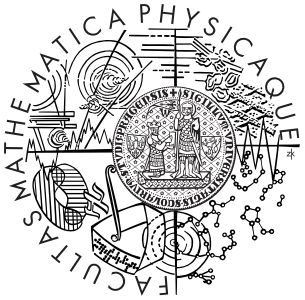
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Kristýna Ivanková

Vícerozměrné extrémy

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Daniel Hlubinka, Ph.D.

Studijní program: matematika

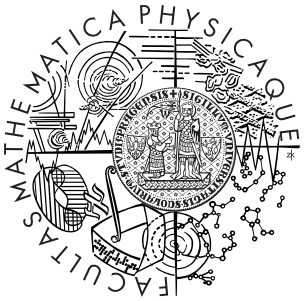
Studijní obor: pravděpodobnost, matematická statistika a ekonometrie

Studijní plán: ekonometrie

2008

Charles University in Prague
Faculty of Mathematics and Physics

DIPLOMA THESIS



Kristýna Ivanková

Multivariate Extremes

Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Daniel Hlubinka, Ph.D.

Programme of study: Mathematics

Field of study: Probability, Mathematical Statistics and Econometrics

Plan of study: Econometrics

2008

Děkuji doc. RNDr. Danielu Hlubinkovi, Ph.D., svému vedoucímu práce, za zadání zajímavého tématu, poskytnutí dat a vedení k samostatnosti. Děkuji Chrisovi Ferrovi a Alecu Stephensonovi za konzultaci ke knihovně `evd` a profesoru Sidney Resnickovi za zprostředkování kontaktu na paní Heffernanovou.

Děkuji ♡ Honzíkovi za korektury angličtiny a hlavně za jeho trpělivost a podporu. ^--^

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 12. prosince 2008 Kristýna Ivanková

I hereby thank my supervisor doc. RNDr. Daniel Hlubinka, Ph.D. for the assignment of an interesting topic, for supplying the dataset and for his guidance towards independence. I thank Chris Ferro and Alec Stephenson for a consultation on the library `evd` and Professor Sidney Resnick for negotiating contact with Dr. Heffernan.

I declare that I've written my diploma thesis on my own and with the exclusive use of cited sources. I give consent to lending the work out and its public exposure.

In Prague, December 12, 2008 Kristýna Ivanková

Contents

1 Introduction **8**

2 One-dimensional maxima **10**

2.1 Limit laws for maxima 11

2.2 Maximum domains of attraction 12

2.3 The GEV distribution and the GPD 13

3 Classical multivariate methods **17**

3.1 Multivariate maxima 18

3.2 Point processes 19

3.3 Dependence structure 21

3.4 Modelling of the measure H 22

3.5 Models 23

3.5.1 Dirichlet model 23

3.5.2 Logistic model 24

3.5.3 Asymmetric logistic model 25

3.5.4 Bilogistic model 25

3.5.5 Negative logistic models 26

3.5.6 Polynomial Pickands dependence functions 26

3.6 Estimation 27

3.6.1 Point process method 27

3.6.2 The censored-likelihood method 29

4 A conditional approach for multivariate extreme values 31

4.1 Marginal fitting 31

4.2 Dependence structure modelling 32

4.3 Estimation 33

4.4 Sampling 34

5 Simulation study 36

5.1 Univariate Peaks Over Threshold 36

5.2 Bivariate censored likelihood 38

5.3 The conditional approach of Heffernan and Tawn 41

5.4 Discussion 42

6 Application 43

6.1 Marginal estimation 44

6.2 Dependence structure estimation 46

6.2.1 Classical methods 46

6.2.2 The conditional approach 48

6.3 Model comparison 50

6.4 Implementation details 51

7 Conclusion 56

Bibliography 57

Appendix: Proof of Fisher-Tippett theorem 59

Název práce: Vícerozměrné extrémy
Autor: Kristýna Ivanková
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: doc. RNDr. Daniel Hlubinka, Ph.D.
e-mail vedoucího: hlubinka@karlin.mff.cuni.cz

Abstrakt: Tato práce je věnována různým přístupům k modelování vícerozměrných extremálních událostí. Nejdříve shrneme teorii pro jednorozměrná data – Fisherovu-Tippettovu větu a zobecněné Paretovo rozdělení.
Dále se zabýváme rozšířením této teorie do více rozměrů. K vyjádření závislosti mezi složkami odpovídajícími jednotlivým rozměrům zavedeme spektrální míru a bodové procesy. Závislostní strukturu popíšeme několika parametrickými modely, pro něž uvedeme způsoby odhadu parametrů.
Výše zmíněnou klasickou metodiku srovnáme s novým semiparametrickým podmíněným přístupem, který nám umožňuje modelovat závislostní strukturu i mimo sdružený chvost rozdělení. Oblasti použití obou postupů ověříme simulační studií a aplikací na reálná data.

Klíčová slova: zobecněné Paretovo rozdělení, vícerozměrné extrémy, metoda špiček nad prahem, Poissonův proces, extremální závislostní struktura, podmíněné rozdělení

Title: Multivariate Extremes
Author: Kristýna Ivanková
Department: Department of Probability and Mathematical Statistics
Supervisor: doc. RNDr. Daniel Hlubinka, Ph.D.
Supervisor's e-mail address: hlubinka@karlin.mff.cuni.cz

Abstract: This work considers various approaches for modelling multivariate extremal events. First we review theory in the univariate case—the Fisher-Tippett theorem and the generalized Pareto distribution.
We proceed with an extension to the multivariate case using the spectral measure and point processes for modelling dependence between components, ending with a review of parametric dependence models and ways to fit them to data.
We compare these classical methods to a new semi-parametric conditional approach. Finally, we apply the discussed methods in a simulation and on a dataset, compare the results and highlight classes of problems that the various approaches are suitable to.

Keywords: GPD, multivariate extremes, Peaks Over Threshold, Poisson process, extremal dependence structure, conditional distribution

Chapter 1

Introduction

Extreme value theory is an important, evolving field of statistics. Its origin can be traced to the publication of Fisher-Tippett theorem for limit laws in 1928, but its application to modelling genuine physical phenomena became first possible as late as in the 1950's. Properties of sample extremes were developed into a coherent theory in the 1970's.

It should be noted that standard methods often underestimate tails (since one of the standard generalizations is the assumption of normality) and can treat genuine extremal values as outliers, which justifies the need for specialized methods.

Extreme values are, by definition, rare, and there is often a need to estimate much higher values than previously observed. Therefore, extrapolation is crucial for extremal event modelling. Extreme value theory provides us with limit models, comparable in form to central limit theory, that make such extrapolation arguments tenable. These models are based on asymptotic assumptions, but they were established on rational foundations and methods derived from them have proven their worth in many fields, including insurance business, financial markets, hydrology, meteorology and ecology.

Given such a short history of the field, it's not surprising that multivariate extreme value theory is still in the development stage.

This work strives to compare multivariate parametric dependence structure models [3], [10], [11] with a semi-parametric approach of Heffernan and Tawn [5], discuss their applications and assess their utilization for modelling extremes.

In Chapter 2, we develop limit laws for maxima of iid random variables (the Fisher-Tippett theorem, proved in the Appendix), obtaining three classes of extreme value distributions and joining them into the GEV distribution family. From there we derive the Generalized Pareto distribution (GPD), introduce the excess distribution function and approximate its tail with the GPD tail, which forms the basis of the Peaks Over Threshold method.

We start Chapter 3 with the extension of maxima to d dimensions and derive the classical approach dealing with marginal and dependence structure estimation with the use of Poisson processes and the GPD. We introduce various parametric models for the dependence structure and outline procedures how to create new ones. In closing, we present the Censored likelihood, an approach that solves some shortcomings of Poisson process-based estimation.

The classical methods estimate tails of all components using GPD fitting, so their models are valid only in the joint tail. In Chapter 4 we review a semi-parametric approach that creates d separate models, each modelling the $(d-1)$-dimensional conditional distribution given that the last component is extreme. These models combined can describe the whole area that is extreme in at least one component. We describe model fitting for this procedure and discuss its advanages and shortcomings.

Chapter 5 is a simulation study: using random samples from known distributions, we explore the behaviour of previously discussed methods in practice. Finally, in Chapter 6 we take approaches from Chapters 3 and 4 and apply them on a homogenous dataset of newborn children.

Chapter 2

One-dimensional maxima

Let us start by reviewing the theory for one-dimensional extremes. We follow [8], the proof of Fisher-Tippett theorem is taken from [4].

Consider $X, X_1, X_2, \ldots$ to be a sequence of iid non-degenerate random variables, their distribution function denoted F. We shall study the behaviour of their maxima

$$M_1 = X_1, \;\; M_n = \max_i\{X_i\}, \;\; 1 \leq i \leq n, \; n \geq 2.$$

The distribution function of M_n is

$$P(M_n \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) = F^n(x), \;\; x \in \mathbb{R}, \;\; n \in \mathbb{N}.$$

Observe that when discussing extremes, we can restrict ourselves only on maxima since $\min f(x) = -\max -f(x)$.

We'll refer to $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$ as the *right endpoint of F*. It can be easily shown that the sequence (M_n) converges almost surely to it:

$$M_n \overset{a.s.}{\to} x_F \text{ as } n \to \infty.$$

In the next section we'll see that under an affine transformation there exist limit distributions for maxima (similar to the CLT for summation). We'll need a few definitions:

We'll write

$$X \overset{d}{=} Y$$

if two random variables X, Y have the same distribution.

X and Y will *belong* or *be of the same type* if they are only an affine transformation of each other, i.e. there exist constants $a \in \mathbb{R}$ and $b > 0$ such that

$$X \stackrel{d}{=} bY + a. \tag{2.1}$$

A non-degenerate random variable X will be called *max-stable* if it satisfies the identity

$$\max(X_1, \ldots, X_n) \stackrel{d}{=} c_n X + d_n, \quad n \geq 2 \tag{2.2}$$

for iid $X, X_1, \ldots, X_n$ and appropriate constants $c_n > 0$, $d_n \in \mathbb{R}$.

We can rewrite the previous identity by using an affinely transformed maximum M_n as

$$\frac{1}{c_n}(M_n - d_n) \stackrel{d}{=} X. \tag{2.3}$$

We'll use equivalent definitions of *belonging to the same type* and *max-stability* also for distributions and distribution functions corresponding to random variables satisfying these definitions.

We'll refer to the centring constants d_n and the scaling constants c_n jointly as *normalising constants*.

2.1 Limit laws for maxima

The Fisher-Tippett theorem specifies the form of the limit distribution for normalised maxima. The following supplemental theorem (Theorem 3.2.2 from [8]) specifies their only possible limit laws:

Theorem 1
The class of max-stable distributions coincides with the class of all possible non-degenerate limit laws for properly normalised maxima of iid random variables.

With the help of Theorem 1, the Fisher-Tippett theorem (Theorem 3.2.3 from [8]) needs to deal only with three families of max-stable distributions.

Theorem 2 — Fisher-Tippett theorem: Limit laws for maxima
Let (X_n) be a sequence of iid random variables. If there exist normalising

constants $c_n > 0$, $d_n \in \mathbb{R}$ and some non-degenerate distribution function G such that

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} G, \tag{2.4}$$

then G is an extreme value distibution, i.e. belongs to the same type as one of the following distribution functions (called the standard extreme value distributions):

$$\text{Weibull:} \quad \Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^\alpha\}, & x \le 0, \\ 1, & x > 0, \end{cases} \quad \alpha > 0.$$

$$\text{Gumbel:} \quad \Lambda(x) = \exp\{-e^{-x}\}, \quad x \in \mathbb{R}.$$

$$\text{Fréchet:} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \le 0, \\ \exp\{-x^{-\alpha}\}, & x > 0, \end{cases} \quad \alpha > 0.$$

The proof is given in Appendix.

Random variables having a (standard) extreme value distribution are called (*standard*) *extremal random variables*.

Suppose that $X > 0$. Then the correspondence between the three distributions is

$$X \text{ has df } \Phi_\alpha \quad \Leftrightarrow \quad \log X^\alpha \text{ has df } \Lambda \quad \Leftrightarrow \quad -X^{-1} \text{ has df } \Psi_\alpha.$$

According to Theorem 1 we can see that for an extremal random variable X relation (2.3) must be satisfied. The normalizing constants from (2.2) can then be written as $M_n \stackrel{d}{=} n^{-1/\alpha}X$ for the Weibull case, $M_n \stackrel{d}{=} n^{1/\alpha}X$ for the Fréchet case and $M_n \stackrel{d}{=} X + \log n$ for the Gumbel case.

2.2 Maximum domains of attraction

Having the three limit classes defined, it's often necessary for a given random variable X to determine which of the extreme value distributions is the limit for its maxima.

We say that the random variable X belongs to the *maximum domain of attraction of the extreme value distribution G* if there exist constants $c_n > 0$, $d_n \in \mathbb{R}$ such that the Fisher-Tippett prerequisites (2.4) hold. We write $X \in \text{MDA}(G)$.

We can again state an equivalent definition for its distribution or its distribution function.

Because extreme value distribution functions are continuous on $\mathbb{R}$, (2.4) is equivalent to

$$\lim_{n\to\infty} P(M_n \leq c_n x + d_n) = \lim_{n\to\infty} F^n(c_n x + d_n) = G(x), \quad x \in \mathbb{R}.$$

We denote $u_n = u_n(x) = c_n x + d_n$ and the *tail of distribution function* F by $\overline{F} = 1 - F$.

For given $\tau \in [0, \infty]$ and a sequence (u_n) of real numbers, the following holds:

$$n\overline{F}(u_n) \to \tau \quad \Longleftrightarrow \quad P(M_n \leq u_n) \to e^{-\tau}. \tag{2.5}$$

This is called the *Poisson approximation* in literature. The next theorem (Proposition 3.3.2 from [8]) is its direct consequence. We'll need it in the proof of the Fisher-Tippett theorem.

Theorem 3 — Characterisation of MDA(G)
The distribution function F belongs to the maximum domain of attraction of the extreme value distribution G with normalising constants $c_n > 0$, $d_n \in \mathbb{R}$ if and only if

$$\lim_{n\to\infty} n\overline{F}(c_n x + d_n) = -\log G(x), \quad x \in \mathbb{R}.$$

When $G(x) = 0$, the limit is interpreted as ∞.

For the complete characterization of MDAs for each of the three extreme value distribution classes we direct the interested reader to chapter 3 of [8].

2.3 The GEV distribution and the GPD

The usual approach in recent publications is to define a single-parameter family which encompasses all standard extreme value distribution types.

We'll denote the *standard generalised extreme value (GEV) distribution* by G_ξ. Its parameter $\xi \in \mathbb{R}$ spans all of the three standard EVDs:

$$G_\xi = \begin{cases} \Psi_{-1/\xi} & \text{if } \xi < 0, \\ \Lambda & \text{if } \xi = 0, \\ \Phi_{1/\xi} & \text{if } \xi > 0. \end{cases}$$

The parametrization of the distribution function G_ξ has the form

$$G_\xi(x) = \begin{cases} \exp\left\{-(1+\xi x)^{-1/\xi}\right\} & \text{if } \xi \neq 0,\ 1+\xi x > 0, \\ \exp\left\{-\exp\{-x\}\right\} & \text{if } \xi = 0. \end{cases}$$

We can treat G_0 to be the limit of G_ξ for $\xi \to 0$. In that case, we can simplify and write

$$G_\xi(x) = \exp\left\{-(1+\xi x)^{-1/\xi}\right\}, \quad 1+\xi x > 0,$$

for all $\xi \in \mathbb{R}$.

We can extend the parametric space with location $\mu \in \mathbb{R}$ and scale $\psi > 0$. We get such a family $G_{\xi;\mu,\psi}$ by replacing x with $(x-\mu)/\psi$. This extended family is also called GEV.

Let X be a random variable with distribution function F. It will be useful to define the *excess distribution function of* F over the fixed threshold u as

$$F_u(x) = P(X - u \leq x | X > u) = \frac{F(x+u) - F(u)}{\overline{F}(u)}, \quad x \geq 0, \quad u < x_F.$$

The function $e(u) = E(X - u | X > u)$ is called the *mean excess function* of X.

Let now $F \in \mathrm{MDA}(G_\xi)$. Then there exists a positive measurable function a such that

$$\lim_{u \to {x_F}^-} P\left(\frac{X-u}{a(u)} > x \middle| X > u\right) = \begin{cases} (1+\xi x)^{-1/\xi} & \text{if } \xi \neq 0, \\ e^{-x} & \text{if } \xi = 0, \end{cases} \tag{2.6}$$

for $1 + \xi x > 0$.

Equation (2.6) motivates the introduction of the *standard generalised Pareto distribution (GPD)* having the distribution

$$H_\xi(x) = \begin{cases} 1 - (1+\xi x)^{-1/\xi}, & \text{if } \xi \neq 0, \\ 1 - e^{-x}, & \text{if } \xi = 0, \end{cases}$$

for $x \geq 0$ if $\xi \geq 0$ and $0 \leq 1 + \xi x \leq 1$ if $\xi < 0$.

The GPD describes limit distributions of scaled excesses over a high threshold.

We can define the related location-scale family $H_{\xi;\nu,\beta}$ in an analogous way as in the GEV. We also refer to $H_{\xi;\nu,\beta}$ as GPD. And finally, we can also consider H_0 to be the limit of H_ξ as $\xi \to 0$ and simplify the notation.

Properties of the GPD

Pickands-Balkema-de Haan Theorem [12]: For every $\xi \in \mathbb{R}$, $F \in \text{MDA}(G_\xi)$ if and only if there exists a positive measurable function β so that

$$\lim_{u \to x_F^-} \sup_{0<x<x_F-u} \left| F_u(x) - H_{\xi;0,\beta(u)}(x) \right| = 0.$$

From this we can conclude that (for sufficiently large u) the GPD can approximate the excess distribution function F_u. For some function β to be estimated from the data, we can use

$$\overline{F}_u(x) = P(X - u > x | X > u) \approx \overline{H}_{\xi;0,\beta(u)}(x), \quad x > 0.$$

or alternatively

$$P(X > x | X > u) \approx \overline{H}_{\xi;u,\beta(u)}(x) \quad \text{for } x > u. \tag{2.7}$$

Approximation of the excess distribution function with the GPD has been coined the term "peaks over threshold (POT) method" by hydrologists. We will use it extensively in subsequent chapters.

From Poisson approximation (2.5) we know that the number of exceedances of a high threshold roughly follows the Poisson distribution. Moreover, it can be shown that the number of exceedances and the excesses are independent in an asymptotic sense.

Assume that N is Poi(λ), independent of the iid sequence (X_n) with distribution $H_{\xi;0,\beta}$. Then

$$P(M_N \leq x) = \exp\left(-\lambda \left(1 + \xi \frac{x}{\beta} \right)^{-1/\xi} \right) = G_{\xi;\mu,\psi}(x),$$

where $\mu = \beta\xi^{-1}(\lambda^\xi - 1)$ and $\psi = \beta\lambda^\xi$.

When estimating the distribution tail using the GPD, one needs to select the threshold u [2]. It has to be large enough for the asymptotic assumptions to work, but we also need to include as much data as possible to decrease variance. From these contradictory conditions it follows that the selection of u is highly subjective.

A widely-used method uses the fact that the mean excess function behaves linearly in u in regions that can be approximated with a GPD:

$$e(u) = E(X - u | X > u) = \frac{\beta + \xi u}{1 - \xi}. \tag{2.8}$$

We can pick the lowest threshold that satisfies the linearity assumption.

Another method uses the fact that a GPD fit over a threshold should be equally valid for a higher threshold. It creates many models for various thresholds and picks the lowest one whose parameters remain approximately stationary for higher thresholds.

We should mention that there exists an another commonly-used approach for extremal statistics called Block Maxima. This approach fits maxima over a few related samples (e.g. concurrent in a time series) with a GEV distribution. The subjective part of this method is choosing the block size (equivalent to choosing the threshold in POT). One of the drawbacks of the method is that it doesn't use the available data very effectively (the block size needs to be large for limit assumptions to hold and extreme events in one block cancel out).

Chapter 3

Classical multivariate methods

We'll describe the parametric approach of Coles and Tawn [3] and Tawn [10], [11] extending the univariate POT method. A comprehensive study of this approach can be found in [1], chapters 8 and 9. Non-parametric methods also exist, but there is seldom enough data to support their use in practice. For further references, see [2], p. 167.

Marginal and dependence structure are usually estimated separately. We provide an introduction into point process theory, which allows us to estimate both structures jointly.

As opposed to marginals, the dependence structure doesn't admit a finite-dimensional parametrization. Usually a well-chosen, simplified submodel is used in practice. We'll introduce some of these models and describe their behaviour.

For the estimation of parameters we'll use the maximum likelihood method. This method has several advantages: every relevant information can be included and put into relation with the rest, it adapts easily to model changes and estimation is simple and fast. The uncertainity of estimates is often a by-product of the method or can be computed quickly when needed.

The Block Maxima approach mentioned at the end of the previous chapter won't be discussed for multivariate data: since we define vector maxima to be component-wise, the input of the method often doesn't correspond to any observations.

3.1 Multivariate maxima

Extension from one dimensional to d-dimensional maxima is non-trivial. We must find an appropriate relation for ordering multivariate observations. Let us introduce a marginal ordering: for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ the relation $\boldsymbol{x} \leq \boldsymbol{y}$ means that $x_j \leq y_j$ for all $1 \leq j \leq d$. We'll use the convention that vector operations are applied componentwise.

Consider an independent sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of d-dimensional random vectors from a d-variate distribution function F. The sample maximum $\boldsymbol{M}_n$ is defined as a vector of component-wise maxima, i.e. $M_{n,j} = \max(X_{1,j}, \ldots, X_{n,j})$, $1 \leq j \leq d$. Observe that this maximum doesn't have to match any of the observations.

The following computational approach for multivariate maxima is similar to the univariate case.

The distribution function of $\boldsymbol{M}_n$ can be written as

$$P(\boldsymbol{M}_n \leq \boldsymbol{x}) = P(\boldsymbol{X}_1 \leq \boldsymbol{x}, \ldots, \boldsymbol{X}_n \leq \boldsymbol{x}) = F^n(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d.$$

Let $\boldsymbol{c}_n$ and $\boldsymbol{d}_n$ be sequences of vectors, $\boldsymbol{c}_n > \boldsymbol{0}$, and G a d-variate distribution function with non-degenerate margins G_j such that

$$\boldsymbol{c}_n^{-1}(\boldsymbol{M}_n - \boldsymbol{d}_n) \xrightarrow{d} G, \quad n \to \infty. \tag{3.1}$$

Then we call G a *multivariate extreme value distribution function* and F belongs to the maximum domain of attraction of G. Margins G_j, $1 \leq j \leq d$, are univariate extreme value distribution functions and the corresponding margins F_j of F are in their maximum domains of attraction — if (3.1) holds, all marginal sequences converge in distribution and this convergence is characterized by the Fisher-Tippett theorem. Also, like in the univariate case, the distribution function G is max-stable.

It will be advantageous to transform marginal components to unit Fréchet margins with distribution function $\Phi_1(z) = e^{-1/z}$, $z > 0$. This allows us to reduce notational complexity and enables us to treat marginal and dependence structures separately. We shall indicate that we work in unit Fréchet margins by writing coordinates as $\boldsymbol{z}$ instead of $\boldsymbol{x}$.

The following result ([14], Proposition 5.15) is the multi-dimensional extension of Theorem 3: A distribution function $F(\boldsymbol{x})$ belongs to the MDA of a multivariate extreme value distribution if and only if

$$\lim_{n\to\infty} \frac{\log F_*(n\boldsymbol{z})}{\log F_*(n\boldsymbol{1})} = \lim_{n\to\infty} \frac{\overline{F}_*(n\boldsymbol{z})}{\overline{F}_*(n\boldsymbol{1})} = \frac{-\log G(\boldsymbol{z})}{-\log G(\boldsymbol{1})} \quad \text{for every } \boldsymbol{z} \in \mathbb{R}_+^d, \tag{3.2}$$

where F_* is the transformation of F to unit Fréchet margins and G is a multivariate extreme value distribution function with unit Fréchet margins.

We'll adopt the notation $V(\boldsymbol{z}) = -\log G(\boldsymbol{z})$.

3.2 Point processes

Let us proceed with the characterization of the dependence structure among the components of a max-stable distribution $G(\boldsymbol{x})$. The application of Poisson point process theory allows us to estimate marginals and dependence simultaneously. We'll use its multi-dimensional generalization (from chapter 7 of [2]) as opposed to the usual one-dimensional definition where the single coordinate is interpreted as time.

Let $\{X_i, 1 \leq i \leq n\}$ represent the location of points occurring randomly in a state space S which is Euclidean or a subset of it. A *point process* N counts the number of points in regions R of S,

$$N(R) = \sum_{i=1}^{n} \mathbf{1}(X_i \in R), \quad R \subseteq S.$$

The expected number of points in a set R is given by the *intensity measure* $\Lambda(R) = E[N(R)]$. If the intensity measure has a density function $\lambda : S \to [0, \infty)$, i.e. $\Lambda(R) = \int_R \lambda(x)\mathrm{d}x$, then λ is called the *intensity function* of the process.

A point process N with intensity measure Λ is said to be a *Poisson process* if for all $m \in \mathbb{N}$ and all disjoint sets $R_1, \ldots, R_m$ the random variables $N(R_1), \ldots, N(R_m)$ are independent and if for each set R with finite $\Lambda(R)$ is $N(R)$ a Poisson random variable with mean $\Lambda(R)$, i.e.

$$P(N(R) = k) = \frac{\Lambda^k(R)}{k!}\, e^{-\Lambda(R)}.$$

Observe that samples from a Poisson process are order- and location-independent.

A Poisson process is called *non-homogenous* if its intensity function λ is non-constant. Assuming unit Fréchet margins, it can be shown that the point process $N_n = \{\boldsymbol{Z}_i/n : 1 \leq i \leq n\}$ on $\mathbb{R}_+^d$ converges in distribution to a non-homogenous Poisson process N on $\mathbb{R}_+^d$.

Let us transform the points into pseudopolar coordinates

$$r_i = \frac{1}{n}\sum_{k=1}^{d} Z_{i,k}, \quad w_{i,j} = Z_{i,j}\Big/\sum_{k=1}^{d} Z_{i,k}\ , \quad 1 \le i \le n,\ 1 \le j \le d, \quad (3.3)$$

where $Z_{i,j}$ is the j-th component of $\boldsymbol{Z}_i$. The intensity measure Λ of the limiting process N can be factorized in these coordinates as

$$\Lambda(\mathrm{d}r \times \mathrm{d}\boldsymbol{w}) = \frac{\mathrm{d}r}{r^2}\,\mathrm{d}H(\boldsymbol{w}) \quad (3.4)$$

for some positive measure H of the angular components defined on the unit simplex

$$S_d = \left\{(w_1, \ldots, w_d) : \textstyle\sum_{j=1}^{d} w_j = 1,\ w_j \ge 0,\ 1 \le j \le d\right\}.$$

H is called the *spectral measure* of an extreme value distribution. The only constraints H must satisfy (resulting from the assumption of unit Fréchet margins) are

$$\int_{S_d} w_j\,\mathrm{d}H(\boldsymbol{w}) = 1, \quad 1 \le j \le d, \quad (3.5)$$

therefore no finite-dimensional parametrization for H exists.

Using (3.4), we can obtain Pickands's representation of limiting distributions of componentwise maxima [13].

Consider the family of regions $R_{\boldsymbol{z}} = \mathbb{R}_+^d \setminus \{(0, z_1] \times \cdots \times (0, z_d]\}$. Then

$$\lim_{n\to\infty} P(\boldsymbol{Z}_i/n \notin R_{\boldsymbol{z}},\ 1 \le i \le n) = P(N(R_{\boldsymbol{z}}) = 0) = e^{-\Lambda(R_{\boldsymbol{z}})},$$

where

$$\Lambda(R_{\boldsymbol{z}}) = \int_{R_{\boldsymbol{z}}} \frac{\mathrm{d}r}{r^2}\,\mathrm{d}H(\boldsymbol{w}) = \int_{S_d}\int_{\min_{1\le j\le d} z_j/w_j}^{\infty} \frac{1}{r^2}\,\mathrm{d}r\,\mathrm{d}H(\boldsymbol{w}) = \int_{S_d} \max_{1\le j\le d} \frac{w_j}{z_j}\,\mathrm{d}H(\boldsymbol{w}).$$

Since $P(\boldsymbol{Z}_i/n \notin R_{\boldsymbol{z}},\ 1 \le i \le n) = P(M_{n,j}/n \le z_j,\ 1 \le j \le d)$, any limit distribution with unit Fréchet margins is of the form $G(\boldsymbol{z}) = e^{-V(\boldsymbol{z})}$, where $V(\boldsymbol{z}) = \Lambda(R_{\boldsymbol{z}})$.

3.3 Dependence structure

It will be advantageous to express dependence models in the terms of the *stable tail dependence function* $l(\boldsymbol{v}) = V(1/v_1, \ldots, 1/v_d), \quad \boldsymbol{v} \in \mathbb{R}^d_+$. We can construct a max-stable distribution function (in non-transformed coordinates) using

$$G(\boldsymbol{x}) = \exp\left\{-l(-\log G_1(x_1), \ldots, -\log G_d(x_d))\right\}, \quad \boldsymbol{x} \in \mathbb{R}^d.$$

Let us describe some useful properties of l.

- homogenity property: $l(s\boldsymbol{v}) = sl(\boldsymbol{v})$ for $0 < s < \infty$
  This is a consequence of the max-stability of G: $G^s(s\boldsymbol{z}) = G(\boldsymbol{z})$.
- $\max\{v_1, \ldots, v_d\} \leq l(\boldsymbol{v}) \leq v_1 + \ldots + v_d$
  The lower bound corresponds to *complete dependence*, $G(\boldsymbol{x}) = \min\{G_1(x_1), \ldots, G_d(x_d)\}$ and the upper bound corresponds to *independence*, $G(\boldsymbol{x}) = G_1(x_1) \cdots G_d(x_d)$.
- l can be expressed by means of the spectral measure H:

$$l(\boldsymbol{v}) = \int_{S_d} \max_{1 \leq j \leq d} \{w_j v_j\} \, \mathrm{d}H(\boldsymbol{w}), \quad \boldsymbol{v} \in \mathbb{R}^d_+. \tag{3.6}$$

  It can be shown that l must be convex.

For the bivariate case, the simplex S_2 is a line and can be thus parametrized with a single parameter $w \in [0, 1]$; $w = w_1 = z_1/(z_1 + z_2) = v_2/(v_1 + v_2)$. In this case, the stable tail dependence function l can be simplified using the notion of the *Pickands dependence function* [13]

$$A(w) = l(1 - w, w).$$

We can rewrite l by means of A as

$$l(v_1, v_2) = (v_1 + v_2)\, A\left(\frac{v_2}{v_1 + v_2}\right)$$

for $v_1 \geq 0$, $v_2 \geq 0$ and $v_1 + v_2 > 0$.

This concludes our characterization of the dependence structure using the stable tail dependence function. Since there is a wealth of different dependence modelling approaches, we'll introduce one more.

The *extremal coefficient* θ is a summary measure which gives a rough picture of the full dependence structure by means of a single number.

If the vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ follows a d-variate extreme value distribution with identically distributed margins, the extremal coefficient θ for $X_1, \ldots, X_d$ is given by

$$P(\max(X_1, \ldots, X_d) < x) = (P(X_1 < x))^{\theta}, \quad x \in \mathbb{R},$$

where θ is independent of the value of x [16].

Since for n iid random variables $X_1, \ldots, X_n$

$$P(\max(X_1, \ldots, X_n) < x) = (P(X_1 < x))^{n}, \quad x \in \mathbb{R},$$

the extremal coefficient θ can be interpreted as the number of independent variables involved in a d-variate distribution and takes values in $[1, d]$, where $\theta = 1$ refers to complete dependence. In the bivariate case, $\theta = 2\,A(1/2)$, where A is the Pickands dependence function.

3.4 Modelling of the measure H

When creating models for the dependence structure (recall the reasons why they're necessary), we come to the realization that it's not obvious how to construct a spectral measure satisfying (3.5). Coles and Tawn [3] have developed two procedures describing how to parametrize new models. Both approaches create a model for H by expressing its density h.

Throughout this section we'll assume that $G(\boldsymbol{z})$ is absolutely continuous.

First we'll construct the densities of H from partial derivatives of $V(\boldsymbol{z})$. This creates a dependence model that doesn't need all components to be extreme simultaneously, allowing us to utilize more data during estimation. The drawback of this method is that it can only be applied to explicitly stated multivariate extreme value distributions, i.e. when we already have a closed-form model of $V(\boldsymbol{z})$.

Let $C = \{i_1, \ldots, i_c\}$ be a subset of coordinate indices in which a particular event is extreme. The coordinates in C span a c-dimensional facet $S_C = \{\boldsymbol{w} \in S_d : w_j = 0 \text{ if } j \notin C\}$ of the simplex S_d. We can now construct models for H by computing its densities h_C:

$$\frac{\partial^c V}{\partial z_{i_1} \cdots \partial z_{i_c}}(\boldsymbol{z}) = -\zeta^{c+1}\, h_C(\zeta z_{i_1}, \ldots, \zeta z_{i_c}), \tag{3.7}$$

where $z_j \in \mathbb{R}_+$ for $j \in C$, $z_j = 0$ for $j \notin C$, and $\zeta = 1 \Big/ \sum_{j \in C} z_j$.

The other approach is to start from a function h^* that doesn't necessarily have to satisfy (3.5) and construct a function h that does. Function l (and V) is then easily obtainable by means of Equation (3.6).

Assume that h^* is a positive function on the unit simplex S_d with finite first moments

$$m_j = \int_{S_d} w_j \, h^*(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w}$$

in every coordinate $j \in \{1, \ldots, d\}$ and denote $\varkappa = 1 \Big/ \sum_{j=1}^{d} m_j w_j$. Then

$$h(\boldsymbol{w}) = \varkappa^{d+1} m_1 \cdots m_d \, h^*(\varkappa m_1 w_1, \ldots, \varkappa m_d w_d) \tag{3.8}$$

gives a density of a spectral measure H with constraints (3.5) satisfied.

In the bivariate case Equation (3.8) reduces to

$$h(w) = \frac{m_1 m_2}{(m_1 w + m_2(1-w))^3} \, h^*\left(\frac{m_1 w}{m_1 w + m_2(1-w)} \right). \tag{3.9}$$

3.5 Models

It's time to describe some parametric dependence models that are used in practice. We'll use the procedures derived in the previous section to convert between spectral densities $h(\boldsymbol{w})$ and stable tail dependence functions $l(\boldsymbol{v})$.

We'll discuss the bivariate case first and extend it to an arbitrary number of dimensions where appropriate (for a thorough discussion see the original source, [3]). Our discussion follows section 9.2.2 from [1].

3.5.1 Dirichlet model

We'll create this model using the previously described approach of (3.8). We start with the probability density function of a Beta distribution

$$h^*(u) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \, u^{\alpha_1 - 1} (1-u)^{\alpha_2 - 1}, \quad 0 < u < 1, \ \alpha_1 > 0, \ \alpha_2 > 0,$$

$u = m_1 w / (m_1 w + m_2(1-w))$ being the transformed coordinate in (3.9).

Substituting $m_j = \alpha_j/(\alpha_1+\alpha_2)$ into (3.9), we arrive at

$$h(w) = \frac{\Gamma(\alpha_1+\alpha_2+1)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\alpha_1^{\alpha_1}\alpha_2^{\alpha_2}\, w^{\alpha_1-1}(1-w)^{\alpha_2-1}}{(\alpha_1 w+\alpha_2(1-w))^{\alpha_1+\alpha_2+1}}, \quad 0<w<1.$$

The stable tail dependence function has the form ([3], section 4.3)

$$l(v_1,v_2) = v_1(1-\mathrm{B}(\alpha_1+1,\alpha_2;q)) + v_2\,\mathrm{B}(\alpha_1,\alpha_2+1;q),$$

where $q = \alpha_1 v_2/(\alpha_1 v_2+\alpha_2 v_1)$, and

$$\mathrm{B}(\alpha_1,\alpha_2;q) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^q t^{\alpha_1-1}\,(1-t)^{\alpha_2-1}\,\mathrm{d}t$$

is the normalized incomplete Beta function.

Complete dependence is obtained when $\alpha_1=\alpha_2$ and $\alpha_1\to\infty$. Independence arises when $\alpha_1\to 0$ or $\alpha_2\to 0$.

The Dirichlet density is the probability density function of the multidimensional extension of the Beta distribution and has the form

$$h^*(\boldsymbol{u}) = \frac{\Gamma(\alpha_1+\cdots+\alpha_d)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_d)}\, u_1^{\alpha_1-1}\cdots u_d^{\alpha_d-1}, \quad \boldsymbol{u}\in S_d,\ \alpha_j>0,\ 1\le j\le d,$$

with $u_j = \varkappa m_j w_j$. We substitute $m_j = \alpha_j/(\alpha_1+\cdots+\alpha_d)$ into (3.8) as in the bivariate case and obtain

$$h(\boldsymbol{w}) = \frac{\Gamma(\alpha_1+\cdots+\alpha_d+1)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_d)} \frac{\alpha_1^{\alpha_1}\cdots\alpha_d^{\alpha_d}\, w_1^{\alpha_1-1}\cdots w_d^{\alpha_d-1}}{(\alpha_1 w_1+\cdots+\alpha_d w_d)^{\alpha_1+\cdots+\alpha_d+1}}, \quad \boldsymbol{w}\in S_d.$$

3.5.2 Logistic model

The stable tail dependence function of the logistic model is

$$l(v_1,v_2) = \left(v_1^{1/\alpha}+v_2^{1/\alpha}\right)^{\alpha}, \quad 0<\alpha\le 1.$$

With the use of (3.7) we can compute its spectral density

$$h(w) = \frac{1-\alpha}{\alpha}\,\left(w\,(1-w)\right)^{1/\alpha-2}\,\left(w^{1/\alpha}+(1-w)^{1/\alpha}\right)^{\alpha-2}.$$

The parameter α measures the strength of dependence between coordinates. Independence corresponds to $\alpha=1$ and complete dependence occurs as $\alpha\to 0$.

The multivariate extension follows naturally:

$$l(\boldsymbol{v}) = \left(v_1^{1/\alpha}+\cdots+v_d^{1/\alpha}\right)^{\alpha}.$$

3.5.3 Asymmetric logistic model

The asymmetric logistic model is Tawn's extension [10] of the logistic model for cases when the components are not interchangeable. Its stable tail dependence function is

$$l(v_1, v_2) = (1-\gamma_1)v_1 + (1-\gamma_2)v_2 + \left((\gamma_1 v_1)^{1/\alpha} + (\gamma_2 v_2)^{1/\alpha}\right)^{\alpha},$$

with parameters $0 < \alpha \leq 1$, $0 \leq \gamma_1 \leq 1$ and $0 \leq \gamma_2 \leq 1$. The spectral density is

$$h(w) = \frac{1-\alpha}{\alpha}(\gamma_1\gamma_2)^{1/\alpha}(w(1-w))^{1/\alpha-2}\left((\gamma_2 w)^{1/\alpha} + (\gamma_1(1-w))^{1/\alpha}\right)^{\alpha-2}.$$

For equal $\gamma = \gamma_1 = \gamma_2$, we obtain a combination of independence with weight $(1-\gamma)$ and the logistic model with weight γ. Independence corresponds to $\alpha = 1$ or $\gamma_1 = 0$ or $\gamma_2 = 0$. Complete dependence is obtained when $\gamma_1 = \gamma_2 = 1$ and $\alpha \to 0$.

This model can also be generalized for more dimensions. Let $\mathcal{C}_d$ be the set of non-empty subsets C of $\{1, \ldots, d\}$. Then

$$l(\boldsymbol{v}) = \sum_{\mathrm{C}\in\mathcal{C}_d}\left(\sum_{j\in\mathrm{C}}(\gamma_{\mathrm{C},j}v_j)^{1/\alpha_{\mathrm{C}}}\right)^{\alpha_{\mathrm{C}}}$$

for $0 < \alpha_{\mathrm{C}} \leq 1$, $\gamma_{\mathrm{C},j} \geq 0$, and $\sum_{\mathrm{C}\in\mathcal{C}_d}\gamma_{\mathrm{C},j} = 1$ for $1 \leq j \leq d$.

3.5.4 Bilogistic model

Another asymmetric generalization of the logistic model was proposed by Smith [15]. It is defined by

$$l(v_1, v_2) = \int_0^1 \max\{(1-\alpha_1)t^{-\alpha_1}v_1, (1-\alpha_2)(1-t)^{-\alpha_2}v_2\}\,\mathrm{d}t,$$

where $0 \leq \alpha_1 < 1$ and $0 \leq \alpha_2 < 1$. The parameters can be interpreted as follows: the sum $\alpha_1 + \alpha_2$ measures the strength of dependence and the difference $\alpha_1 - \alpha_2$ measures asymmetry. The spectral density h can only be computed numerically.

The model turns into complete dependence on $\alpha_1 = \alpha_2$ in the limit $\alpha_1 \to 0$ and into independence when $\alpha_1 \to 1$ or $\alpha_2 \to 1$.

3.5.5 Negative logistic models

The *asymmetric negative logistic model* introduced by Joe [6] resembles the asymmetric logistic model. The bivariate stable tail dependence function has the form

$$l(v_1, v_2) = v_1 + v_2 - \left((\gamma_1 v_1)^{1/\alpha} + (\gamma_2 v_2)^{1/\alpha}\right)^{\alpha}, \tag{3.10}$$

where $\alpha < 0$, $0 \leq \gamma_1 \leq 1$ and $0 \leq \gamma_2 \leq 1$. Independence is reached for $\alpha \to -\infty$ or $\gamma_1 = 0$ or $\gamma_2 = 0$, complete dependence for $\gamma_1 = \gamma_2 = 1$ and $\alpha \to 0$.

Fixing $\gamma_1 = \gamma_2 = 1$, we obtain a reduced symmetric model which is simply called the *negative logistic model*.

The spectral density corresponding to (3.10) has the form

$$h(w) = \frac{\alpha - 1}{\alpha} (\gamma_1 \gamma_2)^{1/\alpha} (w(1-w))^{1/\alpha - 2} \left((\gamma_2 w)^{1/\alpha} + (\gamma_1 (1-w))^{1/\alpha}\right)^{\alpha - 2}.$$

The multivariate version of (3.10) is

$$l(\boldsymbol{v}) = \sum_{j=1}^{d} v_j - \sum_{\mathrm{C} \in \mathcal{C}'_d} (-1)^{|\mathrm{C}|} \left(\sum_{j \in \mathrm{C}} (\gamma_{\mathrm{C},j} v_j)^{1/\alpha_{\mathrm{C}}} \right)^{\alpha_{\mathrm{C}}},$$

where $\mathcal{C}'_d$ is the collection of all subsets C of $\{1, \ldots, d\}$ containing at least two elements. Parameter domains are $\alpha_{\mathrm{C}} < 0$, $\gamma_{\mathrm{C},j} \geq 0$, and $\sum_{\mathrm{C} \in \mathcal{C}'_d} (-1)^{|\mathrm{C}|} \gamma_{\mathrm{C},j} \leq 1$ where $1 \leq j \leq d$.

3.5.6 Polynomial Pickands dependence functions

For the bivariate case, we can also model dependence using Pickands dependence functions. Klüppelberg and May [7] have studied their polynomial subclass

$$A(w) = a_0 + a_1 w + a_2 w^2 + \cdots + a_m w^m, \quad m \in \mathbb{N}. \tag{3.11}$$

A polynomial $A(w)$ is subject to a number of restrictions: we refer the reader to [1] for details.

The spectral density $h(w)$ is defined as $A''(w)$ on $0 < w < 1$.

Complete dependence, $A(w) = \max(w, 1-w)$, is attainable in the limit $m \to \infty$. The only linear solution is $A(w) = 1$ and corresponds to

independence. Most important is the quadratic case called the *mixed model* and the cubic *asymmetric mixed model*.

Mixed model. In the quadratic case a valid Pickand dependence function reduces to

$$A(w) = 1 - aw + aw^2 \tag{3.12}$$

for $0 \leq a \leq 1$. This model can also be obtained from the bivariate asymmetric negative logistic model (3.10) with parameter values $\alpha = -1$ and $\gamma_1 = \gamma_2 = a$.

Asymmetric mixed model. If $m = 3$ in (3.11), the Pickands dependence function takes the form

$$A(w) = 1 - (a_2 + a_3)w + a_2 w^2 + a_3 w^3.$$

The parameters of a valid asymmetric mixed model must satisfy the constraints

$$a_2 \geq 0, \quad a_2 + 3a_3 \geq 0, \quad a_2 + a_3 \leq 1, \quad a_2 + 2a_3 \leq 1.$$

3.6 Estimation

For the estimation of marginal and dependence structure we'll introduce the Point process and censored likelihood methods. These methods fit marginal tails with the GPD, so the models will be valid in the joint tail of the multivariate distribution.

3.6.1 Point process method

The marginal and dependence structure model parameters discussed in the previous sections are estimated using maximum likelihood on the limiting Poisson process N.

We start with a sequence of iid random vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ with unit Fréchet margins $\Phi_1(z_j) = e^{-1/z_j}, z_j > 0$. For a sufficiently large sample size n, the point process $N_n = \{\boldsymbol{Z}_i/n : 1 \leq i \leq n\}$ can be approximated in a region R bounded away from $\mathbf{0}$ by a non-homogenous Poisson process with intensity satisfying equation (3.4). The likelihood over R is

$$L_R(\{\boldsymbol{Z}_i/n\}; \theta) = e^{-n\Lambda(R;\theta)} \prod_{i:\, \boldsymbol{Z}_i/n \in R} \Lambda(\mathrm{d}r_i \times \mathrm{d}\boldsymbol{w}_i; \theta) \tag{3.13}$$

where θ are estimated dependence parameters and r_i and $\boldsymbol{w}_i$ are transformed coordinates defined by (3.3).

For iid random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ on $\mathbb{R}^d$ with arbitrary margins a transformation into unit Fréchet margins is required. We approximate marginal tails above high thresholds u_j by a threshold exceedance model—the generalized Pareto distribution (2.7)

$$P(X_j > x) = P(X_j > u_j)\,(1 + \xi_j(x - u_j)/\beta_j)^{-1/\xi_j}.$$

Points below u_j are dense enough and can follow the empirical distribution function $\tilde{F}_j(x) = \mathbf{1}\{X_j \le x\}/n$. The probability $p_j = P(X_j > u_j)$ can be treated the same way (as $1 - \tilde{F}_j(u_j)$).

The GPD-enhanced margins

$$\hat{F}_j(x; \beta_j, \xi_j) = \begin{cases} 1 - p_j(1 + \xi_j(x - u_j)/\beta_j)^{-1/\xi_j} & \text{for } x > u_j, \\ \tilde{F}_j(x) & \text{for } x \le u_j \end{cases} \tag{3.14}$$

can then be transformed into unit Fréchet margins using

$$Z_{i,j} = \tau_j(X_{i,j};\, \beta_j, \xi_j) = -1/\log \hat{F}_j(X_{i,j};\, \beta_j, \xi_j). \tag{3.15}$$

Since we're fitting the model to samples that are extreme in at least one coordinate, our region will be of the form $R_{\boldsymbol{s}} = \mathbb{R}^d_+ \setminus \{(0, s_1] \times \cdots \times (0, s_d]\}$. Setting $s_j = \tau_j(u_j)/n$ agrees with the GPD threshold and is therefore a natural choice.

Incorporating the transformations (3.15) and $R_{\boldsymbol{s}}$ into (3.13) yields the likelihood function $L_{R_{\boldsymbol{s}}}(\{\boldsymbol{X}_i\}; \theta, \boldsymbol{\beta}, \boldsymbol{\xi})$ as

$$e^{-n\Lambda(R_{\boldsymbol{s}};\theta)} \prod_{i:\, \boldsymbol{Z}_i/n \in R_{\boldsymbol{s}}} \left(\Lambda(\mathrm{d}r_i \times \mathrm{d}\boldsymbol{w}_i; \theta) \prod_{1 \le j \le d} \frac{\partial \tau_j}{\partial x_j}(X_{i,j};\, \beta_j, \xi_j)/n \right).$$

Substituting the factorization (3.4) and omitting constant terms produces

$$e^{-V(\boldsymbol{s};\theta)} \prod_{i:\, \boldsymbol{Z}_i/n \in R_{\boldsymbol{s}}} \left(h(\boldsymbol{w}_i; \theta)\, r_i^{-(d+1)} \prod_{j:\, Z_{i,j}/n > s_j} \beta_j^{-1} p_j^{-\xi_j} Z_{i,j}^2\, e^{1/Z_{i,j}} \left(1 - e^{-1/Z_{i,j}}\right)^{1+\xi_j} \right).$$

These marginal estimates are more precise than univariate GPD analyses as dependence between variables allows transfer of information from one margin to another. However, the method suffers from a number of drawbacks, the most pronounced being that points with one dominant margin pull the statistics to themselves while points with all components large tend to be undervalued (for an explanation we refer the reader to section 9.4.2 of [1]).

3.6.2 The censored-likelihood method

A different generalization of the POT method was developed by Smith [17] and Ledford and Tawn [9]. It solves most of the problems of the point process approach.

The function is again modelled in the region $\{[u_1, \infty) \times \cdots \times [u_j, \infty)\}$ by the GPD parametrization

$$\hat{F}_j(x) = 1 - (1 - \tilde{F}_j(u_j))\,(1 + \xi_j(x - u_j)/\beta_j)^{-1/\xi_j}, \quad x \geq u_j.$$

A transformation to unit Fréchet margins is performed to allow the use of (3.2) for approximating $F(\boldsymbol{z})$ with $G(\boldsymbol{z})$:

$$F(\boldsymbol{z}) = e^{-V(\boldsymbol{z};\theta)}, \quad z_j = -1 \Big/ \log(\hat{F}_j(x)),$$

θ containing dependence parameters to optimize. Observe that this transformation is identical to the one in the Poisson process approach for $x > u_j$. We'll denote s_j the transformed threshold u_j.

The main idea of the method is that it doesn't make sense for points that don't exceed u_j to contribute information to the GPD fit of coordinate j. Every observation $\boldsymbol{X}_i$ is assigned a set $C_i = \{j_1, \ldots, j_k\}$ that holds its extreme component indices, e.g. indices of $\boldsymbol{X}_i$ that exceed their respective thresholds: $C_i = \{j : X_{i,j} \geq u_j\}$. The remaining indices are censored—they are treated as if they haven't been observed at all. The contribution of $\boldsymbol{X}_i$ to the likelihood function is constructed by differentiating the distribution function in the observed components:

$$L(\boldsymbol{X}_i; \theta, \boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{\partial^k F}{\partial x_{j_1} \cdots \partial x_{j_k}}(\max(x_1, u_1), \ldots, \max(x_d, u_d)),$$

For example, in the two-dimensional case, the likelihood contribution $L(\boldsymbol{x}; \theta, \boldsymbol{\beta}, \boldsymbol{\xi})$ will be

$$\begin{cases}
e^{-V(s_1,s_2;\theta)} & \text{if } x_1 < u_1,\ x_2 < u_2, \\
e^{-V(z_1,s_2;\theta)} \frac{\partial}{\partial z_1} V(z_1, s_2; \theta) \frac{\partial z_1}{\partial x_1} & \text{if } x_1 \geq u_1,\ x_2 < u_2, \\
e^{-V(s_1,z_2;\theta)} \frac{\partial}{\partial z_2} V(s_1, z_2; \theta) \frac{\partial z_2}{\partial x_2} & \text{if } x_1 < u_1,\ x_2 \geq u_2, \\
e^{-V(z_1,z_2;\theta)} \left(\frac{\partial}{\partial z_1} \frac{\partial}{\partial z_2} - \frac{\partial^2}{\partial z_1 \partial z_2} \right) V(z_1, z_2; \theta) \frac{\partial z_1}{\partial x_1} \frac{\partial z_2}{\partial x_2} & \text{if } x_1 \geq u_1,\ x_2 \geq u_2.
\end{cases}$$

The whole likelihood is then given by

$$L(\{\boldsymbol{X}_i\}; \theta, \boldsymbol{\beta}, \boldsymbol{\xi}) = \prod_{1 \leq i \leq n} L(\boldsymbol{X}_i; \theta, \boldsymbol{\beta}, \boldsymbol{\xi}).$$

The method's flexibility make it the state-of-the-art POT method for the estimation of the joint tail of a distribution when no special data properties can be exploited.

Chapter 4

A conditional approach for multivariate extreme values

In the article [5] Heffernan and Tawn introduced a semiparametric approach which works under the assumption that at least one of the components of a d-dimensional variable $\boldsymbol{X} = (X_1, \ldots, X_d)$ is extreme. This is less restrictive than the usual extreme value theory approach which instead estimates the joint tail of the distribution.

The motivation for this approach is the fact that the classical methods are unable to estimate probabilities for events that are not extreme in every component. The authors consider this a fundamental problem of the existing framework which can't be redeemed by introducing new dependence models.

We'll adopt the convention (as in Chapter 3) that operations on vectors are applied componentwise.

4.1 Marginal fitting

Consider a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of d-dimensional iid observations from an unknown distribution function $F(\boldsymbol{x})$ of $\boldsymbol{X}$, with components $F_j(x)$, $1 \leq j \leq d$. Analogically to the classical approach, the tail of $F_j(x)$ is fitted with the GPD

$$P(X_j \leq x | X_j > u_j) = H_{\xi_j;u_j,\beta_j}(x), \quad x > u_j; \quad x < u_j - \beta_j/\xi_j \text{ if } \xi_j < 0$$

for a suitably chosen high threshold u_j. Thus, the marginal distributions can be modelled by (3.14).

Marginals are finally transformed to standard Gumbel form $\Lambda(y) = \exp(-e^{-y})$ using the transformation

$$Y_j = t_j(X_j) = -\log(-\log\{\hat{F}_j(X_j)\}).$$

This form has the useful property $P(Y_j > y) \sim e^{-y}$ as $y \to \infty$.

4.2 Dependence structure modelling

Assuming that component k is extreme, we are interested in the conditional distribution of the remaining components. For each k we'll create an asymptotic model of this conditional distibution that is valid under a few theoretical assumptions.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ be a random variable with Gumbel margins and let $\boldsymbol{Y}_{-k}$ denote the distribution of components Y_j, $j \neq k$. We'll assume that their limit conditional distribution (conditionally on the extreme component k) takes the form

$$\lim_{y_k \to \infty} P(\boldsymbol{Y}_{-k} \leq \boldsymbol{a}_{|k}(y_k) + \boldsymbol{b}_{|k}(y_k)\, \boldsymbol{q}_{|k} | Y_k = y_k) = G_{|k}(\boldsymbol{q}_{|k}) \tag{4.1}$$

for some normalizing $\mathbb{R} \to \mathbb{R}^{d-1}$ functions $\boldsymbol{a}_{|k}$ and $\boldsymbol{b}_{|k}$ and that all margins of the limit distribution $G_{|k}$ are non-degenerate.

We introduce standardized variables $\boldsymbol{Q}_{|k} = (\boldsymbol{Y}_{-k} - \boldsymbol{a}_{|k}(y_k))/\boldsymbol{b}_{|k}(y_k)$ to simplify notation. Under the assumption (4.1) it can be shown that

$$\lim_{u_k \to \infty} P(\boldsymbol{Q}_{|k} \leq \boldsymbol{q}_{|k}, Y_k - u_k = y | Y_k > u_k) = e^{-y}\, G_{|k}(\boldsymbol{q}_{|k}), \tag{4.2}$$

so the variables $Y_k - u_k$ and $\boldsymbol{Q}_{|k}$ become independent in the limit provided that $Y_k > u_k$. This property motivates the authors to assume the existence of high thresholds η_k for which

$$P(\boldsymbol{Q}_{|k} \leq \boldsymbol{q}_{|k} | Y_k = y_k) = G_{|k}(\boldsymbol{q}_{|k}) \quad \text{for all } y_k > \eta_k$$

holds, $\boldsymbol{Q}_{|k}$ and Y_k being independent for $Y_k > \eta_k$.

The functions $\boldsymbol{a}_{|k}(y_k)$ and $\boldsymbol{b}_{|k}(y_k)$ are unique up to type (see (2.1)). The authors considered many various dependence structures satisfying (4.1) and encompassed them in a parametric family

$$\begin{aligned} a_{j|k}(y_k) &= A_{j|k}\, y_k + C_{j|k} - D_{j|k} \log(y_k), \\ b_{j|k}(y_k) &= {y_k}^{B_{j|k}}, \end{aligned}$$

where $0 \leq A_{j|k} \leq 1$, $B_{j|k} < 1$ and $0 \leq D_{j|k} \leq 1$ for all $1 \leq j \leq d$, $j \neq k$. The family models positively and negatively associated components differently due to the asymmetry of the Gumbel distribution: $C_{j|k}$ and $D_{j|k}$ are set to zero unless $A_{j|k} = 0$ and $B_{j|k} < 0$.

Classes of dependence
The dependence structure can be categorized into four classes identifying the behaviour of quantiles of the distribution of $(Y_j|Y_k = y_k)$ as $y_k \to \infty$. When $A_{j|k} = 1$ and $B_{j|k} = 0$, the variables Y_j and Y_k are asymptotically dependent. Otherwise they're asymptotically independent and exhibit

positive extremal dependence when $0 < A_{j|k} < 1$ or $B_{j|k} > 0$,

extremal near independence when $A_{j|k} = D_{j|k} = 0$ and $B_{j|k} \leq 0$, or

negative extremal dependence when $A_{j|k} = 0$, $D_{j|k} > 0$ and $B_{j|k} < 0$.

Classes of symmetry
Similarly to the symmetrical models in classical methods, we can derive subclasses with enforced symmetry. The authors consider exchangeability of components Y_j and Y_k in two forms:

weak pairwise extremal exchangeability when dependence structures match, i.e. $a_{j|k}(y) = a_{k|j}(y)$ and $b_{j|k}(y) = b_{k|j}(y)$.

strong pairwise extremal exchangeability when both dependence structures and limit distributions $G_{j|k} = G_{k|j}$ match.

4.3 Estimation

Consider a sample of observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Assuming that they are realizations of iid random variables, we create a marginal fit described in Section 4.1 and transform them to Gumbel margins $\boldsymbol{Y}_{i,j} = t_j(\boldsymbol{X}_{i,j})$, $1 \leq i \leq n$, $1 \leq j \leq d$. Marginal estimates are computed separately from the dependence structure (the authors state that the complexity of joint estimation is unnecessary unless marginal shape parameters vary greatly).

The authors assume (see [5], Section 5.2 for rationale) that the components of the standardized random variables $\hat{\boldsymbol{Q}}_{|k}$ are mutually independent and normally distributed, i.e. $Q_{j|k}$ has the distribution $\mathcal{N}(\mu_{j|k}, \sigma^2_{j|k})$.

The parameter estimates $\hat{\boldsymbol{\mu}}_{|k}$, $\hat{\boldsymbol{\sigma}}_{|k}$ and $\hat{\theta}_{|k} = \{\hat{A}_{j|k}, \hat{B}_{j|k}, \hat{C}_{j|k}, \hat{D}_{j|k}\}$ can be found by maximization of log-likelihood functions $\log L_{|k}(\theta_{|k}, \boldsymbol{\mu}_{|k}, \boldsymbol{\sigma}_{|k})$

having the form

$$-\sum_{i:Y_{i,k}>\eta_k}\sum_{j\neq k}\left(\log(\sigma_{j|k}\,b_{j|k}(Y_{i,k}))+\frac{1}{2}\left(\frac{Y_{i,k}-[a_{j|k}(Y_{i,k})+\mu_{j|k}\,b_{j|k}(Y_{i,k})]}{\sigma_{j|k}\,b_{j|k}(Y_{i,k})}\right)^2\right).$$

Observe that dependence structure margins are fitted separately from each other. Conditional likelihood models $L_{|k}$ are considered to be independent (see 5.3 from [5] for justification), but maximization of the joint likelihood is also possible if the model is chosen so that the parameters influence each other.

4.4 Sampling

We've obtained semi-parametric models for $\boldsymbol{Y}_{-k} = \boldsymbol{a}_{|k}(y_k) + \boldsymbol{b}_{|k}(y_k)\,\boldsymbol{Q}_{|k}$, with $Q_{j|k}$ having the distribution $\mathcal{N}(\mu_{j|k}, \sigma^2_{j|k})$. These models are valid for the case that the k-th coordinate of y is extreme, i.e. $y_k > \eta_k$. Combining these models and using empirical estimators for non-extreme regions enables us to estimate the probability of $\boldsymbol{X} \in R$ for any region R using a Monte Carlo method.

We start with partitioning R into disjoint regions R_0 and R_k, $1 \leq k \leq d$. R_0 is the portion below thresholds over which the extreme models are defined, i.e.

$$R_0 = R \cap \{\boldsymbol{x} \in \mathbb{R}^d : t_j(x_j) \leq \eta_j \text{ for } 1 \leq j \leq d\}$$

and R_k are extreme regions with k-th coordinate dominant in terms of marginal quantiles,

$$R_k = (R \setminus R_0) \cap \{\boldsymbol{x} \in \mathbb{R}^d : \hat{F}_k(x_k) > \hat{F}_j(x_j) \text{ for } 1 \leq j \leq d,\, j \neq k\}.$$

The probability $P(\boldsymbol{X} \in R)$ can be then written as

$$\sum_{k=0}^{d} P(\boldsymbol{X} \in R_k) = P(\boldsymbol{X} \in R_0) + \sum_{k=1}^{d} P(\boldsymbol{X} \in R_k | X_k > l_k) P(X_k > l_k)$$

where every threshold $l_k = \inf_{\boldsymbol{x} \in R_k} x_k$ is chosen for efficiency so that the whole R_k lies above it.

- Since R_0 is not extreme, the empirical estimator for $P(\boldsymbol{X} \in R_0)$ is sufficient.

- For the estimation of $P(X_k > l_k)$ we can use our GPD-enhanced marginal model (3.14).
- $P(\boldsymbol{X} \in R_k | X_k > l_k)$ is obtained by sampling from $\boldsymbol{X}|X_k > l_k$ as the ratio of points that fall into R_k.

Sampling from the fitted conditional distribution $\boldsymbol{X}|X_k > l_k$ can be performed as follows:

1. generate Y_k^* from the Gumbel distribution so that $Y_k^* > t_k(l_k)$.
2. generate Y_j^* from $\mathcal{N}\Big(\hat{a}_{j|k}(Y_k^*) + \hat{\mu}_{j|k}\,\hat{b}_{j|k}(Y_k^*)\,,\ \hat{\sigma}^2_{j|k}\,\hat{b}^2_{j|k}(Y_k^*)\Big)$ for each $j \neq k$.
3. transform $\boldsymbol{Y}^*$ into original coordiantes: $X_j^* = t_j^{-1}(Y_j^*)$ for each j.

There are a few problems with this method. In the frequent case when the marginal GPD fit starts lower than the dependence region determinated by $\boldsymbol{\eta}$, we're left with an extreme section with no dependence structure, but with observations too sparse to make good use of the empirical distribution.

Determining probabilities using sampling may also pose a problem when the region R is extreme in more components. Since the conditional margins are assumed to be independent, the probability that a generated sample falls into such a region decreases rapidly with the number of dimensions. This means that only the roughest estimates can be computed with practical amounts of samples.

These facts make this approach, in a sense, complementary to estimation methods discussed in the previous chapter.

The number of assumptions render the method a little impractical due to the extensive testing that is necessary; furthermore, the authors don't give any guidelines for cases when said assumptions break down. Discussions on the paper have further praised the novel framework and proposed some extension ideas, but questioned the generality of (4.1) and applicability of the method to real-world problems (particularly because of its inability to take time dependencies into account).

Chapter 5

Simulation study

Having gone through theoretical aspects of extreme value theory, we'll now examine how the methods behave in practice when the ground truth is known. First we'll empirically examine robustness and convergence rate of the univariate POT method by fitting GPD models to several known distributions with varying thresholds and sample counts.

With this information acquired, we'll assess the behaviour of the bivariate asymmetric logistic model using censored likelihood, trying different parameter values. The method of Heffernan and Tawn will then be tested by emulating the same distributions and comparing Monte Carlo density estimates with the original models.

5.1 Univariate Peaks Over Threshold

We consider three distributions: unit exponential $F_{\mathrm{E}}(x) = 1 - e^{-x}$, uniform $F_{\mathrm{U}}(x) = x$ on $x \in (0, 1)$ and unit Fréchet $F_{\mathrm{F}}(x) = e^{-1/x}$. According to [2] (page 77), their limit GPD shape parameters are $\xi_{\mathrm{E}} = 0$, $\xi_{\mathrm{U}} = -1$ and $\xi_{\mathrm{F}} = 1$. We've chosen a range of realistic sample sizes (from $n = 10^2$ to 10^4) and thresholds T (from 0%- to 99%-quantiles) and computed maximum likelihood estimations $\hat{\xi}_{n,\mathrm{T}}$ for many random runs m (m spanning from 1000 for $n = 10^2$ to about 100 for $n = 10^4$).

Figures 5.1 to 5.3 summarize our results. Left column depicts absolute differences of ground-truth shape values and their empirical means, ranging from 0 (black) to 1 (white) with higher values clamped to 1. Right column displays variances of $\hat{\boldsymbol{\xi}}_{n,\mathrm{T}}$ transformed to logarithmic scale to better convey range (black $= 6.1 \cdot 10^{-6}$ to white $= 1$).

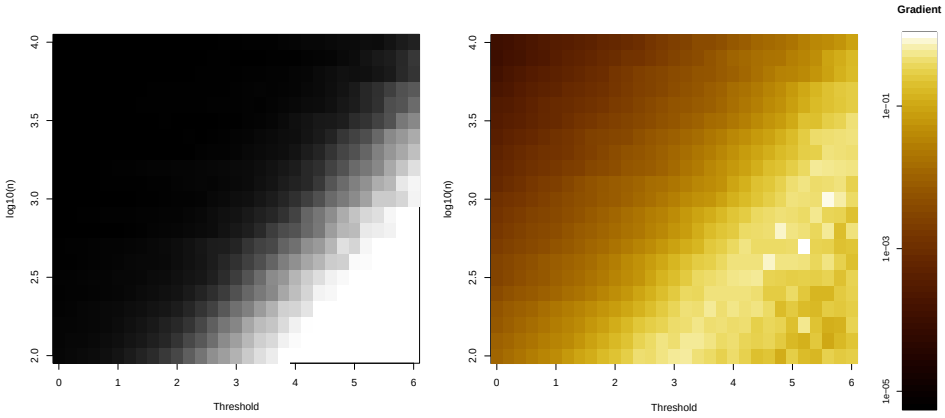


Figure 5.1: GPD shape parameter estimation for the unit exponential distribution: absolute difference of shape mean and ground-truth shape (left), log variance of shape (right).

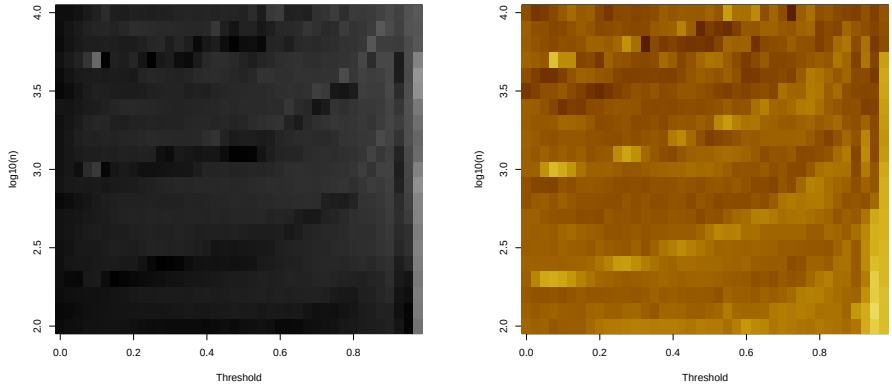


Figure 5.2: Parameter estimation for the uniform distribution on $(0, 1)$.

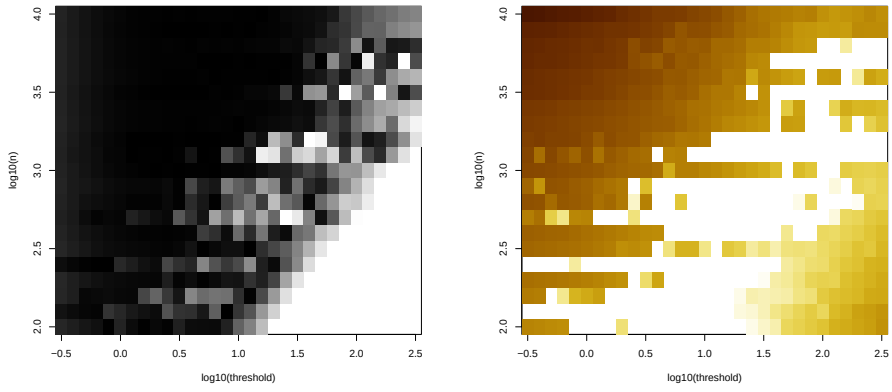


Figure 5.3: Parameter estimation for the unit Fréchet distribution.

These indicators suggest threshold values which can be expected to establish useful results for a given n; they should equal the correct value (i.e. lie in a black area) while keeping the variance as low as possible. Furthermore, they betray the robustness of the procedure: some estimations from Fréchet and uniform simulations didn't converge, showing as points with high deviation. An interesting fact is that—for the discussed distributions—thresholds may be chosen lower than one might expect.

The Fréchet distribution is troublesome—it's difficult to balance variance and stability (impossible for $n < 10^3$). The uniform distribution converges slowly, needing large inputs to produce meaningful fits; also, we've been unable to explain the parabolic "stripes" of higher variance (presumably an artifact of the root-finding algorithm). For these reasons, we've decided to test multivariate dependence structures on exponential margins.

5.2 Bivariate censored likelihood

We've tested the bivariate asymmetric logistic model (see Section 3.5.3)

$$F(z_1, z_2) = \exp\left(-\frac{1-\gamma_1}{z_1} - \frac{1-\gamma_2}{z_2} - \left((\gamma_1/z_1)^{1/\alpha} + (\gamma_2/z_2)^{1/\alpha}\right)^{\alpha}\right)$$

for three parameter sets A, B and C (Table 5.1), each with sample size $n = 3000$. Samples were generated in Fréchet margins and underwent the transformation $x_j = -\log(1 - \exp(-1/z_j))$ to exponential margins. Figure 5.4 (left column) depicts the random samples along with ground-truth density contours. Observe that B has a very large degree of dependence and that C is symmetric.

First we've computed maximum likelihood fits for all models from Section 3.5 and plotted Pickands dependence functions for the asymmetric

	α	γ_1	γ_2
A	0.5	0.7	0.2
B	0.1	0.7	0.2
C	0.5	0.6	0.6

Table 5.1: Parameter values for bivariate simulations from the asymmetric logistic model.

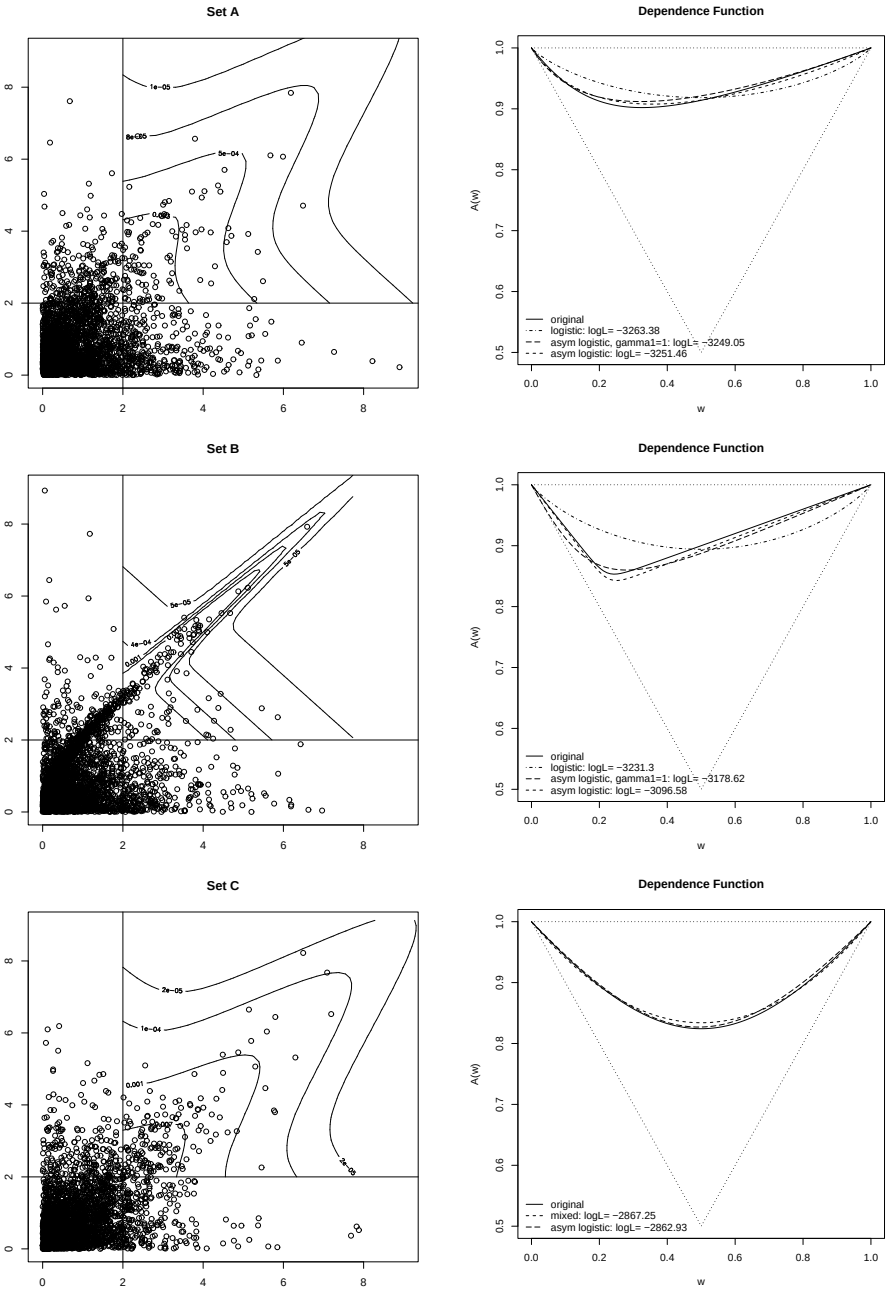


Figure 5.4: *Left*: Generated samples with unit exponential margins and asymmetric logistic dependence structure. *Right*: Corresponding Pickands dependence functions for fitted bivariate dependence models.

logistic model and some of its submodels (Fig. 5.4, right column). Note that the highest likelihood function value isn't always observed for the correct model with all parameters free.

Deviance analysis shows that the best model for case B is the asymmetric logistic model with all parameters free. This is probably because B exhibits strong dependence and asymmetry. On the other hand, the analysis of set C doesn't reject symmetrical submodels. The best single-parameter model is the mixed model; its likelihood function value is very close to the most general asymmetric logistic model. The most interesting result was obtained for set A: a submodel with one fixed parameter was awarded a higher score. Since we could replicate the behaviour in different runs, it could be that the optimization procedure is getting stuck in local minima of the likelihood function (having more variables to optimize).

The negative logistic model likelihoods are nearly equal to the logistic models in all discussed cases. It would seem that model families effectively overlap and that the model choice isn't as important as a good marginal fit.

Finally, for the challenging case B we analyze how sure can we be about the estimates. We generate random samples for various sizes n (see Fig. 5.5) and compute dependence parameters and their confidence intervals (for GPD thresholds $u_1 = u_2 = 1.5$). We can see that parameter values tend to fluctuate for mid- to low sample sizes ($n < 3000$); in fact, some refused to converge at all and re-generations were necessary. However, we can see that the biased fits have $\gamma_1 = 1$: the optimizer probably decided to fix one parameter due to underspecification. Manually constraining two degrees of freedom results in the logistic model (parametrized only by α) and helps to stabilize inconsistencies (gray intervals in Fig. 5.5).

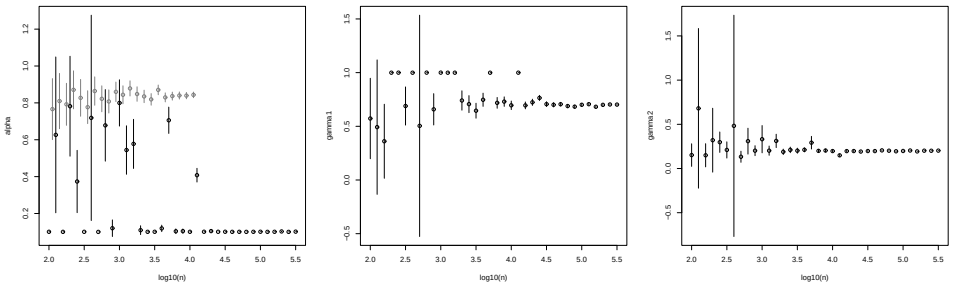


Figure 5.5: Dependence parameter estimates $\hat{\alpha}$, $\hat{\gamma}_1$ and $\hat{\gamma}_2$ for the asymmetric logistic model (black) and $\hat{\alpha}$ for the simpler logistic model (gray)

5.3 The conditional approach of Heffernan and Tawn

We adopt models A, B and C from the previous section and try to reproduce density contours from Figure 5.4. Taking $n = 3000$ again as the sample size, we chose $u_1 = u_2 = 1.5$ for GPD thresholds. We assume there exist thresholds η_k above which the limit model from Section 4.2 is valid. We compute maximum-likelihood estimations of model parameters (described in Section 4.3) for various threshold choices and assume that the parameters stay approximately constant above a well-balanced threshold (before the estimates inevitably start to diverge due to data sparsity). Figure 5.6 shows the behaviour of parameter estimates depending on threshold quantiles for the set B. We've chosen both η's to correspond to 72%-quantiles; thresholds for the remaining models were also chosen visually. Parameter estimates are summarized in Table 5.2.

Contour estimates as shown below (Fig. 5.7) were computed using the

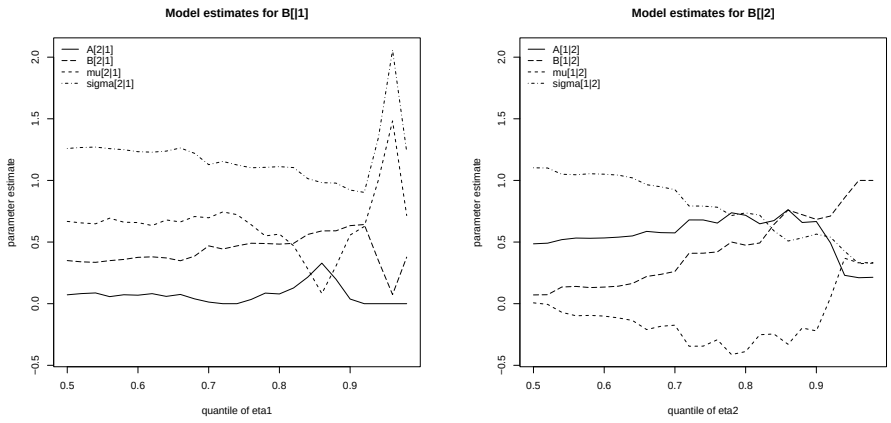


Figure 5.6: Parameter estimates depending on quantiles of η_1 and η_2 for the set B.

$\mathsf{set}_{\|k}$	$\mathsf{A}_{\|1}$	$\mathsf{A}_{\|2}$	$\mathsf{B}_{\|1}$	$\mathsf{B}_{\|2}$	$\mathsf{C}_{\|1}$	$\mathsf{C}_{\|2}$
$t^{-1}(\eta_k)$	1.219	1.182	1.227	1.226	1.229	1.138
$A_{3-k\|k}$	0.136	0.445	0.013	0.574	0.336	0.477
$B_{3-k\|k}$	0.376	0.129	0.469	0.262	0.481	0.463
$\mu_{3-k\|k}$	0.536	−0.003	0.697	−0.174	0.369	0.146
$\sigma_{3-k\|k}$	1.153	1.178	1.128	0.925	0.950	0.904

Table 5.2: Threshold choices and parameter estimates for A, B and C.

procedure described in Section 4.4; the number of Monte Carlo evaluations was around 10^6 for each estimate. Comparison with ground-truth densities (Fig. 5.4) reveals that the fits match the original distributions quite well even for $n = 3000$; the model is, however, unable to represent the independent part of the asymmetric logistic model and has trouble with the high degree of dependence of set B.

5.4 Discussion

First of all, simulations show that the discussed approximations are valid only in the limit and the dataset needs to be well-behaved even for middle-sized samples (1000 observations). Thresholds for fitting the GPD don't need to be chosen too high; depending on the data, even a 70%-quantile might be a safe choice.

Dependence structure analysis showed that once the margins are fitted well, the exact model choice isn't that important and a simpler model will suffice—unless the data show prominent extremal dependence or there's plenty of observations to support a more complex model. Also, we've found that necessary sample sizes for the bivariate case are comparable to the univariate case.

The conditional approach of Heffernan and Tawn can cope well with low amounts of data, but the fitted model (4.1) is quite coarse. This is balanced by the larger region this method can describe.

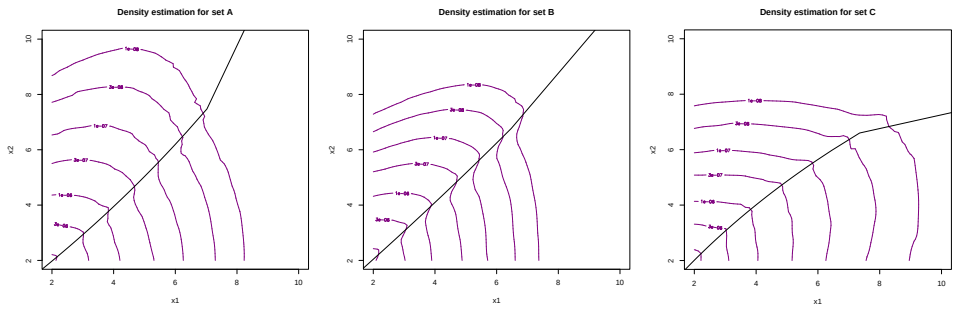


Figure 5.7: Countour plots of Monte Carlo density estimations for the conditional approach

Chapter 6

Application

We now proceed to the application of the described methods on a dataset (Fig. 6.1) consisting of birth weights W and lengths L of 1633 randomly picked Czech children. The dataset was sampled during a short period of about two years and siblings were filtered out, so we've chosen to consider the sample homogenous. Observe that lengths are quantized to centimeters: this fact makes marginal fitting non-trivial.

All computations were carried out using the stastistical program `R` (see

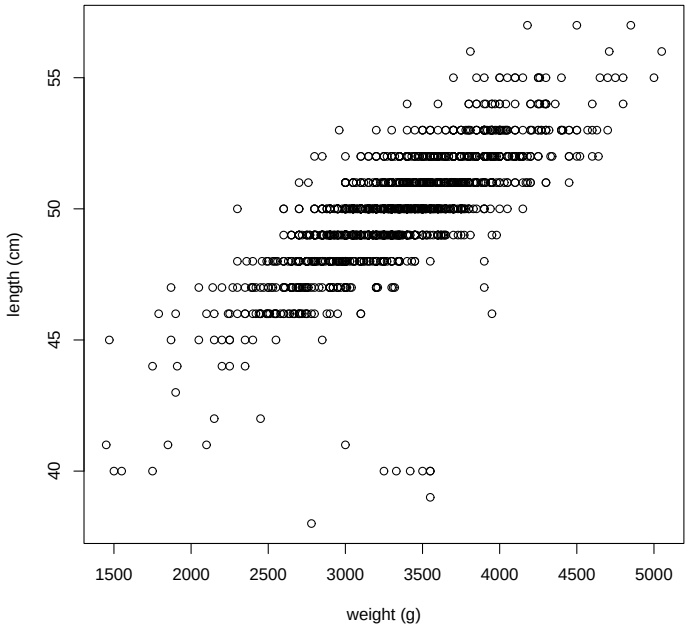


Figure 6.1: The `children` dataset.

Section 6.4 for implementation details).

Throughout this chapter, we will use the abbreviations PP for the Poisson point process method (Section 3.6.1), CL for the censored likelihood method (Section 3.6.2, together called the classical methods) and HT for the conditional approach of Heffernan and Tawn [5] as discussed in Chapter 4.

6.1 Marginal estimation

All considered methods use marginal fitting of the generalised Pareto distribution $H_{\xi_i;u_i,\beta_i}$ (with shape parameter ξ_i and scale parameter β_i, see Section 2.3) over thresholds u_i. Both classical methods include marginal parameters into their likelihood functions; for HT we've chosen to estimate margins and dependence separately (as mentioned in Section 4.3). Some guidelines for choosing u_i are discussed in literature (see [2], p. 78). One approach uses mean excesses over u_i, $E(X_i - u_i | X_i > u_i)$. According to Equation 2.8, empirical means of excesses should behave linearly

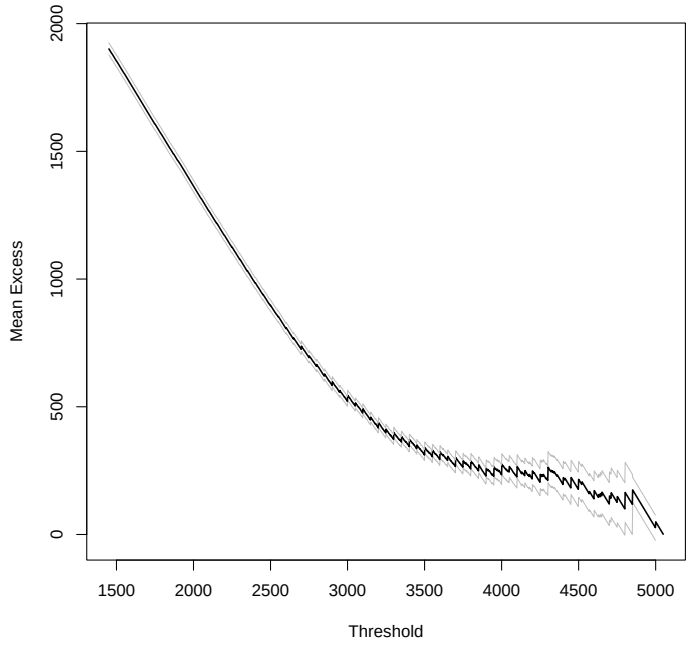


Figure 6.2: Threshold selection for weights using `mrlplot`: mean excesses over thresholds.

above a well-chosen u_i. The plot of u_i against the mean of observations that exceed it is called the *mean residual life plot* (Fig. 6.2).

Another means of threshold estimation uses the fact that shape $\hat{\xi}_i$ and modified scale $\hat{\beta}_i - \hat{\xi}_i u_i$ should be constant above u_i. On the other hand, u_i should be chosen so that the dataset above it isn't too sparse to provide a meaningful fit. For example, Figure 6.3 shows maximum-likelihood fits of shape $\hat{\xi}_w$ (with 95% confidence intervals) for various thresholds u_w. The threshold choice ultimately depends on subjective preferences.

Independently fitted marginal model parameters for the data are summarized in Table 6.1. Observe that both shape parameters are negative. That means that probability densities become zero above $u_i - \beta_i/\xi_i$.

	weight (g)	length (cm)
u_i	3300	50
β_i	492.53	2.4695
ξ_i	−0.24993	−0.33241
$P = 0$ above	5270	57.4

Table 6.1: Independent marginal estimates for the `children` dataset.

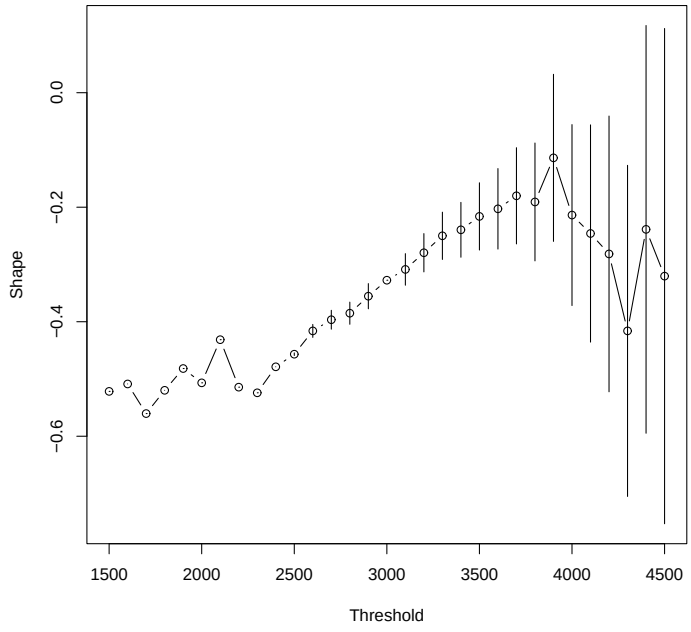


Figure 6.3: Threshold selection for weights using `tcplot`: stability of shape estimates $\hat{\xi}_w$.

6.2 Dependence structure estimation

6.2.1 Classical methods

We've fitted all bivariate dependence models from Section 3.5 with their optimal CL fits and tried to narrow them down to the one that best describes the dependence. See Section 8.3.3 of [2] for an analogous discussion.

Nested models were tested against each other using likelihood ratio tests. The simpler symmetric versions (logistic model, the symmetric negative logistic model, the mixed model and the symmetric version of the Dirichlet model) were shown to suffice for describing the dependence structure. We have further discarded the mixed model because its likelihood was too low when compared to the others.

Next we performed a marginal fit comprasion of both classical methods for the three remaining models by the means of their probability, quantile and density plots. (for instance, Fig. 6.4 shows the best PP model). Since the CL fits clearly follow the empirical distribution better (admittedly due to the fact that the chosen extremal regions contain enough data), we focus only on them.

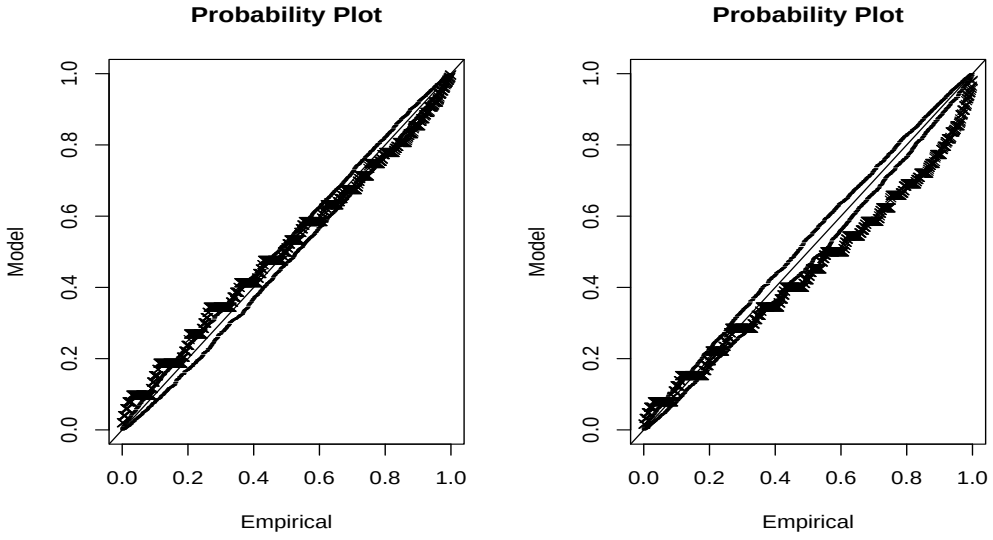


Figure 6.4: Probability plots for the best PP model (symmetric Dirichlet, right) and the corresponding CL fit (left).

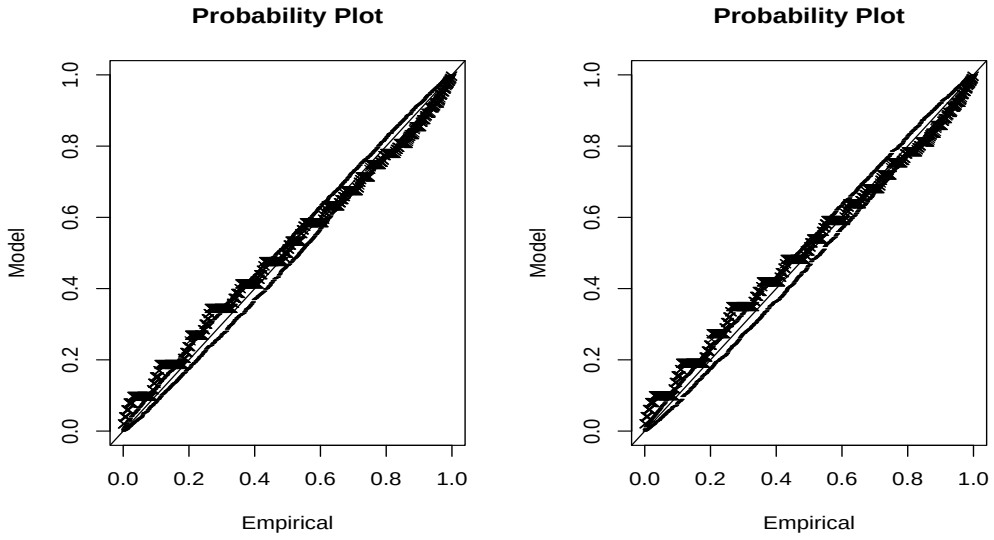


Figure 6.5: Probability plots for CL fits: the logistic model (left) and the negative logistic model (right).

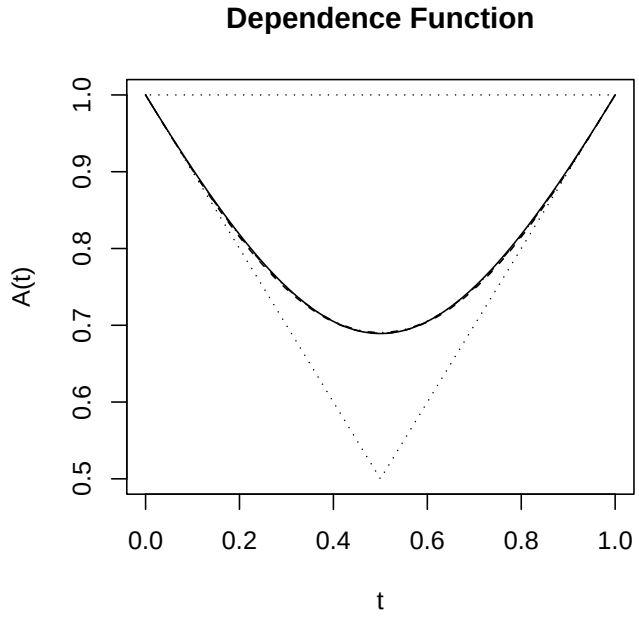


Figure 6.6: Pickands dependence functions for the three CL models, superimposed.

The three models were assessed by means of marginal analysis and their Pickands dependence functions (see Figs 6.4 to 6.6). Observe that the three fits are virtually undistinguishable, so we've decided that we can use any of them. Because of its simplicity, we've chosen the logistic model (see Section 3.5.2 for details).

The final model for the region $\boldsymbol{x} > \boldsymbol{u}$ then becomes

$$F(\boldsymbol{x}) = \exp\left\{-\left[(-\log F_1(x_1))^{-1/\alpha} + (-\log F_2(x_2))^{-1/\alpha}\right]^{\alpha}\right\},$$

with GPD margins given by

$$F_j(x_j) = 1 - p_j(1 + \xi_j(x_j - u_j)/\beta_j)^{-1/\xi_j}.$$

Estimated parameters for the model are given in Table 6.2. The estimated dependence parameter indicates that components are fairly dependent.

The extremal coefficient discussed in Section 3.3 amounts to 1.3784, which also points at a fair degree of dependence between components. But since our dependence model has only a single parameter, this doesn't yield any new information.

p_1	0.5334
p_2	0.3962
β_1	492.2262(19.04)
β_2	2.5132(0.105)
ξ_1	−0.1828(0.027)
ξ_2	−0.3071(0.023)
α	0.4630(0.015)

Table 6.2: Joint marginal and dependence estimates for the logistic distribution (standard errors in parentheses).

6.2.2 The conditional approach

We follow the procedure from Sections 4.2 and 4.3 closely. Figure 6.7 shows the dataset transformed to standard Gumbel margins. The thresholds η_j over which the conditional dependence models are supposed to be valid were found to be satisfactory at 75%-quantiles—raising them didn't improve the stability of the parameter estimates (using a procedure similar plot as in the simulation study). The assumption of independence

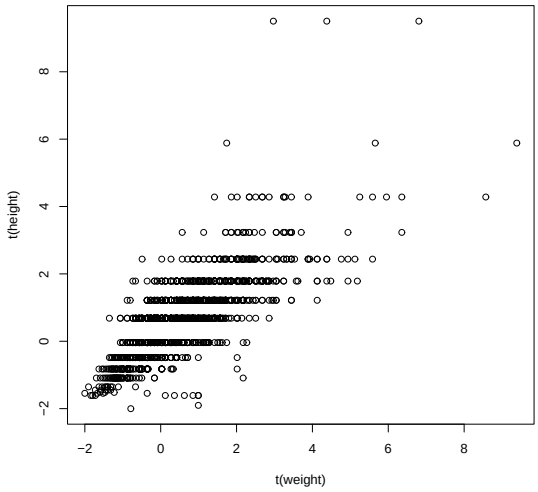


Figure 6.7: The dataset transformed to standard Gumbel margins.

of the extreme component Y_j and the residual $\boldsymbol{Q}_{|j}$ and the assumption of normality of $\boldsymbol{Q}_{|j}$ were found to be valid for our threshold. Since our sample is two-dimensional, there is no internal dependence in $\boldsymbol{Q}_{|j}$ left to be tested.

Parameter values maximizing the log-likelihood are shown in Table 6.3. Models state that conditionally on the extreme component y_k, the other component y_j can be modelled with $\mathcal{N}\Big(A_{j|k}\, y_k + \mu_{j|k}\, {y_k}^{B_{j|k}}\,,\ \sigma^2_{j|k}\, {y_k}^{2\, B_{j|k}}\Big)$. Both conditional distributions exhibit positive extremal dependence and no component exchangeability was detected.

	weight (g)	length (cm)
$t^{-1}(\eta_k)$	3650	51
$\boldsymbol{A}$	$A_{l\|w} = 0.03303$	$A_{w\|l} = 0.69902$
$\boldsymbol{B}$	$B_{l\|w} = 0.75537$	$B_{w\|l} = 0.67455$
$\boldsymbol{\mu}$	$\mu_{l\|w} = 0.99656$	$\mu_{w\|l} = 0.22688$
$\boldsymbol{\sigma}$	$\sigma_{l\|w} = 0.52041$	$\sigma_{w\|l} = 0.64399$
$\log L_{\|k}(\theta_{\|k}, \boldsymbol{\mu}_{\|k}, \boldsymbol{\sigma}_{\|k})$	-167.25	-215.65

Table 6.3: HT parameter estimates for the `childrен` dataset.

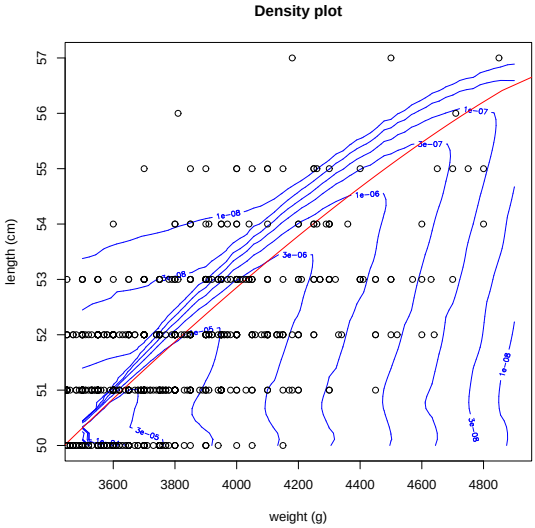


Figure 6.8: Probability density contours for both HT models (along with their boundary).

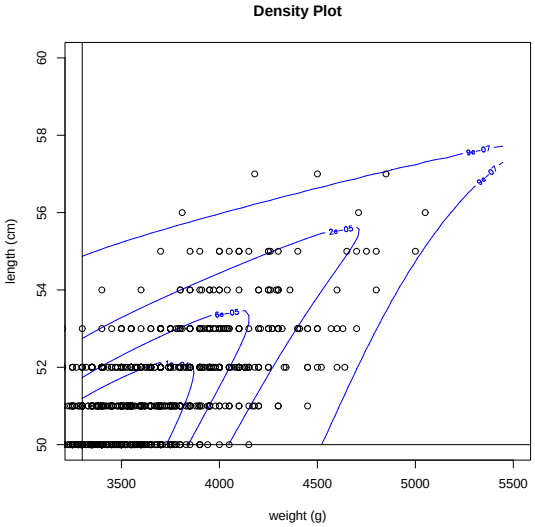


Figure 6.9: Probability density contours for the CL-fitted logistic model.

6.3 Model comparison

To compare the classical and HT models, we display density contour plots for regions where their extreme predictions are valid. For the classical methods this region is an area where both components are extreme (Fig.

6.9). On the other hand, the HT approach creates a different model for every component (Fig. 6.8) that is valid when that component is extreme. This enables us to choose an extreme model for every case when at least one component is extreme.

Density estimates for HT models were computed using the Monte Carlo method discussed in Section 4.4 with appropriate choices of sampling regions R. It should be mentioned that very many samples were needed to obtain smooth estimates in the double-extreme area where the two HT models overlap. Observe that this is precisely the area where the inherently smooth parametric model of the classical methods applies.

Although the classical approach came out symmetric (in normalized Fréchet coordinates), the conditional approach is overfitting quantization instabilities and its parameters are highly asymmetric. The classical approach seems to be more robust in this sense. Another problem might be the estimation incontinuity on the boundary between conditional models (might be remedied by smooth model transition along the boundary).

Finally, we'll illustrate the flexibility of the HT approach by computing a few functionals that could be of interest during dataset interpretation but cannot be satisfactorily produced by other methods.

The first functional will be the full conditional distribution of one component given that the other exceeds a fixed extreme threshold (Figures 6.10 to 6.13). This kind of indicator would be useful in all applications of extreme value theory, since it allows us to model joint extremes of a few selected components only. The classical models can predict only the rate of all extremes occuring at once.

The final indicator g, suggested by the original authors (in Section 6.2.3 of [5]), merges all components of the dataset into a single measure that combines their extreme properties. For an observation $\boldsymbol{Y}$ transformed to standard Gumbel margins, $g(\boldsymbol{Y})$ is computed as the sum of marginal quantiles: $g = \sum_{1 \leq j \leq d} Y_j$. Figure 6.14 plots exceedance probabilities for levels of g for our dataset and Figure 6.15 depicts the 10% largest observations in this criterion.

6.4 Implementation details

There are three libraries dealing with classical extreme value theory available for the statistical package `R`: `POT` by M. Ribatet, `evd` by

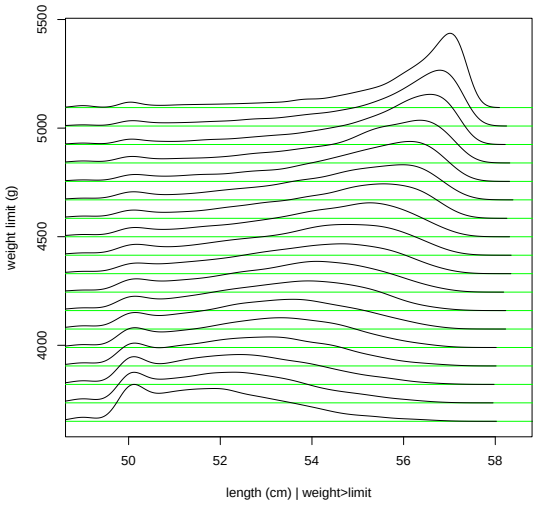


Figure 6.10: Marginal conditional distributions of length when weight exceeds a given limit.

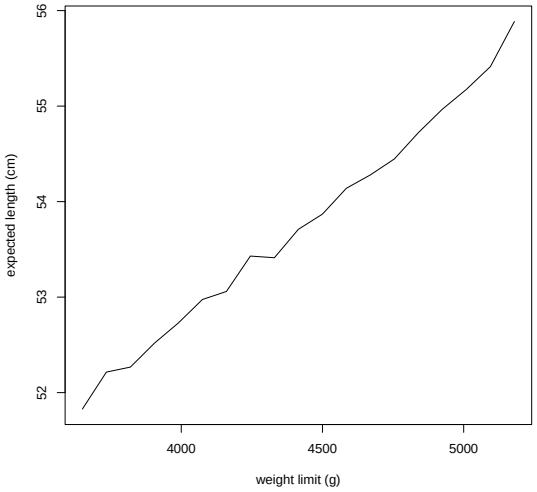


Figure 6.11: Expectations of length when weight exceeds a given limit.

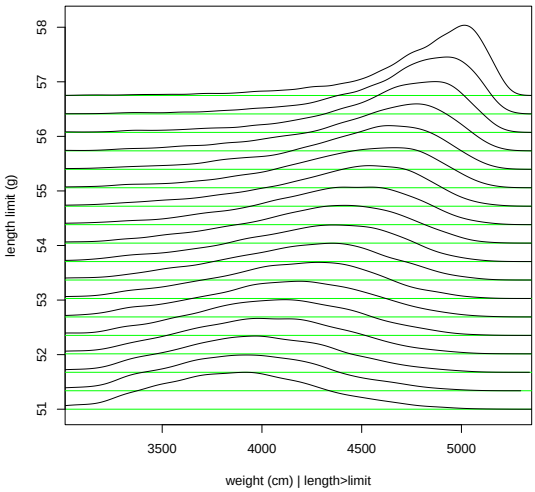


Figure 6.12: Marginal conditional distributions of width when length exceeds a given limit.

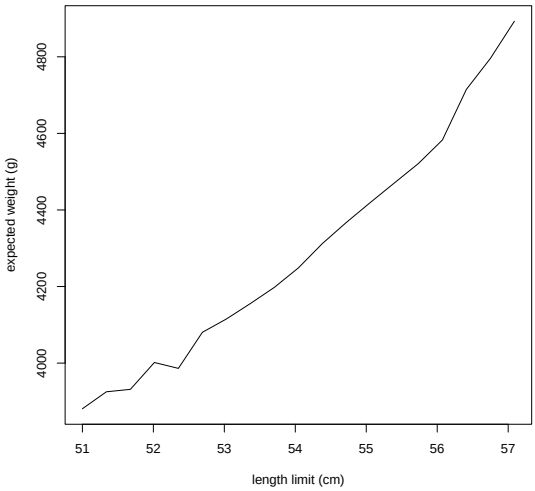


Figure 6.13: Expectations of weight when length exceeds a given limit.

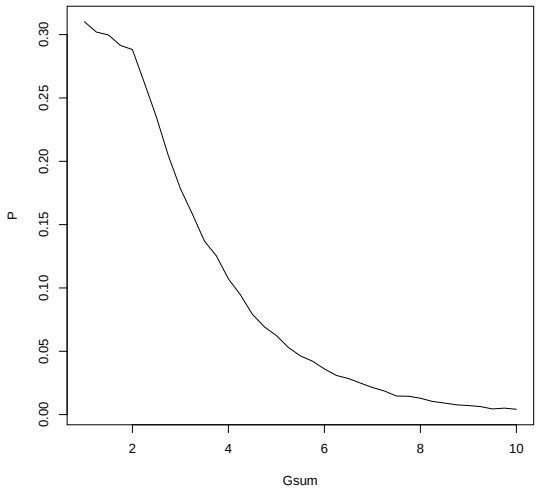


Figure 6.14: Probabilities $P(\boldsymbol{Y} \in R)$ for extreme sets $R : g(\boldsymbol{Y}) \geq Gsum$.

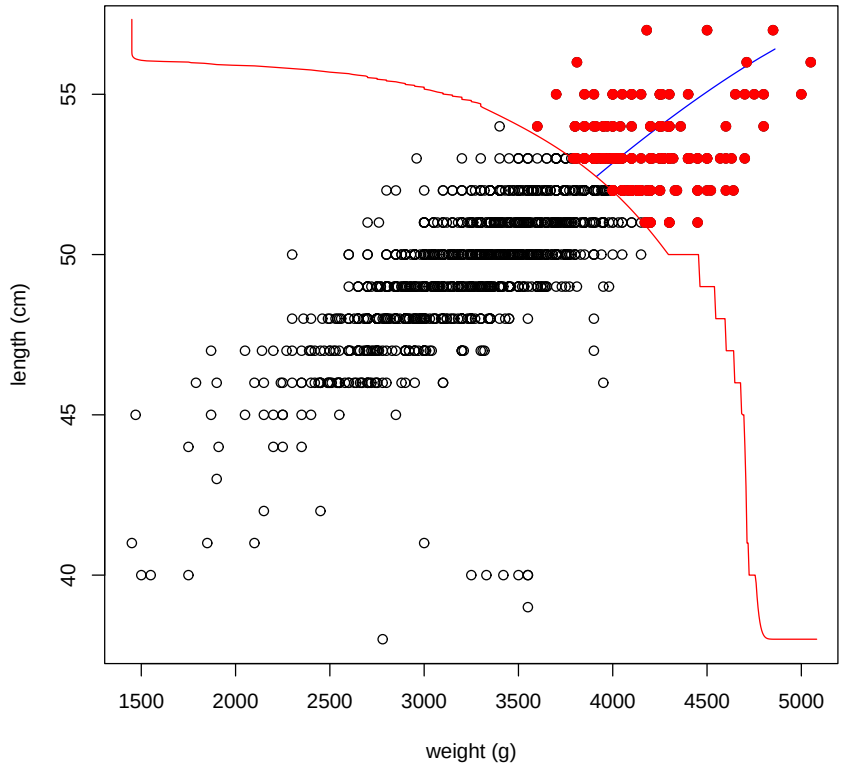


Figure 6.15: The extreme region R: $g(\boldsymbol{Y} \in R) \geq 4.11$ (threshold chosen so that $P(\boldsymbol{Y} \in R) \sim 0.1$). The blue line divides regions with dominant weight and dominant length (R_1 and R_2 as defined in Section 4.4).

A. Stephenson, and `evir`, an `R` port of A. McNeil's library `evis` for the statistical program `S`. Since `evir` is currently orphaned and its functions offer only a subset of the other libraries, we've used `POT` and `evd` in the application and simulation study.

The library `POT` provides many marginal parameter fit metrics and its estimation implementations seem to be more robust (`evd`'s `tcplot` was often unable to converge). It was used for threshold selection and sample generation for simulations.

The library `evd` offers more bivariate dependence models. Parameter values for dependence models of the classical approach correspond to our definitions (with the exception of the asymmetric negative logistic model with `dep` $= -1/\alpha$). The parameter of the Pickands dependence function is inverted, so asymmetry parameters come out swapped.

The authors of `evd` didn't provide Poisson process likelihood maximization for asymmetric models, arguing that there is no consensus on how to deal with point masses arising in these models and that CL makes better fits anyway [18]. Our observations second their reasoning.

On the other hand, Heffernan's and Tawn's approach [5] has no public implementation and the authors were unable to provide us with the original one, so we refer the reader to our own interpretation in `R` (available on the accompanying CD).

Chapter 7

Conclusion

We've summarized univariate and multivariate extreme value theory, described parametric classical methods using Poisson point processes and censored likelihood and compared them with an experimental conditional semi-parameric approach. We tested the methods in a simulation study, applied them to a children dataset and computed several extremal estimates.

The classical methods can answer only queries that are extremal in all components. This could pose a problem when the samples have a high dimension, since samples in the joint extremal region are rare, rendering the model fits inadequate. Also, there exist important applications for which "extreme" means "extreme in any coordiate" and marginal dependencies are anything but negligible.

The semi-parametric method from Heffernan and Tawn provides a means to tackle these kinds of problems. However, its dependence model uses many ad-hoc assumptions that need a more solid theoretical background before it can be brought to widespread use. The fact that functionals are estimated with a Monte Carlo method and need many (relatively slow) sample evaluations for suitable precision levels can also pose a problem.

We've shown that limit assumptions start to hold already for medium amounts of samples. For low data amounts a model simplification is needed to ensure stability of parameter estimates.

Both approaches have shown areas where they shine and where they break down. We believe that the complemental nature of the two methods extends the set of solvable problems in this field.

Bibliography

[1] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: *Statistics of Extremes*, Wiley Series in Probability and Statistics, 2005.

[2] Coles, S. G.: *An Introduction to Statistical Modeling of Extreme Values*, Springer Series in Statistics, 2001.

[3] Coles, S. G., Tawn, J, A.: *Modelling Extreme Multiariate Events*, J. R. Statist. Soc. B, **53**, No. 2 (1991), 377–392.

[4] de Haan, L.: *On Regular Variation and its Application to the Weak Convergence of Sample Extremes*, Mathematical Centre Tract, 32, Amsterdam, 1970.

[5] Heffernan, J. E., Tawn, J, A.: *A conditional approach for multivariate extreme values*, J. R. Statist. Soc. B, **66**, Part 3 (2004), 497–546.

[6] Joe, H.: *Families of min-stable multivariate exponential and multivariate extreme value distributions*, Statistics & Probability Letters, **9** (1990), 75–81.

[7] Klüppelberg, C., May, A.: *The dependence function for bivariate extreme value distribution—a systematic approach*, Unpublished manuscript (1999).

[8] Embrechts, P., Klüpellberg, C., Mikosch, T. *Modeling Extremal Events*. Springer-Verlag (1997).

[9] Ledford, A. W., Tawn, J. A.: *Statistics for near independence in multivariate extreme values*, Biometrika, **83** (1996), 169–187.

[10] Tawn, J, A.: *Bivariate Extreme Value Theory: Models and Estimation*, Biometrika, **75**, No. 3 (Sep. 1988), 397–415.

[11] Tawn, J, A.: *Modelling Multivariate Extreme Value Distributions*, Biometrika, **77**, No. 2 (Jun. 1990), 245–253.

[12] Pickands, J.: *Statistical inference using extreme order statistics*, Ann. Statist., **3** (1975), 119–131.

[13] Pickands, J.: *Multivariate extreme value distributions*, Bulletin of the International Statistical Institute, Proceedings of the 43rd Session (1981), Buenos Aires, 859–878.

[14] Resnick, S.: *Extreme Values, Point Processes and Regular Variation*, Springer Verlag, New York (1987).

[15] Smith, R. L.: *Extreme value theory*, Handbook of Applicable Mathematics, **7** (1990), Wiley.

[16] Smith, R. L.: *Max-stable Processes and Spatial Extremes*, Unpublished manuscript (1990).

[17] Smith, R. L.: *Multivariate threshold methods*, Extreme Value Theory and Applications (1994), 225-248, Kluwer.

[18] Stephenson, A., Ferro, C.: personal correspondence.

Appendix: Proof of Fisher-Tippett theorem

Lemma 1 ([4], Lemma 2.2.3)
Suppose F is a non-degenerate distribution function with $F(0-) = 0$. If there exist positive constants c_n such that

$$F^n(c_n x) = F(x), \quad x \in \mathbb{R}^+, n \in \mathbb{N}.$$

Then there exist constants $\alpha > 0$ and $C > 0$ such that

$$F(x) = \exp\{-(x/C)^\alpha\} \quad \text{for all positive } x.$$

This lemma can be proved by means of Theorem 3. Details can be found in [4].

Proof of Fisher-Tippett theorem: ([4], Proof of Theorem 2.2.1)
By using the limit property of max-stable laws (Theorem 1), the problem reduces to finding all max-stable distributions. We'll denote such a non-degenerate distribution function

$$G(x) = G^n(c_n x + d_n). \tag{A-1}$$

We distinguish three possibilities:

1. Suppose $c_n = 1$ for all n. We define $a_n = e^{d_n}$ for $n \in \mathbb{N}$ and

$$H(y) = \begin{cases} 0 & y \leq 0 \\ G(\log y) & y > 0. \end{cases}$$

   Then, for $y > 0$ and $n \in \mathbb{N}$, $H^n(a_n y) = H(y)$, hence by Lemma 1 the function G is of type Λ.

2. Suppose there exists an $r > 1$ such that $c_r < 1$. We shall prove that G is of type Ψ_α for some positive α. The proof is given in four steps:

- First we prove that

$$G(x) = 1 \quad \text{for } x \geq d_r/(1 - c_r). \tag{A-2}$$

We can see that

$$x \geq \frac{d_r}{1 - c_r} \;\Rightarrow\; c_r x + d_r \leq x \;\Rightarrow\; G(c_r x + d_r) \leq G(x).$$

With (A-1) this gives for these values of x

$$G^r(x) \leq G(x) = G^r(c_r x + d_r) \leq G^r(x).$$

This is true only if $G(x) = 0$ or 1. As G is non-degenerate, (A-2) must hold.

- Now we prove

$$0 < G(x) < 1 \quad \text{for } x < d_r/(1 - c_r). \tag{A-3}$$

Suppose there exists an $x_0 < d_r/(1 - c_r)$ with

$$\begin{cases} G(x) = 1 & x \geq x_0, \\ G(x) < 1 & x < x_0. \end{cases} \tag{A-4}$$

Take

$$x_1 = x_0 - \frac{d_r - x_0\,(1 - c_r)}{2c_r},$$

then we have

$$x_1 < x_0 < c_r x_1 + d_r$$

so that

$$G(x_1) = G^r(c_r x_1 + d_r) \geq G^r(x_0) = 1$$

which contradicts $G(x_1) < 1$. Hence $G(x) < 1$ for $x < d_r/(1 - c_r)$. We can prove $G(x) > 0$ in an analogous way for all x.

- Next we prove $c_n < 1$ for all n. Suppose there exists an s with $c_s = 1$, then for all x

$$G^s(x + d_s) = G(x). \tag{A-5}$$

Take in (A-5) $x = x_2$ with $0 < G(x_2) < 1$, then it follows that $x_2 + d_s > x_2$ or $d_s > 0$. Substitution of

$$\frac{d_r}{1 - c_r} - \frac{d_s}{2}$$

for x in (A-5) gives (see (A-2) and (A-3))

$$1 = G^s\left(\frac{d_r}{1-c_r} - \frac{d_s}{2}\right) = G\left(\frac{d_r}{1-c_r} - \frac{d_s}{2}\right) < 1.$$

Hence $c_s = 1$ is impossible.

Finally suppose there exists an t with $c_t > 1$, then an analogous reasoning as in the first two steps of 2) shows

$$\begin{cases} G(x) = 0 & x < d_t/(1-c_t), \\ 0 < G(x) < 1 & x > d_t/(1-c_t). \end{cases}$$

This contradicts (A-2).

- As (A-2) and (A-3) hold not only for r but for arbitrary $n > 1$, we have

$$\frac{d_n}{1-c_n} = \frac{d_2}{1-c_2}, \quad n = 2, 3, \ldots.$$

We define

$$\tilde{G}(z) = \begin{cases} 0 & z \le 0 \\ G\left(\dfrac{d_2}{1-c_2} - \dfrac{1}{z}\right) & z > 0. \end{cases}$$

Take

$$z = \frac{1}{d_2/(1-c_2) - x} \quad \text{for } x < d_2/(1-c_2),$$

then we have for all $z > 0$

$$\begin{aligned} \tilde{G}^n\left(\frac{z}{c_n}\right) &= G^n\left(\frac{d_n}{1-c_n} - \frac{c_n}{z}\right) = G^n(c_n x + d_n) = G(x) = \\ &= G\left(\frac{d_n}{1-c_n} - \frac{1}{z}\right) = \tilde{G}(z). \end{aligned}$$

Lemma 1 then shows that G is of type Ψ_α for some positive α.

3. Suppose there exists an $r > 1$ such that $c_r > 1$. In an analogous way as in 2) one finds that G is of type Φ_α for some positive α.