and to try to develop a general formulation giving the nature of the penultimate behaviour.

Standard modelling of extremes in time series is considered to be a univariate problem, and corresponds to modelling the marginal tail distribution of the series above a high threshold. It needs a pre-processing step to select approximately independent excesses of the threshold, so that maximum likelihood inference can be used. In Chapter 5, we modelled the extremal dependence in time series, improved on existing inference procedures and gave an assessment of uncertainty for the conditional tail model. We used Bayesian nonparametric tools to construct a semiparametric Dirichlet process mixture. We developed a hierarchical Bayesian procedure incorporating stochastic label-switching moves and a regional adaptive scheme that automates the laborious procedure of choosing appropriate proposal variances. The Bayesian framework has the additional benefit of being able to structure the decay in extremal dependence through time, which the original method of Heffernan and Tawn (2004) did not provide. Unlike standard models for extremes, the conditional approach can capture the decay of extremal dependence strength as we move further into the tail, thus enabling the quantification of risk to be based on subasymptotic measures of extremal dependence. An interesting direction of research would be to adapt the methodology of this chapter to data sets of moderately high dimension, perhaps using the state-dependent proposal distribution discussed in Appendix F and adding more structure in the covariance matrices of the mixture components.

The conditional approach to modelling extremes of Heffernan and Tawn (2004) is appropriate for modelling asymptotically independent and asymptotically dependent data, but it lacks an adequate characterisation of the joint tail distribution, and currently inference is carried out using a wrongly-specified likelihood function. It also fails to guarantee self-consistency of

joint tail probabilities derived from different conditional distributions of the same random vector. Lastly, it does not permit simultaneous inference on the marginal and dependence features, so that uncertainty evaluation relies on bootstrap methods. In Chapter 6, we tackled these various issues by developing two new approaches based on the conditional tail approach, but using a coherent likelihood function where non-extreme observations are censored. These approaches permit efficient inference by combining the marginal and dependence fits. We introduced new constraints under positive or negative association of the extremes, and showed how penalisation can improve the original method of Heffernan and Tawn (2004). The simulation study shows that the new methodology of this chapter is promising, as it performs relatively well compared to the existing incorrect inference even with a strongly simplifying assumption of normality. Using a Dirichlet process mixture would remove the need to make this assumption, and would naturally provide a measure of uncertainty. Further investigation would be needed to study the methodology of this chapter in a multivariate setting. Another potential area of future research is the extension of this multivariate setting to a spatial model that would require a generalised spatial Dirichlet process (Gelfand *et al.*, 2005; Agarwal and Gelfand, 2005; Duan *et al.*, 2007). The model could assume the existence of spatial norming functions $\alpha(s)$ and $\beta(s) > 0$ such that, for all points s in the space of interest,

$$\lim_{u\to\infty} \Pr\left\{ \frac{X(s) - \alpha(s)X(s_0)}{X(s_0)^{\beta(s)}} \leq z(s) \;\middle|\; X(s_0) > u \right\} = H_{s_0}\{z(s)\},$$

where $X(s)$ is the observed spatial process, s_0 is a location where an extreme event is observed, and $H_{s_0}(\cdot)$ is a distribution function with no mass at infinity.

Current methods for modelling extremes of Markov chains assume asymptotic dependence of data at lag 1, thus implicitly assuming asymptotic dependence at all lags. In Chapter 7, we explained the weakness of the approach of Chapter 5 and of the approach of Winter and Tawn (2017), where the same data can be used multiple times in the likelihood function. We developed a model for extremes of first-order Markov chains based on the new bivariate model of Chapter 6, allowing simultaneous inference on the marginal and joint features, and demonstrate how the marginal approach to fitting Dirichlet process mixtures can be used to build an algorithm. We would need to perform simulation studies in order to validate this approach, but ongoing work, under the strong assumptions of Chapter 6, shows promising results. Extending this methodology to higher-order Markov chains would be possible and interesting to develop.

To conclude, this thesis brings new improvements and developments for the modelling of extreme values, combining the flexibility of the conditional tail approach with the adaptability of Bayesian nonparametric inference, thus providing a coherent and efficient framework that naturally enables an assessment of uncertainty. It demonstrates the ability of Bayesian inference, combined with extreme value statistics, to evaluate the risk of rare events, and many paths for extensions and improvements remain to be explored.

Appendices

A Marginal approach to fitting Dirichlet process mixtures

We briefly outline two algorithms from Neal (2000) that constitute examples of the marginal approach to fitting the Dirichlet process mixture model (3.7). Neal presents algorithms numbered from 1 to 8 in order of increasing complexity and efficiency in exploring the parameter space. We focus on Algorithm 5 and Algorithm 8. The former improves on the estimation method suggested by MacEachern and Müller (1998) and introduces an update step on the mixture parameters that is not present in Algorithm 6; Algorithm 8 is the most complex algorithm in Neal's paper and generalises and improves on simpler samplers such as those of MacEachern and Müller (1998) and Bush and MacEachern (1996).

Mixing is improved in Algorithm A.1 by increasing the value of $R \geq 1$. In order to save computational time, it is possible to store the value of the likelihood corresponding to the accepted state and used in the acceptance ratio, so that it can be used in the next sweep of the r-loop. Thus only $R+1$ likelihoods need to be calculated for each $i = 1, \ldots, n$.

Algorithm A.1: Partial Gibbs sampler for the Dirichlet process.

Current state: $\{c_1, \ldots, c_n\}$, $\{\theta_c : c \in \{c_1, \ldots, c_n\}\}$
`# update on the components`
for $i = 1$ **to** n **do**
  for $r = 1$ **to** R **do**
    Sample a candidate $c_i^\star$ according to the probabilities (3.13) and (3.14)
    if $c_i^\star \notin \{c_1, \ldots, c_n\}$ **then**
      Sample a new $\theta_{c_i^\star}$ from P_0
    end
    Compute the acceptance probability $a(c_i^\star, c_i)$ (3.12)
    Set $c_i \hookleftarrow c_i^\star$ with probability $a(c_i^\star, c_i)$, otherwise leave c_i unchanged
  end
end
`# update on the parameters`
foreach $c \in \{c_1, \ldots, c_n\}$ **do**
  Sample a new value for θ_c from the conditional distribution $\theta_c \mid \{x_i : c_i = c\}$
end

Algorithm A.2: Gibbs sampler with auxiliary parameters for the Dirichlet process.

Current state: $\{c_1, \ldots, c_n\}$, $\{\theta_c : c \in \{c_1, \ldots, c_n\}\}$
`# update on the components`
for $i = 1$ **to** n **do**
  Label $\{c_j : j \neq i\}$ from 1 up to k_{-i}
  if $c_i = c_j$ *for some* $j = 1, \ldots, n$ **then**
    Sample values from P_0 for $\theta_c,\ k_{-i} < c \leq k_{-i} + m$
  else if $c_i \neq c_j$ *for all* $j = 1, \ldots, n$ **then**
    Set $c_i \hookleftarrow k_{-i} + 1$
    Sample values from P_0 for $\theta_c,\ k_{-i} + 1 < c \leq k_{-i} + m$
  end
  Sample a new index c_i from the probabilities in (3.15)
  Clear all θ_c corresponding to empty components
  `# update on the parameters`
  forall $c \in \{c_1, \ldots, c_n\}$ **do**
    Sample a new value for θ_c from the conditional distribution $\theta_c \mid \{x_i : c_i = c\}$
  end
end

B Concentration parameter update

The concentration parameter $\gamma > 0$ in the Dirichet process mixture (3.7) controls the prior belief of the expected number of components in the mixture. The smaller γ, the fewer components, and vice versa. In order to lessen the impact of choosing a precise value for γ, a hyperprior can be set on the concentration parameter. Escobar and West (1995) use a gamma hyperprior distribution, for which they provide an update mechanism using an auxiliary variable.

Assuming a continuous hyperprior on γ, we have

$$\Pr(k \mid n) = \mathrm{E}\{\Pr(k \mid \gamma, n)\}, \quad k = 1, \ldots, n,$$

with n the number of observations and k the number of non-empty components in the mixture. From Antoniak (1974), Escobar and West derive

$$\Pr(k \mid \gamma, n) = c_n(k)\, n! \gamma^k \frac{\Gamma(\gamma)}{\Gamma(\gamma + n)}, \quad k = 1, \ldots, n,$$

where $\Gamma(x) = \int_0^\infty s^{x-1} e^{-s} ds$ and

$$c_n(k) = \Pr(k \mid \gamma = 1, n)$$

does not depend on γ.

The posterior distribution for the concentration parameter is

$$\Pr(\gamma \mid k, n) \propto \Pr(\gamma \mid n) \times \Pr(k \mid \gamma, n) = \Pr(\gamma) \times \Pr(k \mid \gamma, n),$$

using a gamma prior distribution for the concentration parameter that does not depend on n, with shape and scale parameters $\eta_1 > 0$ and $\eta_2 > 0$ respectively. Since

$$\frac{\Gamma(\gamma)}{\Gamma(\gamma + n)} = \frac{(\gamma + n)\beta(\gamma + 1, n)}{\gamma \Gamma(n)} \propto \frac{(\gamma + n)\beta(\gamma + 1, n)}{\gamma},$$

where $\beta(\cdot,\cdot)$ is the beta function. We obtain

$$\begin{aligned}\Pr(\gamma \mid k,n) &\propto \Pr(\gamma)\gamma^{k-1}(\gamma+n)\beta(\gamma+1,n)\\ &= \Pr(\gamma)\gamma^{k-1}(\gamma+n)\int_0^1 s^{\gamma}(1-s)^{n-1}\,ds,\end{aligned}$$

which can be interpreted as the marginal probability of

$$\Pr(\gamma,\delta \mid k,n) \propto \Pr(\gamma)\gamma^{k-1}(\gamma+n)\delta^{\gamma}(1-\delta)^{n-1}, \quad \delta\in(0,1).$$

We can now derive the posterior distributions corresponding to $\Pr(\gamma \mid \delta,k,n)$ and $\Pr(\delta \mid \gamma,k,n)$, with

$$\begin{aligned}\Pr(\gamma \mid \delta,k,n) &\propto \gamma^{\eta_1-1}\gamma^{k-1}e^{-\gamma/\eta_2}e^{\gamma\log\delta}(\gamma+n)\\ &= \gamma^{\eta_1+k-2}e^{-\gamma(1/\eta_2-\log\delta)}(\gamma+n)\\ &\propto \gamma^{\eta_1+k-1}e^{-\gamma(1/\eta_2-\log\delta)} + n\gamma^{\eta_1+k-2}e^{-\gamma(1/\eta_2-\log\delta)},\end{aligned}$$

from which we conclude that

$$\gamma \mid (\delta,k,n) \sim \pi_\delta \times \Gamma\left(\eta_1+k, \frac{\eta_2}{1-\eta_2\log\delta}\right) + (1-\pi_\delta)\times\Gamma\left(\eta_1+k-1, \frac{\eta_2}{1-\eta_2\log\delta}\right), \tag{B.1}$$

with

$$\pi_\delta = \frac{\Gamma\left(\eta_1+k\right)}{\left(1/\eta_2-\log\delta\right)^{\eta_1+k}}. \tag{B.2}$$

The posterior distribution for the auxiliary variable δ is given by

$$\Pr(\delta \mid \gamma,k,n) \propto \delta^{\gamma}(1-\delta)^{n-1},$$

thus

$$\delta \mid (\gamma,k,n) \sim \text{Beta}(\gamma+1,n). \tag{B.3}$$

The update mechanism for the concentration parameter γ proceeds by first sampling a new value for δ from (B.3), then drawing a binomial replicate with probability of success (B.2), and finally using δ to sample a new value for γ from the component in the mixture distribution (B.1) selected through the binomial draw.

C List of bond yields and ratings by country

The data from the 53 countries used in Section 3.4 are summarised in Table C.1.

The 3-year bond yield refers to the figures available on the 20th of November 2017 at `https://www.investing.com/rates-bonds/world-government-bonds` on the previous traded day, which is 19 November in most cases, but can be earlier that month for the least liquid bonds.

Moody's latest credit ratings are available at `https://en.wikipedia.org/wiki/List_of_countries_by_credit_rating`. Moody's ratings are Aaa, Aa, A, Baa, Ba, B, Caa, Ca, C, from the least likely to the most likely to default. A number from 1 to 3 is sometimes appended to these ratings and fines down the rating within each of the nine categories, but this additional information was not considered in our example. The ratings reported in Table C.1 are those known on the 20th of November 2017, and most of them date back to the summer of 2017, but some of them date back to one or two years before.

Country	Yield	MCR	Country	Yield	MCR
Argentina	5.73	B	Latvia	0.1	A
Australia	1.956	Aaa	Lithuania	0.05	A
Austria	−0.539	Aa	Malaysia	3.549	A
Belgium	−0.651	Aa	Malta	0.004	A
Botswana	2.2	A	Mauritius	2.538	Baa
Brazil	8.94	Ba	Mexico	7.08	A
Bulgaria	−0.05	Baa	Namibia	9.045	Ba
Canada	1.511	Aaa	Netherlands	−0.656	Aaa
Chile	3.7	Aa	Philippines	4.08	Baa
China	3.771	A	Poland	1.981	A
Croatia	0.518	Ba	Portugal	−0.043	Ba
Czech Republic	0.687	A	Romania	3.391	Baa
Denmark	−0.529	Aaa	Russia	7.4	Ba
Egypt	15.88	B	Slovakia	−0.36	A
Finland	−0.571	Aa	South Africa	8.29	Baa
France	−0.472	Aa	South Korea	2.158	Aa
Germany	−0.637	Aaa	Spain	−0.194	Baa
Hong Kong	1.258	Aa	Sri Lanka	9.94	B
Hungary	0.65	Baa	Switzerland	−0.766	Aaa
India	6.722	Baa	Thailand	1.61	Baa
Indonesia	6.175	Baa	Turkey	13.27	Ba
Ireland	−0.516	A	Uganda	11.679	B
Israel	0.181	A	Ukraine	15.3	Caa
Italy	−0.042	Baa	United Kingdom	0.521	Aa
Japan	−0.158	A	United States	1.831	Aaa
Jordan	4.748	B	Vietnam	4.236	B
Kenya	12.2	B	Vietnam	4.236	B

Table C.1 – List of countries with 3-year sovereign bond yield (percentage) in mid-November 2017 and Moody's credit rating (MCR) at that same time.

D Penultimate approximation in the univariate case

This appendix presents the proof outlined in Smith (1987), with additional details. We are interested in the convergence of the distribution function $F^n(a_n x + b_n)$ towards its limit $G(x)$ with the appropriate norming $a_n > 0$ and b_n, as described in Theorem 2.1. Assuming existence of the density $f(\cdot) = F'(\cdot)$ and its derivative $f'(\cdot)$, we start with the representation of the survival distribution for x large, namely

$$\overline{F}(x) = \exp\left\{-\int_{x_F}^{x} \frac{dt}{h(t)}\right\},$$

where $h(t)$ is the reciprocal hazard function $\{1 - F(x)\}/f(x)$. Assuming $h'(x) \to \xi$ as $x \to x^F$, for some fixed $\xi \in \mathbb{R}$, we can write, for $x > 0$,

$$\begin{aligned}\frac{1 - F\{u + x \times h(u)\}}{1 - F(u)} &= \exp\left\{-\int_{x_F}^{u+x\times h(u)} \frac{dt}{h(t)} + \int_{x_F}^{u} \frac{dt}{h(t)}\right\} \\ &= \exp\left\{-\int_{u}^{u+x\times h(t)} \frac{dt}{h(t)}\right\}.\end{aligned}$$

With the change of variable $t \mapsto u + s \times h(u)$, we obtain

$$\frac{1 - F\{u + x \times h(u)\}}{1 - F(u)} = \exp\left\{-\int_0^x \frac{h(u)}{h\{u + s \times h(u)\}} ds\right\}. \tag{D.1}$$

The mean value theorem guarantees the existence of $y(u, s) = y \in [u, u + s \times h(u)]$, for any $s \in (0, x]$, such that

$$\frac{h\{u + s \times h(u)\}}{h(u)} = 1 + \int_0^s h'\{u + t \times h(u)\} dt = 1 + s \times h'(y),$$

thus there exist $y \in [u, u + x \times h(u)]$ such that

$$\int_0^x \left[\frac{h(u)}{h\{u + s \times h(u)\}} - \frac{1}{1 + s \times h'(y)}\right] ds = 0.$$

We can now make a substitution in (D.1),

$$\begin{aligned}\frac{1-F\{u+x\times h(u)\}}{1-F(u)} &= \exp\left\{-\int_0^x \frac{ds}{1+s\times h'(y)}\right\}\\ &= \exp\left(-\left[\frac{1}{h'(y)}\log\{1+s\times h'(y)\}\right]_{s=0}^{s=x}\right)\\ &= \{1+x\times h'(y)\}^{-1/h'(y)}.\end{aligned}$$

Since, for x close to x^F,

$$\frac{1-F\{u+x\times h(u)\}}{1-F(u)} \approx \frac{\log F\{u+x\times h(u)\}}{\log F(u)},$$

we have, by substitution of $u=b_n$, $h(b_n)=a_n$, and $-\log F(b_n)=1/n$,

$$F^n(a_n x+b_n) = \exp\left\{-(1+\xi_n x)_+^{-1/\xi_n}\right\},$$

with $\xi_n = h'(b_n)\to\xi$ as $n\to\infty$. Smith justifies the approximation of $h'(y)$ by $h'(u)$, by observing that in the case where $\xi=0$, $u+x\times h'(u)$ is much closer to u than to x^F for any fixed x. He also develops an argument justifying the same approximation when $\xi\neq 0$.

E Posterior densities for the semiparametric model

Posterior density for $\boldsymbol{\mu}_k$: The posterior density for $\boldsymbol{\mu}_k = (\mu_{1,k}, \ldots, \mu_{m,k})$ is multivariate Gaussian with independent margins, i.e.,

$$\mu_{j,k} \mid \boldsymbol{X}_j, \boldsymbol{X}_0, \psi^2_{j,k}, \alpha_j, \beta_j \overset{\text{ind}}{\sim} \mathcal{N}\left(M_{(\mu_{j,k})}, S^2_{(\mu_{j,k})}\right), \quad j = 1, \ldots, m, \quad k = 1, \ldots, N,$$

with posterior mean and variance

$$M_{(\mu_{j,k})} = S^2_{(\mu_{j,k})} \left(\frac{1}{\psi^2_{j,k}} \sum_{i \in C_k} \frac{X_{j,i} - \alpha_j X_{0,i}}{X_{0,i}^{\beta_j}} \right), \quad S^2_{(\mu_{j,k})} = \left(\frac{n_k}{\psi^2_{j,k}} + \frac{1}{\psi^2_{(\mu),j}} \right)^{-1},$$

where $\boldsymbol{X}_j = (X_{j,1}, \ldots, X_{j,n})$ are the observations at the jth lag, $C_k = \{i : c_i = k\}$, and $n_k = |C_k|$ is the number of observations in component k; the $\psi^2_{(\mu),j}$ are the variance parameters of the prior for the components' means.

Posterior density for $\boldsymbol{\psi}^2_k$: The multivariate posterior density for the components' variances can be split into independent parts,

$$\psi^2_{j,k} \mid \boldsymbol{X}_j, \boldsymbol{X}_0, \mu_{j,k}, \alpha_j, \beta_j \overset{\text{ind}}{\sim} \text{Inv-Gamma}\left(N_{1,j,k}, N_{2,j,k}\right), \quad j = 1, \ldots, m, \quad k = 1, \ldots, N,$$

with parameters

$$N_{1,j,k} = \frac{n_k}{2} + \nu_{1,j}, \quad N_{2,j,k} = \frac{1}{2} \sum_{i \in C_k} \frac{\left(X_{j,i} - \alpha_j X_{0,i} - \mu_{j,k} X_{0,i}^{\beta_j}\right)^2}{X_{0,i}^{2\beta_j}} + \nu_{2,j}.$$

Posterior density for $\boldsymbol{c}$: The posterior density is such that

$$c_i \mid \boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\psi}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w} \overset{\text{ind}}{\sim} \sum_{k=1}^{N} W_{k,i} \delta_k, \quad i = 1, \ldots, n,$$

where here for convenience $\boldsymbol{X}$, $\boldsymbol{\mu}$ and $\boldsymbol{\psi}^2$ are the matrices with rows $(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_m)$, $(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_N)$, and $(\boldsymbol{\psi}_1^2,\ldots,\boldsymbol{\psi}_N^2)$ respectively; the stick-breaking weights are defined as

$$W_{k,i} = \frac{w_k}{\overline{W}_i} \prod_{j=1}^{m} \left[\frac{1}{X_{0,i}^{\beta_j} \psi_{j,k}} \exp\left\{ -\frac{1}{2} \frac{\left(X_{j,i} - \alpha_j X_{0,i} - \mu_{j,k} X_{0,i}^{\beta_j}\right)^2}{X_{0,i}^{2\beta_j} \psi_{j,k}^2} \right\} \right],$$

with $\overline{W}_i = \sum_{k=1}^{N} W_{k,i}$ $(i = 1,\ldots,n)$ constants that make the weights sum to 1.

Posterior density for $\boldsymbol{w}$: The posterior density for $\boldsymbol{w}$ is generalised Dirichlet,

$$\boldsymbol{w} \mid \boldsymbol{c},\ \gamma \sim \text{GDirichlet}(a_1, b_1, \ldots, a_{N-1}, b_{N-1}),$$

where

$$a_k = 1 + n_k, \quad b_k = \gamma + \sum_{j=k+1}^{N} n_j, \quad k = 1,\ldots,N-1.$$

Posterior density for γ: The posterior density for the concentration parameter γ is

$$\text{Gamma}\left(N + \eta_1 - 1, \frac{\eta_2}{1 - \eta_2 \log w_N}\right) \mathbb{1}_{[\varepsilon,\infty)},$$

with $\varepsilon > 0$, typically $\varepsilon = 0.5$, and $\mathbb{1}$ is the indicator function.

F State-dependent proposal distribution for (α, β)

We suggest a joint proposal distribution for (α, β) in the conditional tail model whose form depends on the current state (α, β). This state-dependent proposal distribution (Roberts and Rosenthal, 2009; Rosenthal, 2011) aims at reducing the number of candidates proposed outside the boundaries defined by the constraints of Keef *et al.* (2013), thus improving the efficiency of the Markov chain Monte Carlo algorithm.

We consider a bivariate Gaussian proposal distribution with fixed standard deviations ψ_α and ψ_β and centred at the current state (α, β). In the following, we need to consider a particular iso-density contour as in Figure F.1, e.g., corresponding to the 2.5% and 97.5% marginal quantiles, defining an ellipse with semi-minor and semi-major axes of lengths $\psi_\alpha q_z(0.975)$ and $\psi_\beta q_z(0.975)$, with $q_z(\cdot)$ the quantile function of a standard Gaussian distribution.

The posterior distribution of (α, β) has a support boundary which cannot be expressed in closed form, and pointwise computations are needed to define an approximate boundary. In Figure F.1, we illustrate how the proposal density can be adapted to maximise the probability mass in the interior of the boundary, showing a particular contour of the proposal density. We use a bisection method to find points on the boundary with arbitrary precision. Given that the current state (α, β) lies in $[0,1] \times [0,1]$ to simplify the discussion, the details of the procedure are shown in Figure F.2 and are as follows:

1. find $P = (\alpha, \beta')$, $\beta' > \beta$, and $Q = (\alpha', \beta)$, $\alpha' > \alpha$, on the boundary ($\bullet$ in the left panel of Figure F.1);

2. the segment $\overline{PQ}$ joining P to Q is approximately tangent to the boundary; compute the perpendicular $\overline{T}$ to $\overline{PQ}$ going through (α, β);

3. find the point R at the intersection of $\overline{T}$ with the boundary;

4. find the point S at the intersection of $\overline{T}$ with a fixed isodensity contour (black ellipse in the right panel of Figure F.1);

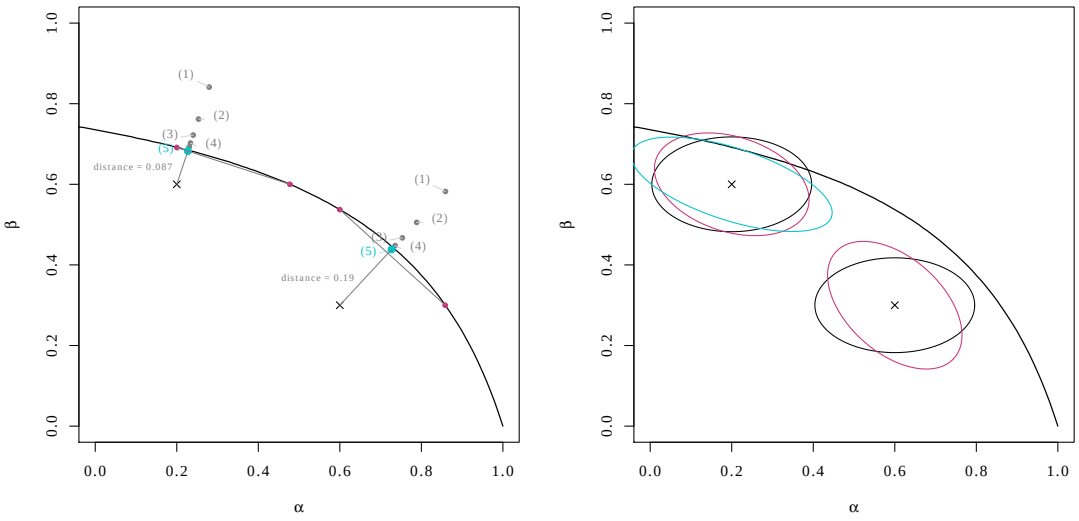

Figure F.1 – Construction of the state-dependent proposal distribution from a bivariate Gaussian distribution with independent margins. Left panel: definition of an approximate tangent to the boundary (•–•) and calculation of the approximate shortest distance (×–•) from the current state (×) to the boundary, where the point on the boundary is found by a bisection method whose successive iterations are shown (•) and numbered; right panel: adapting the bivariate Gaussian proposal distribution by rotation (–) and, if needed, reshaping (–), showing ellipsoids corresponding to the same iso-density curve.

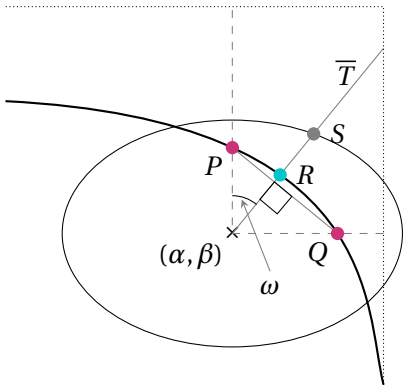

Figure F.2 – Details of the construction of the approximate tangent to the boundary and distance of the current state (α,β) to the boundary. The points P, Q, and S are found using a bisection method along $\overline{T}$ and the dashed lines $(--)$. The dotted segments show the top-right boundary of the set $[0,1]\times[0,1]$ and the ellipse represents the contour of interest of the proposal distribution centred in (α,β).

5. if S is further away from (α,β) than R, denote by d the distance from (α,β) to R and define a new covariance matrix such that the corresponding standard deviations are $q_z(0.975)\psi_\alpha\psi_\beta/d$ and $d/q_z(0.975)$;

6. rotate the new covariance matrix by ω, the angle formed by $\overline{T}$ and the line going through (α,β) and P.

In order to stop the proposal distribution from converging to a degenerate form, we need to define $\delta>0$ and $d'=\max(\delta,d)$ in step 5., so that the probability of sampling a candidate $(\alpha^\star,\beta^\star)$ in the interior of the boundary remains always non-negligible, even when the current state (α,β) is arbitrarily close to the boundary. In order for the state-dependent proposal distribution to be efficient, we assume that the boundary of the support of (α,β) is smooth, so that the approximation $\overline{PQ}$ to the tangent of the boundary is reasonably accurate.

An adaptive scheme (Haario *et al.*, 2001) for the proposal standard deviations ψ_α and ψ_β can be used with the state-dependent proposal distribution. The standard deviations would be adapted depending on the mean acceptance rate computed on a batch of past iterations of the algorithm. A typical condition for the convergence of the sampler is that the size of the adaptation vanishes as the number of iterations increases.

G Proof of Theorem 6.2

In this appendix, we show how to derive an approximation to the integral

$$\int_u^\infty \Phi\left\{\frac{x-\alpha x-\mu x^\beta}{\psi x^\beta}\right\} e^{-x}\,dx, \tag{G.1}$$

for large u, and we define $\mu(x)=\alpha x+\mu x^\beta$ and $\psi(x)=\psi x^\beta$, with $\alpha<1$ since (X,Y) are assumed asymptotically independent.

We can follow the procedure used for the proof of Theorem 6.1 to get

$$1-\Phi\left\{\frac{x-\mu(x)}{\psi(x)}\right\} = \int_{\frac{x-\mu(x)}{\psi(x)}}^\infty \varphi(y)\,dy \tag{G.2}$$

$$\sim \varphi\left\{\frac{x-\mu(x)}{\psi(x)}\right\}\frac{\psi(x)}{x-\mu(x)}, \quad x>u, \tag{G.3}$$

for large u, and because X and Y are assumed asymptotically independent, this ensures that $\{x-\mu(x)\}/\psi(x)$ is strictly increasing in x, so the integral on the right-hand side of (G.2) vanishes when $x\to\infty$. From (G.3), it follows that (G.1) can be approximated as

$$\begin{aligned} &e^{-u}-\int_u^\infty e^{-x}\frac{\psi(x)}{\{x-\mu(x)\}\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left\{\frac{x-\mu(x)}{\psi(x)}\right\}^2\right]dx \\ &=e^{-u}-\int_u^\infty \frac{\psi(x)}{\{x-\mu(x)\}\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left\{\frac{x-\mu(x)+\psi(x)^2}{\psi(x)}\right\}^2\right]\exp\left\{-\mu(x)+\frac{\psi(x)^2}{2}\right\}dx, \end{aligned} \tag{G.4}$$

where the integral can be further expanded by integration by parts, giving

$$\begin{aligned} \int_u^\infty p(x)\varphi\{q(x)\}dx &= \int_u^\infty \frac{p(x)}{\tilde{q}(x)}\varphi\left\{q(x)\right\}\tilde{q}(x)dx \\ &= \left[-\frac{p(x)}{\tilde{q}(x)}\varphi\left\{q(x)\right\}\right]_u^\infty + \int_u^\infty \frac{d}{dx}\left\{\frac{p(x)}{\tilde{q}(x)}\right\}\varphi\left\{q(x)\right\}dx, \end{aligned} \tag{G.5}$$

with

$$\begin{aligned} p(x) &= \frac{\psi(x)}{x-\mu(x)} \exp\left\{-\mu(x) + \frac{\psi(x)^2}{2}\right\}, \\ q(x) &= \frac{x-\mu(x)+\psi(x)^2}{\psi(x)}, \\ \tilde{q}(x) &= \frac{1}{2}\frac{d}{dx} q(x)^2. \end{aligned} \tag{G.6}$$

From (G.5) and using (G.6), we derive the approximation

$$\int_u^\infty p(x)\varphi\{q(x)\}dx \sim \frac{p(u)}{\tilde{q}(u)}\varphi\{q(u)\}, \tag{G.7}$$

for large u, provided $d\{p(x)\tilde{q}(x)^{-1}\}/dx = o\{p(x)\}$ as $x \to \infty$, to ensure that the remaining integral term in (G.5) is of smaller order than the left-hand side of (G.7).

The approximation in (G.7) only holds for $\beta < 1/2$, as we shall see below, and can be written as

$$-\frac{1}{2\sqrt{2\pi}} \frac{\psi^3 u^{3\beta}}{\{u(1-\alpha)-\mu(u)^\beta\}\{u(1-\alpha)-\mu(u)^\beta+\psi^2(u)^{2\beta}\}\{(1-\beta)(1-\alpha)+\beta\psi^2(u)^{2\beta-1}\}} \times \exp\left[-\frac{u^2-2u\{\mu(u)+\psi(u)^2\}+\mu(u)^2}{2\psi(u)^2}\right],$$

where the fraction behaves like

$$-\frac{1}{2\sqrt{2\pi}} \frac{\psi^3}{(1-\alpha)^3(1-\beta)} u^{3\beta-2},$$

which proves the theorem.

In order to show that $dp(x)\tilde{q}(x)^{-1}/dx = o\{p(x)\}$, we first note that

$$\tilde{q}(x) = \frac{\{x-\mu(x)+\psi(x)^2\}\{(1-\beta)(1-\alpha)+\beta\psi^2 x^{2\beta-1}\}}{\psi(x)^2}.$$

From here we can compute the ratio $p(x)/\tilde{q}(x)$ and differentiate it to get

$$
\begin{aligned}
&\frac{d}{dx}\left\{\frac{p(x)}{\tilde{q}(x)}\right\} \\
&= \exp\left\{-\mu(x)+\frac{\psi(x)^2}{2}\right\}\left(\left[\frac{(1-\alpha)(\beta-1)}{\{x-\mu(x)\}^2}-\frac{\alpha+\beta\mu x^{\beta-1}-\beta\psi^2x^{2\beta-1}}{x-\mu(x)}\right]\tilde{q}(x)\psi(x)^2\right. \\
&\quad -\left[\frac{(1-\alpha)(1-\beta)\{1-\alpha-\beta\mu x^{\beta-1}+2\beta\psi(x)^2\}}{x-\mu(x)}\right. \\
&\quad +\frac{2\beta\psi^2x^{2\beta-1}\{\beta(1-\alpha)-(3\beta-1)\mu x^{\beta-1}/2+(2\beta-1/2)\psi^2x^{2\beta-1}\}}{x-\mu(x)} \\
&\quad \left.\left.-\frac{\left(1-\alpha-\mu x^{\beta-1}+\psi^2x^{2\beta-1}\right)\{(1-\alpha)(1-\beta)+\beta\psi^2x^{2\beta-1}\}2\beta}{x-\mu(x)}\right]\right) \\
&\quad \times\frac{\psi(x)^3}{\{x-\mu(x)+\psi(x)^2\}^2\{(1-\beta)(1-\alpha)+\beta\psi^2x^{2\beta-1}\}^2}.
\end{aligned}
\tag{G.8}
$$

We conclude from expression (G.8) that convergence to 0 only holds for $\beta < 1/2$, with rate $O(x^{4\beta-2})$. For $\beta \geq 1/2$ we cannot use the approximation (G.7) and we are left with the integral form (G.4).

H Numerical approach to estimating joint tail probabilities

A standard approach to evaluation of the integral of Theorem 6.2, needed in (6.23) and (6.25), is by simulation of (x, y) pairs, with x above the conditional threshold u, and computation of the proportion of these pairs falling into the set of interest, namely $\{(x, y) \in \mathbb{R}^2 : x > u, y \leq u\}$, as detailed in Heffernan and Tawn (2004). Monte Carlo integration can help reduce the large computational cost incurred by this standard approach; comparison of RMSE for probabilities of extreme sets with data simulated from a logistic distribution shows that the standard approach needs at least 10^4 simulations to reach an accuracy similar to a Monte Carlo-type integration using less than 100 replicates (Winter, 2015, Chap. 2).

In his Ph.D. thesis, Winter treats the residuals $\widehat{z}_i = (y_i - \widehat{\alpha} x_i)/x_i^{\widehat{\beta}}$ $(i = 1, \ldots, n)$ as a given fixed sample of the Gaussian distribution on the left-hand side of (H.1). He then evaluates the exponential density on the left-hand side of (H.1) using the order statistics $\widehat{z}_{(1)} < \cdots < \widehat{z}_{(n)}$, allowing for a computationally efficient procedure to not include residuals that would not contribute to the integral. For further details and an illustration of this methodology, see Winter (2015, Section 2.6.5). This approach relies on an informal argument and is shown to work well in practice. We develop here a more formal argument and show that more careful attention is needed, depending on the exact values of $\widehat{\alpha}$ and $\widehat{\beta}$.

Winter's approach corresponds to a Monte Carlo integration with fixed samples, as we can write

$$\int_{u^{\mathrm{L}}}^{\infty} \overline{\Phi}\left(\frac{u^{\mathrm{L}} - \alpha x - \mu x^{\beta}}{\psi x^{\beta}}\right) e^{-x} dx = \int_{u^{\mathrm{L}}}^{\infty} \int_{\frac{u^{\mathrm{L}} - \alpha x - \mu x^{\beta}}{\psi x^{\beta}}}^{\infty} \varphi(y) e^{-x} dy dx \tag{H.1}$$

$$= \int_{\inf_{x \geq u^{\mathrm{L}}} \left(\frac{u^{\mathrm{L}} - \alpha x - \mu x^{\beta}}{\psi x^{\beta}}\right)}^{\infty} \left\{e^{-A(y)} - e^{-B(y)}\right\} \varphi(y) dy, \tag{H.2}$$

where $A(y), B(y)$ are the roots in x of $y = (u^{\mathrm{L}} - \alpha x - \mu x^{\beta})/(\psi x^{\beta})$, except in some special cases, with

$$A(y) \equiv u^{\mathrm{L}}, \quad y \geq \frac{u^{\mathrm{L}}(1 - \alpha) - (u^{\mathrm{L}})^{\beta} \mu}{(u^{\mathrm{L}})^{\beta} \psi},$$

and

$$\alpha = \beta = 0 \implies A(y) \equiv u^{\mathrm{L}},$$
$$\alpha \geq 0 \text{ or } \beta = 1 \implies B(y) \equiv +\infty.$$

We can derive the following forms for the lower bound of integral (H.2),

$$\inf_{x \geq u^{\mathrm{L}}} \left(\frac{u^{\mathrm{L}} - \alpha x - \mu x^{\beta}}{\psi x^{\beta}} \right) = \begin{cases} -\dfrac{\alpha + \mu}{\psi}, & \beta = 1 \text{ or } \alpha = 0, \\ -\infty, & \alpha > 0,\ \beta < 1, \\ \dfrac{u^{\mathrm{L}}(1-\alpha) - (u^{\mathrm{L}})^{\beta}\mu}{(u^{\mathrm{L}})^{\beta}\psi}, & \alpha < 0,\ \beta < 1,\ u^{\mathrm{L}} > \dfrac{u^{\mathrm{L}}\beta}{\alpha(\beta-1)}, \\ -\dfrac{u^{\mathrm{L}}\{\alpha(\beta-1)\}^{\beta} + (\beta-1)(u^{\mathrm{L}}\beta)^{\beta}\mu}{(\beta-1)(u^{\mathrm{L}}\beta)^{\beta}\psi}, & \text{otherwise.} \end{cases}$$

In Winter's approach, the case $\alpha < 0$ is considered only when u is large enough so that $(u^{\mathrm{L}} - \alpha x - \mu x^{\beta})/(\psi x^{\beta})$ can be considered monotonic, but no condition is given that ensures monotonicity; samples from $\varphi(y)$ are fixed and given by the model fit.

As the integration involved in Φ is handled by R, it is easier to consider the left-hand side of (H.2) in our approach to estimate $\overline{F}(u^{\mathrm{L}}, u^{\mathrm{L}})$ in (6.23). A Monte Carlo approach is easy to implement, and typically a shifted exponential with density $\exp\{-(x - u^{\mathrm{L}})\}$ can be used as an instrumental distribution to avoid sampling replicates outside the limits of the integral. This importance sampling approach boils down to computing

$$\frac{1}{R} \sum_{r=1}^{R} \overline{\Phi} \left\{ \frac{u^{\mathrm{L}} - \alpha x^{(r)} - (x^{(r)})^{\beta}\mu}{(x^{(r)})^{\beta}\psi} \right\} e^{-u^{\mathrm{L}}}, \tag{H.3}$$

with the $x^{(r)}$ sampled from the shifted exponential distribution. Because the integral is one-dimensional, a more efficient technique is Riemannian simulation (see for example Philippe (1997) or Robert and Casella (2004, Chap. 3)),

$$\sum_{r=1}^{R-1} \left(x^{(r+1)} - x^{(r)}\right) \overline{\Phi} \left\{ \frac{u^{\mathrm{L}} - \alpha x^{(r)} - (x^{(r)})^{\beta}\mu}{(x^{(r)})^{\beta}\psi} \right\} e^{-x^{(r)}} = \sum_{r=1}^{R-1} \left(x^{(r+1)} - x^{(r)}\right) d(x^{(r)}), \tag{H.4}$$

which has a variance of order $O(R^{-2})$, compared to $O(R^{-1})$ for the classical Monte Carlo integration estimator in (H.3). Yakowitz *et al.* (1978) show that we can improve the Riemannian estimator by symmetrising (H.4), giving

$$\sum_{r=1}^{R-1} \left(x^{(r+1)} - x^{(r)}\right) \frac{d(x^{(r)}) + d(x^{r+1})}{2},$$

which has variance of order $O(R^{-4})$. Riemann sum estimators handle importance sampling nicely, as they do not involve evaluation of the instrumental distribution.

In this appendix, we have illustrated approaches to estimating integrals of the type of the left-hand side of (H.2) in a bivariate setup for simplicity, but as mentioned in Section 6.3.1, a quadratic approximation is much faster and accurate in two dimensions. In higher dimensions however, Monte Carlo estimation can be more efficient and the methods described here can be used to get reliable estimates of the integral in (H.2).

I State-dependent proposal distribution for batch updates

In Chapter 5, we were able to construct a Gibbs sampler by truncating the infinite sum of the stick-breaking representation (5.14), so that the likelihood function could be expressed as the product of a finite number of Gaussian densities. In Chapter 7, the likelihood function (7.6) is analytically intractable, so that Gibbs sampling is not possible. In this appendix, we detail an attempt at taking advantage of the Gibbs sampler of Chapter 5 in the context of Chapter 7.

To simplify notation, we assume strong exchangeability of (X_t, X_{t+1}) with Laplace marginal distribution, but the argument follows the same lines without exchangeability and with a general marginal distribution. Since we use the setup of Ishwaran and Zarepour (2000), we deal with a mixture of Gaussian distributions with K components, where K is sufficiently large for the truncated stick-breaking representation to well approximate the Dirichlet process, and we consider the means μ_k and the variances ψ_k^2 $(k = 1, \ldots, K)$ of the components. We sample the means and variances using blocked updates, following the terminology of Ishwaran and Zarepour (2000), but here we need a Metropolis within Gibbs approach as the posterior distributions have no closed form. A standard kernel proposal distribution is possible but would be inefficient. We develop a proposal distribution whose candidates are more likely to be accepted.

We suggest using the posterior distributions derived in the context of Chapter 5 for the conditional models $M_{|t}$, where $X_{t+1} \mid X_t > x$ with x extreme, and $M_{|t+1}$, where $X_t \mid X_{t+1} = x$ with x extreme, as state-dependent proposal distributions. This simplifies the acceptance ratio and the computations involved in the Metropolis update, and proposes parameter candidates that improve mixing.

As in Chapter 5, we take independent normal priors for the means of the mixture components, so that the joint prior is

$$\pi_\mu(\boldsymbol{\mu}) = \pi_\mu(\mu_1, \ldots, \mu_K) \propto \prod_{k=1}^{K} \frac{1}{\psi_\mu} \varphi\left(\frac{\mu_k}{\psi_\mu}\right),$$

with $\psi_\mu > 0$ a standard deviation reflecting our prior knowledge, and $\varphi(\cdot)$ the standard normal density function.

Writing $\mathscr{D}_{ij} = \{(x_t, x_{t+1}) \in R_{ij}\}$ and $\mathscr{T}_{ij} = \{t \in 1,\ldots,n : (x_t, x_{t+1}) \in R_{ij}\}$, the joint posterior density for points in $\mathscr{D}_{11} \cup \mathscr{D}_{10}$ is

$$\pi_\mu^{\text{Gibbs}}\left(\boldsymbol{\mu} \,\middle|\, \mathscr{D}_{11} \cup \mathscr{D}_{10}\right) \propto \prod_{\mathscr{T}_{11} \cup \mathscr{T}_{10}} f_{t+1|t}\left(x_{t+1} \mid x_t, \mu_t\right) \pi_\mu\left(\boldsymbol{\mu}\right), \tag{I.1}$$

according to model $M_{|t}$, and similarly the joint posterior density for points in $\mathscr{D}_{11} \cup \mathscr{D}_{01}$ is

$$\pi_\mu^{\text{Gibbs}}\left(\boldsymbol{\mu} \mid \mathscr{D}_{11} \cup \mathscr{D}_{01}\right) \propto \prod_{\mathscr{T}_{11} \cup \mathscr{T}_{01}} f_{t|t+1}\left(x_t \,\middle|\, x_{t+1}, \mu_t\right) \pi_\mu\left(\boldsymbol{\mu}\right), \tag{I.2}$$

according to model $M_{|t+1}$.

The posterior densities (I.1) and (I.2) are used as proposal densities for the model of Chapter 7, so that, for a current state $\boldsymbol{\mu}$ and a candidate state $\boldsymbol{\mu}^\star$, the acceptance ratio is

$$\begin{aligned} e^{-|x_1|} \frac{\prod_{i,j\in\{0,1\}} \prod_{\mathscr{T}_{ij}} f_{t,t+1}\left(x_t, x_{t+1} \,\middle|\, \boldsymbol{\mu}^\star\right) e^{|x_t|} \pi_\mu\left(\boldsymbol{\mu}^\star\right)}{\prod_{i,j\in\{0,1\}} \prod_{\mathscr{T}_{ij}} f_{t,t+1}\left(x_t, x_{t+1} \,\middle|\, \boldsymbol{\mu}\right) e^{|x_t|} \pi_\mu\left(\boldsymbol{\mu}\right)} \\ \times \frac{\pi_\mu^{\text{Gibbs}}\left(\boldsymbol{\mu} \mid \mathscr{D}_{11} \cup \mathscr{D}_{10}\right) \pi_\mu^{\text{Gibbs}}\left(\boldsymbol{\mu} \,\middle|\, \mathscr{D}_{11} \cup \mathscr{D}_{01}\right)}{\pi_\mu^{\text{Gibbs}}\left(\boldsymbol{\mu}^\star \,\middle|\, \mathscr{D}_{11} \cup \mathscr{D}_{10}\right) \pi_\mu^{\text{Gibbs}}\left(\boldsymbol{\mu}^\star \mid \mathscr{D}_{11} \cup \mathscr{D}_{01}\right)}, \end{aligned} \tag{I.3}$$

where

$$\begin{aligned} \prod_{i,j\in\{0,1\}} \prod_{\mathscr{T}_{ij}} f_{t,t+1}\left(x_t, x_{t+1} \,\middle|\, \boldsymbol{\mu}\right) \propto \left\{F_{t,t+1}\left(u, u; \boldsymbol{\mu}\right)\right\}^{n_{00}} \prod_{\mathscr{T}_{|t}} f_{t+1|t}\left(x_{t+1} \,\middle|\, x_t; \mu_t\right) e^{-|x_t|} \\ \times \prod_{\mathscr{T}_{|t+1}} f_{t|t+1}\left(x_t \,\middle|\, x_{t+1}; \mu_t\right) e^{-|x_{t+1}|}, \end{aligned} \tag{I.4}$$

with $\mathscr{T}_{|t} = \{t : (x_t, x_{t+1}) \in R_{|t}\}$ and $\mathscr{T}_{|t+1} = \{t : (x_t, x_{t+1}) \in R_{|t+1}\}$. From (I.4), we conclude that the exponential terms cancel in (I.3), the conditional density terms cancel with the same terms in the proposal densities (I.1) and (I.2) for data points in $R_{|t}$ and $R_{|t+1}$, and the prior density terms cancel with the same terms in (I.1) and (I.2), giving the simplified acceptance ratio

$$e^{-|x_1|} \frac{\left\{F_{t,t+1}\left(u, u \,\middle|\, \boldsymbol{\mu}^\star\right)\right\}^{n_{00}}}{\left\{F_{t,t+1}\left(u, u \,\middle|\, \boldsymbol{\mu}\right)\right\}^{n_{00}}} \frac{\prod_{\mathscr{T}_{11} \cap \mathscr{T}_{|t+1}} f_{t+1|t}\left(x_{t+1} \mid x_t; \mu_t\right) \prod_{\mathscr{T}_{11} \cap \mathscr{T}_{|t}} f_{t|t+1}\left(x_t \mid x_{t+1}; \mu_t\right)}{\prod_{\mathscr{T}_{11} \cap \mathscr{T}_{|t+1}} f_{t+1|t}\left(x_{t+1} \mid x_t; \mu_t^\star\right) \prod_{\mathscr{T}_{11} \cap \mathscr{T}_{|t}} f_{t|t+1}\left(x_t \mid x_{t+1}; \mu_t^\star\right)}. \tag{I.5}$$

A similar argument can be used to simplify the acceptance ratio for the variances ψ_k^2 ($k = 1,\ldots,K$) of the mixture components.

The posterior distribution for the weights of the components $\boldsymbol{w} = (w_1,\ldots,w_K)$ needs care, as these are used in the censored contributions in the likelihood (7.6). The posterior density

function for the weights in $M_{|t}$ and M_{t+1} is

$$\pi_w^{\text{Gibbs}}(\boldsymbol{w} \mid \boldsymbol{c}) \propto f_{c|w}(\boldsymbol{c} \mid \boldsymbol{w})\, \pi_w(\boldsymbol{w}), \tag{I.6}$$

where $\boldsymbol{c} = (c_1, \ldots, c_n)$ is a vector of auxiliary variables storing the component indices for each pair of data points, $f_{c|w}(\boldsymbol{c} \mid \boldsymbol{w})$ is a multinomial density, and $\pi_w(\boldsymbol{w})$ is generalised Dirichlet. With (I.6) and for current and candidate states $\boldsymbol{w}$ and $\boldsymbol{w}^\star$ respectively, the acceptance ratio for the model of Chapter 7 is

$$\frac{\left\{F_{t,t+1}\left(u,u \mid \boldsymbol{w}^\star\right)\right\}^{n_{00}} f_{c|w}\left(\boldsymbol{c} \mid \boldsymbol{w}^\star\right) \pi_w\left(\boldsymbol{w}^\star\right) \pi_w^{\text{Gibbs}}(\boldsymbol{w} \mid \boldsymbol{c})}{\left\{F_{t,t+1}(u,u \mid \boldsymbol{w})\right\}^{n_{00}} f_{c|w}(\boldsymbol{c} \mid \boldsymbol{w})\, \pi_w(\boldsymbol{w})\, \pi_w^{\text{Gibbs}}(\boldsymbol{w}^\star \mid \boldsymbol{c})} = \frac{\left\{F_{t,t+1}\left(u,u \mid \boldsymbol{w}^\star\right)\right\}^{n_{00}}}{\left\{F_{t,t+1}(u,u \mid \boldsymbol{w})\right\}^{n_{00}}}. \tag{I.7}$$

In the bivariate context of this appendix, the integrals involved in (I.5) and (I.7) can be estimated accurately and quickly using quadratic approximations. In higher-dimensional setups, a pseudo-marginal approach can be used, involving Monte Carlo integration at each iteration, termed Monte Carlo within Metropolis by O'Neill *et al.* (2000), or the data augmentation approach of Beaumont (2003), which does not yield approximation of the posterior distribution (Andrieu and Roberts, 2009). Doucet *et al.* (2015) give general guidance for dealing with approximation of likelihood functions within Metropolis algorithms.

Bibliography

Agarwal, D. K. and Gelfand, A. E. (2005) Slice sampling for simulation based fitting of spatial data models. *Statistics and Computing* **15**, 61–69.

Anderson, C. W. (1971) *Contributions to the Asymptotic Theory of Extreme Values*. Ph.D. thesis, University of London.

Andrieu, C. and Roberts, G. O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* **37**, 697–725.

Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

Asadi, P., Davison, A. C. and Engelke, S. (2015) Extremes on river networks. *The Annals of Applied Statistics* **9**, 2023–2050.

Badoux, A., Andres, N., Techel, F. and Hegg, C. (2016) Natural hazard fatalities in Switzerland from 1946 to 2015. *Natural Hazards and Earth System Sciences* **16**, 2747–2768.

Ballani, F. and Schlather, M. (2011) A construction principle for multivariate extreme value distributions. *Biometrika* **98**, 633–645.

Barnett, V. D. (1976) The ordering of multivariate data (with discussion). *Journal of the Royal Statistical Society Series A* **139**, 318–355.

Barron, A. R., Schervish, M. J. and Wasserman, L. (1999) The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics* **27**, 536–561.

Basrak, B., Davis, R. A. and Mikosch, T. (2002) A characterization of multivariate regular variation. *The Annals of Applied Probability* **12**, 908–920.

Beaumont, M. A. (2003) Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160.

Beirlant, J., Goegebeur, Y., Teugels, J. L. and Segers, J. (2004) *Statistics of Extremes: Theory and Applications*. Chichester: John Wiley & Sons.

Bennett, O. and Hartwell-Naguib, S. (2014) Flood defence spending in England. `https://researchbriefings.parliament.uk/ResearchBriefing/Summary/SN05755`. UK Parliament: standard note SN/SC/5755.

Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1987) *Regular Variation*. Cambridge: Cambridge University Press.

Blackwell, D. (1973) Discreteness of Ferguson selections. *The Annals of Statistics* **1**, 356–358.

Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via Polya urn schemes. *The Annals of Statistics* **1**, 353–355.

Bofinger, E. and Bofinger, V. J. (1965) The correlation of maxima in samples drawn from a bivariate normal distribution. *Australian Journal of Statistics* **7**, 57–61.

Bofinger, V. J. (1970) The correlation of maxima in several bivariate non-normal distributions. *Australian Journal of Statistics* **12**, 1–7.

Bortot, P. and Gaetan, C. (2014) A latent process model for temporal extremes. *Scandinavian Journal of Statistics* **41**, 606–621.

Bortot, P. and Tawn, J. A. (1998) Models for the extremes of Markov chains. *Biometrika* **85**, 851–867.

Bush, C. A. and MacEachern, S. N. (1996) A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

Cai, J.-J., Einmahl, J. H. J., de Haan, L. and Zhou, C. (2015) Estimation of the marginal expected shortfall: the mean when a related variable is extreme. *Journal of the Royal Statistical Society Series B* **77**, 417–442.

de Carvalho, M. and Ramos, A. (2012) Bivariate extreme statistics, II. *REVSTAT* **10**, 83–107.

Chavez-Demoulin, V. and Davison, A. C. (2012) Modelling time series extremes. *REVSTAT* **10**, 109–133.

Chavez-Demoulin, V., Davison, A. C. and McNeil, A. J. (2005) Estimating value-at-risk: a point process approach. *Quantitative Finance* **5**, 227–234.

Chavez-Demoulin, V., Embrechts, P. and Sardy, S. (2014) Extreme-quantile tracking for financial time series. *Journal of Econometrics* **181**, 44–52.

Cheng, L., Gilleland, E., Heaton, M. J. and AghaKouchak, A. (2014) Empirical Bayes estimation for the conditional extreme value model. *Stat* **3**, 391–406.

Clayton, D. G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151.

Coles, S. G. (1993) Regional modelling of extreme storms via max-stable processes. *Journal of the Royal Statistical Society Series B* **55**, 797–816.

Coles, S. G. (2001) *An Introduction to Statistical Modeling of Extreme Value*. London: Springer.

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 Pauli, F. (2002) Models and inference for uncertainty in extremal dependence. *Biometrika* **89**, 183–196.

Coles, S. G. and Tawn, J. A. (1991) Modelling extreme multivariate events. *Journal of the Royal Statistical Society Series B* **53**, 377–392.

Coles, S. G. and Tawn, J. A. (1994) Statistical methods for multivariate extremes: an application to structural design (with discussion). *Journal of the Royal Statistical Society Series C* **43**, 1–48.

Coles, S. G. and Tawn, J. A. (1996a) Modelling extremes of the areal rainfall process. *Journal of the Royal Statistical Society Series B* **58**, 329–347.

Coles, S. G. and Tawn, J. A. (1996b) A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society Series C* **45**, 463–478.

Coles, S. G. and Walshaw, D. (1994) Directional modelling of extreme wind speed. *Journal of the Royal Statistical Society Series C* **43**, 139–157.

Connor, R. J. and Mosimann, J. E. (1969) Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association* **64**, 194–206.

Cooley, D., Davis, R. A. and Naveau, P. (2010) The pairwise beta distribution: a flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis* **101**, 2103–2117.

Daley, D. J. and Vere-Jones, D. (2003) *An Introduction to the Theory of Point Processes: Elementary Theory*. Second edition. New York: Springer.

Das, B. and Resnick, S. I. (2011) Conditioning on an extreme component: model consistency with regular variation on cones. *Bernoulli* **17**, 226–252.

Davis, R. A. and Mikosch, T. (2009) The extremogram: a correlogram for extreme events. *Bernoulli* **15**, 977–1009.

Davison, A. C. and Smith, R. L. (1990) Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society Series B* **52**, 393–442.

Dey, D., Müller, P. and Sinha, D. (eds) (1998) *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer.

Doucet, A., Pitt, M. K., Deligiannidis, G. and Kohn, R. (2015) Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102**, 295–313.

Duan, J. A., Guindani, M. and Gelfand, A. E. (2007) Generalized spatial Dirichlet process models. *Biometrika* **94**, 809–825.

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* **58**, 25–45.

Eastoe, E. F. and Tawn, J. A. (2012) Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. *Biometrika* **99**, 43–55.

Einmahl, J., Einmahl, J. H. J. and de Haan, L. (2017) Limits to human life span through extreme value theory. Technical report, Center for Economic Research, Tilburg.

Einmahl, J. H. J., de Haan, L. and Piterbarg, V. I. (2001) Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics* **29**, 1401–1423.

Einmahl, J. H. J. and Segers, J. (2009) Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics* **37**, 2953–2989.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance*. Berlin: Springer.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Escobar, M. D. and West, M. (1998) *Computing Nonparametric Hierarchical Models*, pp. 1–22. New York: Springer.

Fawcett, L. and Walshaw, D. (2006a) A hierarchical model for extreme wind speeds. *Journal of the Royal Statistical Society Series C* **55**, 631–646.

Fawcett, L. and Walshaw, D. (2006b) Markov chain models for extreme wind speeds. *Environmetrics* **17**, 795–809.

Fawcett, L. and Walshaw, D. (2007) Improved estimation for temporally clustered extremes. *Environmetrics* **18**, 173–188.

Fawcett, L. and Walshaw, D. (2012) Estimating return levels from serially dependent extremes. *Environmetrics* **23**, 272–283.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Ferro, C. A. T. and Segers, J. (2003) Inference for clusters of extreme values. *Journal of the Royal Statistical Society Series B* **65**, 545–556.

Fisher, R. A. and Tippett, L. H. C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society* **24**, 180–190.

Flegal, J. M., Hughes, J., Vats, D. and Dai, N. (2017) *mcmcse: Monte Carlo standard errors for MCMC*. R package version 1.3-2.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005) Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

Genest, C., Ghoudi, K. and Rivest, L.-P. (1995) A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82**, 543–552.

Genz, A. and Bretz, F. (2009) *Computation of Multivariate Normal and t Probabilities*. Heidelberg: Springer.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leich, F., Scheipl, F. and Hothorn, T. (2014) *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-2.

Geyer, C. J. (2011) Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*, eds S. Brooks, A. Gelman, G. Jones and X.-L. Meng. Chapman & Hall/CRC Press.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999) Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics* **27**, 143–158.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000) Convergence rates of posterior distributions. *The Annals of Statistics* **28**, 500–531.

Ghosal, S. and van der Vaart, A. W. (2017) *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Gnedenko, B. V. (1943) Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics* **44**, 423–453.

Gomes, M. I. (1984) Penultimate limiting forms in extreme value theory. *Annals of the Institute of Statistical Mathematics* **36**, 71–85.

Gomes, M. I. (1994) Penultimate behaviour of the extremes. In *Extreme Value Theory and Applications*, eds J. Galambos, J. Lechner and E. Simiu, pp. 403–418.

Gomes, M. I. and Pestana, D. D. (1987) Nonstandard domains of attraction and rates of convergence. In *New Perspectives in Theoretical and Applied Statistics*, eds M. L. Puri, J. P. Vilaplana and W. Wertz, pp. 467–477. John Wiley & Sons.

Gong, L. and Flegal, J. M. (2016) A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* **25**, 684–700.

Gosh, J. K. and Ramamoorthi, R. V. (2003) *Bayesian Nonparametrics*. New York: Springer.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Gumbel, E. J. (1960) Bivariate exponential distributions. *Journal of the American Statistical Association* **55**, 698–707.

Gumbel, E. J. and Goldstein, N. (1964) Analysis of empirical bivariate extremal distributions. *Journal of the American Statistical Association* **59**, 794–816.

Gummer, B. and Leasom, A. (2016) National flood resilience review. `http://www.gov.uk/government/publications`.

de Haan, L. and Ferreira, A. (2006) *Extreme Value Theory: An Introduction*. New York: Springer.

de Haan, L. and de Ronde, J. (1998) Sea and wind: multivariate extremes at work. *Extremes* **1**, 7–45.

Haario, H., Saksman, E. and Tamminen, J. (2001) An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242.

Hall, P. and Tajvidi, N. (2000) Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli* **6**, 835–844.

Hartmann, P., Straetmans, S. and de Vries, C. G. (2004) Asset market linkages in crisis periods. *The Review of Economics and Statistics* **86**, 313–326.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Heffernan, J. E. and Resnick, S. I. (2007) Limit laws for random vectors with an extreme component. *The Annals of Applied Probability* **17**, 537–571.

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* **66**, 497–546.

Hilal, S., Poon, S.-H. and Tawn, J. A. (2011) Hedging the Black Swan: conditional heteroskedasticity and tail dependence in S&P500 and VIX. *Journal of Banking and Finance* **35**, 2374–2387.

Hilal, S., Poon, S.-H. and Tawn, J. A. (2014) Portfolio risk assessment using multivariate extreme value methods. *Extremes* **17**, 531–556.

Hilker, N., Badoux, A. and Hegg, C. (2008) Unwetterschäden in der Schweiz im Jahre 2007. *Wasser Energie Luft* **2**, 115–123.

Hilker, N., Badoux, A. and Hegg, C. (2009) The Swiss flood and landslide damage database 1972–2007. *Natural Hazards and Earth System Sciences* **9**, 913–925.

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (eds) (2010) *Bayesian Nonparametrics*. Cambridge: Cambridge University Press.

Hooker, G. and Vidyashankar, A. N. (2014) Bayesian model robustness via disparities. *TEST* **23**, 556–584.

Hsing, T., Hüsler, J. and Leadbetter, M. R. (1988) On the exceedance point process for stationary sequences. *Probability Theory and Related Fields* **78**, 97–112.

Huser, R. and Davison, A. C. (2014) Space-time modelling of extreme events. *Journal of the Royal Statistical Society Series B* **76**, 439–461.

Hüsler, J. and Reiss, R.-D. (1989) Maxima of normal random vectors: between independence and complete dependence. *Statistics and Probability Letters* **7**, 283–286.

Hyndman, R. J. (1996) Computing and graphing highest density regions. *The American Statistician* **50**, 120–126.

Ishwaran, H. and James, L. F. (2001) Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Ishwaran, H. and James, L. F. (2002) Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* **11**, 508–532.

Ishwaran, H. and James, L. F. (2003) Some further developments for stick-breaking priors: finite and infinite clustering and classification. *The Indian Journal of Statistics* **65**, 577–592.

Ishwaran, H. and Zarepour, M. (2000) Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87**, 371–390.

Joe, H. (2014) *Dependence Modeling with Copulas*. Boca Raton: Chapman & Hall/CRC Press.

Joe, H., Smith, R. L. and Weissman, I. (1992) Bivariate threshold methods for extremes. *Journal of the Royal Statistical Society Series B* **54**, 171–183.

Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006) Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association* **52**, 93–100.

Kass, R. E., Carlin, B. P., Gelman, A. and Neal, R. M. (1998) Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician* **52**, 93–100.

Katz, R. W., Parlange, M. B. and Naveau, P. (2002) Statistics of extremes in hydrology. *Advances in Water Resources* **25**, 1287–1304.

Keef, C., Papastathopoulos, I. and Tawn, J. A. (2013) 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., Svensson, C. and Tawn, J. A. (2009a) Spatial dependence in extreme river flows and precipitation for Great Britain. *Journal of Hydrology* **378**, 240–252.

Keef, C., Tawn, J. A. and Svensson, C. (2009b) Spatial risk assessment for extreme river flows. *Journal of the Royal Statistical Society Series C* **58**, 601–618.

Kish, L. (1965) *Survey Sampling*. New York: John Wiley & Sons.

Kotz, S. and Nadarajah, S. (2000) *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press.

Kratz, M. F. and Rootzén, H. (1997) On the rate of convergence for extremes of mean square differentiable stationary normal processes. *Journal of Applied Probability* **34**, 908–923.

Kulik, R. and Soulier, P. (2015) Heavy tailed time series with extremal independence. *Extremes* **18**, 273–299.

Leadbetter, M. R. (1974) On extreme values in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **28**, 289–303.

Leadbetter, M. R. (1976) Weak convergence of high level exceedances by a stationary sequence. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **34**, 11–15.

Leadbetter, M. R. (1983) Extremes and local dependence in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **65**, 291–306.

Leadbetter, M. R. (1995) On high level exceedance modeling and tail inference. *Journal of Statistical Planning and Inference* **45**, 247–260.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983) *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer.

Leadbetter, M. R., Weissman, I., de Haan, L. and Rootzén, H. (1989) On clustering of high values in statistically stationary series. In *Proceedings of the 4th International Meeting on Statistical Climatology*, pp. 217–222. Wellington: New Zealand Meteorological Service.

Ledford, A. W. and Tawn, J. A. (1996) Statistics for near independence in multivariate extreme values. *Biometrika* **83**, 169–187.

Ledford, A. W. and Tawn, J. A. (1997) Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society Series B* **59**, 475–499.

Ledford, A. W. and Tawn, J. A. (2003) Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society Series B* **65**, 521–543.

Lehmann, E. L. (1966) Some concepts of dependence. *The Annals of Mathematical Statistics* **37**, 1137–1153.

Liu, Y. and Tawn, J. A. (2014) Self-consistent estimation of conditional multivariate extreme value distributions. *Journal of Multivariate Analysis* **127**, 19–35.

Lugrin, T., Davison, A. C. and Tawn, J. A. (2016) Bayesian uncertainty management in temporal dependence of extremes. *Extremes* **19**, 491–515.

MacEachern, S. N. (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics* **23**, 727–741.

MacEachern, S. N. and Müller, P. (1998) Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error. *The Annals of Statistics* **20**, 712–736.

Marshall, A. W. and Olkin, I. (1983) Domains of attraction of multivariate extreme value distributions. *The Annals of Probability* **11**, 168–177.

von Mises, R. (1936) La distribution de la plus grande de *n* valeurs. *Revue Mathématique de l'Union Interbalkanique* **1**, 141–160. (Reproduced in Selected Papers of Richard von Mises (1964), *American Mathematical Society*, Providence, **2**, 271–294).

Mitra, A. and Resnick, S. I. (2013) Modeling multiple risks: hidden domain of attraction. *Extremes* **16**, 507–538.

Morgenstern, D. (1956) Einfache Beispiele zweidimensionaler Verteilungen. *Mitteilungsblatt für Mathematische Statistik* **8**, 234–235.

Müller, P., Erkanli, A. and West, M. (1996) Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

Neal, R. M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

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* **66**, 93–120.

O'Brien, G. L. (1987) Extreme values for stationary and Markov sequences. *The Annals of Probability* **15**, 281–291.

Oesting, M., Schlather, M. and Friederichs, P. (2017) Statistical post-processing of forecasts for extremes using bivariate Brown–Resnick processes with an application to wind gusts. *Extremes* **20**, 309–332.

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M. and Mollison, D. (2000) Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series C* **49**, 517–542.

Papaspiliopoulos, O. (2008) A note on posterior sampling from Dirichlet mixture models. Technical report, Coventry.

Papaspiliopoulos, O. and Roberts, G. O. (2008) Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169–186.

Papastathopoulos, I. (2016) Conditional independence and conditioned limit laws. *Statistics and Probability Letters* **112**, 1–4.

Papastathopoulos, I., Strokorb, K., Tawn, J. A. and Butler, A. (2017) Extreme events of Markov chains. *Advances in Applied Probability* **49**, 134–161.

Papastathopoulos, I. and Tawn, J. A. (2016) Conditioned limit laws for inverted max-stable processes. *Journal of Multivariate Analysis* **150**, 214–228.

Peng, L. and Qi, Y. (2004) Discussion of "A conditional approach for multivariate extreme values", by J. E. Heffernan and J. A. Tawn. *Journal of the Royal Statistical Society Series B* **66**, 541–542.

Pfister, C. (2009) Die Katastrophenlücke des 20. Jahrhunderts und der Verlust traditionalen RisikobewusstseinsThe. *GAIA-Ecological Perspectives for Science and Society* **18**, 239–246.

Philippe, A. (1997) Processing simulation output by Riemann sums. *Journal of Statistical Computation and Simulation* **59**, 295–314.

Pickands, J. (1975) Statistical inference using extreme order statistics. *The Annals of Statistics* **3**, 119–131.

Pickands, J. (1981) Multivariate extreme value distributions. *Bulletin of the International Statistical Institute* **49**, 859–878.

Pickands, J. (1986) The continuous and differentiable domains of attraction of the extreme-value distributions. *The Annals of Probability* **14**, 996–1004.

Piessens, R., Doncker-Kapenga, E. D., Überhuber, C. W. and Kahaner, D. K. (1983) *QUADPACK: A Subroutine Package for Automatic Integration.* Berlin: Springer.

Pitman, J. and Yor, M. (1997) The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**, 855–900.

Poon, S.-H., Rockinger, M. and Tawn, J. A. (2003) Modelling extreme-value dependence in international stock market. *Statistica Sinica* **13**, 929–953.

Porteous, I., Ihler, A., Smyth, P. and Welling, M. (2006) Gibbs sampling for (coupled) infinite mixture models in the stick-breaking representation. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, pp. 385–392. Arlington, VA: AUAI Press.

R Core Team (2017) *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Reich, B. J., Shaby, B. A. and Cooley, D. (2014) A hierarchical model for serially-dependent extremes: a study of heat waves in the Western US. *Journal of Agricultural, Biological, and Environmental Statistics* **19**, 119–135.

Resnick, S. I. (1987) *Extreme Values, Regular Variation and Point Processes*. New York: Springer.

Resnick, S. I. (2002) Hidden regular variation, second order regular variation and asymptotic independence. *Extremes* **5**, 303–336.

Resnick, S. I. (2007) *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. New York: Springer.

Ribatet, M., Ouarda, T. B. M. J., Sauquet, E. and Gresillon, J.-M. (2009) Modeling all exceedances over a high threshold using an extremal dependence structure: inferences on several flood characteristics. *Water Resources Research* **45**.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society Series B* **57**, 731–792.

Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods*. Second edition. New York: Springer.

Robert, C. Y. (2013) Automatic declustering of rare events. *Biometrika* **100**, 587–606.

Roberts, G. O. and Rosenthal, J. S. (2009) Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* **18**, 349–367.

Rohrbeck, C., Eastoe, E. F., Frigessi, A. and Tawn, J. A. (2018) Extreme value modelling of water-related insurance claims. *The Annals of Applied Statistics* In print.

Rootzén, H. and Zholud, D. (2017) Human life is unlimited – but short. *Extremes* **20**, 713–728.

Rosenthal, J. S. (2011) Optimal Proposal Distributions and Adaptive MCMC. In *MCMC Handbook*, eds S. Brooks, A. Gelman, G. Jones and X.-L. Meng, pp. 93–111. Chapman & Hall/CRC Press.

Sang, H. and Gelfand, A. E. (2009) Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics* **16**, 407–426.

Segers, J. (2003) Functionals of clusters of extremes. *Advances in Applied Probability* **35**, 1028–1045.

Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Sharkey, P. and Tawn, J. A. (2017) A Poisson process reparameterisation for Bayesian inference for extremes. *Extremes* .

Shemyakin, A. (2014) Hellinger distance and non-informative priors. *Bayesian Analysis* **9**, 923–938.

Sibuya, M. (1960) Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics* **11**, 195–210.

Sisson, S. A. and Coles, S. G. (2003) Modelling dependence uncertainty in the extremes of Markov chains. *Extremes* **6**, 283–300.

Sklar, A. (1959) Fonctions de répartition à *n* dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* **8**, 229–231.

Smith, R. L. (1987) Approximations in extreme value theory. Technical report, University of North Carolina.

Smith, R. L. (1989) Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science* **4**, 367–393.

Smith, R. L. (1990) Max-stable processes and spatial extremes. Unpublished manuscript.

Smith, R. L., Tawn, J. A. and Coles, S. G. (1997) Markov chain models for threshold exceedances. *Biometrika* **84**, 249–268.

Smith, R. L., Tawn, J. A. and Yuen, H. K. (1990) Statistics of multivariate extremes. *International Statistical Review* **58**, 47–58.

Smith, R. L. and Weissman, I. (1994) Estimating the extremal index. *Journal of the Royal Statistical Society Series B* **56**, 515–528.

Spicher, B. (2017) Nationale Plattform Naturgefahren PLANAT. `http://www.planat.ch/de/wissen/chronik`. Accessed: 2017-04-20.

Stephenson, A. and Tawn, J. A. (2005) Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika* **92**, 213–227.

Süveges, M. (2007) Likelihood estimation of the extremal index. *Extremes* **10**, 41–55.

Süveges, M. and Davison, A. C. (2012) A case study of a "Dragon-King": the 1999 Venezuelan catastrophe. *The European Physical Journal Special Topics* **205**, 131–146.

Swiss Federal Institute for Forest, Snow and Landscape Research WSL (2016) Swiss flood and landslide damage database. `https://www.bafu.admin.ch/bafu/de/home/themen/naturgefahren/fachinformationen/schaeden-und-lehren-aus-naturereignissen/schaeden-durch-naturgefahren-seit-1972.html`. Accessed: 10-01-2018.

Tawn, J. A. (1988) Bivariate extreme value theory: models and estimation. *Biometrika* **75**, 397–415.

Tiago de Oliveira, J. (1980) Bivariate extremes; foundations and statistics. In *Multivariate Analysis V*, ed. P. R. Krishnaiah, pp. 349–366. Amsterdam: North-Holland.

de Valk, C. (2016) Approximation and estimation of very small probabilities of multivariate extreme events. *Extremes* **19**, 687–717.

Wadsworth, J. L. and Tawn, J. A. (2012) Dependence modelling for spatial extremes. *Biometrika* **99**, 253–272.

Wadsworth, J. L. and Tawn, J. A. (2013) A new representation for multivariate tail probabilities. *Bernoulli* **19**, 2689–2714.

Walker, S. G. (2007) Sampling the Dirichlet mixture model with slices. *Communications in Statistics* **36**, 45–54.

Winter, H. C. (2015) *Extreme Value Modelling of Heatwaves*. Ph.D. thesis, Lancaster University.

Winter, H. C. and Tawn, J. A. (2016) Modelling heatwaves in central France: a case-study in extremal dependence. *Journal of the Royal Statistical Society Series C* **65**, 345–365.

Winter, H. C. and Tawn, J. A. (2017) kth-order Markov extremal models for assessing heatwave risks. *Extremes* **20**, 393–415.

Wu, Y. and Hooker, G. (2013) Hellinger distance and Bayesian non-parametrics: hierarchical models for robust and efficient Bayesian inference. Unpublished manuscript.

Yakowitz, S., Krimmel, J. E. and Szidarovszky, F. (1978) Weighted Monte Carlo integration. *SIAM Journal on Numerical Analysis* **15**, 1289–1300.

Yun, S. (2000) The distribution of cluster functionals of extreme events in a dth-order Markov chain. *Journal of Applied Probability* **37**, 29–44.

Thomas Lugrin

thomas.lugrin@alumni.epfl.ch
www.thomaslugrin.com
ch.linkedin.com/in/thomaslugrin/

Education

2013-2018 **PhD candidate in Mathematics, EPFL & Lancaster University, UK**

2011-2013 **Master in Mathematical Engineering, EPFL**
Advanced courses in Statistics and Probability, Biostatistics, Numerical Analysis, Financial Mathematics; Master's thesis in Lancaster University.

2008-2011 **Bachelor in Mathematics, EPFL**

2005-2008 **Maturité (secondary diploma), Gymnase de Burier, La Tour-de-Peilz**
Enhanced level of mathematics; Latin orientation.

Conferences

Sept. 2017 **STOR-i Forum, Lancaster, UK *(Talk)***
"Modelling time series extremes using the conditional tail approach."

Feb. 2017 **Extreme Value Analysis, Delft, Netherlands *(Talk)***
"Modelling extremes of Markov chains."

Feb. 2017 **EPFL Extremes Workshop, Lausanne *(Talk)***
"Modelling extremes of Markov chains."

July 2016 **STOR-i Extremes Workshop, Lancaster, UK *(Talk)***
"A new approach to multivariate extremes." *(cancelled)*

June 2016 **International Society for Bayesian Analysis, Forte Village, Italy *(Invited talk)***
"Uncertainty in time series extremes modelling." *(cancelled)*

July 2015 **60th World Statistics Congress (ISI), Rio, Brazil *(Invited talk)***
"Modelling the clustering of extreme river flows for hydrological risk assessment."

June 2015 **Extreme Value Analysis, Ann Arbor, MI, USA *(Poster)***
"Uncertainty in extremal dependence."

Aug. 2014 **21st International Conference on Computational Statistics, Geneva *(Talk)***
"Bayesian semiparametrics for modelling the clustering of extreme values."

Aug. 2014 **Joint Statistical Meetings, Boston, MA, USA *(Invited talk)***
"Bayesian semiparametrics for modelling the clustering of extreme values."

July 2014 **High-dimensional and Multivariate Extremes, Bristol, UK *(Poster)***
"Bayesian semiparametrics for modelling the clustering of extreme values."

June 2014 **Young Researchers' Conference, Geneva *(Talk)***
"Introduction to Bayesian statistics."

May 2014 **Extremes Reading Group, Skipton, UK *(Talk)***
"Automatic declustering of rare events."

Computing

Scientific R, Matlab
Programming C++ (incl. Qt), C, MySQL, VBScript, Python

Office Office suite, LaTeX
Web HTML, CSS, PHP

Languages

French (mother tongue), **English** (very good), **German** (good), **Albanian** (intermediate)