Modelling extreme-value dependence in high dimensions using threshold exceedances

Anna Kiriliouk

A thesis submitted to the
Université catholique de Louvain
in partial fulfillment of the requirements for the degree of

Doctor of Sciences

Thesis committee:
Johan Segers (Université catholique de Louvain), Supervisor
Michel Denuit (Université catholique de Louvain), Supervisor
Philippe Lambert (Université catholique de Louvain)
Pierre Devolder (Université catholique de Louvain)
Tim Verdonck (Katholieke Universiteit Leuven)
John Einmahl (Tilburg University)
Holger Rootzén (Chalmers University)

August 2016

Acknowledgements

First of all, my thanks goes to Johan Segers, for being everything one could wish from a supervisor. He has always been available when I needed him to, and his guidance, support and advice were crucial for the success of this thesis. I feel lucky to have collaborated with someone filled with so many inspiring ideas. I am also grateful to Michel Denuit, my co-supervisor, who was the driving force behind the last chapter of this thesis. Other parts of this thesis are the result of fruitful and pleasant collaborations with John Einmahl and Holger Rootzén, whom I had the chance to visit several times. I am grateful for all their ideas and feedback. Finally, I would like to thank the other members of my thesis committee: Philippe Lambert, Tim Verdonck, and Pierre Devolder, for their constructive comments that improved the contents of this thesis.

The Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA) of the Université catholique de Louvain (UCL) is a very welcoming and supporting work environment. After arriving in Belgium, all members of ISBA made it easy to integrate and to learn French, making my daily thesis life so enjoyable: Adrien, Alain, Aleksandar, Aurélie, Benjamin, Cédric, Florian, François, Gildas, Hélène, Joris, Josefine, Maïlis, Manon, Marco, Michał, Mickaël, Nathalie, Nathan, Nicolas, Samuel, Sylvie, and Vincent, thank you! A special mention goes to Michał: our paths have been crossing since 2008, and his prediction that we would once write a book together was partly realised in our collaboration for a book chapter. I would equally like to thank my friends from the Netherlands, whom unfortunately I don't get to see very often: Abel, Corine, Hanne, Jaap, Mike, Naomie, and Tjapko, thanks!

I am very grateful to my parents, who have always been my greatest inspiration. Ever since, as a child, I saw my mother finishing her PhD, I have always wanted to do one too — and for the next stages of my life, I consider my father to be the example of a researcher that I would like to become.

Most of all, I would like to thank Benjamin for all his love and support. The last year of my thesis was such a happy one because I had you by my side.

Introduction

Extreme-value theory is the branch of statistics concerned with the characterization of *extreme events*. Extreme events are encountered in a large variety of fields, such as hydrology, meteorology, finance, and insurance, but also in less obvious settings like air pollution or athletic records.

Think for instance of the world financial crisis in 2008/2009, often considered to be the worst financial crash since the 1930s. This crisis is sometimes partly blamed on the so-called Gaussian copula model, also referred to as "the formula that killed Wall Street" (Salmon, 2009). This model, introduced in Li (2001), is intended for dependence modelling of financial objects which before were considered extremely difficult to price. It relies entirely on the Gaussian distribution, and thus on the concept of *correlation*, destined to measure dependence between financial instruments around their mean values. However, a Gaussian structure cannot properly account for *tail dependence*, that is, it does not account for the joint occurrence of (very) extreme losses on these financial objects. It is exactly due to strong tail dependence that financial instruments often crash jointly. Moreover, financial crashes tend to be more severe and more frequent over the last decades, which is inevitably due to increasingly refined financial instruments that are strongly connected to one another.

An increase in the frequency and magnitude of (joint) extreme events cannot only be seen in finance, but also in the environment, where extreme weather conditions such as heat waves, floods and hurricanes are often linked to the concept of global warming. Consider for example the 1953 North Sea flood, leading to the death of over 2500 people. More than a third of the Netherlands is below sea level, which is why the country relies heavily on its dykes. In February 1953, dykes were breached all along the coast and almost 2000 people got killed. This was the direct motivation to the construction of the Delta Works, an intricate system of dams and storm surge barriers, and one of the first large scale extreme value projects. The heights of the dykes were calculated such that the probability of exceeding at least one of the dykes is sufficiently low, which in case of the Delta Works is defined to be a one in a ten thousand year event.

Although the two examples above seem to concern quite different situations, estimating the probability of a future joint stock crash or of a dyke breach

requires the same mathematical toolbox. In classical statistical theory, it is often the behaviour of the mean or average that is of interest. However, for heavy-tailed data, to which financial returns usually belong, the second moment or even the mean might be infinite, so that the classical theory based on the normal distribution is no longer relevant. Moreover, we will often want to estimate the probability of an event more severe than the ones we saw in the past: for instance, we are aiming to define the height of a dyke such that it is breached once every ten thousand years on average, with only a hundred years of data available.

The exact definition of an extreme event varies with the application one has in mind: in the above examples, an event is called extreme during the financial crisis if *many* stocks crash at the same time, whereas an extreme flood event is one where *at least one* of the dykes is breached. Moreover, when selecting extreme events, several strategies are possible: one might select the *maxima* over a certain fixed time span, e.g., yearly maximum water heights along the coast, or one might select all values *exceeding a large threshold*, e.g., all stock returns with a weekly loss of at least 10%. These concepts will be made more concrete in the first chapter, which aims at giving an introductory, though far from exhaustive, overview of the field of extreme value theory: only the concepts that are necessary to understand the subsequent chapters are presented, and illustrated with the help of a financial dataset. Section 1.3 is inspired by the book chapter

> Nonparametric estimation of extremal dependence (2016). Kiriliouk, A., Segers, J., and Warchoł, M. In: *Extreme Value Modelling and Risk Analysis: Methods and Applications*, D. Dey and J. Yan (eds), CRC Press.

Chapter 2 proposes a semi-parametric method for estimating the parameters of spatial tail dependence models, based on minimizing the distance between a vector of integrals of parametric pairwise tail dependence functions and the vector of their empirical counterparts. This method provides an alternative to the use of (pairwise) likelihood methods, that are difficult to handle for large dimensions. This chapter is based on the paper

> Einmahl, J.H., Kiriliouk, A., Krajina, A., and Segers, J. (2016a). An M-estimator of spatial tail dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):275–298.

Chapter 3 proposes an estimator which is more flexible than the one from Chapter 2: instead of comparing vectors of integrals, one compares vectors of tail dependence functions evaluated at finitely many points to any nonparametric estimator thereof, allowing to go beyond the pairwise setting and reach even higher dimensions. Moreover, we do not restrict our attention to spatial models, but illustrate the performance of the estimator on a non-differentiable model defined on a directed graph. This chapter is based on

Einmahl, J.H., Kiriliouk, A., and Segers, J. (2016b). A continuous updating weighted least squares estimator of tail dependence in high dimensions. Available at `http://arxiv.org/abs/1601.04826`.

Chapter 5 introduces multivariate generalized Pareto distributions, leading to the multivariate analogue of the peaks-over-threshold method introduced in Subsection 1.2.2. Contrary to the methods proposed in Chapters 2 and 3, we make inference using censored maximum likelihood estimation. An application aimed at the modelling of landslides in northern Sweden is presented. This chapter, which starts with a summary of the results in Rootzén, Segers, and Wadsworth (2016), corresponds to a paper in preparation, jointly with Holger Rootzén, Johan Segers, and Jennifer Wadsworth.

Chapter 6 treats a quite different subject, focusing on the modelling of dependence between joint defaults in credit risk. These defaults are usually modelled as a linear function depending on a latent factor which is assumed to follow a normal distribution. We propose to model them instead as a function of the maximum of two Gumbel-distributed latent variables, and we illustrate the benefits of our model on a classical credit risk data set provided in Standard and Poor's (2001). This chapter is based on the paper

Denuit, M., Kiriliouk, A. and Segers, J. (2015). Max-factor individual risk models with application to credit portfolios. *Insurance: Mathematics and Economics*, 62:162–172.

Finally, the Appendix illustrates the use of the R package tailDepFun, which implements the estimators proposed in Chapters 2 and 3.

Kiriliouk, A. (2016). *tailDepFun: Minimum Distance Estimation of Tail Dependence Models.* R package version 1.0.0.

Table of contents

Chapter 1

Statistics of Extremes

1.1 Preliminaries

Let X be a random variable with distribution function $F(x) = \mathbb{P}[X \leq x]$, denoted $X \sim F$. The *quantile function* is

$$F^{-1}(q) = \inf \{x : F(x) \geq q\}.$$

Let $\overline{F}(x) = 1 - F(x)$ denote the *survival function*. Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) copies of X. The *rank* of X_i among $X_1, \ldots, X_n$ is denoted by $R_{i,n}$ and is defined as

$$R_{i,n} := \sum_{l=1}^{n} \mathbb{1}\{X_l \leq X_i\}. \tag{1.1.1}$$

The *empirical distribution function* $\widetilde{F}_n$ is defined as

$$\widetilde{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}.$$

When focusing on the ranks of a sample, we will usually use a modified version of the empirical distribution function, given by

$$\widehat{F}_n(x) := \frac{1}{n} \left(\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\} - 1 \right). \tag{1.1.2}$$

Let $\xrightarrow{\mathrm{d}}$ denote convergence in distribution and let $\overset{\mathrm{d}}{=}$ denote equality in distribution. Write $[x]_+ = \max(x, 0)$. For a set $\mathcal{A}$, let $\partial\mathcal{A}$ denote its boundary and $|\mathcal{A}|$ its cardinality. Throughout, we use bold symbols for d-variate vectors, e.g., $\boldsymbol{X} = (X_1, \ldots, X_d)$ and $\boldsymbol{a} = (a_1, \ldots, a_d) \in \mathbb{R}^d$. All operations involving vectors should be interpreted componentwise, e.g., $\boldsymbol{aX} = (a_1 X_1, \ldots, a_d X_d)$.

The cumulative distribution function and the density function of a d-variate normal random variable with mean $\mathbf{0}$ and covariance matrix Σ are denoted by $\Phi_d(\,\cdot\,;\Sigma)$ and $\phi_d(\,\cdot\,;\Sigma)$ respectively. Finally, let Δ_{d-1} denote the unit simplex defined by

$$\Delta_{d-1} := \{\boldsymbol{w} \in [0,\infty)^d : w_1 + \cdots + w_d = 1\}.$$

Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^d$ with marginal distribution functions $F_j(x_j) = \mathbb{P}[X_j \leq x_j]$ for $j \in \{1,\ldots,d\}$ and joint distribution function $F(\boldsymbol{x}) = \mathbb{P}[X_1 \leq x_1, \ldots, X_d \leq x_d]$. If the margins $F_1,\ldots,F_d$ are continuous, then the *copula* of $\boldsymbol{X}$ is defined as the joint cumulative distribution function of $F_1(X_1),\ldots,F_d(X_d)$ and is denoted by C:

$$C(\boldsymbol{u}) := \mathbb{P}[F_1(X_1) \leq u_1, \ldots, F_d(X_d) \leq u_d], \qquad \boldsymbol{u} \in [0,1]^d.$$

The copula C, which has uniform margins, is used to describe the dependence structure of $\boldsymbol{X}$. Sklar's theorem (Sklar, 1959) states that every multivariate distribution function F of a random vector $\boldsymbol{X}$ with continuous margins can be expressed in terms of its marginal distribution functions and a unique copula C,

$$F(\boldsymbol{x}) = \mathbb{P}[F_1(X_1) \leq F_1(x_1), \ldots, F_d(X_d) \leq F_d(x_d)] = C(F_1(x_1),\ldots,F_d(x_d)).$$

Conversely, because we assumed the margins to be continuous,

$$C(\boldsymbol{u}) = F\left(F_1^{-1}(u_1),\ldots,F_d^{-1}(u_d)\right), \qquad \boldsymbol{u} \in [0,1]^d.$$

A *point process* is a stochastic rule for the occurrence and position of point events. If $\{\boldsymbol{X}_i\}_{i\geq 1}$ denotes a collection of points occurring in a space $\mathcal{X} \subset \mathbb{R}^d$, then a point process N counts the number of points on subsets of $\mathcal{X}$,

$$N(\mathcal{A}) = \sum_{i\geq 1} \mathbb{1}\{\boldsymbol{X}_i \in \mathcal{A}\}, \qquad \mathcal{A} \subset \mathcal{X}.$$

The best known point process is the *Poisson process*. Let ν be a Borel measure on $\mathcal{X}$, finite for any compact $\mathcal{A} \subset \mathcal{X}$. Then N is a Poisson process on $\mathcal{X}$ with *intensity measure* ν if

1. For $k \in \mathbb{N}$ and $\mathcal{A}_1,\ldots,\mathcal{A}_k \subset \mathcal{X}$ disjoint, $N(\mathcal{A}_1),\ldots,N(\mathcal{A}_k)$ are independent random variables.
2. $N(\mathcal{A})$ has a Poisson distribution with mean $\nu(\mathcal{A})$.

The quantity $\nu(\mathcal{A})$ can be interpreted as the expected number of points of the Poisson process located in $\mathcal{A}$, i.e., $\nu(\mathcal{A}) = \mathbb{E}[N(\mathcal{A})]$. If $(N_i)_{i\geq 1}$ is a sequence of point processes on $\mathcal{A}$, we say that the sequence converges in distribution to a point process N, denoted

$$N_n \xrightarrow{\mathrm{d}} N, \qquad \text{as } n \to \infty,$$

if for every $m \in \mathbb{N}$ and for all bounded sets $\mathcal{A}_1,\ldots,\mathcal{A}_m \subset \mathcal{X}$ with $\mathbb{P}[N(\partial\mathcal{A}_j) = 0] = 1$ for $j = 1,\ldots,m$, the joint distribution of $(N_n(\mathcal{A}_1),\ldots,N_n(\mathcal{A}_m))$ converges in distribution to $(N(\mathcal{A}_1),\ldots,N(\mathcal{A}_m))$.

1.2 Univariate extremes

1.2.1 Generalized extreme-value distribution

Consider the stock prices of IBM between January 1, 1984, and December 31, 2015, downloaded from `http://finance.yahoo.com`, and convert them to weekly negative log-returns: if $P_1, \ldots, P_n$ is a series of stock prices, the negative log-returns are

$$X_i = -\log(P_i/P_{i-1}), \qquad i = 2, \ldots, n.$$

By taking negative log-returns we force (extreme) losses to be in the upper right tail of the distribution function. Figure 1.1 shows the time series of the stock prices and the weekly negative log-returns.

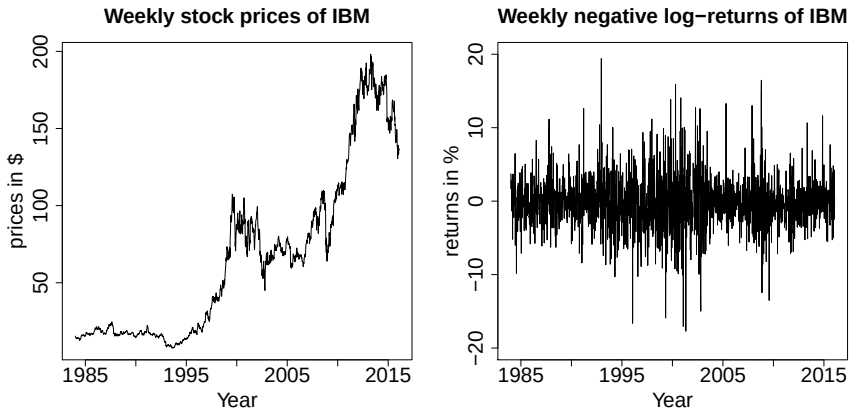


Figure 1.1: IBM weekly stock prices and negative log-returns.

Let $X_1, \ldots, X_n$ represent iid copies of a random variable X with continuous distribution function F. Being interested in the upper right tail means we wish to study high quantiles of F, i.e., values x_q such that $F(x_q) = 1 - q$ for some small q. An intuitive approach would be to estimate F by the empirical distribution function $\widehat{F}_n$ and to calculate $\widehat{F}_n^{-1}(1-q)$. However, we are often interested in events that are more extreme than the ones occurred in the past, so that the empirical distribution function is of no use. What we need is an extreme-value analogue of the central limit theorem, leading to a parametric limiting model for the maximum of a sample instead of the sum.

Consider $M_n = \max(X_1, \ldots, X_n)$. If n is the number of observations in, for instance, a year, then M_n represents the annual maximum. Then, for $F(x) < 1$,

$$\mathbb{P}[M_n \leq x] = F^n(x) \to 0, \qquad \text{as } n \to \infty,$$

so that the limiting distribution is degenerate. If there exist normalizing sequences $a_n > 0$, $b_n \in \mathbb{R}$, such that, as $n \to \infty$,

$$\mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = F^n(a_n x + b_n) \xrightarrow{\mathrm{d}} G(x), \tag{1.2.1}$$

and G is non-degenerate, then F is said to be in the *max-domain of attraction* of the *generalized extreme-value* (GEV) distribution G (Fisher and Tippett, 1928; Gnedenko, 1943). The distribution G has the form

$$G(x) = \exp\left\{-\left[1 + \xi\frac{(x-\mu)}{\sigma}\right]_+^{-1/\xi}\right\}, \qquad x \in \mathbb{R}.$$

The parameters μ and σ represent the location and scale respectively. The parameter ξ is the shape parameter, determining the tail behaviour of the distribution and the support of G:

$\xi > 0$ is the heavy-tailed *Fréchet* case; the support of the GEV distribution is $(\mu - \sigma/\gamma, \infty)$. The Pareto or log-gamma distributions are in the max-domain of attraction of a GEV distribution with $\xi > 0$;

$\xi = 0$ is the light-tailed *Gumbel* case; the support of the GEV distribution is $(-\infty, \infty)$. The exponential distribution is in the max-domain of attraction of a GEV distribution with $\xi = 0$;

$\xi < 0$ is the bounded-tail *Weibull* case; the support of the GEV distribution is $(-\infty, \mu - \sigma/\gamma)$. The beta or uniform distributions are in the max-domain of attraction of a GEV distribution with $\xi < 0$.

Example 1.2.1. If X has a unit Pareto distribution, $F(x) = 1 - 1/x$ for $x > 0$, and we set $b_n = 0$ and $a_n = n$, then

$$\mathbb{P}\left[M_n/n \leq x\right] = F^n(nx) = \left(1 - \frac{1}{nx}\right)^n \xrightarrow{d} \exp(-1/x), \qquad \text{as } n \to \infty,$$

so that the unit Pareto distribution is in the max-domain of attraction of the unit Fréchet distribution.

A related concept is that of max-stability: a distribution G is said to be *max-stable* if, for every $n \in \mathbb{N}$ there exist constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that

$$G^n(a_n x + b_n) = G(x), \qquad x \in \mathbb{R}. \tag{1.2.2}$$

It can be shown that a distribution is max-stable if and only if it is a generalized extreme-value distribution.

Inference on the tail of F can be made by assuming equality for large enough n in expression (1.2.1),

$$\mathbb{P}\left[M_n \leq x\right] \approx G((x - b_n)/a_n) = G_0(x),$$

where G_0 is some other GEV distribution. Then we can proceed as follows. Divide the data, $X_1, \ldots, X_n$, into k blocks of length m, $k \times m = n$, obtaining a sample $\boldsymbol{X}_{m,1}, \ldots, \boldsymbol{X}_{m,k}$. Calculating the maxima for the m consecutive blocks, we obtain a sequence of block maxima $M_{m,1}, \ldots, M_{m,k}$ to which we can fit a generalized extreme-value distribution G. Note that we do not need to estimate the normalization constants a_n and b_n since these are absorbed in the parameters of the GEV distribution. The block size is important for the quality of our estimates: too large a block size will lead to less datapoints and thus to a higher variance, whereas too small a block size will lead to a bad approximation by the limiting model and thus to a higher bias. Often, the block size is guided by the data: for instance, only the yearly maxima might be available. The parameters (μ, σ, ξ) of G can be then estimated by maximum likelihood (Smith, 1985; Prescott and Walden, 1980; Bücher and Segers, 2016) or, for instance, an approach using probability-weighted moments (Hosking et al., 1985; Hosking and Wallis, 2005).

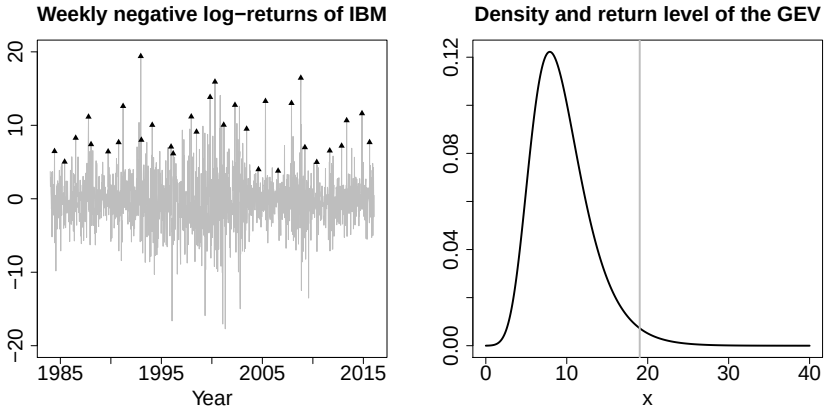


Figure 1.2: Yearly maxima (left) and the density and return level of the estimated GEV distribution (right) for the weekly negative log-returns of IBM.

As an illustration, we extract 32 yearly maxima from the weekly negative log-returns of the IBM stock prices (shown on the left-hand side of Figure 1.2) and estimate the parameters of the GEV distribution by maximum likelihood. We find $\mu = 7.8$ (0.62), $\sigma = 3.0$ (0.46), $\xi = -0.03$ (0.16), with standard errors in parentheses. Suppose that we are interested in the *return level* x_q, which satisfies $G(x_q) = 1 - q$ for small q. We say that $1/q$ is the *return period*; since the annual maximum exceeds x_q in a given year with probability q, we can say that approximately, x_q is expected to be exceeded on average once every $1/q$ years. If we're interested in the fifty-year return level, setting $1/q = 50$ years gives $x_q = 19.1$ (3.0); we can expect a weekly loss of at least 19.1%

once every fifty years, with a standard error of 3%. The right-hand side of Figure 1.2 shows the density function of the estimated GEV distribution with the fifty-year return level.

1.2.2 Peaks-over-thresholds methods

When restricting ourselves to block maxima, although many extreme events could have occurred in the same block, only one event per block is recorded. An alternative is to fix some high threshold $u \in \mathbb{R}$ and to consider all observations above u as extreme. This way, we might waste less data than with the block maxima method. Concretely, we wish to study the limiting distribution of $X-u \mid X > u$. If F is in the max-domain of attraction of a generalized extreme-value distribution G, then the conditional probability of threshold exceedances can be written, for $x > 0$,

$$\mathbb{P}[X \leq x + u \mid X > u] = \frac{F(u+x) - F(u)}{1 - F(u)} \to H(x) := 1 - \left(1 + \frac{\xi x}{\eta_u}\right)_+^{-1/\xi}, \tag{1.2.3}$$

as $u \to \infty$, where ξ is equal to the shape parameter of the GEV distribution and $\eta_u = \sigma + \xi(u - \mu)$ (Balkema and De Haan, 1974; Pickands III, 1975). We call H the *generalized Pareto* (GP) distribution. For $\xi \to 0$, we get $H(x) = 1 - \exp(-x/\eta)$.

Point processes can be used to derive expression (1.2.3) and to link it to the GEV distribution. Define a sequence of point processes on $\mathbb{R}$ as

$$N_n = \left\{ i \in \{1, \ldots, n\} : \frac{X_i - b_n}{a_n} \right\}.$$

This point process can be shown to converge to a Poisson process N, whose intensity measure can be calculated as follows. For a set $\mathcal{A} \subset \mathbb{R}$, we have $\lim_{n\to\infty} \mathbb{P}[N_n(\mathcal{A}) = 0] = \exp\{-\nu(\mathcal{A})\}$ since $N(\mathcal{A})$ follows a Poisson distribution. The intensity measure ν of this point process can then be found by considering regions of the form $\mathcal{A} = (x, \infty)$ for $x > 0$,

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}[N_n(\mathcal{A}) = 0] &= \lim_{n\to\infty} \mathbb{P}\left[\frac{X_1 - b_n}{a_n} \leq x, \ldots, \frac{X_n - b_n}{a_n} \leq x\right] \\ &= \lim_{n\to\infty} \mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = G(x), \end{aligned}$$

so that, for x large enough, the sequence of point processes N_n converges to a Poisson process with intensity

$$\nu(\mathcal{A}) = \left[1 + \xi\left(\frac{x - \mu}{\sigma}\right)\right]^{-1/\xi}.$$

We get for $i = 1, \ldots, n$ and $x > 0$,

$$H(x) = \lim_{n\to\infty} \mathbb{P}\left[\frac{X_i - b_n}{a_n} \leq x \,\middle|\, \frac{X_i - b_n}{a_n} > 0\right] \approx 1 - \frac{\nu\{(x,\infty)\}}{\nu\{(0,\infty)\}} = 1 - \frac{\log G(x)}{\log G(0)}, \tag{1.2.4}$$

which leads directly to (1.2.3).

After having fixed the threshold u, we can fit a GP distribution to the threshold exceedances through maximum likelihood estimation (Davison and Smith, 1990). Note that the choice of the threshold leads to an analogous bias-variance trade-off as the choice of the block size for the block maxima approach. In practice, the threshold is often chosen as a high quantile of the empirical distribution function of our data, although many advanced methods on threshold selection are available: see, for instance, Scarrott and MacDonald (2012) for a review.

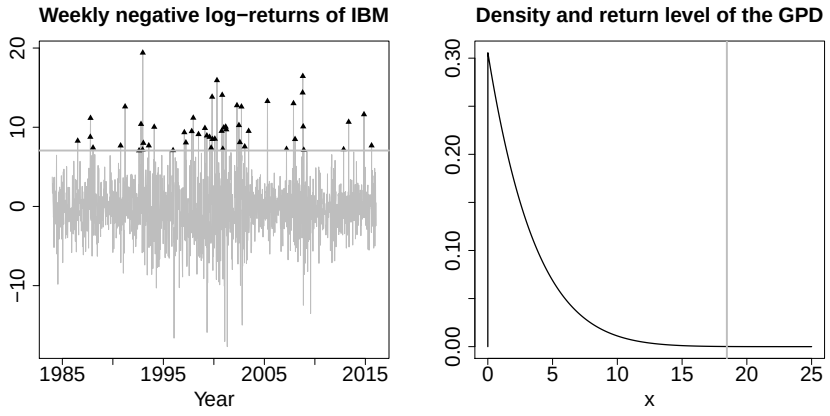


Figure 1.3: Threshold exceedances above $u = 7.06$ (left) and the density and return level of the estimated GP distribution (right) for the weekly negative log-returns of IBM.

When fitting a GP distribution, we assume that our data are independent, which is a reasonable assumption since we chose weekly log-returns. Let u be the 97% quantile of $X_1, \ldots, X_n$, $u = 7.06$. We select the values that are above u, shown on the left-hand side of Figure 1.3, obtaining $k = 51$ threshold exceedances. We estimate the parameters η and ξ by maximum likelihood, obtaining $\widehat{\eta} = 3.27$ (0.67) and $\widehat{\xi} = -0.11$ (0.15), where standard errors are in parentheses. If we estimate again the fifty-year return level, we find that we can expect a loss of at least 18.4% every fifty years with a standard error of

2.2%. Note that the estimated values for ξ and for the return level are very similar to the block maxima method. The right-hand side of Figure 1.3 shows the estimated GP density and the fifty-year return level.

1.3 Multivariate extremes

1.3.1 Tail dependence

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ denote a d-variate random variable in $\mathbb{R}^d$ with joint distribution function F and continuous marginal distribution functions $F_1, \ldots, F_d$. Before studying the dependence structure of $\boldsymbol{X}$, it is useful to eliminate the influence of marginal aspects so that we can focus on the dependence structure only. The probability integral transform produces variables $F_1(X_1), \ldots, F_d(X_d)$ that are uniformly distributed on the interval $(0, 1)$. Large values of X_j correspond to $F_j(X_j)$ being close to unity, whatever the original scale in which X_j was measured. Recall that this is exactly what is done when studying copulas. In the extreme-value set-up, it can be useful to magnify large values, so that we consider a further transformation to unit Pareto margins via $X_j^* = 1/\{1 - F_j(X_j)\}$ for $j = 1, \ldots, d$.

In practice, the marginal distributions are unknown and need to be estimated. This can be done parametrically, by fitting a GEV or GP distribution to block maxima or threshold excesses, after which the parameter estimates are used to transform the marginal variables to a common scale. An alternative is to estimate the margins nonparametrically, transforming them to unit Pareto margins by setting

$$\widehat{X}_{ij}^* \coloneqq \frac{1}{1 - \widehat{F}_{n,j}(X_{ij})}, \qquad i \in \{1, \ldots, n\},\ j \in \{1, \ldots, d\}. \tag{1.3.1}$$

The empirical distribution functions $\widehat{F}_{n,j}$ evaluated at the data are

$$\widehat{F}_{n,j}(X_{ij}) = \frac{R_{ij,n} - 1}{n}, \qquad i \in \{1, \ldots, n\},\ j \in \{1, \ldots, d\}; \tag{1.3.2}$$

where $R_{ij,n}$ is the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$; see (1.1.1) and (1.1.2).

Consider a financial portfolio containing three stocks: JP Morgan, Citibank and IBM. We again download stock prices from `http://finance.yahoo.com` between January 1, 1984, and December 31, 2015, and convert them to weekly negative log-returns. These series of negative log-returns will be denoted by the vectors $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$ for $i = 1, \ldots, n$. We transform them to the unit Pareto scale using expressions (1.3.1) and (1.3.2). Figure 1.4 shows the scatterplots of the weekly negative log-returns of JP Morgan versus Citibank and JP Morgan versus IBM on the unit Pareto scale, plotted on the logarithmic scale for better visibility. We observe that joint occurrences of large losses are more frequent for JP Morgan and Citibank (left) than for JP Morgan and

IBM (right), i.e., JP Morgan and Citibank exhibit a stronger degree of *tail dependence* than JP Morgan and IBM. This is no surprise given the fact that JP Morgan and Citibank are financial institutions while IBM is an IT company.

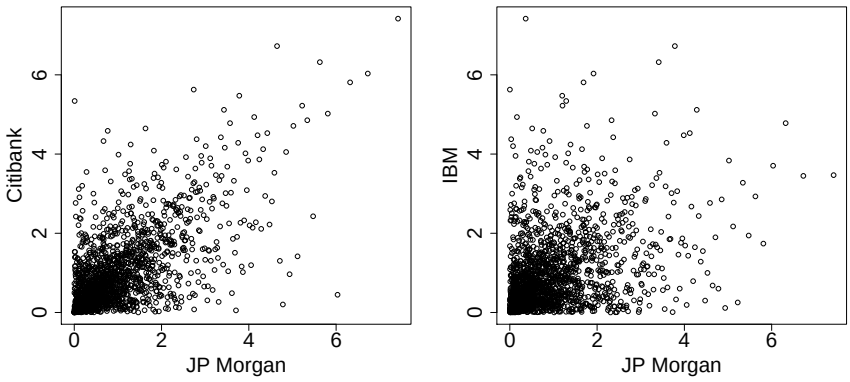


Figure 1.4: Scatterplots of weekly negative log-returns of stock prices of JP Morgan versus Citibank and JP Morgan versus IBM, plotted on the logarithmic scale.

In order to study tail dependence, we zoom in on the joint distribution of $(F_1(X_1), \dots, F_d(X_d))$ in the neighbourhood of its upper endpoint $(1, \dots, 1)$. That is, we look at

$$\begin{aligned} 1 - \mathbb{P}[F_1(X_1) \leq 1 - tx_1, \dots, F_d(X_d) \leq 1 - tx_d] \\ = \mathbb{P}[F_1(X_1) > 1 - tx_1 \text{ or } \cdots \text{ or } F_d(X_d) > 1 - tx_d], \end{aligned} \quad (1.3.3)$$

where $t > 0$ is small and where the numbers $x_1, \dots, x_d \in [0, \infty)$ parametrize the relative distances to the upper endpoints of the d variables. The above probability converges to zero as $t \to 0$. Since the marginal probabilities $\mathbb{P}[F_j(X_j) > 1 - tx_j]$ are equal to tx_j for all $j \in \{1, \dots, d\}$, we get that the probability above is bounded by

$$\begin{aligned} \max(tx_1, \dots, tx_d) &\leq \mathbb{P}[F_1(X_1) > 1 - tx_1 \text{ or } \cdots \text{ or } F_d(X_d) > 1 - tx_d] \\ &\leq tx_1 + \cdots + tx_d. \end{aligned}$$

We divide by t and obtain the *stable tail dependence function*, $\ell : [0, \infty)^d \to [0, \infty)$, defined by

$$\ell(\boldsymbol{x}) \coloneqq \lim_{t \downarrow 0} t^{-1} \mathbb{P}[F_1(X_1) > 1 - tx_1 \text{ or } \cdots \text{ or } F_d(X_d) > 1 - tx_d], \quad (1.3.4)$$

for $\boldsymbol{x} \in [0,\infty)^d$ (Huang, 1992; Drees and Huang, 1998). The existence of the limit in (1.3.4) is an assumption that can be tested for (Einmahl et al., 2006). Every stable tail dependence function ℓ has the following properties:

1. $\max(x_1,\ldots,x_d) \le \ell(\boldsymbol{x}) \le x_1 + \cdots + x_d$. In particular, $\ell(0,\ldots,0,x_j,0,\ldots,0) = x_j$ for $j = 1,\ldots,d$;
2. convexity, that is, $\ell\big(t\boldsymbol{x} + (1-t)\boldsymbol{y}\big) \le t\,\ell(\boldsymbol{x}) + (1-t)\,\ell(\boldsymbol{y})$, for $t \in [0,1]$;
3. order-one homogeneity: $\ell(ax_1,\ldots,ax_d) = a\,\ell(x_1,\ldots,x_d)$, for $a > 0$;

(Beirlant et al., 2004). When $d = 2$, these three properties characterize the class of stable tail dependence functions. When $d \ge 3$, a function satisfying these properties is not necessarily a stable tail dependence function (Molchanov, 2008; Ressel, 2013). For any dimension $d \ge 2$, the collection of d-variate stable tail dependence functions is infinite-dimensional. This poses challenges to inference on tail dependence, especially in higher dimensions. The usual way of dealing with this problem consists of considering parametric models for ℓ, a number of which are presented in Sections 1.3.4 and 1.4.2.

The probability in (1.3.3) represents the situation where *at least one* of the variables is large: for instance, the sea level exceeds a critical height at one or more coastal locations. Alternatively, we might be interested in the situation where *all* variables are large simultaneously. Think of the prices of all stocks in a financial portfolio going down together. The *tail copula*, $R : [0,\infty)^d \to [0,\infty)$, is defined by

$$R(\boldsymbol{x}) \coloneqq \lim_{t \downarrow 0} t^{-1}\mathbb{P}[F_1(X_1) > 1 - tx_1,\ldots,F_d(X_d) > 1 - tx_d], \tag{1.3.5}$$

for $\boldsymbol{x} \in [0,\infty)^d$ (Schmidt and Stadtmüller, 2006). Again, existence of the limit is an assumption. In the bivariate case, the functions ℓ and R are directly related by $R(x_1,x_2) = x_1 + x_2 - \ell(x_1,x_2)$. The difference between ℓ and R is visualized in Figure 1.5 for the log-returns of the stock prices of JP Morgan versus Citibank. From now on, we will focus on the function ℓ because of its direct link to the joint distribution function; see (1.3.9).

A random vector $\boldsymbol{X} \in \mathbb{R}^2$ is said to be *asymptotically independent* if the stable tail dependence function assumes its upper bound, $\ell(x_1,x_2) = x_1 + x_2$, for all $x_1, x_2 \in [0,\infty)$. The opposite case, $\ell(x_1,x_2) < x_1 + x_2$ for some $x_1, x_2 \in [0,\infty)$, is referred to as *asymptotic dependence*. Sibuya (1960) already observed that bivariate normally distributed vectors are asymptotically independent as soon as the correlation is less than unity. We find that in case of asymptotic independence, we cannot rely on the function ℓ to quantify the amount of dependence left above high but finite thresholds.

The best-known measure for the strength of asymptotic dependence of a random vector $\boldsymbol{X} \in \mathbb{R}^2$ is the *tail dependence coefficient* χ, defined as

$$\chi \coloneqq \lim_{u \uparrow 1} \mathbb{P}[F_1(X_1) > u \mid F_2(X_2) > u].$$

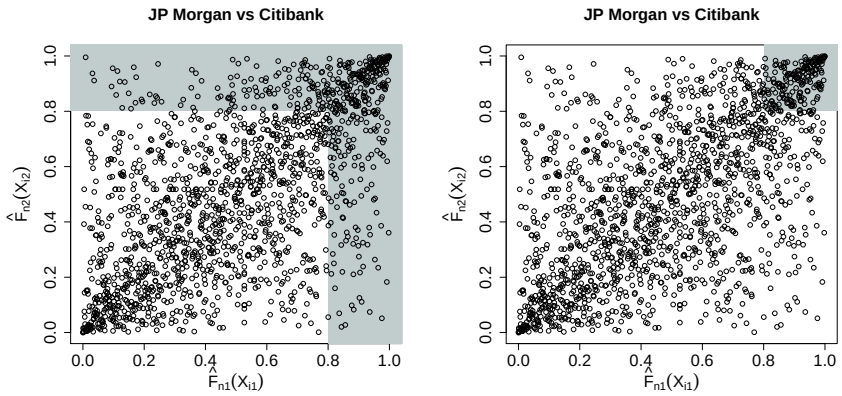


Figure 1.5: Scatterplots of negative weekly log-returns of stock prices of JP Morgan versus Citibank, transformed using ranks via (1.3.2). Left: the region where at least one variable is large, inspiring the definition of the stable tail dependence function ℓ in (1.3.4). Right: the region where both variables are large, inspiring the definition of the tail copula R in (1.3.5).

(Coles et al., 1999). We can write χ as the limit of a function $\chi(u)$ such that $\lim_{u\uparrow 1}\chi(u) = \chi$; the function $\chi(u)$ is

$$\chi(u) := 2 - \frac{\mathbb{P}[F_1(X_1) > u \text{ or } F_2(X_2) > u]}{1-u}. \tag{1.3.6}$$

Setting $1-u = t$ and letting $t \downarrow 0$, we see that $\chi = 2-\ell(1,1) = R(1,1)$. By the properties of ℓ, we have $\ell(x_1, x_2) = x_1 + x_2$ for all $x_1, x_2 \in [0, \infty)$ if and only if $\ell(1,1) = 2$. That is, asymptotic independence is equivalent to $\chi = 0$, whereas asymptotic dependence is equivalent to $\chi \in (0, 1]$. Before fitting a dataset to some parametric extreme-value model, we need to make sure the data are not asymptotically independent, for instance by estimating $\chi(u)$ nonparametrically and investigating its behaviour as $u \to 1$ (see Section 1.3.2). In case the data are asymptotically independent, other types of models are needed; see Ledford and Tawn (1996), Ramos and Ledford (2009), or Wadsworth et al. (2016) among others.

1.3.2 Nonparametric estimation of tail dependence

Given the random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$, our aim is to estimate the stable tail dependence function ℓ in (1.3.4). A straightforward nonparametric estimator can be defined as follows. Let $k := k_n \in \{1, \ldots, n\}$ be such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. Replacing $\mathbb{P}$ by the empirical distribution function,

t by k/n, and $F_1, \ldots, F_d$ by $\widehat{F}_{n,1}, \ldots, \widehat{F}_{n,d}$ as defined in (1.3.2), we obtain the *empirical tail dependence function* (Huang, 1992; Drees and Huang, 1998)

$$\begin{aligned}\widehat{\ell}'(\boldsymbol{x}) &\coloneqq \frac{n}{k}\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\widehat{F}_{n,1}(X_{i1}) > 1 - \frac{kx_1}{n} \text{ or } \cdots \text{ or } \widehat{F}_{n,d}(X_{id}) > 1 - \frac{kx_d}{n}\right\} \\ &= \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}\left\{R_{i1,n} > n + 1 - kx_1 \text{ or } \cdots \text{ or } R_{id,n} > n + 1 - kx_d\right\}.\end{aligned}$$

Under minimal assumptions, the estimator is consistent and asymptotically normal with a convergence rate of $\sqrt{k}$ (Einmahl et al., 2012; Bücher et al., 2014). Alternatively, an estimator with slightly better finite-sample properties (Einmahl et al., 2012) is

$$\widehat{\ell}(\boldsymbol{x}) \coloneqq \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}\left\{R_{i1,n} > n + 1/2 - kx_1 \text{ or } \cdots \text{ or } R_{id,n} > n + 1/2 - kx_d\right\}.$$

To estimate $\chi(u)$ from a sample $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$, simply replace $\mathbb{P}$, F_1 and F_2 by their empirical counterparts in (1.3.6),

$$\widehat{\chi}(u) \coloneqq 2 - \frac{1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\widehat{F}_{n,1}(X_{i1}) \le u, \widehat{F}_{n,2}(X_{i2}) \le u\right\}}{1 - u}. \tag{1.3.7}$$

We will use this estimator in Section 1.3.4 to quantify the amount of dependence between our series of weekly negative log-returns.

A way to visualize the function ℓ or an estimator thereof in two dimensions is via the level sets $\mathcal{D}_c \coloneqq \{(x_1, x_2) : \ell(x_1, x_2) = c\}$ for a range of value of $c > 0$. Note that the level sets are between the lines $x_1 + x_2 = c$ and the elbow curves $\max(x_1, x_2) = c$. Likewise, a plot of the level sets of an estimator of ℓ can be used as a graphical diagnostic of asymptotic (in)dependence; see de Haan and de Ronde (1998) or de Haan and Ferreira (2006, Section 7.2).

We plot the lines $\mathcal{D}_c$ for $c \in \{0.2, 0.4, 0.6, 0.8, 1\}$ and $k = 50$ of $\widehat{\ell}_{n,k}$ for the weekly negative log-returns of JP Morgan versus Citibank and JP Morgan versus IBM in Figure 1.6. The grey lines represent $x_1 + x_2 = c$ and $\max(x_1, x_2) = c$. The level sets for JP Morgan versus IBM resemble the straight lines $x + y = c$ much more closely than the level sets for JP Morgan versus Citibank do. There are several formal approaches to test for asymptotic (in)dependence; see for instance Draisma et al. (2004) or Hüsler and Li (2009).

Other nonparametric estimation methods for tail dependence can be found in Gudendorf and Segers (2011), Gudendorf and Segers (2012), or Marcon et al. (2016), among others. Nonparametric estimators for ℓ can also serve as a stepping stone for semiparametric inference. Assume that $\ell \in \{\ell_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ for some $\Theta \subset \mathbb{R}^p$, a finite-dimensional parametric family. Then one could estimate $\boldsymbol{\theta}$ by minimizing a distance or discrepancy measure between $\widehat{\ell}_{n,k}$ and members of the parametric family. This is the subject of Chapters 2 and 3.

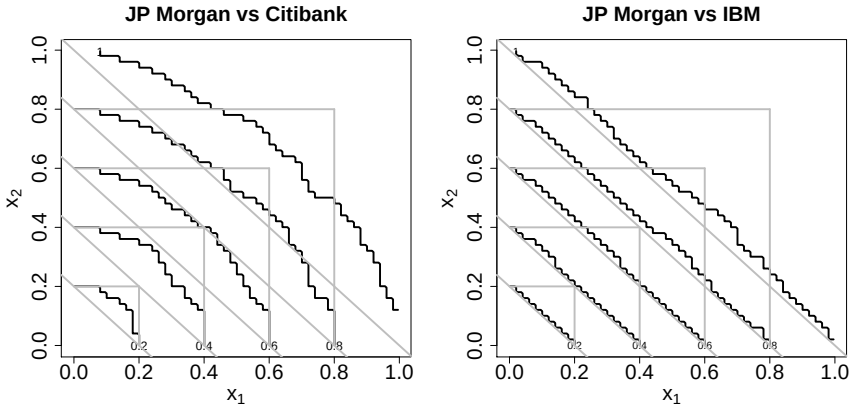


Figure 1.6: Level sets $\mathcal{D}_c$ of $\widehat{\ell}_{n,k}(x_1, x_2)$ for $c \in \{0.2, 0.4, 0.6, 0.8, 1\}$ for the weekly negative log-returns of JP Morgan versus Citibank (left) and JP Morgan versus IBM (right).

1.3.3 Multivariate maxima, point processes and the exponent measure

In our treatment of tail dependence thusfar we have not mentioned taking maxima, although this was our primary approach in the univariate case. We now present the analogue of a univariate generalized extreme-value distribution, and show how it is linked to the stable tail dependence function.

Let $\boldsymbol{M}_n = (M_{n,1}, \ldots, M_{n,d})$ with $M_{n,j} := \max(X_{1,j}, \ldots, X_{n,j})$ for $j = 1, \ldots, d$ denote the componentwise maxima. Note that these do not necessarily correspond to existing data points. On the left-hand side of Figure 1.7 the yearly componentwise block maxima of JP Morgan and Citibank are shown. If there exist sequences of normalizing constants $\boldsymbol{a}_n = (a_{n1}, \ldots, a_{nd}) > \boldsymbol{0}$ and $\boldsymbol{b}_n = (b_{n1}, \ldots, b_{nd}) \in \mathbb{R}^d$ such that,

$$\mathbb{P}\left[\frac{\boldsymbol{M}_n - \boldsymbol{b}_n}{\boldsymbol{a}_n} \leq \boldsymbol{x}\right] = F^n\left(\boldsymbol{a}_n \boldsymbol{x} + \boldsymbol{b}_n\right) \xrightarrow{d} G(\boldsymbol{x}), \qquad \text{as } n \to \infty, \tag{1.3.8}$$

and G is non-degenerate, then G is a *multivariate generalized extreme-value distribution*, and F is said to be in the *max-domain of attraction* of G. This implies that the marginal distribution functions $F_1, \ldots, F_d$ converge to univariate extreme-value distributions as well. Note that G possesses a multivariate version of the max-stability property we saw in (1.2.2).

Recall the transformation of the random variables $X_1, \ldots, X_j$ to the unit Pareto distribution via $X_j^* = 1/\{1 - F_j(X_j)\}$ for $j = 1, \ldots, d$. The existence of ℓ in (1.3.4) is equivalent to the statement that the joint distribution function, F_*,

of the random vector $\boldsymbol{X}^* = (X_1^*, \ldots, X_d^*)$ is in the max-domain of attraction of a d-variate extreme-value distribution, say G_*, with unit Fréchet margins (see Example 1.2.1). The link between ℓ and G_* is given by

$$\begin{aligned}\ell(\boldsymbol{x}) &= \lim_{n\to\infty} n\{1 - F_*(n/x_1, \ldots, n/x_d)\} \\ &= \lim_{n\to\infty} -\log F_*^n(n/x_1, \ldots, n/x_d) \\ &= -\log G_*(1/x_1, \ldots, 1/x_d),\end{aligned}$$

where the first line is obtained by rewriting expression (1.3.4) in unit Pareto random variables. We find that $G_*(\boldsymbol{z}) = \exp\{-\ell(1/z_1, \ldots, 1/z_d)\}$. Since, for "large" $z_1, \ldots, z_d$ we have

$$F_*(\boldsymbol{z}) \approx \exp\left\{-\frac{1}{n}\,\ell\left(\frac{n}{z_1}, \ldots, \frac{n}{z_d}\right)\right\} = \exp\left\{-\ell\left(\frac{1}{z_1}, \ldots, \frac{1}{z_d}\right)\right\}, \qquad (1.3.9)$$

we see that estimating the stable tail dependence function ℓ is key to estimation of the tail of F_*, and thus, after marginal transformation, of F. If, in addition to the existence of ℓ, the marginal distributions $F_1, \ldots, F_d$ are in the max-domains of attraction of the univariate extreme-value distributions $G_1, \ldots, G_d$, then F is in the max-domain of attraction of the extreme-value distribution

$$G(\boldsymbol{x}) = \exp\{-\ell(-\log G_1(x_1), \ldots, -\log G_d(x_d))\}, \qquad \boldsymbol{x} \in \mathbb{R}^d. \qquad (1.3.10)$$

Relation (1.3.8) is equivalent to relation (1.3.4) *and* convergence of the d marginal distributions in (1.3.8). As a consequence, (1.3.4) is substantially weaker than (1.3.8), since it only concerns the copula C corresponding to F,

$$\ell(\boldsymbol{x}) = \lim_{t\downarrow 0} \frac{1 - C(1 - tx_1, \ldots, 1 - tx_d)}{t}, \qquad \boldsymbol{x} \in [0, \infty)^d. \qquad (1.3.11)$$

The class of distribution functions satisfying (1.3.4) is hence much larger than the class of functions satisfying the multivariate max-domain of attraction condition (1.3.8). It contains, for instance, all distributions of the form $F(x) = F_1(x_1) \cdots F_d(x_d)$ with continuous margins, even if some of those margins do not belong to the max-domain of attraction of a univariate GEV distribution. Note also that if F is already an extreme-value distribution, then it is attracted by itself.

As in the univariate case, we can use point processes to express the above results. Suppose that F is in the max-domain of attraction of a multivariate GEV distribution G with margins $G_1, \ldots, G_d$. Let $\boldsymbol{l}$ denote the vector of marginal lower endpoints, i.e., l_j is the lower endpoint of G_j for $j \in \{1, \ldots, d\}$. Consider the point processes N_n on $[\boldsymbol{l}, \boldsymbol{\infty})$, defined as

$$N_n = \left\{ i \in \{1, \ldots, n\} : \max\left(\frac{\boldsymbol{X}_i - \boldsymbol{b}_n}{\boldsymbol{a}_n}, \boldsymbol{l}\right)\right\}. \qquad (1.3.12)$$

Then it can be shown (Resnick, 1987) that this sequence of point processes converges in distribution to a Poisson process N. For $\mathcal{A} = [\boldsymbol{l}, \boldsymbol{\infty}] \setminus [\boldsymbol{l}, \boldsymbol{z}]$, we have, for $\boldsymbol{z} > \boldsymbol{l}$,

$$\lim_{n\to\infty} \mathbb{P}[N(\mathcal{A}) = 0] = \lim_{n\to\infty} \mathbb{P}\left[\max(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \leq \boldsymbol{a}_n \boldsymbol{z} + \boldsymbol{b}_n\right] = G(\boldsymbol{z}).$$

The intensity measure of this limiting Poisson process is usually called the *exponent measure* (Balkema and Resnick, 1977) and denoted by μ: it is defined by

$$\mu([\boldsymbol{l}, \boldsymbol{\infty}) \setminus [\boldsymbol{l}, \boldsymbol{z}]) = -\log G(\boldsymbol{z}), \qquad \boldsymbol{z} \in [\boldsymbol{l}, \boldsymbol{\infty}] \setminus \{\boldsymbol{l}\}. \tag{1.3.13}$$

For a connection with the stable tail dependence function, we recall that $\ell(\boldsymbol{x}) = -\log G^*(\mathbf{1}/\boldsymbol{x})$, where G^* is a GEV distribution with unit Fréchet margins. Considering expression (1.3.13) for G^*, we get $\boldsymbol{l} = \mathbf{0}$ and thus

$$\ell(\boldsymbol{x}) = \mu([\mathbf{0}, \boldsymbol{\infty}) \setminus [\mathbf{0}, \mathbf{1}/\boldsymbol{x}]), \qquad \boldsymbol{x} \in [\mathbf{0}, \boldsymbol{\infty}) \setminus \{\mathbf{0}\}.$$

Note that we can rewrite the stable tail dependence function in terms of the unit Pareto random variable $\boldsymbol{X}^*$ as

$$\ell(\boldsymbol{x}) = \lim_{n\to\infty} n\, \mathbb{P}\left[\boldsymbol{X}^*/n \in [\mathbf{0}, \boldsymbol{\infty}] \setminus [\mathbf{0}, \mathbf{1}/\boldsymbol{x}]\right],$$

so that the exponent measure satisfies $n\,\mathbb{P}[\boldsymbol{X}^*/n \in \cdot\,] \to \mu(\cdot)$ as $n \to \infty$; we recognize the point processes N_n in (1.3.12) with $\boldsymbol{a}_n = n$ and $\boldsymbol{b}_n = \mathbf{0}$. The stable tail dependence functions thus acts as a distribution function for the exponent measure.

In the subsequent chapters, we will sometimes work with a measure Λ obtained from μ after the transformation $\boldsymbol{x} \mapsto \mathbf{1}/\boldsymbol{x}$. The relation between Λ and ℓ is given by

$$\ell(\boldsymbol{x}) = \Lambda\left(\left\{\boldsymbol{w} \in [0, \infty]^d : w_1 \leq x_1 \text{ or } \cdots \text{ or } w_d \leq x_d\right\}\right). \tag{1.3.14}$$

The measure Λ is also called the exponent measure and it is homogeneous, i.e., for all $a > 0$ and Borel sets $\mathcal{A} \subset [\mathbf{0}, \boldsymbol{\infty}] \setminus \{\boldsymbol{\infty}\}$, we have $\Lambda(a\mathcal{A}) = a\Lambda(\mathcal{A})$.

Recall that, contrary to the univariate case, there is no single parametric family characterizing a d-variate extreme-value distribution; any valid stable tail dependence function will lead to a GEV distribution. Consequentially, we will assume a parametric model on ℓ; see Subsections 1.3.4 and 1.4.2 for some examples. Likelihood-based procedures constitute the most common approach to estimate tail dependence parameters. However, in high dimensions likelihood estimation might be difficult because joint densities of parametric models for multivariate GEV distributions are rarely available in $d \geq 3$. This is due to the exponential in the expression for a GEV distribution G (see (1.3.10)), which leads to a combinatorial explosion of the number of terms when taking the derivative: the number of summands of the d-variate density will be equal to the d-th Bell number. Therefore, one usually resorts to composite likelihood

approaches; see Varin et al. (2011) for a recent review. Recently, progress has been made to reduce the number of summands in the density, by including the occurrence times of the componentwise maxima: see for instance Stephenson and Tawn (2005) or Wadsworth and Tawn (2014).

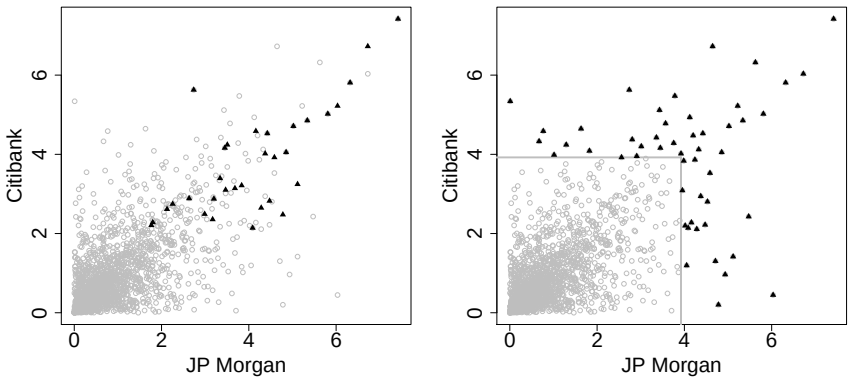


Figure 1.7: Scatterplot of the weekly negative log-returns of stock prices of JP Morgan versus Citibank, illustrating componentwise yearly block maxima (left) and threshold exceedances (right).

Alternatively, one could focus on fitting threshold exceedances instead of componentwise maxima. An observation is considered to be a threshold exceedance if at least one of its components is large, as illustrated on the right-hand side of Figure 1.7, although other definitions of a threshold exceedance are sometimes used as well (Huser et al., 2015). The most common approach is to build a likelihood from the Poisson process approximation (1.3.12) (Coles and Tawn, 1991). To make better use of the data points with some non-extreme components, a threshold censored likelihood is often used, where the components falling below the threshold u are censored at u, i.e., we assume that they fall somewhere between their lower boundary and u rather than using their true values (Ledford and Tawn, 1996; Smith et al., 1997). In Chapter 5, we will use a threshold censored likelihood when estimating the parameters of a multivariate generalized Pareto distribution, which is the multivariate analogue of a univariate GP distribution.

1.3.4 Parametric tail dependence models

We start by presenting a construction tool for parametric tail dependence models, known from Segers (2012) and Falk et al. (2010) among others. Let Y^* be a unit Pareto random variable and let $\boldsymbol{V}$ be a random vector, independ-

ent of Y^*, such that $\mathbb{E}[V_j] = 1$ for all $j = 1, \ldots, d$. Set $\boldsymbol{Y} = (Y_1, \ldots, Y_d) = (Y^*V_1, \ldots, Y^*V_d)$. Notice that the margins $F_1, \ldots, F_d$ of $\boldsymbol{Y}$ are asymptotically unit Pareto, i.e.,

$$\lim_{n\to\infty} n\{1 - F_j(n/x_j)\} = \lim_{n\to\infty} \mathbb{E}[\min(V_j x_j, n)] = x_j, \qquad j \in \{1, \ldots, d\}.$$

Plugging this into expression (1.3.4) for the stable tail dependence function,

$$\begin{aligned}
\ell(\boldsymbol{x}) &= \lim_{n\to\infty} n\left\{1 - \mathbb{P}\left[1 - F_j(X_j) \geq \frac{\mathbb{E}[\min(V_j x_j, n)]}{n},\ j = 1, \ldots, d\right]\right\} \\
&= \lim_{n\to\infty} n\{1 - \mathbb{P}[X_1 \leq n/x_1, \ldots, X_d \leq n/x_d]\} \\
&= \lim_{n\to\infty} n\left\{1 - \mathbb{E}\left[1 - \min\left(\frac{\max(V_1 x_1, \ldots, V_d x_d)}{n}, 1\right)\right]\right\} \\
&= \mathbb{E}[\max(V_1 x_1, \ldots, V_d x_d)] \qquad (1.3.15)
\end{aligned}$$

Note that in Ferreira and de Haan (2014) a similar construction recipe is proposed for so-called *Pareto processes*, but with the constraint $\mathbb{P}[\max_{j=1,\ldots,d} V_j = 1] = 0$ instead of $\mathbb{E}[V_j] = 1$ for $j = 1, \ldots, d$. Expression (1.3.15) can be used to construct a large variety of tail dependence models.

Example 1.3.1 (*Logistic model*)**.** Let $\Gamma(\,\cdot\,)$ denote the gamma function. If, for $j = 1, \ldots, d$,

$$V_j = \frac{A_j}{\Gamma(1 - \alpha^{-1})}, \qquad A_1, \ldots, A_d \overset{\text{iid}}{\sim} \text{Fréchet}(\alpha),$$

with $\alpha > 1$, then we get the *logistic model* (Aulbach et al., 2015), which was introduced already in Gumbel (1960). It has stable tail dependence function

$$\ell(\boldsymbol{x}) = (x_1^{1/\alpha} + \cdots + x_d^{1/\alpha})^\alpha, \qquad \alpha \in (0, 1].$$

The parameter α measures the dependence between the variables, such that $\alpha \downarrow 0$ corresponds to complete dependence and $\alpha = 1$ corresponds to independence. In the bivariate case, the tail dependence coefficient is $\chi = 2 - \ell(1, 1) = 2 - 2^\alpha$. An asymmetric extension of this model is given in Tawn (1990).

Example 1.3.2 (*Dirichlet model*)**.** Another frequently used model is the *Dirichlet model*. Proposed in Coles and Tawn (1991), it can be constructed setting $A_j \sim \text{Gamma}(\alpha_j, 1)$ for $j = 1, \ldots, d$, that is, A_j has density

$$f_j(z) = \frac{z^{\alpha_j - 1} e^{-z}}{\Gamma(\alpha_j)}, \qquad z > 0,\ j \in \{1, \ldots, d\}.$$

Setting $V_j = \alpha_j^{-1} A_j$, it can be shown (Segers, 2012) that

$$\ell(\boldsymbol{x}) = \frac{\Gamma(\sum_{j=1}^d \alpha_j + 1)}{\prod_{j=1}^d \Gamma(\alpha_j)} \int_{\Delta_{d-1}} \max_{j=1,\ldots,d} \left(\frac{x_j v_j}{\alpha_j}\right) \prod_{j=1}^d v_j^{\alpha_j - 1}\, \mathrm{d}v_1 \cdots \mathrm{d}v_{d-1}.$$

Complete dependence is obtained when $\alpha_1 = \alpha_2 \to \infty$, whereas independence is obtained when $\alpha_1 = \alpha_2 \to 0$. In the bivariate case, the tail dependence coefficient is

$$\chi = 1 + \mathrm{Be}\left(\alpha_1 + 1, \alpha_2; \frac{\alpha_1}{\alpha_1 + \alpha_2}\right) - \mathrm{Be}\left(\alpha_1, \alpha_2 + 1; \frac{\alpha_1}{\alpha_1 + \alpha_2}\right),$$

where Be denotes the regularized incomplete Beta function defined by

$$\mathrm{Be}(\alpha_1, \alpha_2; v) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^v w^{\alpha_1 - 1}(1 - w)^{\alpha_2 - 1} \, \mathrm{d}w.$$

Example 1.3.3 (*Max-linear model*)**.** The parameters of the models in Examples 1.3.1 and 1.3.2 can be estimated using maximum likelihood techniques. However, maximum likelihood is not applicable to non-differentiable extreme-value models, such as *max-linear models*. A max-linear model $\boldsymbol{Y}$ with r factors is constructed by setting

$$Y_j = \max_{t=1,\ldots,r} a_{jt} Z_t, \qquad j \in \{1, \ldots, d\}, \tag{1.3.16}$$

where Z_t are independent unit Fréchet random variables and the a_{jt} are non-negative constants such that $\max_{j=1,\ldots,d} a_{jt} > 0$ for every $t \in \{1, \ldots, r\}$. The stable tail dependence function of $\boldsymbol{Y}$ is

$$\ell(\boldsymbol{y}) = \sum_{t=1}^{r} \max_{j=1,\ldots,d} b_{jt} y_j, \qquad \boldsymbol{y} \in [0, \infty)^d,$$

where $b_{jt} = a_{jt} / \sum_{l=1}^{r} a_{jl}$ for $j \in \{1, \ldots, d\}$. Note that $\sum_{t=1}^{r} b_{jt} = 1$ for every $j \in \{1, \ldots, d\}$. We will denote the matrix of coefficients as

$$B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1r} \\ b_{21} & b_{22} & \cdots & b_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ b_{d1} & b_{d2} & \cdots & b_{dr} \end{pmatrix},$$

and we define the parameter vector $\boldsymbol{\theta}$ by stacking the first $(r-1)$ columns in a vector, in decreasing order of their sums. Note that $\boldsymbol{\theta} \in \mathbb{R}^p$ where $p = d \times (r-1)$.

Many other parametric models exist, see for instance Fougères et al. (2009), Cooley et al. (2010), or Ballani and Schlather (2011).

Consider the 32 componentwise yearly block maxima for the JP Morgan versus Citibank and the JP Morgan versus IBM weekly negative log-returns. Before fitting a parametric dependence model, we need to check if our data is asymptotically dependent, since our models are of no use when $\ell(x_1, x_2) = x_1 + x_2$. The left-hand side of Figure 1.8 shows the estimated tail dependence

coefficients $\widehat{\chi}(u)$ for the weekly negative log-returns of JP Morgan versus Citibank and JP Morgan versus IBM, as a function of $u \geq 0.8$; see (1.3.7). We conclude that JP Morgan versus IBM might be asymptotically independent, which is in line with what we saw in Figure 1.6, and we demonstrate the parametric models presented in Examples 1.3.1 and 1.3.2 on the JP Morgan versus Citibank data only.

Let $\boldsymbol{M}_{k,n} = (M_{k,n,1}, M_{k,n,2})$ for $k \in \{1, \ldots, 32\}$ denote the componentwise yearly block maxima of JP Morgan versus Citibank. We first transform the margins to unit Fréchet random variables by fitting univariate GEV distributions to the block maxima $M_{1,n,j}, \ldots, M_{32,n,j}$, obtaining parameter estimates (μ_j, σ_j, ξ_j) for $j \in \{1, 2\}$, and setting

$$\widetilde{X}_{ij} = \left\{1 + \widehat{\xi}_j \left(\frac{M_{i,n,j} - \widehat{\mu}_j}{\widehat{\sigma}_j}\right)\right\}^{1/\widehat{\xi}_j}, \qquad i \in \{1, \ldots, 32\},\, j \in \{1, d\},$$

since the multivariate GEV distribution $G^*(\boldsymbol{z}) = \exp\{-\ell(1/z_1, \ldots, 1/z_d)\}$ is assumed to have unit Fréchet margins. Then we fit a bivariate logistic model and a bivariate Dirichlet model to the data $\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_{32}$. We find $\widehat{\alpha} = 0.41$ (0.08) for the logistic model and $\widehat{\alpha}_1 = 1.52$ (1.28) and $\widehat{\alpha}_2 = 3.9$ (6.1) for the Dirichlet model. The *Akaike information criterion* (AIC) is 254.7 and 258.2 respectively, so that we prefer the logistic model over the Dirichlet model. The tail dependence coefficients obtained by plugging in the parameter estimates are given by $\widehat{\chi} = 0.68$ for the logistic model and $\widehat{\chi} = 0.64$ for the Dirichlet model, which is in line with the left-hand side of Figure 1.8.

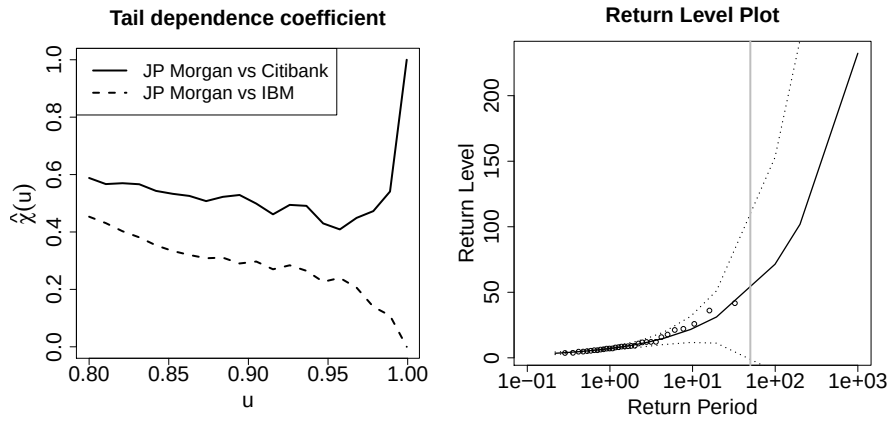


Figure 1.8: Estimators $\widehat{\chi}(u)$ for the JP Morgan versus Citibank and the JP Morgan versus IBM weekly negative log-returns (left) and the return-level plot for a joint crash of JP Morgan and Citibank.

Suppose now we are interested in the scenario where both stock returns are large simultaneously, i.e., we are interested in the event that the *minimum* of the two componentwise maxima is large. We set $Z_k = \min(M_{k,n,1}, M_{k,n,2})$ for $k \in \{1, \ldots, 32\}$ and we notice that we can use univariate methods on $Z_1, \ldots, Z_{32}$ to calculate returns levels. The right-hand side of Figure 1.8 shows a return level plot for this quantity. The fifty-year return level is 50.1 with a standard error of 23.4, i.e., we expect that both JP Morgan and Citibank have a weekly loss of at least 50.1% once every fifty years. The plot illustrates that extrapolating beyond the range of the data will necessarily lead to enormous confidence intervals.

1.4 Spatial extremes

1.4.1 Max-stable processes

Max-stable processes arise in the study of componentwise maxima of random processes rather than of random vectors. This is of interest in the spatial setting, where wave heights, precipitation amounts or temperatures occur continuously over a certain geographical region. Let $\mathcal{S}$ be a compact subset of $\mathbb{R}^2$, and let $\mathbb{C}(\mathcal{S})$ denote the space of continuous, real-valued functions on $\mathcal{S}$, equipped with the supremum norm $\|f\|_\infty = \sup_{\boldsymbol{s}\in\mathcal{S}} |f(\boldsymbol{s})|$ for $f \in \mathbb{C}(\mathcal{S})$. In the applications that we have in mind, $\mathcal{S}$ will represent the region of interest. Consider independent copies $\{X_i(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathcal{S}}$ for $i \in \{1, \ldots, n\}$ of a process $\{X(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathcal{S}}$ in $\mathbb{C}(\mathcal{S})$. Then X is in the max-domain of attraction of the *max-stable process* Y if there exist sequences of continuous functions $a_n(\boldsymbol{s}) > 0$ and $b_n(\boldsymbol{s})$ such that

$$\left\{ \frac{\max_{i=1,\ldots,n} X_i(\boldsymbol{s}) - b_n(\boldsymbol{s})}{a_n(\boldsymbol{s})} \right\}_{\boldsymbol{s}\in\mathcal{S}} \xrightarrow{w} \{Y(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathcal{S}}, \qquad \text{as } n \to \infty, \tag{1.4.1}$$

where $\xrightarrow{w}$ denotes weak convergence in $\mathbb{C}(\mathcal{S})$; see de Haan and Lin (2001) for a full characterization of max-domain of attraction conditions for the case $\mathcal{S} = [0, 1]$.

A max-stable process Y is called *simple* and denoted by Y^* if its marginal distribution functions are all unit Fréchet. Its finite-dimensional distributions $\mathbb{P}[Y(\boldsymbol{s}_1) \leq y_1, \ldots, Y(\boldsymbol{s}_d) \leq y_d]$ are d-dimensional multivariate extreme-value distributions. In de Haan (1984) it was shown that a process Y on $\mathcal{S}$ with unit Fréchet margins is simple max-stable if and only if it has the representation

$$Y^*(\boldsymbol{x}) = \max_{i\in\mathbb{N}} \xi_i V_i(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathcal{S}, \tag{1.4.2}$$

where $\{\xi_i\}_{i\geq 1}$ denote the points of a Poisson process on $(0, \infty)$ with intensity measure $\xi^{-2}\,\mathrm{d}\xi$, i.e., $\{\xi_i^{-1}\}_{i\geq 1}$ are the points of a Poisson process with unit rate, and $\{V_i\}_{i\geq 1}$ are independent replicates of a nonnegative process $V \in \mathbb{C}(\mathcal{S})$ with mean one.

Smith (1990) offers the following intuition on the above formula. Imagine a rainfall storm on the region $\mathcal{S}$, such that ξ_i represents the magnitude of a storm and $\xi_i V_i(\boldsymbol{s})$ represents the amount of rainfall for this storm at a location $\boldsymbol{s} \in \mathcal{S}$. Then max-stable processes are the pointwise maxima over the storms $\{(\xi_i, \{V_i(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}) : i \geq 1\}$. The distribution function of $Y(\boldsymbol{s})$ is

$$\mathbb{P}[Y^*(\boldsymbol{s}) \leq y(\boldsymbol{s})] = \exp\left\{-\mathbb{E}\left[\sup_{\boldsymbol{s}\in\mathcal{S}}\left(\frac{V(\boldsymbol{s})}{y(\boldsymbol{s})}\right)\right]\right\}, \qquad \boldsymbol{s} \in \mathcal{S}, \tag{1.4.3}$$

see Schlather (2002). Since

$$\mathbb{P}[Y^*(\boldsymbol{s}_1) \leq y_d, \ldots, Y^*(\boldsymbol{s}_d) \leq y_d] = \exp\left\{-\ell(1/y_1, \ldots, 1/y_d)\right\},$$

we recognize in (1.4.3) the continuous analogue of the construction device presented in (1.3.15). Note that we assumed $\mathcal{S} \subset \mathbb{R}^2$ for convenience only: we could have set $\mathcal{S} \subset \mathbb{R}^3$, for instance when considering spatial processes in three dimensions or when studying space-time processes, where the third coordinate represents time.

Some terminology used in the analysis of spatial extremes comes from the field of geostatistics. There, the process $X(\boldsymbol{s})$ is often modelled as the sum of a non-random mean function and a zero-mean Gaussian process $\epsilon(\boldsymbol{s})$. A Gaussian process ϵ is called *second-order stationary* if

$$\mathrm{Cov}[\epsilon(\boldsymbol{s}_1), \epsilon(\boldsymbol{s}_2)] = \mathrm{Cov}[\epsilon(\boldsymbol{s}_1 + \boldsymbol{h}), \epsilon(\boldsymbol{s}_2 + \boldsymbol{h})], \qquad \boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{h} \in \mathbb{R}^2,$$

and it is *isotropic* if its covariance function C is a function of distance only, i.e.,

$$\mathrm{Cov}[\epsilon(\boldsymbol{s}), \epsilon(\boldsymbol{s} + \boldsymbol{h})] = C(\|\boldsymbol{h}\|), \qquad \boldsymbol{s}, \boldsymbol{h} \in \mathbb{R}^2.$$

Finally, any second-order stationary and isotropic random field is characterized by its *semi-variogram*

$$\gamma(\boldsymbol{h}) = \frac{\mathrm{Var}\left[\epsilon(\boldsymbol{x} + \boldsymbol{h}) - \epsilon(\boldsymbol{x})\right]}{2} = \frac{\mathbb{E}\left[(\epsilon(\boldsymbol{x} + \boldsymbol{h}) - \epsilon(\boldsymbol{x}))^2\right]}{2}, \qquad \boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^2.$$

See for more background Davison and Gholamrezaee (2012) or Cooley et al. (2012a).

Inference on max-stable processes is similar to inference on multivariate GEV distributions, since in practice the number of observed locations d is finite and brings us back to the ordinary multivariate setting. The reason that we need the stochastic processes framework is that one will usually want to extrapolate beyond the locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d$, for instance, to make weather predictions on sites with no measurement stations.

It is common to transform the marginal distributions to unit Fréchet random variables, and to fit a max-stable model to componentwise block maxima using likelihood methods. Multivariate models such as the ones presented in Examples 1.3.2 and 1.3.3 are rarely used in very high dimensions because of

the quickly growing number of parameters, whereas the spatial dependence models presented in Example 1.4.1 and 1.4.2 are designed to have a limited number of parameters. This is important since spatial data are often gathered from hundreds of measuring stations. Thus, it is even more important that the estimation method we use is adequate in very high dimensions, which is why one is usually limited to composite (pairwise or triplewise) likelihoods, used in either a frequentist (Padoan et al., 2010; Davison et al., 2012; Huser and Davison, 2013) or a Bayesian setting (Reich and Shaby, 2012; Cooley et al., 2012b). Alternatively, Yuen and Stoev (2014) propose an M-estimator based on finite-dimensional cumulative distribution functions. Very recently, higher-order likelihood inference has been introduced as well (Castruccio et al., 2016; Thibaud et al., 2015). Note that joint likelihood inference, i.e., estimation of the dependence structure and the marginal parameters simultaneously, is rarely feasible due to the high number of parameters to estimate.

Estimation using threshold exceedances is only a recent topic within the context of spatial extremes, and is usually done using a threshold censored likelihood similar to the multivariate case (Wadsworth and Tawn, 2014; Thibaud et al., 2015; Thibaud and Opitz, 2015); for more examples, see Section 2.1.

1.4.2 Parametric models for spatial tail dependence

Although in theory many processes $V \in \mathbb{C}(\mathcal{S})$ with mean one lead to valid models, whose stable tail dependence function is obtained by calculating expression (1.4.3), only a handful of models are used in practice since calculation of expression (1.4.3) needs to be feasible for dimensions much higher than two. We present the two best-known examples.

Example 1.4.1 (*Gaussian extreme-value process*)**.** We will start by characterizing the *Gaussian extreme-value process* (Smith, 1990), which we will call simply the *Smith model*. Let $\{(\xi_i, U_i), i \geq 1\}$ denote the points of a Poisson process on $(0, \infty) \times \mathbb{R}^2$ with intensity measure $\xi^{-2}\,\mathrm{d}\xi\,\mathrm{d}u$. Then

$$Y^*(\boldsymbol{s}) = \max_{i \geq 1} \xi_i \, \phi_2(\boldsymbol{s} - U_i; \Sigma), \qquad \boldsymbol{s} \in \mathcal{S},$$

i.e., the process V is a deterministic Gaussian density function. From an expression similar to (1.4.3) one can calculate the finite-dimensional distributions; see for instance Schlather (2002). The pairwise stable tail dependence function, denoted ℓ_{uv} to indicate that it corresponds to the locations $\boldsymbol{s}_u, \boldsymbol{s}_v \in \mathcal{S}$, is given by

$$\ell_{uv}(x_u, x_v) = x_u \Phi\left(\frac{a_{uv}}{2} + \frac{1}{a_{uv}} \log \frac{x_u}{x_v}\right) + x_v \Phi\left(\frac{a_{uv}}{2} + \frac{1}{a_{uv}} \log \frac{x_v}{x_u}\right),$$

where

$$a_{uv} = \sqrt{(\boldsymbol{s}_u - \boldsymbol{s}_v)^T \Sigma^{-1} (\boldsymbol{s}_u - \boldsymbol{s}_v)}, \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

The stable tail dependence function for $d > 2$ is given in Genton et al. (2011). This model is isotropic if $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} = 0$.

The three plots on the left of Figure 1.9 show a simulation from the Smith model, plotted on the logarithmic scale for better visibility. The difference between the upper and the middle plot is the addition of anisotropy. The bottom plot shows how decreasing the parameter values affects the model. Note that complete dependence corresponds to $a_{uv} \to 0$ whereas asymptotic independence corresponds to $a_{uv} \to \infty$, so that dependence decreases as the distance between locations increases. Figure 1.9 clearly shows the deterministic (Gaussian) shape of this process, which is not very realistic for data applications.

Example 1.4.2 (*Brown–Resnick process*)**.** Kabluchko et al. (2009) extend a model originally proposed in Brown and Resnick (1977) and define the *Brown–Resnick* process as

$$Y^*(\boldsymbol{s}) = \max_{i \in \mathbb{N}} \xi_i \exp\{\epsilon_i(\boldsymbol{s}) - \gamma(\boldsymbol{s})\}, \qquad \boldsymbol{s} \in \mathcal{S},$$

where $\{\epsilon_i(\,\cdot\,)\}_{i \geq 1}$ are independent copies of a Gaussian process with stationary increments, $\epsilon(0) = 0$, variance $2\gamma(\,\cdot\,)$, and semi-variogram $\gamma(\,\cdot\,)$. Note that $Y(\boldsymbol{s})$ is a stationary process even if $\epsilon(\boldsymbol{s})$ is not. Kabluchko et al. (2009) show that the process with $\gamma(\boldsymbol{s}) = (\|\boldsymbol{s}\|/\rho)^\alpha$ is the only limit of (rescaled) maxima of stationary and isotropic Gaussian random fields; here $\rho > 0$ and $0 < \alpha \leq 2$.

Since isotropy may not be a reasonable assumption for many spatial applications, Blanchet and Davison (2011) and Engelke et al. (2015) introduce a transformation matrix V defined by

$$V \coloneqq V(\beta, c) \coloneqq \begin{bmatrix} \cos\beta & -\sin\beta \\ c\sin\beta & c\cos\beta \end{bmatrix}, \qquad \beta \in [0, \pi/2),\ c > 0, \tag{1.4.4}$$

and a transformed space $\mathcal{S}' = \{V^{-1}\boldsymbol{s} : \boldsymbol{s} \in \mathcal{S}\}$, so that an isotropic process on $\mathcal{S}$ is transformed to an anisotropic process on $\mathcal{S}'$. For $\boldsymbol{s}' \in \mathcal{S}'$ the anisotropic Brown–Resnick process is

$$Z_V(\boldsymbol{s}') \coloneqq Z(V\boldsymbol{s}') = \max_{i \in \mathbb{N}} \xi_i \exp\{\epsilon_i(V\boldsymbol{s}') - \gamma(V\boldsymbol{s}')\}, \tag{1.4.5}$$

whose semi-variogram is defined by

$$\gamma_V(\boldsymbol{s}') \coloneqq \gamma(V\boldsymbol{s}') = \left[\boldsymbol{s}'^T \frac{V^T V}{\rho^2} \boldsymbol{s}'\right]^{\alpha/2}.$$

The pairwise stable tail dependence function ℓ_{uv}, corresponding to locations $\boldsymbol{s}'_u, \boldsymbol{s}'_v \in \mathcal{S}$, is given by

$$\ell_{uv}(x_u, x_v) = x_u \Phi\left(\frac{a_{uv}}{2} + \frac{1}{a_{uv}} \log\frac{x_u}{x_v}\right) + x_v \Phi\left(\frac{a_{uv}}{2} + \frac{1}{a_{uv}} \log\frac{x_v}{x_u}\right),$$

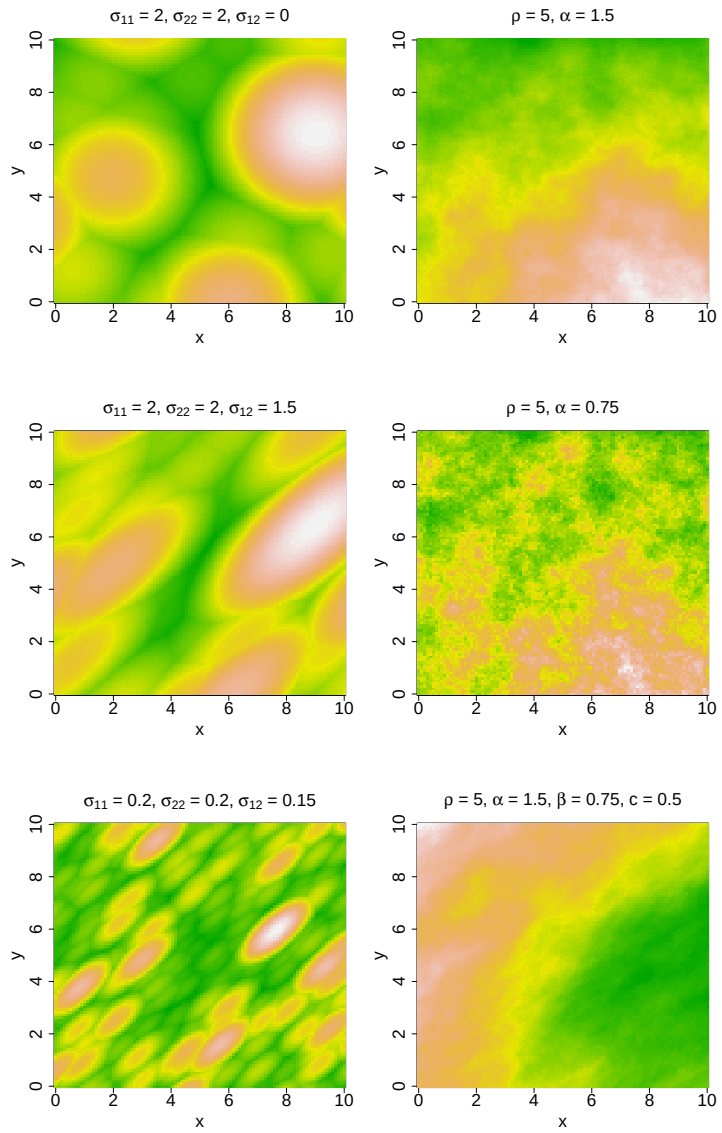


Figure 1.9: Simulations from the Smith model (left) and the Brown–Resnick process (right) for different sets of parameter values.

where $a_{uv} \coloneqq \sqrt{2\gamma_V(\boldsymbol{s}'_u - \boldsymbol{s}'_v)}$. Observe that the choice $\alpha = 2$ leads to

$$a_{uv}^2 = 2\gamma(V(\boldsymbol{s}'_u - \boldsymbol{s}'_v)) = (\boldsymbol{s}'_u - \boldsymbol{s}'_v)^T \Sigma^{-1} (\boldsymbol{s}'_u - \boldsymbol{s}'_v), \text{ for some } \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix},$$

where Σ represents any valid 2×2 covariance matrix. The Smith model is thus a special (smooth) case of the Brown–Resnick process.

The right-hand plots of Figure 1.9 show a simulation of the Brown–Resnick process. The upper and middle plot illustrate the difference when decreasing the shape parameter α; the closer α is to zero, the less smooth the process is. The bottom plot shows a simulation where we used an anisotropy matrix V to transform the underlying space. We see that the shape of the storms looks less smooth and therefore more realistic than the one of the Smith model.

Other parametric max-stable processes are the *Schlather model*, also called the extremal Gaussian process (Schlather, 2002) and the *extremal-t model* (Nikoloulopoulos et al., 2009; Opitz, 2013). Recently, multivariate max-stable processes have been proposed in Genton et al. (2015), allowing to model multiple variables at once that have been observed on a set of locations: for instance, rainfall, temperature and wind. They derive a multivariate generalization of the Brown–Resnick model.

Chapter 2

An M-estimator of spatial tail dependence

Abstract

Tail dependence models for distributions attracted to a max-stable law are fitted using observations above a high threshold. To cope with spatial, high-dimensional data, a rank-based M-estimator is proposed relying on bivariate margins only. A data-driven weight matrix is used to minimize the asymptotic variance. Empirical process arguments show that the estimator is consistent and asymptotically normal. Its finite-sample performance is assessed in simulation experiments involving popular max-stable processes perturbed with additive noise. An analysis of wind speed data from the Netherlands illustrates the method. This chapter is based on Einmahl, Krajina, Kiriliouk and Segers (2016a).

2.1 Introduction

Max-stable random processes have become the standard for modelling extremes of environmental quantities, such as wind speed, precipitation, or snow depth. In such a context, data are modelled as realizations of spatial processes, observed at a finite number of locations. The statistical problem then consists of modelling the joint tail of a multivariate distribution. This problem can be divided into two separate issues: modelling the marginal distributions and modelling the dependence structure. A popular practice is to transform the marginals into an appropriate form and to fit a max-stable model to componentwise block maxima using composite likelihood methods; see Subsections 1.3.3 and 1.4.1

As opposed to block maxima, more information can be extracted from the data by using all data vectors of which at least one component is large. Threshold-based methods are relatively new in the context of spatial extremes. A first example is de Haan and Pereira (2006), where several one- and two-dimensional models for spatial extremes are proposed. Another parametric

model for spatial tail dependence is introduced in Buishand et al. (2008). In Huser and Davison (2014), a pairwise censored likelihood is used to analyse space-time extremes. Another study of space-time extremes can be found in Davis et al. (2013), where asymptotic normality of the pairwise likelihood estimators of the parameters of a Brown–Resnick process is proven for a jointly increasing number of spatial locations and time points. A numerical study comparing two distinct approaches for composite likelihoods can be found in Bacro and Gaetan (2014).

Although all the above approaches are pairwise methods, higher-order inference methods are starting to be developed as well. In Wadsworth and Tawn (2014), a censored Poisson process likelihood is considered in order to simplify the likelihood expressions in the Brown–Resnick process and in Engelke et al. (2015), the distribution of extremal increments of processes that are in the max-domain of attraction of the Brown–Resnick process is investigated. In Bienvenüe and Robert (2014), a censored likelihood procedure is used to fit high-dimensional extreme-value models for which the tail dependence function has a particular representation. Finally, in Castruccio et al. (2016), extensive simulations are presented using higher-order composite likelihoods.

The aim of this paper is to propose a new method for fitting multivariate tail dependence models to high-dimensional data arising for instance in spatial statistics. No likelihoods come into play as our approach relies on the stable tail dependence function, and the method is threshold-based in the sense that a data point is considered to be extreme if the rank of at least one component is sufficiently high. The only assumption is that the copula corresponding to the underlying distribution is attracted to a parametrically specified multivariate extreme-value distribution, see (1.3.11).

By reducing the data to their ranks, the tails of the univariate marginal distributions need not be estimated. Indeed, the marginal distributions are not even required to be attracted to an extreme-value distribution. Another advantage of the rank-based approach is that the estimator is invariant under monotone transformations of the margins, notably for Box–Cox type of transformations.

Our starting point is Einmahl et al. (2012), where an M-estimator for a parametrically modelled stable tail dependence function in dimension d is derived. However, that method crucially relies on d-dimensional integration, which becomes intractable in high dimensions. This is why we consider tail dependence functions of pairs of variables only. Our estimator is constructed as the minimizer of the distance between a vector of integrals of parametric pairwise tail dependence functions and the vector of their empirical counterparts. The asymptotic variance of the estimator can be minimized by replacing the Euclidean distance by a quadratic form based on a weight matrix estimated from the data. In the simulation studies we will consider models in dimensions up to 100.

We show that our estimator is consistent under minimal assumptions and

asymptotically normal under an additional condition controlling the growth of the threshold. In our analysis, we take into account the variability stemming from the rank transformation, the randomness of the threshold, the random weight matrix and, in particular, the fact that the max-stable model is only an approximation in the tail.

A point worth noticing is the generality of our methodology. Where many studies focus on a specific parametric (tail) model, ours is generic and makes weak assumptions only. It does not require estimation of the tails of the marginal distributions, which can be a cumbersome task if the number of variables is large. Moreover, it allows for estimation of non-differentiable max-stable models, e.g., spectrally discrete max-stable models (Wang and Stoev, 2011).

For our approach a common, continuous distribution is required. The method does not apply to count data, for instance, and care must be taken with environmental variables that exhibit yearly seasonality or a trend, for instance due to global warming. In our case study, we study data on wind speeds in the Netherlands over a relatively short time period and limited to the summer months only.

This chapter is organized as follows. Section 2.2 contains the definition of the pairwise M-estimator and the main theoretical results on consistency and asymptotic normality, as well as the practical aspects of the choice of the weight matrix. In Section 2.3 the anisotropic Brown–Resnick process and the Smith model are recalled, and we present several simulation studies: two for a large number of locations, illustrating the computational feasibility of the estimator in high dimensions, and one for a smaller number of locations, presenting the benefits of the weight matrix. In addition, we compare the performance of our estimator to the one proposed in Engelke et al. (2015). Section 2.4 contains comparisons between our pairwise M-estimator and the estimator proposed in Einmahl et al. (2012). Finally, in Section 2.5 we present an application to wind speed data from the Netherlands. Proofs are deferred to Appendix 2.A. In Appendix 2.B, technical details on the computation of the asymptotic variance of the estimator are presented. The wind speed data and the programs that were used for the simulation studies are implemented in the R package tailDepFun (Kiriliouk, 2016); see also Chapter 4.

2.2 M-estimator

2.2.1 Set-up

Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})$, $i \in \{1, \ldots, n\}$, be independent random vectors in $\mathbb{R}^d$ with continuous distribution function F and marginal distribution functions $F_1, \ldots, F_d$. Suppose that the distribution function of $(1/\{1 - F_j(X_{1j})\})_{j=1,\ldots,d}$ is in the max-domain of attraction of the extreme-value distribution with unit Fréchet margins $G^*(\boldsymbol{z}) = \exp\{-\ell(1/z_1, \ldots, 1/z_d)\}$, $\boldsymbol{z} \in (0, \infty)^d$, which us equivalent to the existence of the stable tail dependence function $\ell : [0, \infty)^d \to$

$[0, \infty)$,

$$\ell(\boldsymbol{x}) \coloneqq \lim_{t \downarrow 0} t^{-1} \mathbb{P}[F_1(X_1) > 1 - t x_1 \text{ or } \cdots \text{ or } F_d(X_d) > 1 - t x_d], \qquad (2.2.1)$$

for $\boldsymbol{x} \in [0, \infty)^d$; see Subsection 1.3.3.

From now on we will only assume relation (2.2.1), making no assumptions on the marginal distributions $F_1, \ldots, F_d$ except for continuity. Recall that this is even weaker than the assumption that F belongs to the max-domain of attraction of a max-stable distribution.

We assume that ℓ belongs to some parametric family $\{\ell(\,\cdot\,; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, with $\Theta \subset \mathbb{R}^p$. Let $\boldsymbol{\theta}_0$ denote the true parameter vector, that is, $\boldsymbol{\theta}_0$ is the unique point in Θ such that $\ell(\boldsymbol{x}) = \ell(\boldsymbol{x}; \boldsymbol{\theta}_0)$ for all $\boldsymbol{x} \in [0, \infty)^d$. The goal is to estimate the parameter vector $\boldsymbol{\theta}_0$.

Let $\mathcal{S}$ be a compact subset of $\mathbb{R}^2$, representing a spatial region of interest. Consider independent copies $\{X_i(\boldsymbol{s})\}_{\boldsymbol{s} \in \mathcal{S}}$ for $i \in \{1, \ldots, n\}$ of a process $\{X(\boldsymbol{s})\}_{\boldsymbol{s} \in \mathcal{S}}$ in $\mathbb{C}(\mathcal{S})$. Then X is in the max-domain of attraction of the max-stable process Y if there exist sequences of continuous functions $a_n(\boldsymbol{s}) > 0$ and $b_n(\boldsymbol{s})$ such that (1.4.1) holds. Although our interest lies in the underlying stochastic processes X_i, data are always obtained on a finite subset of $\mathcal{S}$ only, i.e., at fixed locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d$. As a consequence, statistical inference is based on a sample of d-dimensional random vectors. The finite-dimensional distributions of Y are multivariate extreme-value distributions. This brings us back to the ordinary, multivariate setting.

2.2.2 Estimation

Recall the empirical tail dependence function defined in Subsection 1.3.2 as

$$\widehat{\ell}_{n,k}(\boldsymbol{x}) \coloneqq \frac{1}{k} \sum_{i=1}^{n} \mathbb{1} \left\{ R_{i1}^n > n + \frac{1}{2} - k x_1 \text{ or } \cdots \text{ or } R_{id}^n > n + \frac{1}{2} - k x_d \right\}.$$

For the estimator to be consistent, we need $k = k_n \in \{1, \ldots, n\}$ to depend on n in such a way that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.

Let $\ell = \ell(\,\cdot\,; \boldsymbol{\theta}_0)$, and let $\boldsymbol{g} = (g_1, \ldots, g_q)^T : [0,1]^d \to \mathbb{R}^q$ with $q \geq p$ denote a column vector of integrable functions. In Einmahl et al. (2012), an M-estimator of $\boldsymbol{\theta}_0$ is defined by

$$\widehat{\boldsymbol{\theta}}_n' \coloneqq \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \sum_{m=1}^{q} \left(\int_{[0,1]^d} g_m(\boldsymbol{x}) \left\{ \widehat{\ell}_{n,k}(\boldsymbol{x}) - \ell(\boldsymbol{x}; \boldsymbol{\theta}) \right\} \mathrm{d}\boldsymbol{x} \right)^2. \qquad (2.2.2)$$

Under suitable conditions, the estimator $\widehat{\boldsymbol{\theta}}_n'$ is consistent and asymptotically normal. The use of ranks via the empirical stable tail dependence function permits to avoid having to fit a model to the (tails of the) marginal distributions.

However, the approach is ill-adapted to the spatial setting, where data are gathered from dozens of locations. In high dimensions, the computation of

$\widehat{\boldsymbol{\theta}}'_n$ becomes infeasible due to the presence of d-dimensional integrals in the objective function in (2.2.2).

Akin to composite likelihood methods, we opt for a pairwise approach, minimizing over quadratic forms of vectors of two-dimensional integrals. Let q represent the number of pairs of locations that we wish to take into account, so that $p \leq q \leq d(d-1)/2$ and let $\mathcal{P}$ denote the sequence of pairs we consider, e.g., $(u, v) \in \mathcal{P}$. In the spatial setting, the indices u and v correspond to locations $\boldsymbol{s}_u$ and $\boldsymbol{s}_v$ respectively.

The bivariate margins of the stable tail dependence function $\ell(\,\cdot\,;\boldsymbol{\theta})$ and its empirical counterpart are given by

$$\begin{aligned}\ell_{uv}(x_u, x_v; \boldsymbol{\theta}) &\coloneqq \ell(0, \ldots, 0, x_u, 0, \ldots, 0, x_v, 0, \ldots, 0; \boldsymbol{\theta}),\\ \widehat{\ell}_{n,k,uv}(x_u, x_v) &\coloneqq \widehat{\ell}_{n,k}(0, \ldots, 0, x_u, 0, \ldots, 0, x_v, 0, \ldots, 0),\end{aligned}$$

respectively. Define $L : \Theta \to \mathbb{R}^q$ by

$$L(\boldsymbol{\theta}) \coloneqq \left(\int_{[0,1]^2} \ell_{uv}(x_u, x_v; \boldsymbol{\theta}) \, \mathrm{d}x_u \, \mathrm{d}x_v \right)_{(u,v)\in\mathcal{P}}. \tag{2.2.3}$$

Consider the random $q \times 1$ column vector

$$\widehat{\boldsymbol{L}}_{n,k} \coloneqq \left(\int_{[0,1]^2} \widehat{\ell}_{n,k,uv}(x_u, x_v) \, \mathrm{d}x_u \, \mathrm{d}x_v \right)_{(u,v)\in\mathcal{P}},$$

and set $D_{n,k}(\boldsymbol{\theta}) \coloneqq L(\boldsymbol{\theta}) - \widehat{\boldsymbol{L}}_{n,k}$. Let $\widehat{\Omega}_n \in \mathbb{R}^{q\times q}$ be a symmetric, positive definite, possibly random matrix. Define

$$f_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta}) \coloneqq D_{n,k}(\boldsymbol{\theta})^T \, \widehat{\Omega}_n \, D_{n,k}(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \Theta.$$

The pairwise M-estimator of $\boldsymbol{\theta}_0$ is defined as

$$\widehat{\boldsymbol{\theta}}_n \coloneqq \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} f_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta}) = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \left\{ D_{n,k}(\boldsymbol{\theta})^T \, \widehat{\Omega}_n \, D_{n,k}(\boldsymbol{\theta}) \right\}. \tag{2.2.4}$$

The simplest choice for $\widehat{\Omega}_n$ is just the $q \times q$ identity matrix I_q, yielding

$$\begin{aligned} f_{n,k,I_q}(\boldsymbol{\theta}) &= \|D_{n,k}(\boldsymbol{\theta})\|^2 \\ &= \sum_{(u,v)\in\mathcal{P}} \left(\int_{[0,1]^2} \left\{ \widehat{\ell}_{n,k,uv}(x_u, x_v) - \ell_{uv}(x_u, x_v; \theta) \right\} \mathrm{d}x_u \, \mathrm{d}x_v \right)^2. \end{aligned} \tag{2.2.5}$$

Note the similarity of this objective function with the one for the original M-estimator in equation (2.2.2). The role of the matrix $\widehat{\Omega}_n$ is to be able to assign data-driven weights to quantify the size of the vector of discrepancies $D_{n,k}(\boldsymbol{\theta})$ via a generalized Euclidian norm. As we will see in Section 2.2.3, a judicious choice of this matrix will allow to minimize the asymptotic variance.

2.2.3 Asymptotic results and choice of the weight matrix

We show consistency and asymptotic normality of the rank-based pairwise M-estimator. Moreover, we provide a data-driven choice for $\widehat{\Omega}_n$ which minimizes the asymptotic covariance matrix of the limiting normal distribution. Results for the construction of confidence regions and hypothesis tests are presented as well.

Recall the exponent measure Λ, defined in (1.3.14). Let W_Λ be a mean-zero Gaussian process, indexed by the Borel sets of $[0,\infty]^d \setminus \{\boldsymbol{\infty}\}$ and with covariance function

$$\mathbb{E}[W_\Lambda(\mathcal{A}_1)\, W_\Lambda(\mathcal{A}_2)] = \Lambda(\mathcal{A}_1 \cap \mathcal{A}_2),$$

where $\mathcal{A}_1$, $\mathcal{A}_2$ are Borel sets in $[0,\infty]^d \setminus \{\boldsymbol{\infty}\}$. For $\boldsymbol{x} \in [0,\infty)^d$, define

$$\begin{aligned} W_\ell(\boldsymbol{x}) &= W_\Lambda(\{\boldsymbol{w} \in [0,\infty]^d \setminus \{\boldsymbol{\infty}\} : w_1 \le x_1 \text{ or } \cdots \text{ or } w_d \le x_d\}), \\ W_{\ell,j}(x_j) &= W_\ell(0,\ldots,0,x_j,0,\ldots,0), \qquad j = 1,\ldots,d. \end{aligned}$$

Let $\dot{\ell}_j$ be the partial derivative of ℓ with respect to x_j, and define

$$B(\boldsymbol{x}) \coloneqq W_\ell(\boldsymbol{x}) - \sum_{j=1}^{d} \dot{\ell}_j(\boldsymbol{x})\, W_{\ell,j}(x_j), \qquad \boldsymbol{x} \in [0,\infty)^d.$$

For every pair $(u,v) \in \mathcal{P}$, put

$$B_{uv}(x_u, x_v) \coloneqq B(0,\ldots,0,x_u,0,\ldots,0,x_v,0,\ldots,0).$$

Also define the mean-zero random column vector

$$\widetilde{B} \coloneqq \left(\int_{[0,1]^2} B_{uv}(x_u, x_v)\, \mathrm{d}x_u\, \mathrm{d}x_v \right)_{(u,v)\in\mathcal{P}}.$$

The law of $\widetilde{B}$ is zero-mean Gaussian and its covariance matrix $\Gamma(\boldsymbol{\theta}_0) \in \mathbb{R}^{q\times q}$ depends on $\boldsymbol{\theta}_0$ via the model assumption $\ell = \ell(\,\cdot\,;\boldsymbol{\theta}_0)$. Write (u,v) and (u',v') for the i-th and j-th element of $\mathcal{P}$ respectively. Then we can obtain the (i,j)-th entry of $\Gamma(\boldsymbol{\theta})$ by

$$\begin{aligned} \Gamma_{ij}(\boldsymbol{\theta}) &= \mathbb{E}[\widetilde{B}_{uv}\widetilde{B}_{u'v'}] \\ &= \int_{[0,1]^4} \mathbb{E}\left[B_{uv}(x_u,x_v)\, B_{u'v'}(x_{u'},x_{v'})\right] \mathrm{d}x_u\, \mathrm{d}x_v\, \mathrm{d}x_{u'}\, \mathrm{d}x_{v'}. \end{aligned} \tag{2.2.6}$$

Assuming $\boldsymbol{\theta}$ is an interior point of Θ and L is differentiable in $\boldsymbol{\theta}$, let $\dot{L}(\boldsymbol{\theta}) \in \mathbb{R}^{q\times p}$ denote the total derivative of L at $\boldsymbol{\theta}$.

Theorem 2.2.1 (Existence, uniqueness and consistency)**.** *Let $\{\ell(\,\cdot\,;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset \mathbb{R}^p$, be a parametric family of d-variate stable tail dependence functions and let $\mathcal{P}$ be a sequence of q distinct pairs, with $p \le q \le d(d-1)/2$, such that the map L in (2.2.3) is a homeomorphism from Θ to $L(\Theta)$. Let the d-variate distribution function F have continuous margins and stable tail dependence function $\ell(\,\cdot\,;\boldsymbol{\theta}_0)$ for some interior point $\boldsymbol{\theta}_0 \in \Theta$. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be an iid sample from F. Let $k = k_n \in \{1, \ldots, n\}$ satisfy $k \to \infty$ and $k/n \to 0$, as $n \to \infty$. Assume also that*

(C1) L is twice continuously differentiable on a neighbourhood of $\boldsymbol{\theta}_0$ and $\dot{L}(\boldsymbol{\theta}_0)$ is of full rank;

(C2) there exists a symmetric, positive definite matrix Ω such that $\widehat{\Omega}_n \xrightarrow{p} \Omega$ entry-wise.

Then with probability tending to one, the minimizer $\widehat{\boldsymbol{\theta}}_n$ of $f_{n,k,\widehat{\Omega}_n}$ exists and is unique. Moreover,

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0, \qquad \text{as } n \to \infty.$$

Theorem 2.2.2 (Asymptotic normality)**.** *If in addition to the assumptions of Theorem 2.2.1*

(C3) $t^{-1}\mathbb{P}[1 - F_1(X_{11}) \le t x_1 \text{ or } \cdots \text{ or } 1 - F_d(X_{1d}) \le t x_d] - \ell(\boldsymbol{x};\boldsymbol{\theta}_0) = O(t^{\alpha})$ uniformly in $\boldsymbol{x} \in \Delta_{d-1}$ as $t \downarrow 0$ for some $\alpha > 0$;

(C4) $k = o(n^{2\alpha/(1+2\alpha)})$ and $k \to \infty$ as $n \to \infty$,

then

$$\sqrt{k}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, M(\boldsymbol{\theta}_0))$$

where, for $\boldsymbol{\theta} \in \Theta$ such that $\dot{L}(\boldsymbol{\theta})$ is of full rank,

$$M(\boldsymbol{\theta}) \coloneqq \big(\dot{L}(\boldsymbol{\theta})^T\,\Omega\,\dot{L}(\boldsymbol{\theta})\big)^{-1}\,\dot{L}(\boldsymbol{\theta})^T\,\Omega\,\Gamma(\boldsymbol{\theta})\,\Omega\,\dot{L}(\boldsymbol{\theta})\,\big(\dot{L}(\boldsymbol{\theta})^T\,\Omega\,\dot{L}(\boldsymbol{\theta})\big)^{-1}. \quad (2.2.7)$$

The proofs of Theorems 2.2.1 and 2.2.2 are deferred to Appendix 2.A.

An asymptotically optimal choice for the random weight matrix $\widehat{\Omega}_n$ would be one for which the limit Ω minimizes the asymptotic covariance matrix $M(\boldsymbol{\theta}_0)$ with respect to the positive semi-definite partial ordering on the set of symmetric matrices. This minimization problem shows up in other contexts as well, and its solution is well-known: provided $\Gamma(\boldsymbol{\theta})$ is invertible, the minimum is attained at $\Omega = \Gamma(\boldsymbol{\theta})^{-1}$, the matrix $M(\boldsymbol{\theta})$ simplifying to

$$M_{\mathrm{opt}}(\boldsymbol{\theta}) = \big(\dot{L}(\boldsymbol{\theta})^T\,\Gamma(\boldsymbol{\theta})^{-1}\,\dot{L}(\boldsymbol{\theta})\big)^{-1}, \quad (2.2.8)$$

see for instance Abadir and Magnus (2005, page 339). However, this choice of the weight matrix requires the knowledge of $\boldsymbol{\theta}_0$, which is unknown. One possible solution consists of computing the optimal weight matrix evaluated at a preliminary estimator of $\boldsymbol{\theta}_0$.

For $\boldsymbol{\theta} \in \Theta$, let $H_{\boldsymbol{\theta}}$ be the *spectral measure* related to $\ell(\,\cdot\,;\boldsymbol{\theta})$ (de Haan and Resnick, 1977; Resnick, 1987), a finite measure defined on the unit simplex Δ_{d-1} which satisfies

$$\ell(\boldsymbol{x};\boldsymbol{\theta}) = \int_{\Delta_{d-1}} \max_{j=1,\ldots,d} \{w_j x_j\}\, H_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{w}), \qquad \boldsymbol{x} \in [0,\infty)^d.$$

Corollary 2.2.3 (Optimal weight matrix)**.** *In addition to the assumptions of Theorem 2.2.2, assume the following:*

(C5) for all $\boldsymbol{\theta}$ in the interior of Θ, the matrix $\Gamma(\boldsymbol{\theta})$ in (2.2.6) *has full rank;*

(C6) the mapping $\boldsymbol{\theta} \mapsto H_{\boldsymbol{\theta}}$ is weakly continuous at $\boldsymbol{\theta}_{\mathbf{0}}$.

Assume $\widehat{\boldsymbol{\theta}}_n^{(0)}$ converges in probability to $\boldsymbol{\theta}_0$ and let $\widehat{\boldsymbol{\theta}}_n$ be the pairwise M-estimator with weight matrix $\widehat{\Omega}_n = \Gamma(\widehat{\boldsymbol{\theta}}_n^{(0)})^{-1}$. Then, with M_{opt} as in (2.2.8)*, we have*

$$\sqrt{k}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, M_{\mathrm{opt}}(\boldsymbol{\theta}_0)), \qquad n \to \infty.$$

For any choice of the positive definite matrix Ω in (2.2.7)*, the difference $M(\boldsymbol{\theta}_0) - M_{\mathrm{opt}}(\boldsymbol{\theta}_0)$ is positive semi-definite.*

In view of Corollary 2.2.3, we propose the following two-step procedure:

1. Compute the pairwise M-estimator $\widehat{\boldsymbol{\theta}}_n^{(0)}$ with the weight matrix equal to the identity matrix, i.e., by minimizing f_{n,k,I_q} in (2.2.5).

2. Calculate the pairwise M-estimator $\widehat{\boldsymbol{\theta}}_n$ by minimizing $f_{n,k,\widehat{\Omega}_n}$ with $\widehat{\Omega}_n = \Gamma(\widehat{\boldsymbol{\theta}}_n^{(0)})^{-1}$.

We will see in Section 2.3.2 that this choice of $\widehat{\Omega}_n$ indeed reduces the estimation error.

Calculating $M(\boldsymbol{\theta})$ can be a challenging task. The matrix $\Gamma(\boldsymbol{\theta})$ can become quite large since for a d-dimensional model, the maximal number of pairs is $d(d-1)/2$. In practice we will choose a smaller number of pairs: we will see in Section 2.3.2 that this may even have a positive influence on the quality of our estimator. Appendix 2.B contains details on the calculation and implementation of the matrix $\Gamma(\boldsymbol{\theta})$.

A natural competitor of the two-step procedure could be a one-step procedure where the weight matrix $\Gamma(\boldsymbol{\theta})^{-1}$ is recalculated within the minimisation routine. This resembles, but is substantially different from a continuously updating generalised method of moments (Hansen et al., 1996). Rather than as in equation (2.2.4), the pairwise M-estimator of $\boldsymbol{\theta}_0$ would be defined as the minimizer of the function

$$\boldsymbol{\theta} \mapsto L_{n,k}(\boldsymbol{\theta})^T\, \Gamma(\boldsymbol{\theta})^{-1}\, L_{n,k}(\boldsymbol{\theta}).$$

Calculation of $\Gamma(\boldsymbol{\theta})$ being time-consuming however, such an approach would be computationally unwieldy.

Finally, we present results that can be used for the construction of confidence regions and hypothesis tests.

Corollary 2.2.4. *If the assumptions from Corollary 2.2.3 are satisfied, then*

$$k(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T M(\widehat{\boldsymbol{\theta}}_n)^{-1}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_p, \qquad \text{as } n \to \infty.$$

Let $r < p$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \Theta$ with $\boldsymbol{\theta}_1 \in \mathbb{R}^{p-r}$ and $\boldsymbol{\theta}_2 \in \mathbb{R}^r$. Suppose we want to test $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_2^*$ against $\boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_2^*$. Write $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n})$ and let $M_2(\boldsymbol{\theta})$ be the $r \times r$ matrix corresponding to the lower right corner of $M(\boldsymbol{\theta})$.

Corollary 2.2.5. *If the assumptions from Corollary 2.2.3 are satisfied and if* $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*) \in \Theta$ *for some* $\boldsymbol{\theta}_1$*, then*

$$k(\widehat{\boldsymbol{\theta}}_{2n} - \boldsymbol{\theta}_2^*)^T M_2(\widehat{\boldsymbol{\theta}}_{1n}, \boldsymbol{\theta}_2^*)^{-1}(\widehat{\boldsymbol{\theta}}_{2n} - \boldsymbol{\theta}_2^*) \xrightarrow{d} \chi^2_r.$$

We will not prove these corollaries here, since their proofs are straightforward extensions of those in Einmahl et al. (2012, Corollary 4.3; Corollary 4.4).

2.3 Spatial models

2.3.1 Theory and definitions

Consider a Brown–Resnick process on $\mathcal{S} \subset \mathbb{R}^2$. Recall that an isotropic process on $\mathcal{S}$ is equivalent to an anisotropic process on $\mathcal{S}' = \{V^{-1}\boldsymbol{s} : \boldsymbol{s} \in \mathcal{S}\}$, where V is as in (1.4.4). Recall that the pairwise stable tail dependence function for a pair (u, v), corresponding to locations $\boldsymbol{s}'_u, \boldsymbol{s}'_v \in \mathcal{S}'$, is given by

$$\ell_{uv}(x_u, x_v) = x_u \Phi\left(\frac{a_{uv}}{2} + \frac{1}{a_{uv}} \log \frac{x_u}{x_v}\right) + x_v \Phi\left(\frac{a_{uv}}{2} + \frac{1}{a_{uv}} \log \frac{x_v}{x_u}\right),$$

where $a_{uv} \coloneqq \sqrt{2\gamma_V(\boldsymbol{s}'_u - \boldsymbol{s}'_v)}$ and the semi-variogram γ_V is defined by

$$\gamma_V(\boldsymbol{s}') \coloneqq \gamma(V\boldsymbol{s}') = \left[\boldsymbol{s}'^T \frac{V^T V}{\rho^2} \boldsymbol{s}'\right]^{\alpha/2}, \qquad \alpha \in (0, 2],\ \rho > 0,$$

for $\boldsymbol{s}' \in \mathcal{S}'$. We will present simulation studies for processes in the domain of attraction, in the sense of (2.2.1), of both the Smith model ($\alpha = 2$) and the anisotropic Brown–Resnick process. The parameter vectors are $\boldsymbol{\theta} = (\alpha, \rho, \beta, c)$ and $\boldsymbol{\theta} = (\sigma_{11}, \sigma_{22}, \sigma_{12})$ respectively; see Subsection 1.4.2. To calculate the weight matrix $\Gamma(\boldsymbol{\theta})^{-1}$, we will need to compute integrals over the four-dimensional margins of the stable tail dependence function, see (2.2.6) and Appendix 2.B.

In Huser and Davison (2013) the following representation is given for $\ell(\boldsymbol{x};\boldsymbol{\theta})$ for general d. If Z_V is defined as in (1.4.5) then for $\boldsymbol{s}_{u_1},\ldots,\boldsymbol{s}_{u_d} \in \mathcal{S} \subset \mathbb{R}^2$,

$$\ell_{u_1,\ldots,u_d}(\boldsymbol{x}) = \sum_{j=1}^{d} x_j \Phi_{d-1}(\zeta^{(j)}(\mathbf{1}/\boldsymbol{x}); \Upsilon^{(j)}),$$

where $\zeta^{(j)}(\mathbf{1}/\boldsymbol{x}) \in \mathbb{R}^{d-1}$ and

$$\zeta^{(j)}(\boldsymbol{x}) = (\zeta_1^{(j)}(x_1,x_j),\ldots,\zeta_{j-1}^{(j)}(x_{j-1},x_j),\zeta_{j+1}^{(j)}(x_{j+1},x_j),\ldots,\zeta_d^{(j)}(x_d,x_j)),$$
$$\zeta_j^{(j)}(x_i,x_j) = \sqrt{\frac{\gamma_V(s_{u_j}-s_{u_i})}{2}} + \frac{\log(x_i/x_j)}{\sqrt{2\gamma_V(s_{u_j}-s_{u_i})}},$$

and $\Upsilon^{(j)} \in \mathbb{R}^{(d-1)\times(d-1)}$ is the correlation matrix with entries

$$\Upsilon_{ik}^{(j)} = \frac{\gamma_V(s_{u_j}-s_{u_i}) + \gamma_V(s_{u_j}-s_{u_k}) - \gamma_V(s_{u_i}-s_{u_k})}{2\sqrt{\gamma_V(s_{u_j}-s_{u_i})\gamma_V(s_{u_j}-s_{u_k})}},$$

for $i,k = 1,\ldots,d$; $i,k \neq j$.

2.3.2 Simulation studies

In order to study the performance of the pairwise M-estimator when the underlying distribution function F satisfies (2.2.1) for a function ℓ corresponding to the max-stable models described before, we generate random samples from Brown–Resnick processes and Smith models perturbed with additive noise. If $\boldsymbol{Y} = (Y_1,\ldots,Y_d)$ is a max-stable process observed at d locations, then we consider

$$X_j = Y_j + \epsilon_j, \qquad j = 1,\ldots,d,$$

where ϵ_j are independent half normally distributed random variables, corresponding to the absolute value of a normally distributed random variable with standard deviation $1/2$. All simulations are done in R (R Core Team, 2015). Realizations of Y are simulated using the SpatialExtremes package (Ribatet, 2015).

Perturbed max-stable processes on a large grid.

Assume that we have $d = 100$ locations on a 10×10 unit distance grid. We simulate 500 samples of size $n = 500$ from the perturbed Smith model with parameters

$$\Sigma = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.5 \end{bmatrix},$$

and from a perturbed anisotropic Brown–Resnick process with parameters $\alpha = 1$, $\rho = 3$, $\beta = 0.5$ and $c = 0.5$. Instead of estimating ρ, β, and c directly, we

estimate the three parameters of the matrix

$$\mathcal{T} = \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{12} & \tau_{22} \end{bmatrix} = \rho^{-2} V(\beta, c)^T V(\beta, c).$$

In practice, this parametrization, which is in line with the one of the Smith model, often yields better results. We study the bias and root mean squared error (RMSE) for $k \in \{25, 50, 75, 100\}$. We compare the estimators for two sets of pairs: one containing all pairs ($q = 4950$) and one containing only pairs of neighbouring locations ($q = 342$). Although the first option may sound like a time-consuming procedure, estimation of the parameters for one sample takes about 20 seconds for the Smith model and less than two minutes for the anisotropic Brown–Resnick process. We let the weight matrix Ω be the $q \times q$ identity matrix, since for so many pairs a data-driven computation of the optimal weight matrix is too time-consuming. Figure 2.1 shows the bias, standard deviation and RMSE of $(\sigma_{11}, \sigma_{22}, \sigma_{12})$ for the Smith model. We see that great improvements are achieved by using only pairs of neighbouring locations and that the thus obtained estimator performs well. Using all pairs causes the parameters to have a large positive bias, which translates into a high RMSE. In general, distant pairs often lead to less dependence and hence less information about ℓ and its parameters. Observe that small values of k are preferable, i.e. $k = 25$ or $k = 50$.

Figures 2.2 and 2.3 show the bias, standard deviation and RMSE of the pairwise M-estimators of the parameters (α, ρ, β, c) of the anisotropic Brown–Resnick process. We see again that using only pairs of neighbouring locations improves the quality of estimation. The corresponding estimators perform well for the estimation of α, β, and c. The lesser performance when estimating ρ seems to be inherent to the Brown–Resnick process and appears regardless of the estimation procedure: see for example Engelke et al. (2015) or Wadsworth and Tawn (2014), who both report a positive bias of ρ for small sample sizes. Compared to those for the Smith model, the values of k for which the estimation error is smallest are higher, i.e., $k = 50$ or $k = 75$.

A perturbed Brown–Resnick process on a small grid with optimal weight matrix.

We consider $d = 12$ locations on an equally spaced unit distance 4×3 grid. We simulate 500 samples of size $n = 1000$ from an anisotropic Brown–Resnick process with parameters $\alpha = 1.5$, $\rho = 1$, $\beta = 0.25$ and $c = 1.5$. We study the bias, standard deviation, and RMSE for $k \in \{25, 75, 125\}$. In Figures 2.4 and 2.5, three estimation methods are compared: one involving all pairs ($q = 66$), one involving only pairs of neighbouring locations ($q = 29$), and one using optimal weight matrices chosen according to the two-step procedure described after Corollary 3.3, based on the 29 pairs of neighbouring locations. In line with Corollary 3.3, the weighted estimators have lower (or equal) standard deviation

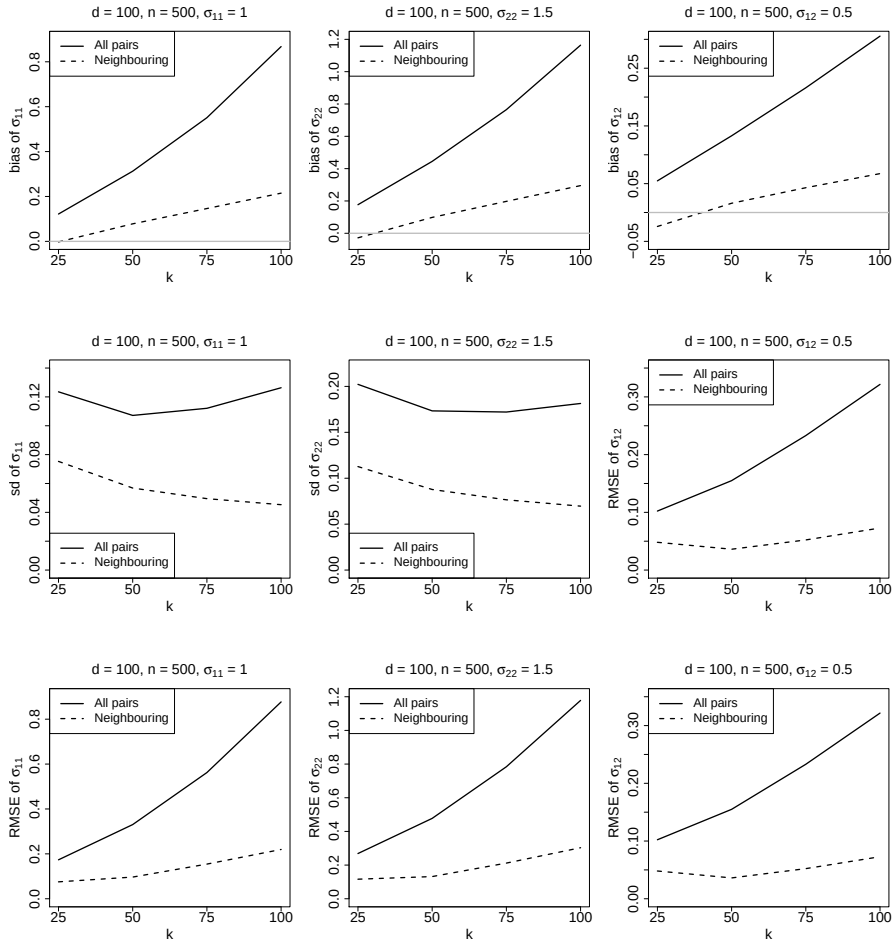


Figure 2.1: Bias, standard deviation and RMSE for estimators of $\sigma_{11} = 1$ (left), $\sigma_{22} = 1.5$ (right), and $\sigma_{12} = 0.5$ (bottom) for the perturbed 100-dimensional Smith model with identity weight matrix; 500 samples of $n = 500$.

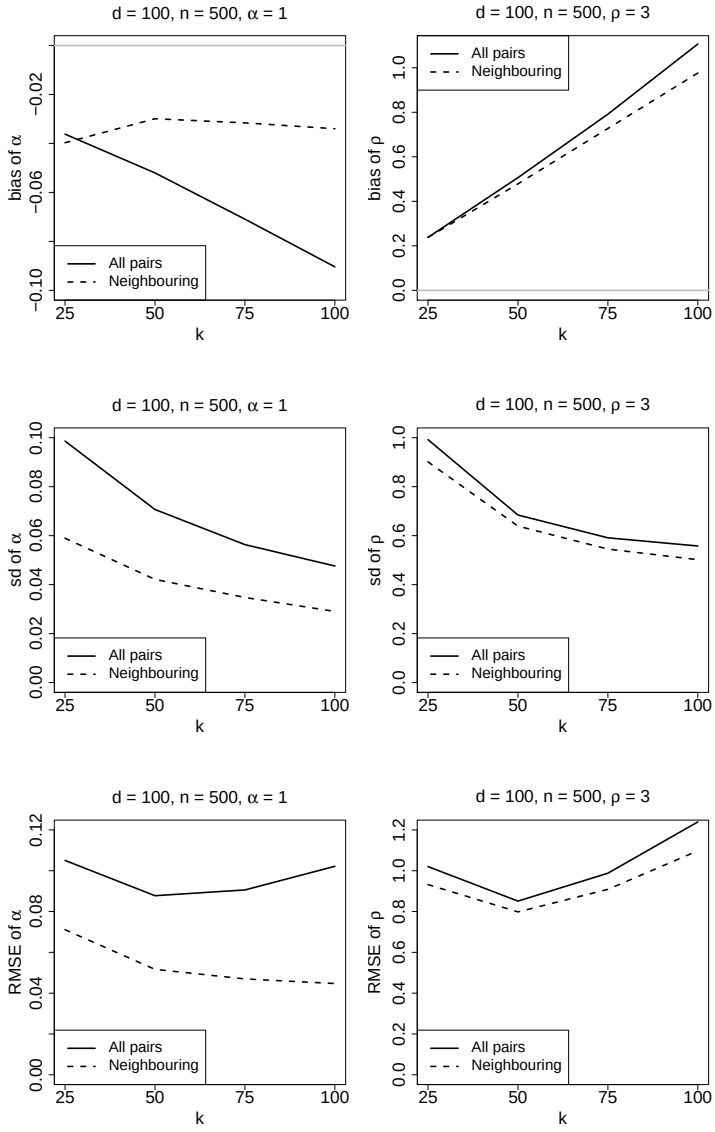


Figure 2.2: Bias, standard deviation and RMSE for estimators of $\alpha = 1$ (left) and $\rho = 3$ (right) for the perturbed 100-dimensional Brown–Resnick process with identity weight matrix; 500 samples of $n = 500$.

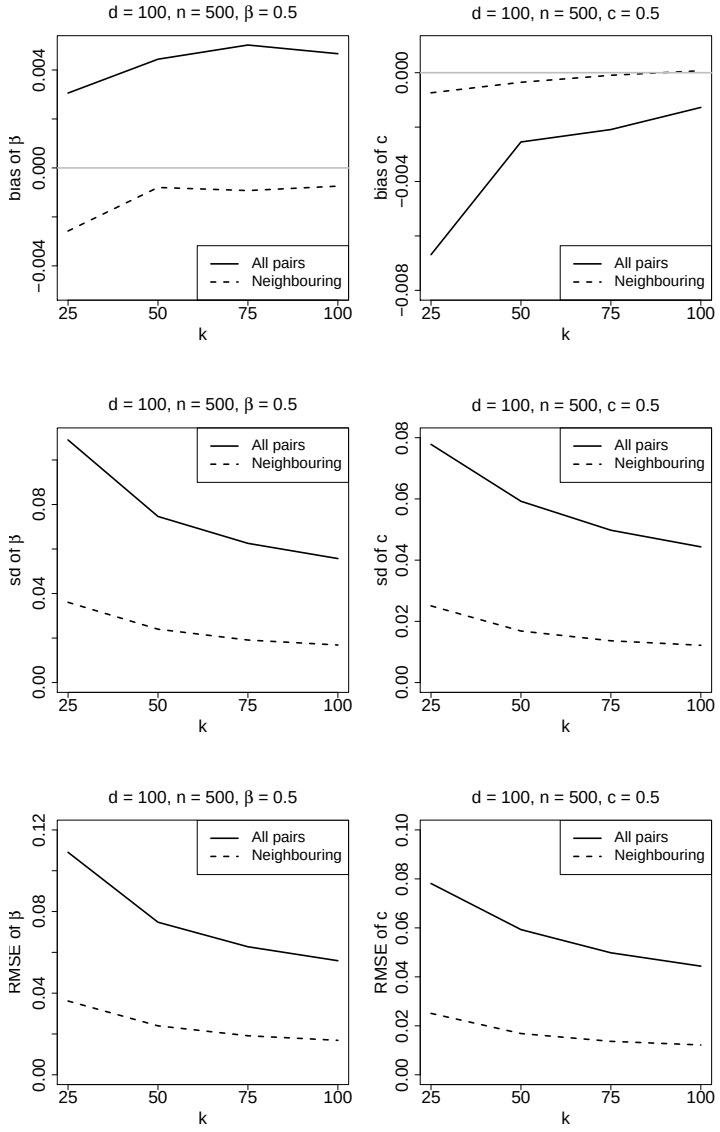


Figure 2.3: Bias, standard deviation and RMSE for estimators of $\beta = 0.5$ (left) and $c = 0.5$ (right) for the perturbed 100-dimensional Brown–Resnick process with identity weight matrix; 500 samples of $n = 500$.

(and RMSE) than the unweighted estimators. The difference is clearest for low k for α and ρ.

Comparison with Engelke et al. (2015).

To compare the pairwise M-estimator with the one from Engelke et al. (2015), we consider the setting used in the simulation study of the latter paper: we simulate 500 samples of size 8000 of the *univariate* Brown–Resnick process on an equidistant grid on the interval $[0, 3]$ with step size 0.1. The parameters of the model are $(\alpha, \rho) = (1, 1)$. We estimate the unknown parameters for $k = 500$ and $q = 140$ pairs, so that the locations of the selected pairs are at most a distance of 0.5 apart. We use the identity weight matrix, since in this particular setting the weight matrix is very large and, as far as we could tell from some preliminary experiments, it leads to only a small reduction in estimation error. Asymptotically we see a reduction of the standard deviations of about 13% for α and 3% for ρ. In Figure 2.6 below, the results are presented in the form of boxplots, to facilitate comparison with Figure 2 in Engelke et al. (2015). Our procedure turns out to perform equally well for the estimation of α and only slightly worse when estimating ρ. It is to be kept in mind that, whereas the method in Engelke et al. (2015) is tailor-made for the Brown–Resnick process, our method is designed to work for general parametric models.

Discussion.

We have seen in the 100-dimensional simulation study that the computation of the unweighted pairwise estimator is fast even for a large number of pairs. However, calculating an entry for the optimal weight matrix takes about 15 seconds on a standard computer. Since we have to calculate $q(q + 1)/2$ entries of the weight matrix, this method gets more time-consuming when the number of pairs q is large.

We also noticed that for large dimensions, a relatively small sample size of $n = 500$ is sufficient to obtain good results. However, the smaller the dimension, the larger the sample size needs to be, i.e., a decrease of information in space must be compensated by an increase of information in "time". We have observed that the choice of the starting value hardly affects the outcome of the optimisation procedure, unless the dimension is less than five. More guidelines and rules-of-thumb for practical use of the estimator can be found in the reference manual and vignette of the tailDepFun package (Kiriliouk, 2016); see Chapter 4.

Another interesting feature is that, for both the Smith model and the Brown–Resnick process, considering only neighbouring pairs leads to better results than considering all pairs. As the distance between two locations increases, they become tail independent, so that including pairs of distant locations adds little information about the model parameters.

Finally, to assess the quality of the normal approximation to the sampling distribution of the estimator, we have conducted simulation experiments for

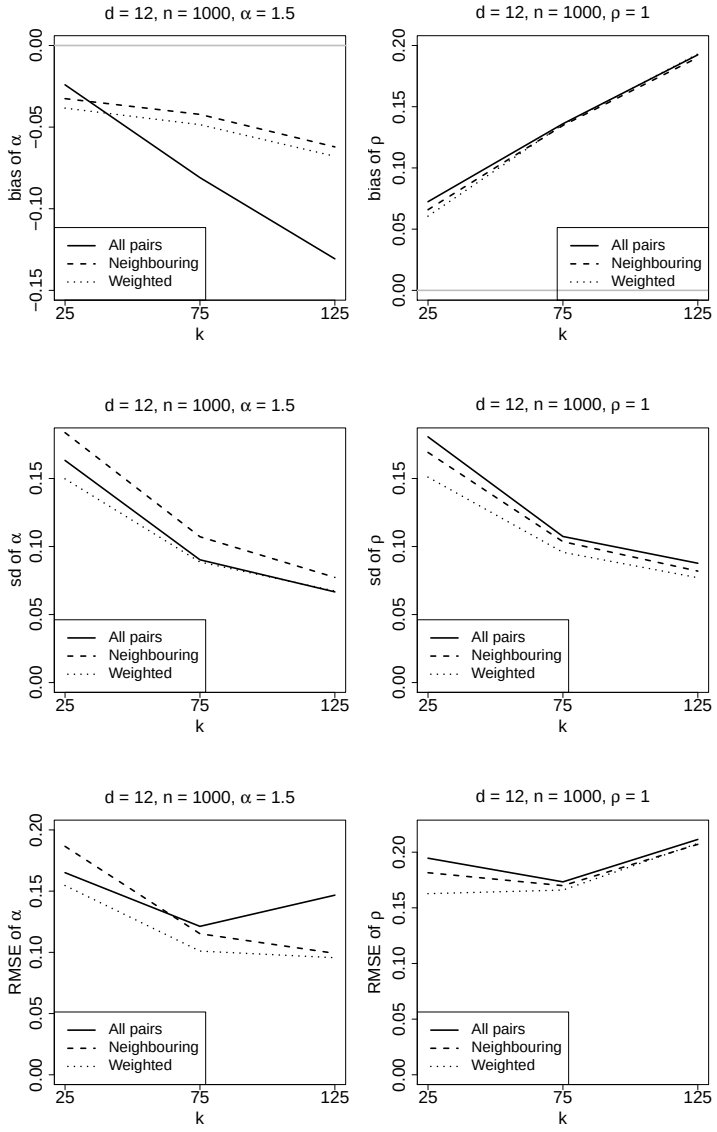


Figure 2.4: Bias, standard deviation and RMSE for estimators of $\alpha = 1.5$ (left) and $\rho = 1$ (right) for the perturbed 12-dimensional Brown–Resnick process; 500 samples of $n = 1000$.

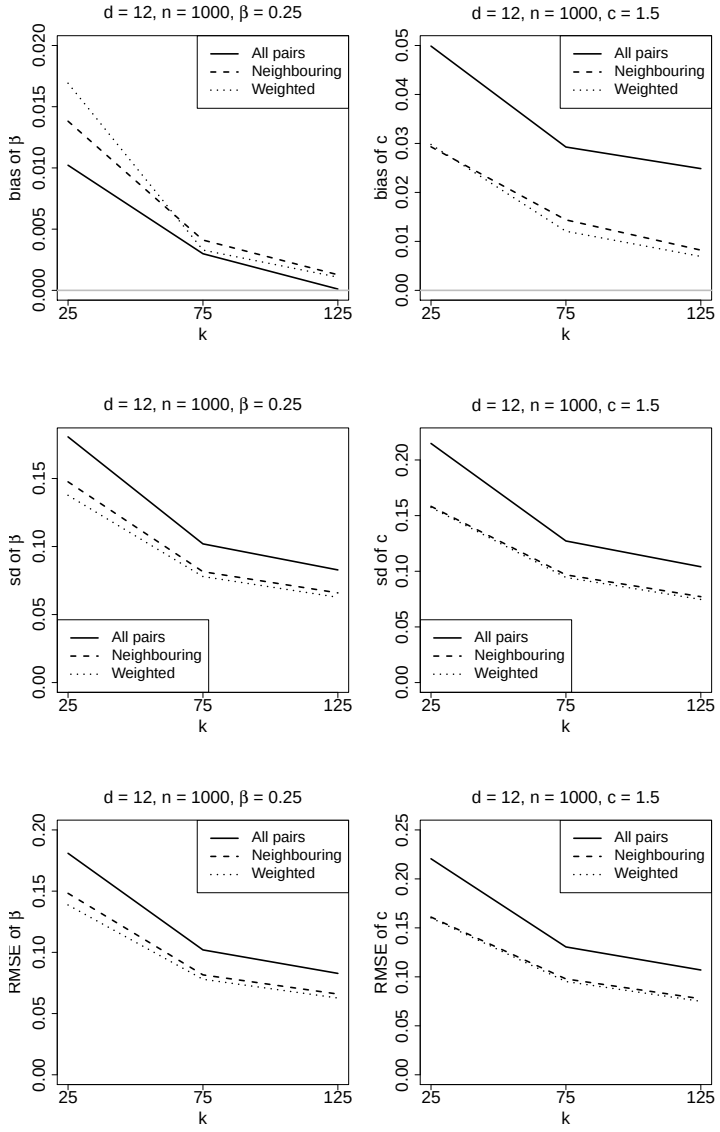


Figure 2.5: Bias, standard deviation and RMSE for estimators of $\beta = 0.25$ (left) and $c = 1.5$ (right) for the perturbed 12-dimensional Brown–Resnick process; 500 samples of $n = 1000$.

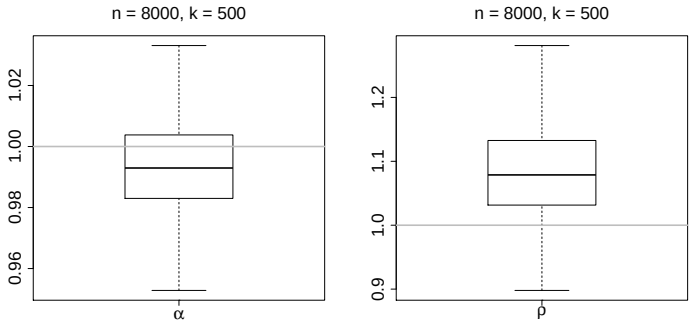


Figure 2.6: Boxplots of estimators of $\alpha = 1$ (left) and $\rho = 1$ (right) for a univariate Brown–Resnick process on the interval $[0, 3]$ with $d = 31$ and $q = 140$; 500 samples of $n = 8000$, $k = 500$.

the Smith model. For sample sizes $n = 5000$ and $n = 10\,000$, multivariate normality was not rejected for any of the values of k we considered.

2.4 Efficiency comparisons

2.4.1 Finite-sample comparisons

A natural question that arises is whether the quality of estimation decreases when making the step from the d-dimensional estimator $\widehat{\boldsymbol{\theta}}'_n$ in (2.2.2) to the pairwise estimator $\widehat{\boldsymbol{\theta}}_n$ in (2.2.4). We will demonstrate for the multivariate logistic model and the Smith model that this is not the case, necessarily in a dimension where $\widehat{\boldsymbol{\theta}}'_n$ can be computed. Recall that the d-dimensional logistic model has stable tail dependence function

$$\ell(\boldsymbol{x}; \theta) = \left(x_1^{1/\theta} + \cdots + x_d^{1/\theta} \right)^{\theta}, \qquad \theta \in [0, 1].$$

We simulate 200 samples of size $n = 1500$ from the logistic model in dimension $d = 5$ with parameter value $\theta_0 = 0.5$ and we assess the quality of our estimates via the bias, standard deviation and RMSE for $k \in \{40, 80, \ldots, 320\}$. The dashed lines in the left panels of Figure 2.7 show the the M-estimator of Einmahl et al. (2012) with the function $g \equiv 1$. The results are the same as in Einmahl et al. (2012, Figure 1). The solid lines show the bias and RMSE for the pairwise M-estimator with $q = 10$ and $\widehat{\Omega}_n = I_q$. We see that the pairwise estimator performs somewhat better in terms of bias and also has the lower minimal RMSE, for $k = 160$. Note that we only show results for the pairwise estimator

with identity weight matrix since using the optimal weight matrix has no effect on the estimator.

Next, consider the Smith model with $d = 4$ locations on an equally spaced unit distance 2×2 grid. We simulate 200 samples of size $n = 5000$ from an isotropic Smith model with parameter value $\theta_0 = \sigma = 2$, i.e., $\Sigma = \sigma I_2$. The right panels of Figure 2.7 show the bias, standard deviation and RMSE for $k \in \{100, \ldots, 600\}$ for the four-dimensional M-estimator with $q = 5$ weight functions, given by $g_m(\boldsymbol{x}) = x_m$ for $m = 1, \ldots, 4$ and $g_m \equiv 1$ for $m = 5$, the pairwise M-estimator with identity weight matrix, and the pairwise M-estimator with optimal weight matrix. We see clearly that the pairwise weighted method is the best one in terms of both bias and RMSE.

2.4.2 Asymptotic variances

Another question is whether the asymptotic variance increases when switching to the pairwise estimator. First, we consider the Smith model on the line with d equidistant locations, i.e.,

$$a_{uv}^2 = \frac{(s_u - s_v)^2}{\sigma}, \qquad s_u, s_v \in \{1, \ldots, d\}.$$

The upper panels of Figure 2.8 show values for the asymptotic variances of a number of estimators when $\sigma \in \{0.5, 1, 1.5, 2\}$ and $d \in \{4, 6\}$. For the d-dimensional estimator $\widehat{\boldsymbol{\theta}}'_n$, we used $g \equiv 1$ as before, and thus $q = 1$; the formula for the asymptotic variance is given in (4.6) in Einmahl et al. (2012). For the pairwise estimator, we computed the asymptotic variance in (2.2.7) in two cases: first, neighbouring pairs only and identity weight matrix, and second, all pairs and the optimal weight matrix. Throughout, both pairwise estimators have a slightly lower asymptotic variance than the d-dimensional estimator.

When the dimension, d, is large, say 100, the method from Einmahl et al. (2012) involves intractable, high-dimensional integrals. For the sake of comparison, we construct a computationally tractable variant of the logistic model that mimics the property of the Smith model that tail dependence vanishes as the distance between locations increases.

Consider d locations in r "regions", every region containing d/r locations. Within all regions, assume a logistic stable tail dependence function, with a common value of $\theta_0 \in [0, 1]$ for all regions; locations in different regions are assumed to be tail independent. The right panel of Figure 2.8 shows the asymptotic variances of a number of estimators for $\theta_0 \in \{0.1, 0.2, \ldots, 0.9\}$, $d = 100$, and $r = 20$. For the d-dimensional estimator, we used again $g \equiv 1$ and $q = 1$. For the pairwise estimator, we used all 10 pairs in each of the 20 regions, yielding $q = 200$ pairs in total; because of symmetry, the optimal weight matrix produces the same asymptotic variance as the identity weight matrix. For most of the parameter values, using the pairwise estimator entails only a modest increase in asymptotic variance. For some parameter values, it even leads to a small decrease.

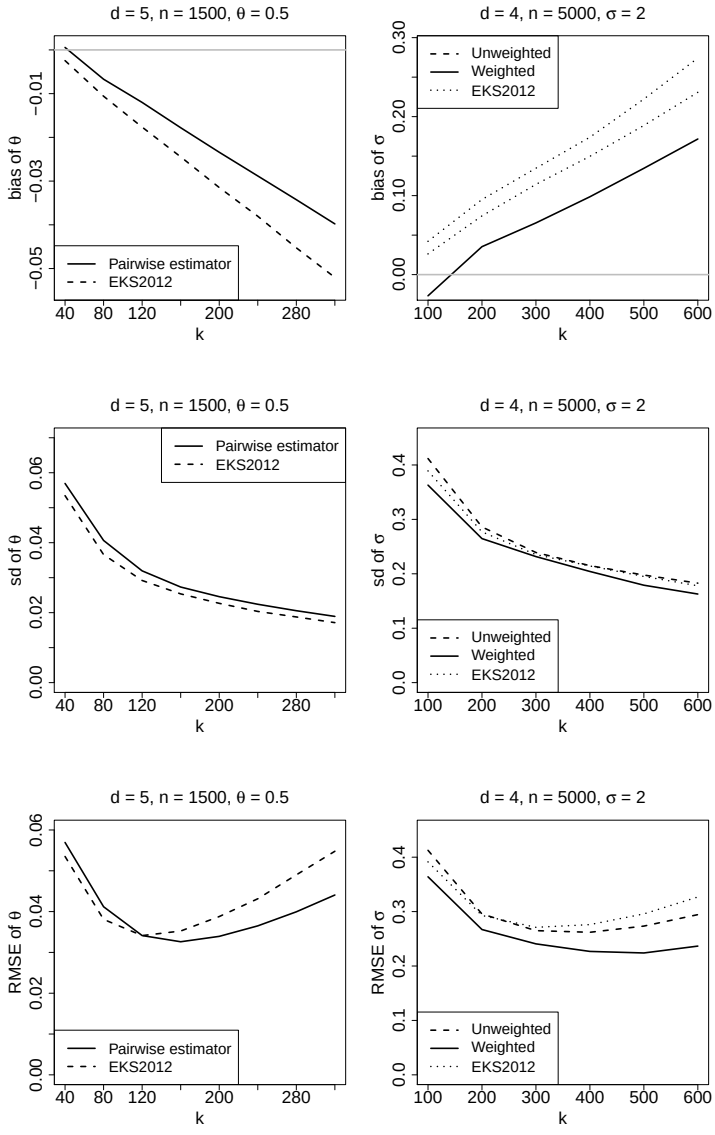


Figure 2.7: Left: bias, standard deviation and RMSE for estimators of $\theta_0 = 0.5$ for the logistic model; 200 samples of $n = 1500$. Right: bias, standard deviation and RMSE for estimators of $\sigma = 2$ for the Smith model; 200 samples of $n = 5000$.

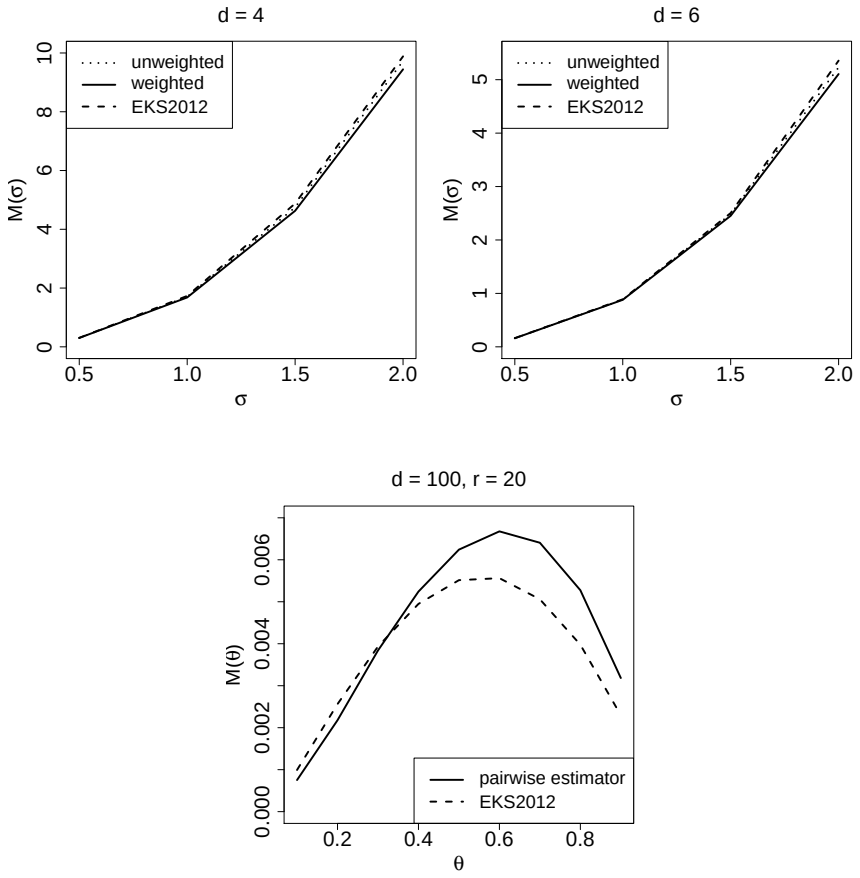


Figure 2.8: Top: asymptotic variance $M(\sigma)$ for $\sigma \in \{0.5, 1, 1.5, 2\}$ and $d \in \{4, 6\}$ for the d-dimensional Smith model on the line; the pairwise estimator with identity weight matrix (unweighted), the pairwise estimator with optimal weight matrix (weighted) and the d-dimensional M-estimator from Einmahl et al. (2012). Bottom: asymptotic variance $M(\theta_0)$ for $\theta_0 \in \{0.1, \ldots, 0.9\}$, $d = 100$ and $r = 20$ for the logistic model; the pairwise estimator with identity weight matrix and the d-dimensional M-estimator from Einmahl et al. (2012).

2.5 Application: speeds of wind gusts

Using extreme-value theory to estimate the frequency and magnitude of extreme wind events or to estimate the return levels for (extremely) long return periods is not a novelty in the fields of meteorology and climatology. Numerous research papers published in the last 20–25 years are applying methods from extreme-value theory to treat those estimation problems, see, for example, Karpa and Naess (2013); Ceppi et al. (2008); Palutikof et al. (1999) and the references therein. However, until very recently, all statistical approaches were univariate. In Engelke et al. (2015) and Oesting et al. (2015), for instance, Brown–Resnick processes are used to model wind speed data.

We consider a data set from the Royal Netherlands Meteorological Institute (KNMI), consisting of the daily maximal speeds of wind gusts, which are measured in 0.1 m/s. The data are observed at 35 weather stations in the Netherlands, over the time period from January 1, 1990 to May 16, 2012. The data set is freely available from `http://www.knmi.nl/climatology/daily_data/selection.cgi`. Due to the strong influence of the sea on the wind speeds in the coastal area, we only consider the inland stations, of which we removed three stations with more than 1000 missing observations. The thus obtained 22 stations and the remaining amount of missing data per station are shown in the left panel of Figure 2.9. We aggregate the daily maxima to three-day maxima in order to minimize temporal dependence and we also restrict our observation period to the summer season (June, July and August) to obtain more or less equally distributed data. To treat the missing data, if at least one of the observations for the three-day maximum is present, we define this to be a valid three-day maximum, thus ignoring these missing observations. We consider a three-day maximum missing only if all three constituting daily maxima are missing. In this way only a few data are missing. We use the "complete deletion approach" for these data and obtain a data set with $n = 672$ observations. This data set is available from the tailDepFun package.

Using the R package extremogram, we first study the univariate sample extremograms to verify if our series of three-day maximal wind speeds exhibit temporal dependence in the extremes (Davis and Mikosch, 2009). For none of the stations we found any significant temporal dependence.

Consider the stable tail dependence function corresponding to the Brown–Resnick process; see Section 2.3.1. It is frequently argued, see e.g. Engelke et al. (2015) or Ribatet (2013), that an anisotropic model is needed to describe the spatial tail dependence of wind speeds. Using Corollary 2.2.5 we first test, based on the $q = 29$ pairs of stations that are at most 50 kilometres apart, if the isotropic process suffices for the above data. In the reparametrization introduced in Section 2.3.2, the case $\tau_{11} = \tau_{22}$ and $\tau_{12} = 0$ corresponds to isotropy. The test statistic

$$k\,(\widehat{\tau}_{11} - \widehat{\tau}_{22}, \widehat{\tau}_{12})\, M_2\,(\widehat{\alpha}, \widehat{\tau}_{11} + \widehat{\tau}_{22}, 0, 0)^{-1}\,(\widehat{\tau}_{11} - \widehat{\tau}_{22}, \widehat{\tau}_{12})^T$$

is computed for $k = 60$. We obtain a value of 0.180, leading to a p-value of 0.914 against the χ^2_2-distribution (Corollary 2.2.5), so we can not reject the null hypothesis. Although the stable tail dependence function corresponding to the more complicated anisotropic Brown–Resnick process is usually assumed for this type of data, the test result shows that the more simple isotropic Brown–Resnick process suffices for the Dutch inland summer season wind speeds.

The estimate of the parameter vector (α, ρ) corresponding to the isotropic Brown–Resnick process is obtained for $k = 60$, with $q = 29$ pairs and using the optimal weight matrix chosen according to the two-step procedure described after Corollary 2.2.3. The estimates, with standard errors in parentheses, are $\widehat{\alpha} = 0.398$ (0.020) and $\widehat{\rho} = 0.372$ (0.810). We also see that the Smith model would not fit these data well since α is much smaller than 2.

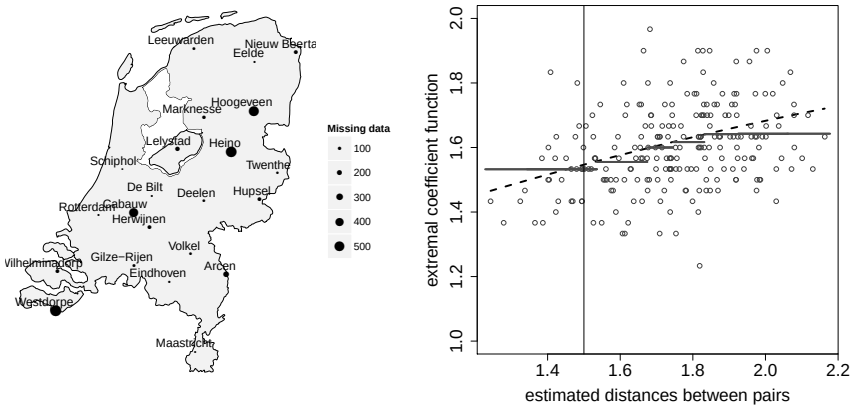


Figure 2.9: KNMI weather stations (left). Estimates of the extremal coefficient function (right).

To visually assess the goodness-of-fit, we compare the nonparametric and the Brown–Resnick model based estimates of the pairwise extremal coefficients function, $\ell(1, 1)$. Instead of presenting them as a function of the actual distance between stations, we exploit the simple expression $\ell(1, 1) = 2\Phi\left(a_{uv}/2\right)$ for the extremal coefficient function of the Brown-Resnick process, see Section 2.3.1.

In the right panel of Figure 2.9, the following are depicted:

- the 231 nonparametric estimates of the extremal coefficient function $\ell(1, 1)$, based on all pairs of stations (circles),
- 6 per-bin averages of the nonparametric estimates of $\ell(1, 1)$ (solid line), and
- the extremal coefficient function values computed from the model (dashed line),

against the estimated distances

$$\widehat{a}_{uv} = \sqrt{2\widehat{\gamma}(\mathbf{s}_u - \mathbf{s}_v)} = \sqrt{2}\left(\frac{\|\mathbf{s}_u - \mathbf{s}_v\|}{\widehat{\rho}}\right)^{\widehat{\alpha}/2}.$$

The vertical line in the plot represents the 50 km threshold. It is more in line with our M-estimator, which uses integration over $[0,1]^2$, to focus on the center $(1/2, 1/2)$ instead of the vertex $(1,1)$ of the unit square. Hence, we use the homogeneity of ℓ to replace $\ell(1,1)$ with $2\ell(1/2,1/2)$ and then estimate the latter with $2\widehat{\ell}_{n,k}(1/2,1/2)$. The nonparametric estimates of $\ell(1,1)$ in the figure are obtained in this way. We see that the estimated $\ell(1,1)$ of the Brown–Resnick process is quite close to the average $2\widehat{\ell}_{n,k}(1/2,1/2)$ per-bin, supporting the adequacy of the model.

2.A Proofs

The notations are as in Section 2.2. Let $\widehat{\Theta}_{n,k}$ denote the (possibly empty) set of minimizers of the function

$$f_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta}) = D_{n,k}(\boldsymbol{\theta})^T\, \widehat{\Omega}_n\, D_{n,k}(\boldsymbol{\theta}) =: \|D_{n,k}(\boldsymbol{\theta})\|^2_{\widehat{\Omega}_n}.$$

Write δ_0 for the Dirac measure concentrated at zero. Write (u,v) for the m-th element of the sequence of pairs $\mathcal{P}$, for $m \in \{1,\dots,q\}$. Let $\boldsymbol{\upsilon} = (\upsilon_1,\dots,\upsilon_q)$ denote a column vector of measures on $\mathbb{R}^d$ whose m-th element is defined as

$$\begin{aligned}\upsilon_m(\mathrm{d}\boldsymbol{x}) &= \upsilon_m(\mathrm{d}x_1 \times \cdots \times \mathrm{d}x_d) = \upsilon_{m1}(\mathrm{d}x_1) \times \cdots \times \upsilon_{md}(\mathrm{d}x_d)\\ &\coloneqq \mathrm{d}x_u\, \mathrm{d}x_v \prod_{j \neq u,v} \delta_0(\,\mathrm{d}x_j),\end{aligned}$$

so that υ_{mj} is the Lebesgue measure if $j = u$ or $j = v$, and υ_{mj} is the Dirac measure at zero for $j \neq u, v$. Using the measures υ_m allows us to write

$$D_{n,k}(\theta) = \left(\int_{[0,1]^d} \left\{\widehat{\ell}_{n,k}(\boldsymbol{x}) - \ell(\boldsymbol{x};\theta)\right\}\, \upsilon_m(\mathrm{d}\boldsymbol{x})\right)_{m=1}^{q} = \widehat{\boldsymbol{L}}_{n,k} - L(\boldsymbol{\theta}).$$

Lemma 2.A.1. *If $0 < \lambda_{n,1} \leq \cdots \leq \lambda_{n,q}$ and $0 < \lambda_1 \leq \cdots \leq \lambda_q$ denote the ordered eigenvalues of the symmetric matrices $\widehat{\Omega}_n$ and $\Omega \in \mathbb{R}^{q\times q}$, respectively, then, as $n \to \infty$,*

$$\widehat{\Omega}_n \xrightarrow{p} \Omega \qquad \text{implies} \qquad (\lambda_{n,1},\dots,\lambda_{n,q}) \xrightarrow{p} (\lambda_1,\dots,\lambda_q).$$

Proof of Lemma 2.A.1. The convergence $\widehat{\Omega}_n \xrightarrow{p} \Omega$ elementwise implies $\|\widehat{\Omega}_n - \Omega\| \xrightarrow{p} 0$ for any matrix norm $\|\cdot\|$ on $\mathbb{R}^{q\times q}$. If we take the spectral norm $\|\Omega\|$

(i.e., $\|\Omega\|^2$ is the largest eigenvalue of $\Omega^T\Omega$), then Weyl's perturbation theorem (Jiang, 2010, page 145) states that

$$\max_{i=1,\ldots,q} |\lambda_{n,i} - \lambda_i| \leq \|\widehat{\Omega}_n - \Omega\|,$$

so that the desired result follows immediately. □

By the diagonalization of $\widehat{\Omega}_n$ in terms of its eigenvectors and eigenvalues, the norm $\|\cdot\|_{\widehat{\Omega}_n}$ is equivalent to the Euclidian norm $\|\cdot\|$ in the sense that

$$\lambda_{n,1}\|D_{n,k}(\boldsymbol{\theta})\|^2 \leq \|D_{n,k}(\boldsymbol{\theta})\|^2_{\widehat{\Omega}_n} \leq \lambda_{n,q}\|D_{n,k}(\boldsymbol{\theta})\|^2.$$

Proof of Theorem 2.2.1. Let $\varepsilon_0 > 0$ be such that the closed ball $B_{\varepsilon_0}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \varepsilon_0\}$ is a subset of Θ; such an ε_0 exists since $\boldsymbol{\theta}_0$ is an interior point of Θ. Fix $\varepsilon > 0$ such that $0 < \varepsilon \leq \varepsilon_0$. We first show that

$$\mathbb{P}[\widehat{\Theta}_n \neq \varnothing \text{ and } \widehat{\Theta}_{n,k} \subset B_\varepsilon(\theta_0)] \to 1, \qquad n \to \infty. \tag{2.A.1}$$

Because L is a homeomorphism, there exists $\delta > 0$ such that for $\boldsymbol{\theta} \in \Theta$, $\|L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_0)\| \leq \delta$ implies $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \varepsilon$. Equivalently, for every $\boldsymbol{\theta} \in \Theta$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \varepsilon$ we have $\|L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_0)\| > \delta$. Define the event

$$A_n = \left\{ \left\| L(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{L}}_{n,k} \right\| < \frac{\delta\sqrt{\lambda_{n,1}}}{(1+\sqrt{\lambda_{n,1}})\max\left(1, \sqrt{\lambda_{n,q}}\right)} \right\}.$$

If $\boldsymbol{\theta} \in \Theta$ is such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \varepsilon$, then on the event A_n, we have

$$\begin{aligned}
\|D_{n,k}(\boldsymbol{\theta})\|_{\widehat{\Omega}_n} &\geq \sqrt{\lambda_{n,1}}\,\|D_{n,k}(\boldsymbol{\theta})\| \\
&= \sqrt{\lambda_{n,1}} \left\| L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta}) - \left(L(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{L}}_{n,k} \right) \right\| \\
&\geq \sqrt{\lambda_{n,1}} \left(\|L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta})\| - \|L(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{L}}_{n,k}\| \right) \\
&> \sqrt{\lambda_{n,1}} \left(\delta - \frac{\delta\sqrt{\lambda_{n,1}}}{1+\sqrt{\lambda_{n,1}}} \right) = \frac{\sqrt{\delta\lambda_{n,1}}}{1+\sqrt{\lambda_{n,1}}}.
\end{aligned}$$

It follows that on A_n,

$$\begin{aligned}
\inf_{\theta:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\varepsilon} \|D_{n,k}(\boldsymbol{\theta})\|_{\widehat{\Omega}_n} &\geq \frac{\delta\sqrt{\lambda_{n,1}}}{1+\sqrt{\lambda_{n,1}}} > \sqrt{\lambda_{n,q}} \left\| L(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{L}}_{n,k} \right\| \\
&\geq \left\| L(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{L}}_{n,k} \right\|_{\widehat{\Omega}_n} \geq \inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\varepsilon} \left\| L(\boldsymbol{\theta}) - \widehat{\boldsymbol{L}}_{n,k} \right\|_{\widehat{\Omega}_n}.
\end{aligned}$$

The infimum on the right-hand side is actually a minimum since L is continuous and $B_\varepsilon(\boldsymbol{\theta}_0)$ is compact. Hence on A_n the set $\widehat{\Theta}_{n,k}$ is non-empty and $\widehat{\Theta}_{n,k} \subset B_\varepsilon(\boldsymbol{\theta}_0)$.

To show (2.A.1), it remains to be shown that $\mathbb{P}[A_n] \to 1$ as $n \to \infty$. Uniform consistency of $\widehat{\ell}_{n,k}$ for $d = 2$ was shown in Huang (1992); see also de Haan and Ferreira (2006, page 237). The proof for $d > 2$ is a straightforward extension. By the continuous mapping theorem, it follows that $\widehat{\boldsymbol{L}}_{n,k}$ is consistent for $L(\boldsymbol{\theta}_0)$. By Lemma 2.A.1, $\lambda_{n,m}$ is consistent for λ_m for all $m \in \{1, \ldots, q\}$. This yields $\mathbb{P}[A_n] \to 1$ and thus (2.A.1).

Next we wish to prove that, with probability tending to one, $\widehat{\Theta}_{n,k}$ has exactly one element, i.e., the function $f_{n,k,\widehat{\Omega}_n}$ has a unique minimizer. To do so, we will show that there exists $\varepsilon_1 \in (0, \varepsilon_0]$ such that, with probability tending to one, the Hessian of $f_{n,k,\widehat{\Omega}_n}$ is positive definite on $B_{\varepsilon_1}(\boldsymbol{\theta}_0)$ and thus $f_{n,k,\widehat{\Omega}_n}$ is strictly convex on $B_{\varepsilon_1}(\boldsymbol{\theta}_0)$. In combination with (2.A.1) for $\varepsilon \in (0, \varepsilon_1]$, this will yield the desired conclusion.

For $\boldsymbol{\theta} \in \Theta$, define the symmetric $p \times p$ matrix $\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$ by

$$\big(\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)\big)_{i,j} \coloneqq 2 \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right)^T \Omega \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right) - 2 \left(\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_i}\right) \Omega \big(L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta})\big)$$

for $i, j \in \{1, \ldots, p\}$. The map $\boldsymbol{\theta} \mapsto \mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$ is continuous and

$$\mathcal{H}(\boldsymbol{\theta}_0) \coloneqq \mathcal{H}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0) = 2\, \dot{L}(\boldsymbol{\theta}_0)^T\, \Omega\, \dot{L}(\boldsymbol{\theta}_0),$$

is a positive definite matrix. Let $\|\cdot\|$ denote a matrix norm. By an argument similar to that in the proof of Lemma 2.A.1, there exists $\eta > 0$ such that every symmetric matrix $A \in \mathbb{R}^{p \times p}$ with $\|A - \mathcal{H}(\boldsymbol{\theta}_0)\| \leq \eta$ has positive eigenvalues and is therefore positive definite. Let $\varepsilon_1 \in (0, \varepsilon_0]$ be sufficiently small such that the second-order partial derivatives of L are continuous on $B_{\varepsilon_1}(\boldsymbol{\theta}_0)$ and such that $\|\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0) - \mathcal{H}(\boldsymbol{\theta}_0)\| \leq \eta/2$ for all $\boldsymbol{\theta} \in B_{\varepsilon_1}(\boldsymbol{\theta}_0)$.

Let $\mathcal{H}_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta}) \in \mathbb{R}^{p \times p}$ denote the Hessian matrix of $f_{n,k,\widehat{\Omega}_n}$. Its (i,j)-th element is

$$\begin{aligned}
\big(\mathcal{H}_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta})\big)_{ij} &= \frac{\partial^2}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_i} \Big[D_{n,k}(\boldsymbol{\theta})^T\, \widehat{\Omega}_n\, D_{n,k}(\boldsymbol{\theta})\Big] \\
&= \frac{\partial}{\partial \boldsymbol{\theta}_j} \left[-2 D_{n,k}(\boldsymbol{\theta})^T\, \widehat{\Omega}_n \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right] \\
&= 2 \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right)^T \widehat{\Omega}_n \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right) - 2 \left(\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_i}\right) \widehat{\Omega}_n\, D_{n,k}(\boldsymbol{\theta}).
\end{aligned}$$

Since $D_{n,k}(\boldsymbol{\theta}) = \widehat{\boldsymbol{L}}_{n,k} - L(\boldsymbol{\theta})$ and since $\widehat{\boldsymbol{L}}_{n,k}$ converges in probability to $L(\boldsymbol{\theta}_0)$, we obtain

$$\sup_{\boldsymbol{\theta} \in B_{\varepsilon_1}(\boldsymbol{\theta}_0)} \|\mathcal{H}_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta}) - \mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)\| \xrightarrow{p} 0, \qquad n \to \infty. \tag{2.A.2}$$

By the triangle inequality, it follows that

$$\mathbb{P}\left[\sup_{\boldsymbol{\theta} \in B_{\varepsilon_1}(\boldsymbol{\theta}_0)} \|\mathcal{H}_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta}) - \mathcal{H}(\boldsymbol{\theta}_0)\| \leq \eta\right] \to 1, \qquad n \to \infty.$$

In view of our choice for η, this implies that, with probability tending to one, $\mathcal{H}_{n,k}(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta} \in B_{\varepsilon_1}(\boldsymbol{\theta}_0)$, as required. □

Proof of Theorem 2.2.2. First note that, as $n \to \infty$,

$$\sqrt{k}\, D_{n,k}(\boldsymbol{\theta}_0) \xrightarrow{d} \widetilde{B}, \qquad \text{where } \widetilde{B} \sim \mathcal{N}_q(\mathbf{0}, \Gamma(\boldsymbol{\theta}_0)).$$

This follows directly from Einmahl et al. (2012, Proposition 7.3) by replacing $g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ with $\upsilon(\mathrm{d}\boldsymbol{x})$. Also, from $(C2)$ and Slutsky's lemma, we have

$$\begin{aligned}\sqrt{k}\, \nabla f_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta}_0) &= -2\sqrt{k}\, D_{n,k}(\boldsymbol{\theta}_0)^T\, \widehat{\Omega}_n\, \dot{L}(\boldsymbol{\theta}_0) \\ &\xrightarrow{d} -2\, \widetilde{B}^T\, \Omega\, \dot{L}(\boldsymbol{\theta}_0) \sim \mathcal{N}_p\big(\mathbf{0},\ 4\, \dot{L}(\boldsymbol{\theta}_0)^T\, \Omega\, \Gamma(\boldsymbol{\theta}_0)\, \Omega\, \dot{L}(\boldsymbol{\theta})\big).\end{aligned}$$

Since $\widehat{\boldsymbol{\theta}}_n$ is a minimizer of $\widehat{f}_{k,n}$ we have $\nabla f_{n,k,\widehat{\Omega}_n}(\widehat{\boldsymbol{\theta}}_n) = 0$. Applying the mean value theorem to the function $t \mapsto \nabla f_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta}_0 + t(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0))$ at $t = 0$ and $t = 1$ yields

$$0 = \nabla f_{n,k,\widehat{\Omega}_n}(\widehat{\boldsymbol{\theta}}_n) = \nabla f_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta}_0) + \mathcal{H}_{n,k,\widehat{\Omega}_n}(\widetilde{\boldsymbol{\theta}}_n)\, (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

where $\widetilde{\boldsymbol{\theta}}_n$ is a random vector on the segment connecting $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_n$. As $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$, we have $\widetilde{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ too. By (2.A.2) and continuity of $\boldsymbol{\theta} \mapsto \mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$, it then follows that $\mathcal{H}_{n,k,\widehat{\Omega}_n}(\widetilde{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathcal{H}(\boldsymbol{\theta}_0)$. Putting these facts together, we conclude that

$$\sqrt{k}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\big(\mathcal{H}_{n,k,\widehat{\Omega}_n}(\widetilde{\boldsymbol{\theta}}_n)\big)^{-1}\, \sqrt{k}\, \nabla f_{n,k,\widehat{\Omega}_n}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_p\big(0, M(\boldsymbol{\theta}_0)\big),$$

as required. □

Proof of Corollary 2.2.3. Assumption (C6) implies that the map $\boldsymbol{\theta} \mapsto \Gamma(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_0$ (Einmahl et al., 2008, Lemma 7.2). Further, $\Gamma(\widehat{\boldsymbol{\theta}}_n^{(0)})^{-1}$ converges in probability to $\Gamma(\boldsymbol{\theta}_0)^{-1}$, because of the continuous mapping theorem and the fact that $\widehat{\boldsymbol{\theta}}_n^{(0)}$ is a consistent estimator of $\boldsymbol{\theta}_0$. Finally, the choice $\Omega_{\mathrm{opt}} = \Gamma(\boldsymbol{\theta})^{-1}$ in (2.2.7) leads to the minimal matrix $M_{\mathrm{opt}}(\boldsymbol{\theta})$ in (2.2.8); see for example Abadir and Magnus (2005, page 339). □

2.B Calculating the asymptotic variance matrix

Let $(u, v), (u', v') \in \mathcal{P}$ denote the i-th and j-th element of the sequence $\mathcal{P}$ respectively. Let $\dot{\ell}_{uv,1}(x_u, x_v)$ and $\dot{\ell}_{uv,2}(x_u, x_v)$ denote the partial derivatives of $\ell_{uv}(x_u, x_v)$ with respect to x_u and x_v respectively and define

$$W_{\ell,uv} = W_\Lambda(\{w_u, w_v \in [0,\infty]^2 \setminus \{(\infty,\infty)\} : w_u \le x_u \text{ or } w_v \le x_v\}).$$

Note that

$$B_{uv}(x_u, x_v) = W_{\ell,uv}(x_u, x_v) - \dot{\ell}_{uv,1}(x_u, x_v) W_{\ell,u}(x_u) - \dot{\ell}_{uv,2}(x_u, x_v) W_{\ell,v}(x_v).$$

The (i,j)-th entry of $\Gamma(\boldsymbol{\theta}) \in \mathbb{R}^{q\times q}$ from (2.2.6) is given by

$$\int_{[0,1]^4} \mathbb{E}[B_{uv}(x_u, x_v) B_{u'v'}(x_{u'}, x_{v'})] \, \mathrm{d}\boldsymbol{x} = \int_{[0,1]^4} (T_1 - T_2 - T_3 + T_4 + T_5) \, \mathrm{d}\boldsymbol{x}, \quad (2.B.1)$$

for $\boldsymbol{x} = (x_u, x_v, x_{u'}, x_{v'})$ where

$$\begin{aligned}
T_1 &= \mathbb{E}[W_{\ell,uv}(x_u, x_v) W_{\ell,u'v'}(x_{u'}, x_{v'})], \\
T_2 &= \dot{\ell}_{u'v',1}(x_{u'}, x_{v'}) \mathbb{E}[W_{\ell,uv}(x_u, x_v) W_{\ell,u'}(x_{u'})] \\
&\quad + \dot{\ell}_{u'v',2}(x_{u'}, x_{v'}) \mathbb{E}[W_{\ell,uv}(x_u, x_v) W_{\ell,v'}(x_{v'})], \\
T_3 &= \dot{\ell}_{uv,1}(x_u, x_v) \mathbb{E}[W_{\ell,u}(x_u) W_{\ell,u'v'}(x_{u'}, x_{v'})] \\
&\quad + \dot{\ell}_{uv,2}(x_u, x_v) \mathbb{E}[W_{\ell,v}(x_v) W_{\ell,u'v'}(x_{u'}, x_{v'})], \\
T_4 &= \dot{\ell}_{uv,1}(x_u, x_v) \dot{\ell}_{u'v',1}(x_{u'}, x_{v'}) \mathbb{E}[W_{\ell,u}(x_u) W_{\ell,u'}(x_{u'})] \\
&\quad + \dot{\ell}_{uv,2}(x_u, x_v) \dot{\ell}_{u'v',2}(x_{u'}, x_{v'}) \mathbb{E}[W_{\ell,v}(x_v) W_{\ell,v'}(x_{v'})], \\
T_5 &= \dot{\ell}_{uv,1}(x_u, x_v) \dot{\ell}_{u'v',2}(x_{u'}, x_{v'}) \mathbb{E}[W_{\ell,u}(x_u) W_{\ell,v'}(x_{v'})] \\
&\quad + \dot{\ell}_{uv,2}(x_u, x_v) \dot{\ell}_{u'v',1}(x_{u'}, x_{v'}) \mathbb{E}[W_{\ell,v}(x_v) W_{\ell,u'}(x_{u'})].
\end{aligned}$$

Suppose (u, v, u', v') are all different and define the sets

$$\mathcal{A}_{ij}(z_i, z_j) = \{\boldsymbol{w} \in [0,\infty]^d \setminus \{\boldsymbol{\infty}\} : w_i \le z_i \text{ or } w_j \le z_j\}.$$

Then

$$\begin{aligned}
\mathbb{E}[W_{\ell,uv}(x_u, x_v) W_{\ell,u'v'}(x_{u'}, x_{v'})] &= \mathbb{E}[W_\Lambda(A_{uv}(x_u, x_v)) W_\Lambda(A_{u'v'}(x_{u'}, x_{v'}))] \\
&= \Lambda(A_{uv}(x_u, x_v) \cap A_{u'v'}(x_{u'}, x_{v'})) \\
&= \Lambda(A_{uv}(x_u, x_v)) + \Lambda(A_{u'v'}(x_{u'}, x_{v'})) \\
&\quad - \Lambda(A_{uv}(x_u, x_v) \cup A_{u'v'}(x_{u'}, x_{v'})) \\
&= \ell_{uv}(x_u, x_v) + \ell_{u'v'}(x_{u'}, x_{v'}) \\
&\quad - \ell_{uvu'v'}(x_u, x_v, x_{u'}, x_{v'}).
\end{aligned}$$

Similar calculations for the other terms yield

$$\begin{aligned}
T_1 &= \ell_{uv}(x_u, x_v) + \ell_{u'v'}(x_{u'}, x_{v'}) - \ell_{uvu'v'}(x_u, x_v, x_{u'}, x_{v'}), \\
T_2 &= \dot{\ell}_{u'v',1}(x_{u'}, x_{v'}) [\ell_{uv}(x_u, x_v) + x_{u'} - \ell_{uvu'}(x_u, x_v, x_{u'})]
\end{aligned}$$

$$
\begin{aligned}
&\quad + \dot{\ell}_{u'v',2}(x_{u'}, x_{v'})[\ell_{uv}(x_u, x_v) + x_{v'} - \ell_{uvv'}(x_u, x_v, x_{v'})],\\
T_3 &= \dot{\ell}_{uv,1}(x_u, x_v)[\ell_{u'v'}(x_{u'}, x_{v'}) + x_u - \ell_{uu'v'}(x_u, x_{u'}, x_{v'})]\\
&\quad + \dot{\ell}_{uv,2}(x_u, x_v)[\ell_{u'v'}(x_{u'}, x_{v'}) + x_v - \ell_{vu'v'}(x_v, x_{u'}, x_{v'})],\\
T_4 &= \dot{\ell}_{uv,1}(x_u, x_v)\dot{\ell}_{u'v',1}(x_{u'}, x_{v'})[x_u + x_{u'} - \ell_{uu'}(x_u, x_{u'})]\\
&\quad + \dot{\ell}_{uv,2}(x_u, x_v)\dot{\ell}_{u'v',2}(x_{u'}, x_{v'})[x_v + x_{v'} - \ell_{vv'}(x_v, x_{v'})],\\
T_5 &= \dot{\ell}_{uv,1}(x_u, x_v)\dot{\ell}_{u'v',2}(x_{u'}, x_{v'})[x_u + x_{v'} - \ell_{uv'}(x_u, x_{v'})]\\
&\quad + \dot{\ell}_{uv,2}(x_u, x_v)\dot{\ell}_{u'v',1}(x_{u'}, x_{v'})[x_v + x_{u'} - \ell_{vu'}(x_v, x_{u'})].
\end{aligned}
$$

Integrating directly over $T_1, \ldots, T_5$ is very slow, so we would like to simplify as many terms as possible. Introduce the notations

$$
\begin{aligned}
I(u, v) &\coloneqq \int_0^1 \int_0^1 \ell_{uv}(x_u, x_v)\,\mathrm{d}x_u\,\mathrm{d}x_v,\\
I(u, v; x_u) &\coloneqq \int_0^1 \ell_{uv}(x_u, x_v)\,\mathrm{d}x_v,\\
I_u(u, v; x_u) &\coloneqq \int_0^1 \frac{\partial \ell_{uv}(x_u, x_v)}{\partial x_u}\,\mathrm{d}x_v.
\end{aligned}
$$

Now we can write the four-dimensional integrals in (2.B.1) as:

$$
\begin{aligned}
\int_{[0,1]^4} T_1 &= I(u, v) + I(u', v')\\
&\quad - \int_{[0,1]^4} \ell_{uvu'v'}(x_u, x_v, x_{u'}, x_{v'})\,\mathrm{d}x_u\,\mathrm{d}x_v\,\mathrm{d}x_{u'}\,\mathrm{d}x_{v'},\\
\int_{[0,1]^4} T_2 &= I(u, v)[2I(u', v'; 1) - 1] + 2I(u', v'; 1) - 2I(u', v')\\
&\quad - \int_{[0,1]^3} I_{u'}(u', v'; x_{u'})\ell_{uvu'}(x_u, x_v, x_{u'})\,\mathrm{d}x_{u'}\,\mathrm{d}x_u\,\mathrm{d}x_v\\
&\quad - \int_{[0,1]^3} I_{v'}(u', v'; x_{v'})\ell_{uvv'}(x_u, x_v, x_{v'})\,\mathrm{d}x_{v'}\,\mathrm{d}x_u\,\mathrm{d}x_v,\\
\int_{[0,1]^4} T_3 &= I(u', v')[2I(u, v; 1) - 1] + 2I(u, v; 1) - 2I(u, v)\\
&\quad - \int_{[0,1]^3} I_u(u, v; x_u)\ell_{uu'v'}(x_u, x_{u'}, x_{v'})\,\mathrm{d}x_u\,\mathrm{d}x_{u'}\,\mathrm{d}x_{v'}\\
&\quad - \int_{[0,1]^3} I_v(u, v; x_v)\ell_{vu'v'}(x_v, x_{u'}, x_{v'})\,\mathrm{d}x_v\,\mathrm{d}x_{u'}\,\mathrm{d}x_{v'},\\
\int_{[0,1]^4} T_4 &= [I(u, v) - I(u, v; 1)][1 - 2I(u', v'; 1)]\\
&\quad + [I(u', v') - I(u', v'; 1)][1 - 2I(u, v; 1)]
\end{aligned}
$$

$$
\begin{aligned}
&- \int_{[0,1]^2} I_u(u,v;x_u) I_{u'}(u',v';x_{u'}) \ell_{uu'}(x_u,x_{u'}) \,\mathrm{d}x_u \,\mathrm{d}x_{u'} \\
&- \int_{[0,1]^2} I_v(u,v;x_v) I_{v'}(u',v';x_{v'}) \ell_{vv'}(x_v,x_{v'}) \,\mathrm{d}x_v \,\mathrm{d}x_{v'}, \\
\int_{[0,1]^4} T_5 &= [I(u,v) - I(u,v;1)][1 - 2I(u',v';1)] \\
&+ [I(u',v') - I(u',v';1)][1 - 2I(u,v;1)] \\
&- \int_{[0,1]^2} I_u(u,v;x_u) I_{v'}(u',v';x_{v'}) \ell_{uv'}(x_u,x_{v'}) \,\mathrm{d}x_u \,\mathrm{d}x_{v'} \\
&- \int_{[0,1]^2} I_v(u,v;x_v) I_{u'}(u',v';x_{u'}) \ell_{vu'}(x_v,x_{u'}) \,\mathrm{d}x_v \,\mathrm{d}x_{u'}.
\end{aligned}
$$

For the Brown–Resnick process, we can compute the integrals $I(u,v)$, $I(u,v;x_u)$ and $I_u(u,v;x_u)$ analytically. To calculate $I(u,v)$, we make the change of variables $\log(x_u/x_v) = 2z_1$ and $\log(x_u x_v) = 2z_2$, so that $\mathrm{d}x_u \,\mathrm{d}x_v = 2\exp(2z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2$ and the area we integrate over is the area between the lines $z_2 = z_1$ and $z_2 = -z_1$ for $z_2 < 0$. We obtain, for $a = a_{uv} = \sqrt{2\gamma(V(\boldsymbol{s}_u - \boldsymbol{s}_v))}$

$$
\begin{aligned}
I(u,v) &= \int_{-\infty}^{\infty} \int_{-\infty}^{-|z_1|} \left\{ e^{z_2+z_1} \Phi\left(\frac{a}{2} + \frac{2z_1}{a}\right) \right. \\
&\qquad \left. + e^{z_2-z_1} \Phi\left(\frac{a}{2} - \frac{2z_1}{a}\right) \right\} 2e^{2z_2} \,\mathrm{d}z_2 \,\mathrm{d}z_1 \\
&= \Phi(a/2) + \frac{e^{a^2}\Phi(-3a/2)}{3}.
\end{aligned}
$$

The other two integrals are given by

$$
\begin{aligned}
I(u,v;x_u) &= \frac{1}{2}\Phi\left(\frac{a}{2} - \frac{\log x_u}{a}\right) + x_u \Phi\left(\frac{a}{2} + \frac{\log x_u}{a}\right) \\
&\qquad + \frac{x_u^2 e^{a^2}}{2} \Phi\left(-\frac{3a}{2} - \frac{\log x_u}{a}\right), \\
I_u(u,v;x_u) &= \Phi\left(\frac{a}{2} + \frac{\log x_u}{a}\right) + x_u e^{a^2} \Phi\left(-\frac{3a}{2} - \frac{\log x_u}{a}\right).
\end{aligned}
$$

These functions are implemented in the R package tailDepFun (Kiriliouk, 2016). For multivariate integration, the package cubature was used. The four-dimensional stable tail dependence function, which can be written as a sum of three-dimensional normal distribution functions as in Section 4.1, is computed using the package mvtnorm.

Chapter 3

A continuous updating weighted least squares estimator of tail dependence in high dimensions

Abstract

Likelihood-based procedures are a common way to estimate tail dependence parameters, although they can be hard to compute in high dimensions. Moreover, they are not applicable to non-differentiable models such as those arising from max-linear structural equation models. An adaptive weighted least squares procedure matching nonparametric estimates of the stable tail dependence function with the corresponding values of a parametrically specified proposal yields a novel minimum-distance estimator. The estimator is easy to calculate and applies to a wide range of sampling schemes and tail dependence models. In large samples, it is asymptotically normal with an explicit and estimable covariance matrix. The minimum distance obtained forms the basis of a goodness-of-fit statistic whose asymptotic distribution is chi-square. Extensive Monte Carlo simulations confirm the excellent finite-sample performance of the estimator and demonstrate that it is a strong competitor to currently available methods. The estimator is then applied to disentangle sources of tail dependence in European stock markets. This chapter is based on Einmahl, Kiriliouk and Segers (2016b).

3.1 Introduction

Extreme-value analysis has been applied to measure and manage financial and actuarial risks, assess natural hazards stemming from heavy rainfall, wind storms, and earthquakes, and control processes in the food industry, internet traffic, aviation, and other branches of human activity. Multivariate data gives rise to tail dependence, represented here by the stable tail dependence function

ℓ. Estimating this tail dependence function is the subject of this chapter; fitting tail dependence models for spatial phenomena observed at finitely many sites constitutes an interesting special case.

In high(er) dimensions, the class of tail dependence functions becomes rather unwieldy, and therefore we follow the common route of modelling it parametrically. Note that this is far from imposing a fully parametric model on the data generating process. In particular, we only assume a domain-of-attraction condition at the copula level.

Recall that likelihood-based procedures are not applicable to models involving non-differentiable stable tail dependence functions, such as the ones arising in max-linear models (Wang and Stoev, 2011; Einmahl et al., 2012). Recently, these models seem to have gained in popularity. In Gissibl and Klüppelberg (2015), max-linear structural equation models are introduced, allowing one to model extremes on directed acyclic graphs. While the max-linear model in Wang and Stoev (2011) is based on a random vector with independent, unit Fŕechet distributed components, in Falk et al. (2015) a generalized max-linear model is introduced, which is based on a random vector with any type of dependence structure. Other related work includes a spatial generalization of the max-linear model in Strokorb et al. (2015).

Non-differentiability is not the only motivation for the use of the stable tail dependence function. Even for smooth and simple models like the logistic one, likelihoods can be hard to compute when the dimension is very high. This is why current likelihood methods are usually based on composite likelihoods, relying on pairs or triples of variables only, not exploiting information from higher-dimensional tuples.

It is the goal of this chapter to estimate the true parameter vector $\boldsymbol{\theta}_0$ of the stable tail dependence function ℓ and to assess the goodness-of-fit of the parametric model. The parameter estimator is obtained by comparing, at finitely many points in the domain of ℓ, some initial, typically nonparametric, estimator of the latter with the corresponding values of the parametrically specified proposals, and retaining the parameter value yielding the best match. The method is generic in the sense that it applies to many parametric models, differentiable or not, and to many initial estimators, not only the usual empirical tail dependence function but also, for instance, bias-corrected versions thereof (Fougères et al., 2015; Beirlant et al., 2016). Further, the method avoids integration or differentiation of functions of many variables and can therefore handle joint dependence between many variables simultaneously, more easily than the likelihood methods mentioned earlier and the pairwise M-estimator approach in Chapter 2. This feature is particularly interesting for inferring on higher-order interactions, going beyond mere distance-based dependence models such as those frequently employed for spatial extremes. Finally, in those situations where likelihood methods are applicable, the new estimator is a strong competitor.

The distance between the initial estimator and the parametric candidates

is measured through weighted least squares. The weight matrix may depend on the unknown parameter $\boldsymbol{\theta}$ and is hence estimated simultaneously. The construction of the estimator bears some similarity with the continuous updating generalized method of moments (Hansen et al., 1996); the present estimator, however, is substantially different and does not use moments. The flexible estimation procedure is related to that in Chapter 2, but the continuous updating procedure is new in multivariate extreme-value statistics.

We show that the weighted least squares estimator for the tail dependence function is consistent and asymptotically normal, provided that the initial estimator enjoys these properties too, as is the case for the empirical tail dependence function and its recently proposed bias-corrected variations. The asymptotic covariance matrix is a function of the unknown parameter and can thus be estimated by a plug-in technique. We also provide novel goodness-of-fit tests for the parametric tail dependence model based on a comparison between the nonparametric and the parametric estimators. Under the null hypothesis that the tail dependence model is correctly specified, the test statistics are asymptotically chi-square distributed.

This chapter is organized as follows. In Section 3.2 we present the estimator, the goodness-of-fit statistic, and their asymptotic distributions. Section 3.3 reports on a Monte Carlo simulation study involving a variety of models, as well as a finite-sample comparison of our estimator with estimators based on composite likelihoods. An application to European stock market data is presented in Section 3.4, where we try to disentangle sources of tail dependence stemming from the country of origin (Germany versus France) and the economic sector (chemicals versus insurance), fitting a structural equation model. All proofs are deferred to the appendix.

3.2 Inference on tail dependence parameters

3.2.1 Set-up

Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})$, $i \in \{1, \ldots, n\}$, be random vectors in $\mathbb{R}^d$ with cumulative distribution function F and marginal distribution functions $F_1, \ldots, F_d$. Recall that the stable tail dependence function is defined as

$$\ell(\boldsymbol{x}) \coloneqq \lim_{t \downarrow 0} t^{-1} \, \mathbb{P}[1 - F_1(X_{11}) \leq t x_1 \text{ or } \cdots \text{ or } 1 - F_d(X_{1d}) \leq t x_d], \quad (3.2.1)$$

for $\boldsymbol{x} \in [0, \infty)^d$, provided the limit exists, as we will assume throughout. Existence of the limit is a necessary, but not sufficient, condition for F to be in the max-domain of attraction of a d-variate generalized extreme-value distribution. Henceforth we assume that ℓ belongs to a parametric family $\{\ell(\,\cdot\,; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ with $\Theta \subset \mathbb{R}^p$. Let $\boldsymbol{\theta}_0$ denote the true parameter vector, that is, let $\boldsymbol{\theta}_0$ denote the unique point in Θ such that $\ell(\boldsymbol{x}) = \ell(\boldsymbol{x}; \boldsymbol{\theta}_0)$ for all $\boldsymbol{x} \in [0, \infty)^d$. Our aim is to estimate the parameter $\boldsymbol{\theta}_0$ and to test the goodness-of-fit of the model.

Extremal coefficients constitute a popular summary measure of tail dependence (de Haan, 1984; Smith, 1990; Schlather and Tawn, 2003). For non-empty $J \subset \{1, \dots, d\}$, let $\boldsymbol{e}_J \in \mathbb{R}^d$ be defined by

$$(\boldsymbol{e}_J)_j \coloneqq \begin{cases} 1 & \text{if } j \in J, \\ 0 & \text{if } j \in \{1, \dots, d\} \setminus J. \end{cases} \tag{3.2.2}$$

The extremal coefficients are defined by

$$\ell_J \coloneqq \ell(\boldsymbol{e}_J) = \lim_{t \downarrow 0} t^{-1} \mathbb{P}\left[\max_{j \in J} F_j(X_{1j}) \geq 1 - t\right]. \tag{3.2.3}$$

The extremal coefficients $\ell_J \in [1, |J|]$ can be interpreted as assigning to each subset J the effective number of tail independent variables among $(X_{1j})_{j \in J}$. Note that for $J = \{j_1, j_2\}$, the extremal coefficient $\ell(\boldsymbol{e}_J)$ is equal to $2 - \chi_{j_1, j_2}$, where χ_{j_1, j_2} denotes the tail dependence coefficient corresponding to the variables X_{1j_1}, X_{1j_2}; see (1.3.6).

Comparing initial and parametric estimators of the extremal coefficients is a special case of the inference method that we propose. In fact, Smith (1990) already proposes an estimator based on pairwise ($|J| = 2$) extremal coefficients; see also de Haan and Pereira (2006) and Oesting et al. (2015).

3.2.2 Continuous updating weighted least squares estimator

Let $\widetilde{\ell}_{n,k}$ denote an initial estimator of ℓ based on $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$; some possibilities will be described in Subsection 3.2.5. The estimators $\widetilde{\ell}_{n,k}$ that we will consider depend on an intermediate sequence $k = k_n \in (0, n]$, that is,

$$k \to \infty \quad \text{and } k/n \to 0, \qquad \text{as } n \to \infty. \tag{3.2.4}$$

The sequence k will determine the tail fraction of the data that we will use for inference, see for instance Subsection 3.2.5.

Let $\boldsymbol{c}_1, \dots, \boldsymbol{c}_q \in [0, \infty)^d$, with $\boldsymbol{c}_m = (c_{m1}, \dots, c_{md})$ for $m = 1, \dots, q$, be q points in which we will evaluate ℓ and $\widetilde{\ell}_{n,k}$. Consider the $q \times 1$ column vectors

$$\widehat{\boldsymbol{L}}_{n,k} \coloneqq \big(\widetilde{\ell}_{n,k}(\boldsymbol{c}_m)\big)_{m=1}^q,$$

$$L(\boldsymbol{\theta}) \coloneqq \big(\ell(\boldsymbol{c}_m; \boldsymbol{\theta})\big)_{m=1}^q, \tag{3.2.5}$$

$$D_{n,k}(\boldsymbol{\theta}) \coloneqq \widehat{L}_{n,k} - L(\boldsymbol{\theta}), \tag{3.2.6}$$

where $\boldsymbol{\theta} \in \Theta$. The points $\boldsymbol{c}_1, \dots, \boldsymbol{c}_q$ need to be chosen in such a way that the map $L : \Theta \to \mathbb{R}^q$ is one-to-one, i.e., $\boldsymbol{\theta}$ is identifiable from the values of $\ell(\boldsymbol{c}_1; \boldsymbol{\theta}), \dots, \ell(\boldsymbol{c}_q; \boldsymbol{\theta})$. In particular, we will assume that $q \geq p$, where p is the dimension of the parameter space Θ. Since $\ell(a e_{\{j\}}) = a$ for any stable

tail dependence function ℓ, any $a \in [0, \infty)$ and any $j \in \{1, \ldots, d\}$, we will choose the points $\boldsymbol{c}_m$ in such a way that each point has at least two positive coordinates.

For $\boldsymbol{\theta} \in \Theta$, let $\Omega(\boldsymbol{\theta})$ be a symmetric, positive definite $q \times q$ matrix with ordered eigenvalues $0 < \lambda_1(\boldsymbol{\theta}) \leq \cdots \leq \lambda_q(\boldsymbol{\theta})$ and define

$$f_{n,k}(\boldsymbol{\theta}) := \|D_{n,k}(\boldsymbol{\theta})\|^2_{\Omega(\boldsymbol{\theta})} := D^T_{n,k}(\boldsymbol{\theta})\, \Omega(\boldsymbol{\theta})\, D_{n,k}(\boldsymbol{\theta}). \tag{3.2.7}$$

Our continuous updating weighted least squares estimator for $\boldsymbol{\theta}_0$ is defined as

$$\widehat{\boldsymbol{\theta}}_{n,k} := \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} f_{n,k}(\boldsymbol{\theta}) = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \left\{ D_{n,k}(\boldsymbol{\theta})^T\, \Omega(\boldsymbol{\theta})\, D_{n,k}(\boldsymbol{\theta}) \right\}. \tag{3.2.8}$$

The set of minimizers could be empty or could have more than one element. The present notation, suggesting that there exists a unique minimizer, will be justified in Theorem 3.2.1. If all points $\boldsymbol{c}_m$ are chosen as e_{J_m} in (3.2.2) for some collection $J_1, \ldots, J_q$ of q different subsets of $\{1, \ldots, d\}$, each subset having at least two elements, then we will refer to our estimator as an extremal coefficients estimator.

We will address the optimal choice of $\Omega(\boldsymbol{\theta})$ below. The simplest choice for $\Omega(\boldsymbol{\theta})$ is the identity matrix I_q, yielding an ordinary least squares estimator

$$\widehat{\boldsymbol{\theta}}_{n,k} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \sum_{m=1}^{q} \left(\widetilde{\ell}_{n,k}(\boldsymbol{c}_m) - \ell(\boldsymbol{c}_m; \boldsymbol{\theta}) \right)^2. \tag{3.2.9}$$

This special case of our estimator is similar to the estimator proposed in Nolan et al. (2015) in the more specific context of fitting max-stable distributions to a random sample from such a distribution.

3.2.3 Consistency and asymptotic normality

If L is differentiable at an interior point $\boldsymbol{\theta} \in \Theta$, its total derivative will be denoted by $\dot{L}(\boldsymbol{\theta}) \in \mathbb{R}^{q \times p}$. Differentiability of the map $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta})$ is a basic smoothness condition on the model; we do not assume differentiability of the map $\boldsymbol{x} \mapsto \ell(\boldsymbol{x}; \boldsymbol{\theta})$.

Theorem 3.2.1 (Existence, uniqueness and consistency)**.** *Let $\{\ell(\,\cdot\,; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, with $\Theta \subset \mathbb{R}^p$, be a parametric family of d-variate stable tail dependence functions. Let $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0, \infty)^d$ be $q \geq p$ points such that the map $L : \boldsymbol{\theta} \mapsto (\ell(\boldsymbol{c}_m; \boldsymbol{\theta}))_{m=1}^q$ is a homeomorphism from Θ to $L(\Theta)$. Let the true d-variate distribution function F have stable tail dependence function $\ell(\,\cdot\,; \boldsymbol{\theta}_0)$ for some interior point $\boldsymbol{\theta}_0 \in \Theta$. Assume that L is twice continuously differentiable on a neighbourhood of $\boldsymbol{\theta}_0$ and that $\dot{L}(\boldsymbol{\theta}_0)$ is of full rank; also assume that $\Omega : \Theta \to \mathbb{R}^{q \times q}$ is twice continuously differentiable on a neighbourhood of $\boldsymbol{\theta}_0$. Assume $\lambda_1 := \inf_{\boldsymbol{\theta} \in \Theta} \lambda_1(\boldsymbol{\theta}) > 0$. Finally assume, for $m = 1, \ldots, q$, and for a positive sequence $k = k_n$ satisfying* (3.2.4)*,*

$$\widetilde{\ell}_{n,k}(\boldsymbol{c}_m) \xrightarrow{p} \ell(\boldsymbol{c}_m; \boldsymbol{\theta}_0), \qquad \text{as } n \to \infty. \tag{3.2.10}$$

Then with probability tending to one, the minimizer $\widehat{\boldsymbol{\theta}}_{n,k}$ *in* (3.2.8) *exists and is unique. Moreover,*

$$\widehat{\boldsymbol{\theta}}_{n,k} \xrightarrow{p} \boldsymbol{\theta}_0, \qquad \text{as } n \to \infty.$$

Theorem 3.2.2 (Asymptotic normality)**.** *If in addition to the assumptions of Theorem 3.2.1, the estimator* $\widetilde{\ell}_{n,k}$ *satisfies*

$$\sqrt{k}\, D_{n,k}(\boldsymbol{\theta}_0) = \Big(\sqrt{k}\,\Big\{\widetilde{\ell}_{n,k}(\boldsymbol{c}_m) - \ell(\boldsymbol{c}_m;\boldsymbol{\theta}_0)\Big\}\Big)_{m=1}^{q} \xrightarrow{d} \mathcal{N}_q\big(\mathbf{0}, \Gamma(\boldsymbol{\theta}_0)\big), \quad (3.2.11)$$

as $n \to \infty$*, for some* $q \times q$ *covariance matrix* $\Gamma(\boldsymbol{\theta}_0)$*, then, as* $n \to \infty$*,*

$$\sqrt{k}\,(\widehat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_0) = \big(\dot{\boldsymbol{L}}^T \Omega \dot{\boldsymbol{L}}\big)^{-1} \dot{\boldsymbol{L}}^T \Omega \sqrt{k}\, D_{n,k}(\boldsymbol{\theta}_0) + o_p(1) \quad (3.2.12)$$

$$\xrightarrow{d} \mathcal{N}_p\big(\mathbf{0}, M(\boldsymbol{\theta}_0)\big), \quad (3.2.13)$$

where the $p \times p$ *covariance matrix* $M(\boldsymbol{\theta}_0)$ *is defined by*

$$M(\boldsymbol{\theta}_0) := (\dot{\boldsymbol{L}}^T \Omega \dot{\boldsymbol{L}})^{-1}\, \dot{\boldsymbol{L}}^T \Omega\, \Gamma(\boldsymbol{\theta}_0)\, \Omega \dot{\boldsymbol{L}}\, (\dot{\boldsymbol{L}}^T \Omega \dot{\boldsymbol{L}})^{-1},$$

and the matrices $\dot{\boldsymbol{L}}$ *and* Ω *are evaluated at* $\boldsymbol{\theta}_0$*.*

Provided $\Gamma(\boldsymbol{\theta}_0)$ is invertible, we can choose Ω in such a way that the asymptotic covariance matrix $M(\boldsymbol{\theta}_0)$ is minimal, say $M_{\mathrm{opt}}(\boldsymbol{\theta}_0)$, i.e., the difference $M(\boldsymbol{\theta}_0) - M_{\mathrm{opt}}(\boldsymbol{\theta}_0)$ is positive semi-definite. The minimum is attained at $\Omega(\boldsymbol{\theta}_0) = \Gamma(\boldsymbol{\theta}_0)^{-1}$ and the matrix $M(\boldsymbol{\theta}_0)$ becomes simply

$$M_{\mathrm{opt}}(\boldsymbol{\theta}_0) = \big(\dot{\boldsymbol{L}}(\boldsymbol{\theta}_0)^T\, \Gamma(\boldsymbol{\theta}_0)^{-1}\, \dot{\boldsymbol{L}}(\boldsymbol{\theta}_0)\big)^{-1}, \quad (3.2.14)$$

see for instance Abadir and Magnus (2005, page 339). Now extend the covariance matrix $\Gamma(\boldsymbol{\theta}_0)$ to the whole parameter space Θ by letting the map $\boldsymbol{\theta} \mapsto \Gamma(\boldsymbol{\theta})$ be such that $\Gamma(\boldsymbol{\theta})$ is an invertible covariance matrix and $\Gamma^{-1} : \Theta \to \mathbb{R}^{q \times q}$ satisfies the assumptions on Ω.

Corollary 3.2.3 (Optimal weight matrix)**.** *If in addition to the assumptions of Theorem 3.2.2,* $\widehat{\boldsymbol{\theta}}_{n,k}$ *is the estimator based on the weight matrix* $\Omega(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}$*, then, with* M_{opt} *as in* (3.2.14)*, we have*

$$\sqrt{k}(\widehat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_p\big(\mathbf{0}, M_{\mathrm{opt}}(\boldsymbol{\theta}_0)\big), \qquad \text{as } n \to \infty. \quad (3.2.15)$$

The asymptotic covariance matrices M and M_{opt} in (3.2.12) and (3.2.15), respectively, depend on the unknown parameter vector $\boldsymbol{\theta}_0$ through the matrices $\dot{\boldsymbol{L}}(\boldsymbol{\theta})$, $\Omega(\boldsymbol{\theta})$ and $\Gamma(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. If these matrices vary continuously with $\boldsymbol{\theta}$, then it is a standard procedure to construct confidence regions and hypothesis tests, cf. Einmahl et al. (2012, Corollaries 4.3 and 4.4).

3.2.4 Goodness-of-fit testing

It is of obvious importance to be able to test the goodness-of-fit of the parametric family of tail dependence functions that we intend to use. The basis for such a test is $D_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k})$, the difference vector between the initial and parametric estimators of $\ell(\boldsymbol{c}_m)$ at the estimated value of the parameter.

Corollary 3.2.4. *Under the assumptions of Theorem 3.2.2, we have, as* $n \to \infty$,

$$\begin{aligned}\sqrt{k}\, D_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) &= (I_q - P(\boldsymbol{\theta}_0))\, \sqrt{k}\, D_{n,k}(\boldsymbol{\theta}_0) + o_p(1) \\ &\xrightarrow{d} \mathcal{N}_q\big(\mathbf{0}, (I_q - P(\boldsymbol{\theta}_0))\, \Gamma(\boldsymbol{\theta}_0)\, (I_q - P(\boldsymbol{\theta}_0))^T\big), \qquad (3.2.16)\end{aligned}$$

where $P(\boldsymbol{\theta}_0) := \dot{L}(\dot{L}^T \Omega \dot{L})^{-1}\, \dot{L}^T \Omega$ *has rank* p, $I_q - P(\boldsymbol{\theta}_0)$ *has rank* $q - p$ *and the matrices* $\dot{L}$ *and* Ω *are evaluated at* $\boldsymbol{\theta}_0$.

The easiest case in which (3.2.16) can be exploited is when $\Gamma(\boldsymbol{\theta})$ is invertible and $\Omega(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}$. Then it suffices to consider the minimum attained by the criterion function $f_{n,k}$ in (3.2.7), i.e., the test statistic is just $f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) = \min_{\boldsymbol{\theta}\in\Theta} f_{n,k}(\boldsymbol{\theta})$. Observe that it is important here that we allow Ω to depend on $\boldsymbol{\theta}$.

Corollary 3.2.5. *Let* $q > p$. *If the assumptions of Corollary 3.2.3 are satisfied, in particular if* $\Omega(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}$, *then*

$$k\, f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) \xrightarrow{d} \chi^2_{q-p}, \qquad \text{as } n \to \infty.$$

If $\Omega(\boldsymbol{\theta})$ is different from $\Gamma(\boldsymbol{\theta})^{-1}$, for instance when $\Gamma(\boldsymbol{\theta})$ is not invertible, a goodness-of-fit test can still be based upon (3.2.16) by considering the spectral decomposition of the limiting covariance matrix. For convenience, we suppress the dependence on $\boldsymbol{\theta}$. Let

$$(I_q - P)\, \Gamma\, (I_q - P)^T = V D V^T,$$

where $V = (v_1, \ldots, v_q)$ is an orthogonal $q \times q$ matrix, $V^T V = I_q$, the columns of which are orthonormal eigenvectors, and D is diagonal, $D = \operatorname{diag}(\kappa_1, \ldots, \kappa_q)$, with $\kappa_1 \geq \cdots \geq \kappa_q = 0$ the corresponding eigenvalues, at least p of which are zero, the rank of $I_q - P$ being $q - p$. Let $s \in \{1, \ldots, q-p\}$ be such that $\kappa_s > 0$ and consider the $q \times q$ matrix

$$A := V_s D_s^{-1} V_s^T,$$

where $D_s = \operatorname{diag}(\kappa_1, \ldots, \kappa_s)$ is an $s \times s$ diagonal matrix and where $V_s = (v_1, \ldots, v_s)$ is a $q \times s$ matrix having the first s eigenvectors as its columns.

Corollary 3.2.6. *If the assumptions of Theorem 3.2.2 hold and if* $s \in \{1, \ldots, q-p\}$ *is such that, in a neighbourhood of* $\boldsymbol{\theta}_0$, $\kappa_s(\boldsymbol{\theta}) > 0$ *and the matrix* $A(\boldsymbol{\theta})$ *depends continuously on* $\boldsymbol{\theta}$, *then*

$$k\, D_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k})^T\, A(\widehat{\boldsymbol{\theta}}_{n,k})\, D_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) \xrightarrow{d} \chi^2_s, \qquad \text{as } n \to \infty.$$

Remark 3.2.1. If $\Gamma(\boldsymbol{\theta})$ is invertible for all $\boldsymbol{\theta}$, then we can set $s = q - p$ and $\Omega(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}$. The difference between the two test statistics in Corollaries 3.2.5 and 3.2.6 then converges to zero in probability, i.e., the two tests are asymptotically equivalent under the null hypothesis.

3.2.5 Choice of the initial estimator

Our estimator in (3.2.8) is flexible enough to allow for various initial estimators, perhaps based on exceedances over high thresholds or rather on vectors of componentwise block maxima extracted from a multivariate time series (Bücher and Segers, 2014). Here we will focus on the former case, and more specifically on the empirical tail dependence function and a variant thereof.

For simplicity, we assume that the random vectors $\boldsymbol{X}_i$, $i \in \{1, \ldots, n\}$, are not only identically distributed but also independent, so that they are a random sample from F. Recall that R_{ij}^n denotes the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$ for $j = 1, \ldots, d$; see (1.1.1). For convenience, assume that F is continuous.

Empirical stable tail dependence function

Recall the definition of the empirical tail dependence function $\widehat{\ell}'_{n,k}$ in Section 1.3.2 and a slight modification thereof, allowing for better finite-sample properties,

$$\widehat{\ell}_{n,k}(\boldsymbol{x}) \coloneqq \frac{1}{k} \sum_{i=1}^{n} \mathbb{1} \left\{ R_{i1}^n > n + \frac{1}{2} - k x_1 \text{ or } \cdots \text{ or } R_{id}^n > n + \frac{1}{2} - k x_d \right\}.$$

By Einmahl et al. (2012, Theorem 4.6), the estimator $\widehat{\ell}_{n,k}$ satisfies (3.2.11) under conditions controlling the rate of convergence in (3.2.1) and the growth rate of the intermediate sequence $k = k_n$. The first-order partial derivatives $\dot{\ell}_j(\boldsymbol{x}; \boldsymbol{\theta}_0)$ of $\boldsymbol{x} \mapsto \ell(\boldsymbol{x}; \boldsymbol{\theta}_0)$ are assumed to exist and to be continuous in neighbourhoods of the points $\boldsymbol{c}_m$ for which $c_{mj} > 0$ for $j = 1, \ldots, d$.

In this case, the entries of the matrix $\Gamma(\boldsymbol{\theta})$ in (3.2.11), for $\boldsymbol{\theta}$ in the interior of Θ, are given by

$$\Gamma_{i,j}(\boldsymbol{\theta}) = \mathbb{E}[B(\boldsymbol{c}_i)\, B(\boldsymbol{c}_j)], \qquad i, j \in \{1, \ldots, q\}. \tag{3.2.17}$$

with

$$B(\boldsymbol{c}_i) \coloneqq W_\ell(\boldsymbol{c}_i) - \sum_{j=1}^{d} \dot{\ell}_j(\boldsymbol{c}_i) W_\ell(c_{ij}\, \boldsymbol{e}_j),$$

and with $(W_\ell(\boldsymbol{x}) : \boldsymbol{x} \in [0, \infty)^d)$ a zero-mean Gaussian process with covariance function

$$\mathbb{E}[W_\ell(\boldsymbol{x}_1)\, W_\ell(\boldsymbol{x}_2)] = \ell(\boldsymbol{x}_1) + \ell(\boldsymbol{x}_2) - \ell\left(\max(\boldsymbol{x}_1, \boldsymbol{x}_2)\right).$$

For points $\boldsymbol{c}_i$ of the form $\boldsymbol{e}_J$ in (3.2.2), the expectation in (3.2.17) can be calculated as follows: for non-empty subsets J and K of $\{1,\dots,d\}$,

$$\begin{aligned}\mathbb{E}[B(\boldsymbol{e}_J)\,B(\boldsymbol{e}_K)] = \ell_J + \ell_K - \ell_{J\cup K} &- \sum_{j\in J} \dot{\ell}_{j,J}\,(1+\ell_K-\ell_{\{j\}\cup K}) \\ &- \sum_{k\in K} \dot{\ell}_{k,K}\,(\ell_J+1-\ell_{J\cup\{k\}}) \\ &+ \sum_{j\in J}\sum_{k\in K} \dot{\ell}_{j,J}\dot{\ell}_{k,K}\,(2-\ell_{\{j,k\}}),\end{aligned}$$

where $\ell_J := \ell(\boldsymbol{e}_J;\boldsymbol{\theta}_0)$ and $\dot{\ell}_{j,J} := \dot{\ell}_j(\boldsymbol{e}_J;\boldsymbol{\theta}_0)$.

Bias-corrected estimator

A drawback of $\widehat{\ell}_{n,k}$ is its possibly quickly growing bias as k increases. Recently, two bias-corrected estimators have been proposed. We consider here the kernel-type estimator of Beirlant et al. (2016), which is partly based on (the one in) Fougères et al. (2015).

Consider first a rescaled version of $\widehat{\ell}'_{n,k}$, defined as $\widehat{\ell}_{n,k,a}(\boldsymbol{x}) := a^{-1}\widehat{\ell}'_{n,k}(a\boldsymbol{x})$ for $a>0$. Then define the weighted average

$$\breve{\ell}_{n,k}(\boldsymbol{x}) := \frac{1}{k}\sum_{j=1}^{k} K(a_j)\,\widehat{\ell}_{n,k,a_j}(\boldsymbol{x}), \quad a_j := \frac{j}{k+1},\ j\in\{1,\dots,k\}, \tag{3.2.18}$$

where K is a kernel function, i.e., a positive function on $(0,1)$ such that $\int_0^1 K(u)\,\mathrm{d}u = 1$.

In addition to (3.2.1), we assume there exist a positive function α on $(0,\infty)$ tending to 0 as $t\downarrow 0$ and a non-zero function M on $[0,\infty)^d$ such that for all $\boldsymbol{x}\in[0,\infty)^d$,

$$M(x) \;=\; \lim_{t\downarrow 0}\frac{1}{\alpha(t)}\left[\frac{\mathbb{P}\left[\exists j\in\{1,\dots,d\}: 1-F_j(X_{1j})\le t x_j\right]}{t} - \ell(\boldsymbol{x})\right] \tag{3.2.19}$$

Moreover, we assume a third-order condition on ℓ (Beirlant et al., 2016, equation (3)). In Beirlant et al. (2016, Theorem 1) the asymptotic distribution of $\breve{\ell}_{n,k}$ in (3.2.18) is derived under these three assumptions and for intermediate sequences $k=k_n$ growing faster than the ones considered above. A non-zero asymptotic bias term arises and the idea is to estimate and remove it, thereby obtaining a possibly more accurate estimator.

In order to achieve this bias reduction, the rate function, α, and its index of regular variation, β, need to be estimated. Consider another intermediate sequence $k_1 = k_{1,n}$ such that $k/k_1\to 0$ as $n\to\infty$. The bias-corrected estimator is then defined as

$$\overline{\ell}_{n,k,k_1}(\boldsymbol{x}) := \frac{\breve{\ell}_{n,k}(\boldsymbol{x}) - (k_1/k)^{\widehat{\beta}_{k_1}(\boldsymbol{x})}\widehat{\alpha}_{k_1}(x)\frac{1}{k}\sum_{j=1}^{k} K(a_j)a_j^{-\widehat{\beta}_{k_1}(\boldsymbol{x})}}{\frac{1}{k}\sum_{j=1}^{k}K(a_j)},$$

where $\widehat{\alpha}_{k_1}$ and $\widehat{\beta}_{k_1}$ are the estimators of α and β defined in Beirlant et al. (2016). Under the mentioned conditions, asymptotic normality as in (3.2.11) holds, where the limiting random vector is equal in distribution to $\int_0^1 K(u)u^{-1/2}\,\mathrm{d}u$ times the one corresponding to $\widehat{\ell}_{n,k}$. Here, the growth rate of k can be taken faster than when using $\widehat{\ell}_{n,k}$.

A simple choice for K is a power kernel, i.e, $K(t) = (\tau+1)t^\tau$ for $t \in (0,1)$ and $\tau > -1/2$. Then $\int_0^1 K(u)u^{-1/2}\,\mathrm{d}u = (2+\tau)/(1+2\tau)$. Note that this factor tends to 1 if $\tau \to \infty$. In practice, we take $\tau = 5$ as recommended in Beirlant et al. (2016).

3.3 Simulation studies

We conduct simulation studies for data in the max-domain of attraction of the logistic model, the Brown–Resnick process and the max-linear model. For each model, we report the empirical bias, standard deviation, and root mean squared error (RMSE) of our estimators. We also study the finite-sample performance of the goodness-of-fit statistic of Corollary 3.2.5. All simulations were done in the R statistical software environment (R Core Team, 2015) and all functions used in this chapter are part of the R package tailDepFun; see Chapter 4.

3.3.1 Logistic model

Recall that the d-dimensional logistic model has stable tail dependence function

$$\ell(\boldsymbol{x};\theta) = \left(x_1^{1/\theta} + \cdots + x_d^{1/\theta}\right)^\theta, \qquad \theta \in [0,1].$$

The domain-of-attraction condition (3.2.1) holds for instance if F has continuous margins and its copula is Archimedean with generator $\phi(t) = 1/(t^\theta+1)$, also known as the outer power Clayton copula (Hofert et al., 2015).

Comparison with likelihood methods

In Huser et al. (2015), a comprehensive comparison of likelihood estimators for θ has been performed based on random samples from this copula. We compare those results to our extremal coefficients estimator, i.e., the weighted least squares estimator based on points $\boldsymbol{c}_m$ of the form $\boldsymbol{e}_J$, with J ranging in the collection

$$\mathcal{Q}_a := \big\{J \subset \{1,\ldots,d\} : |J| = a\big\} \tag{3.3.1}$$

for $a \in \{2,3\}$. Moreover, we let $\Omega(\theta)$ be the identity matrix, since by exchangeability of the model, a weighting procedure can bring no improvements.

Following Huser et al. (2015, Section 4.2), we simulated 10 000 random samples of size $n = 10\,000$ from the outer power Clayton copula. For the

likelihood-based estimators, the margins are standardized to the unit Pareto scale via the rank transformation

$$X_{ij}^* := \frac{n}{n + 1/2 - R_{ij}^n}, \qquad i \in \{1, \ldots, n\},\ j \in \{1, \ldots, d\}.$$

We take $d \in \{2, 5, 10, 15, 20, 25, 30\}$ and $\theta_0 \in \{0.3, 0.6, 0.9, 0.95\}$ as in Huser et al. (2015, Section 4.2). Note that in the likelihood setting, this is a very demanding experiment, and three of the ten likelihood-based estimators considered in Huser et al. (2015) are only computed for $d \in \{2, 5, 10\}$. In Huser et al. (2015), threshold probabilities are set to 0.98, corresponding to $k = 200$ in our set-up.

Figures 3.1 and 3.2 show the bias, standard deviation and RMSE of three estimators based on the empirical tail dependence function: the two extremal coefficient estimators mentioned above and the pairwise M-estimator from Einmahl et al. (2016a) as implemented in the R package tailDepFun (Kiriliouk, 2016). As the tuple size changes from pairs to triples, the absolute bias increases but the standard deviation decreases. When dependence is strong, $\theta_0 = 0.3$, the gains in variance offset the losses in bias and the estimator based on $\mathcal{Q}_3$ performs best. Note also that when the dependence is not too weak, the estimators based on extremal coefficients perform better than the pairwise M-estimator of Einmahl et al. (2016a). Finally, our estimation procedures have almost constant RMSE as the dimension increases, in line with the pairwise composite likelihood methods studied in Huser et al. (2015).

Comparing these results to the ten likelihood-based estimators in Huser et al. (2015, Figure 4), we see that our estimators are strong competitors in the sense that they rank highly when comparing RMSEs, and are not dominated by one of the likelihood-based estimators. More precisely, for $\theta_0 = 0.3$, only the likelihood estimators based on the Poisson process representation (Coles and Tawn, 1991) and the multivariate generalized Pareto distribution outperform our estimators; for $\theta_0 = 0.6$, the same two likelihood estimators outperform ours, but only for $d \geq 15$; finally, for $\theta_0 = 0.9$ and $\theta_0 = 0.95$ only the pairwise censored likelihood estimator (Huser and Davison, 2014) has a smaller RMSE than our estimators.

Coverage probabilities

We are interested in the 95% coverage probabilities of our estimators, i.e., in the proportion of time that the 95% confidence interval contains the true parameter value θ_0. To this end, we simulate 1000 replications from the five-dimensional logistic model with sample size $n = 1000$, for a range of parameter values $\theta_0 \in \{0.3, 0.6, 0.9, 0.95\}$ and values $k \in \{50, 100, 150, 200\}$. Table 3.1 shows the results for the weighted least squares estimator for $\mathcal{Q}_3$ based on the empirical tail dependence function and on the bias-corrected tail dependence function; results for $\mathcal{Q}_2$, not shown here, are slightly worse. We see that the coverage probabilities decrease dramatically for the empirical tail dependence

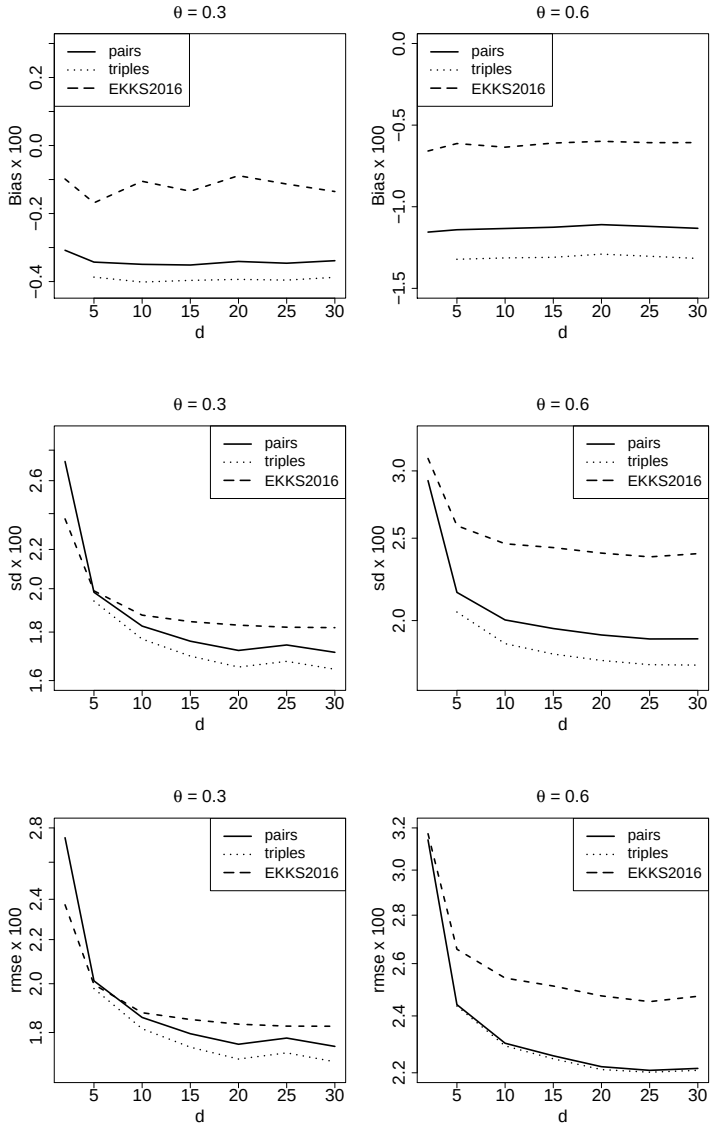


Figure 3.1: Bias, standard deviation and RMSE for estimators of $\theta_0 = 0.3$ (left) and $\theta_0 = 0.6$ (right) for the outer power Clayton copula; 10 000 samples of size $n = 10\,000$. Standard errors and RMSEs are displayed on a logarithmic scale.

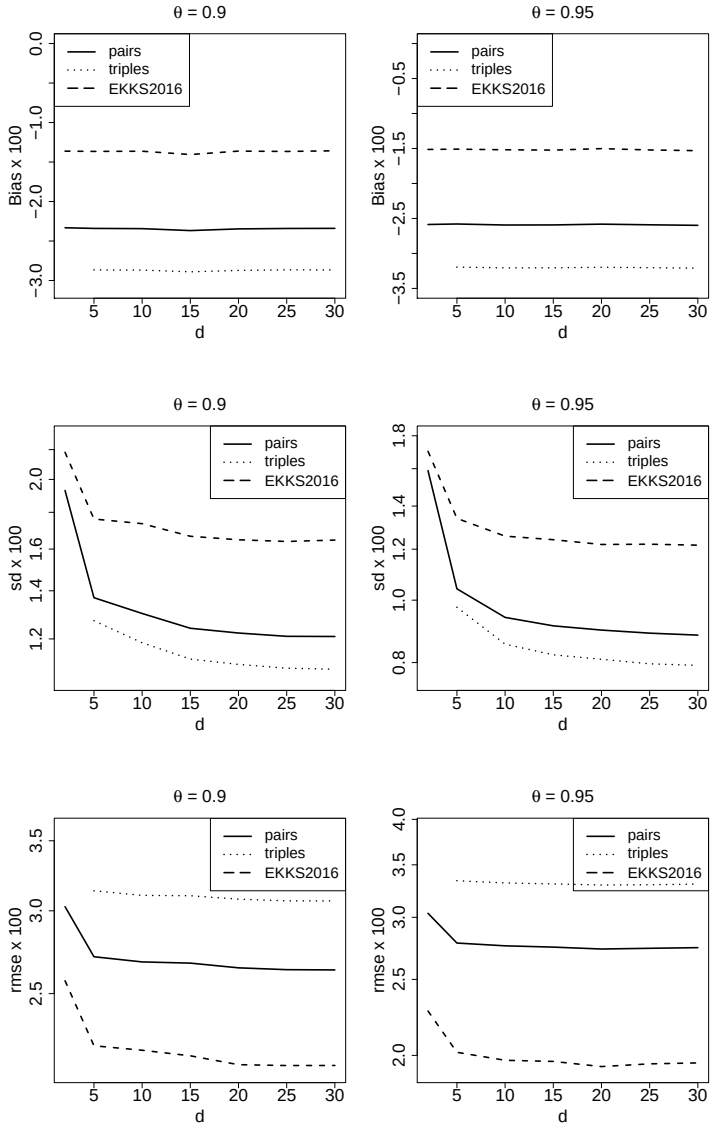


Figure 3.2: Bias, standard deviation and RMSE for estimators of $\theta_0 = 0.9$ (left) and $\theta_0 = 0.95$ (right) for the outer power Clayton copula; 10 000 samples of size $n = 10\,000$. Standard errors and RMSEs are displayed on a logarithmic scale.

function as k increases. The bias-corrected tail dependence function seems to do a good job correcting the bias for larger k, leading to reasonable coverage probabilities.

	Empirical				Bias-corrected			
k	50	100	150	200	50	100	150	200
$\theta_0 = 0.3$	95.0	80.7	53.2	23.4	93.9	84.6	78.2	73
$\theta_0 = 0.6$	97.3	75.7	19.3	0.3	97.6	94.9	91.7	88.8
$\theta_0 = 0.9$	98.0	41.2	0.1	0.0	99.3	97.8	97.3	95.9
$\theta_0 = 0.95$	98.5	26.6	0.0	0.0	99.1	98.7	97.4	96.2

Table 3.1: Observed coverage probabilities for the weighted least squares estimator based on the empirical tail dependence function (left) and the bias-corrected tail dependence function (right) for the logistic model in $d = 5$ for nominal coverage probabilities of 95%; 1000 samples of $n = 1000$.

3.3.2 Brown–Resnick process

Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d$ represent d locations in $\mathcal{S} \subset \mathbb{R}^2$. Recall the definition of the Brown–Resnick process from Subsections 1.4.2 and 2.3.1. From Huser and Davison (2013), we obtain the following representation for the extremal coefficients ℓ_J in (3.2.3),

$$\ell_J = \sum_{j \in J} \Phi_{d-1}(\boldsymbol{\zeta}^{(j)}; \Upsilon^{(j)}), \qquad J \subset \{1, \ldots, d\},\, J \neq \varnothing,$$

where $\boldsymbol{\zeta}^{(j)} = (\zeta_1^{(j)}, \ldots, \zeta_{j-1}^{(j)}, \zeta_{j+1}^{(j)}, \ldots, \zeta_d^{(j)})$ with $\zeta_i^{(j)} = \sqrt{\gamma(\boldsymbol{s}_j - \boldsymbol{s}_i)/2}$, and where $\Upsilon^{(j)}$ is a $(d-1) \times (d-1)$ correlation matrix with entries given by

$$\Upsilon_{ik}^{(j)} = \frac{\gamma(\boldsymbol{s}_j - \boldsymbol{s}_i) + \gamma(\boldsymbol{s}_j - \boldsymbol{s}_k) - \gamma(\boldsymbol{s}_i - \boldsymbol{s}_k)}{2\sqrt{\gamma(\boldsymbol{s}_j - \boldsymbol{s}_i)\,\gamma(\boldsymbol{s}_j - \boldsymbol{s}_k)}}, \qquad i, k \in \{1, \ldots, d\} \setminus \{j\}.$$

We simulate 300 random samples of size $n = 1000$ from the Brown–Resnick process on a 3×4 unit distance grid using the R package SpatialExtremes (Ribatet, 2015). To arrive at a more realistic estimation problem, we perturb the samples thus obtained with additive noise, i.e., if $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id})$ is an observation from the Brown–Resnick process, then we set $X_{ij} = Y_{ij} + |\epsilon_{ij}|$ for $i = 1, \ldots, n$ and $j = 1, \ldots, d$, where ϵ_{ij} are independent $\mathcal{N}(0, 1/4)$ random variables.

We estimate the parameter vector $\boldsymbol{\theta}_0 = (\alpha, \rho) = (1, 1)$ using the extremal coefficient estimator based on the subset of $\mathcal{Q}_2$ in (3.3.1) consisting of pairs of neighbouring locations, i.e., locations that are at most a distance $\sqrt{2}$ apart.

This leads to $q = 29$ pairs. Including pairs of locations that are further away tends to drastically increase the bias (Einmahl et al., 2016a).

Figure 3.3 shows the bias, standard deviation and RMSE for three estimators: the estimator based on the empirical tail dependence function with $\Omega(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}$ (solid lines), the estimator based on the bias-corrected tail dependence function with $\Omega(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}$ (dotted lines), and the pairwise M-estimator from Einmahl et al. (2016a) (dashed lines). We see that for the estimation of the shape parameter $\alpha = 1$ it is better to use one of the estimators based on the empirical stable tail dependence function, whereas for the scale parameter $\rho = 1$ the bias-corrected estimator performs better.

To show the feasibility of the estimation procedure in high dimensions, we simulate 300 samples of size $n = 1000$ from the perturbed Brown–Resnick process on a 10×15 unit-distance grid ($d = 150$), using again $(\alpha, \rho) = (1, 1)$ and selecting pairs of neighbouring locations only, yielding $q = 527$ pairs in total. Figure 3.4 show the bias, standard deviation and RMSE for the estimator based on the empirical tail dependence function with $\Omega(\boldsymbol{\theta}) = I_q$ (solid lines), the estimator based on the bias-corrected tail dependence function with $\Omega(\boldsymbol{\theta}) = I_q$ (dotted lines), and the pairwise M-estimator from Einmahl et al. (2016a) (dashed lines). Compared to $d = 12$ above, the estimation of α has improved whereas the estimation quality of ρ stays roughly the same.

3.3.3 Max-linear models on directed acyclic graphs

Recall the definition of a max-linear model from Subsection 1.3.4; it has stable tail dependence function

$$\ell(\boldsymbol{x}) = \sum_{t=1}^{r} \max_{j=1,\dots,d} b_{jt} x_j, \qquad x \in [0, \infty)^d, \tag{3.3.2}$$

where the factor loadings b_{jt} are non-negative constants such that $\sum_{t=1}^{r} b_{jt} = 1$ for every $j \in \{1, \dots, d\}$ and all column sums of the $d \times r$ matrix $B := (b_{jt})_{j,t}$ are positive. Since the rows of B sum up to one, it has only $d \times (r-1)$ free elements. Further structure may be added to the coefficient matrix B, leading to parametric models whose parameter dimension is lower than $d \times (r-1)$; see below. Even then, the map L in (3.2.5) induced by restricting the points $\boldsymbol{c}_m$ to be of the form $\boldsymbol{e}_J$ in (3.2.2) is typically not one-to-one. Therefore, we really need more general choices of the points $\boldsymbol{c}_m$ in the definition of the estimator.

In Gissibl and Klüppelberg (2015), a link is established between max-linear models and structural equation models, from which graphical models based on directed acyclic graphs (DAGs) can be constructed. A max-linear structural equation model is defined via

$$Y_j = \max_{k \in \mathrm{pa}(j)} u_{kj} Y_k \vee u_j Z_j, \qquad j = 1, \dots, d,$$

where $\mathrm{pa}(j) \subset \{1, \dots, d\}$ denotes the set of parents of node j in the graph, $u_{kj} > 0$ for all $k \in \mathrm{pa}(j) \cup \{j\}$ and $u_j > 0$ for all $j \in \{1, \dots, d\}$. We let $Z_1, \dots, Z_d$ be

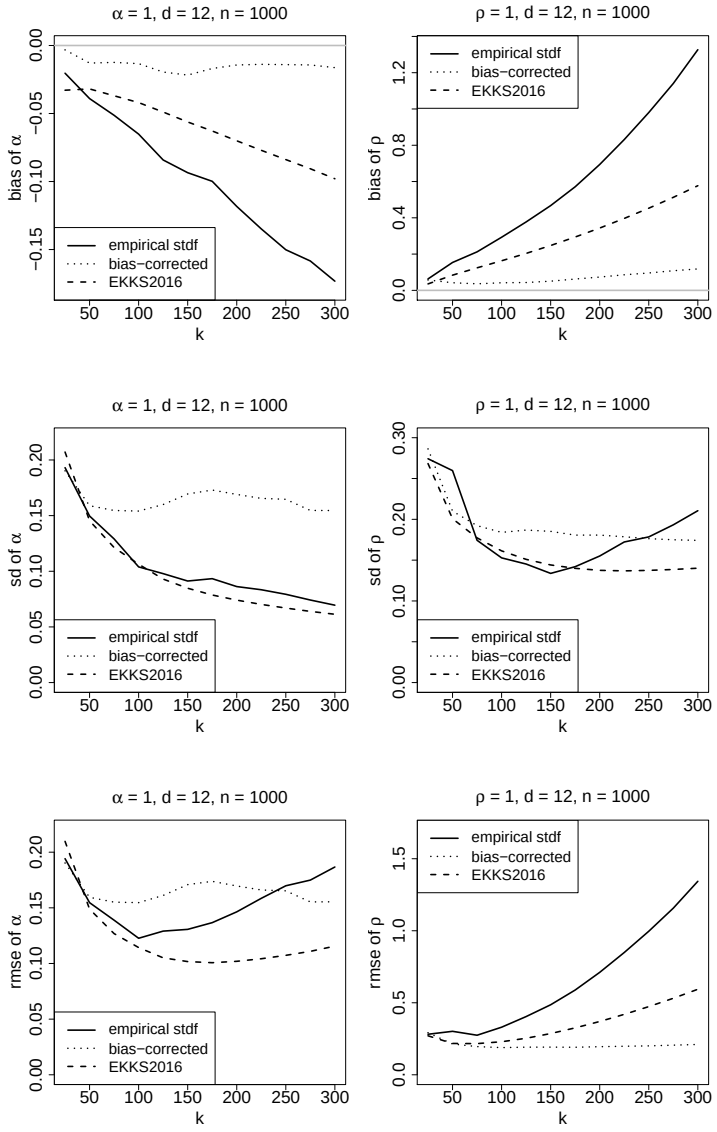


Figure 3.3: Bias, standard deviation and RMSE for estimators of $\alpha = 1$ (left) and $\rho = 1$ (right) for the perturbed 12-dimensional Brown–Resnick process; 300 samples of size $n = 1000$.

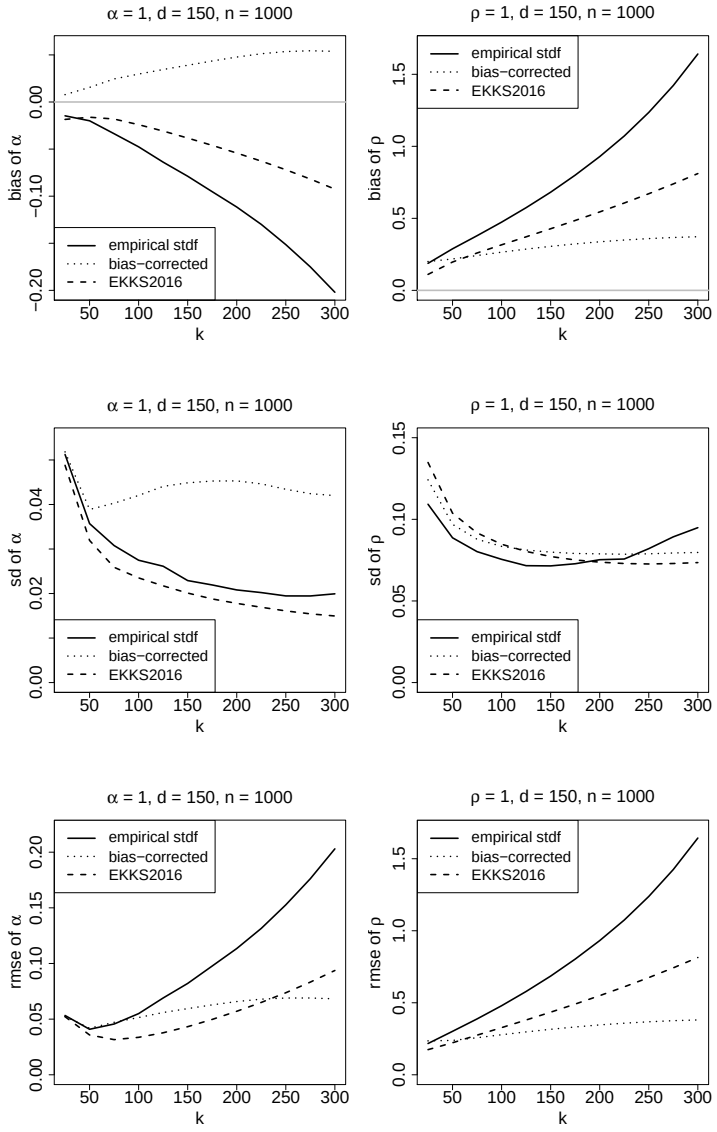
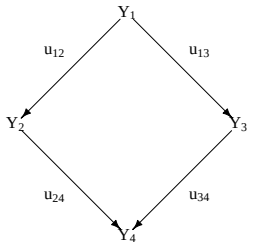


Figure 3.4: Bias, standard deviation and RMSE for estimators of $\alpha = 1$ (left) and $\rho = 1$ (right) for the perturbed 150-dimensional Brown–Resnick process; 300 samples of size $n = 1000$.

independent unit Fréchet random variables. A max-linear structural equation model can then be written as a max-linear model with parameters determined by the paths of the corresponding graph.

We focus on the four-dimensional model corresponding to the following directed acyclic graph (Gissibl and Klüppelberg, 2015, Example 2.1):



$$
\begin{aligned}
Y_1 &= u_1 Z_1,\\
Y_2 &= u_{12} Y_1 \vee u_2 Z_2 = u_{12} u_1 Z_1 \vee u_2 Z_2,\\
Y_3 &= u_{13} Y_1 \vee u_3 Z_3 = u_{13} u_1 Z_1 \vee u_3 Z_3,\\
Y_4 &= u_{24} Y_2 \vee u_{34} Y_3 \vee u_4 Z_4\\
&= (u_{24} u_{12} u_1 \vee u_{34} u_{13} u_1) Z_1 \vee u_{24} u_2 Z_2\\
&\qquad \vee u_{34} u_3 Z_3 \vee u_4 Z_4.
\end{aligned}
$$

Note that we rewrote the max-linear structural equation model to resemble the construction (1.3.16). If we require $Y_1, \ldots, Y_4$ to be unit Fréchet, we need to set $u_1 = 1$ and the matrix of factor loadings becomes

$$
B = \begin{pmatrix}
1 & 0 & 0 & 0\\
u_{12} & u_2 & 0 & 0\\
u_{13} & 0 & u_3 & 0\\
u_{12}u_{24} \vee u_{13}u_{34} & u_2 u_{24} & u_3 u_{34} & u_4
\end{pmatrix},
$$

where the diagonal elements u_j for $j \in \{2, 3, 4\}$ are such that the row sums are equal to one. The parameter vector is then given by $\boldsymbol{\theta} = (u_{12}, u_{13}, u_{24}, u_{34})$.

We conduct a simulation study based on 300 samples of size $n = 1000$ from the four-dimensional model with tail dependence function (3.3.2) and B as above, with parameter vector $\boldsymbol{\theta}_0 = (0.3, 0.8, 0.4, 0.55)$. As before, we put $X_{ij} = Y_{ij} + |\epsilon_{ij}|$, with $(Y_{i1}, \ldots, Y_{id})$ as above and ϵ_{ij} independent $\mathcal{N}(0, 1/4)$ random variables. The estimators are based on the $q = 72$ points $\boldsymbol{c}_m$ on the grid $\{0, 1/2, 1\}^4$ having at least two positive coordinates.

Figures 3.5 and 3.6 show the bias, standard deviation and RMSE for the estimator based on the empirical tail dependence function with $\Omega(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}$ (solid lines), the estimator based on the bias-corrected tail dependence function with $\Omega(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}$ (dotted lines) and the pairwise M-estimator from Einmahl et al. (2016a) (dashed lines). The difference between the pairwise M-estimator and our estimators based on the empirical tail dependence function is negligible. The estimators based on the empirical tail dependence function perform better than the ones based on the bias-corrected version, especially for the parameters u_{13} and u_{24}.

Remark 3.3.1. For the weight matrix, we actually defined $\Omega(\boldsymbol{\theta})$ as $(\Gamma(\boldsymbol{\theta}) + aI_q)^{-1}$ for some small $a > 0$. The reason for applying such a Tikhonov cor-

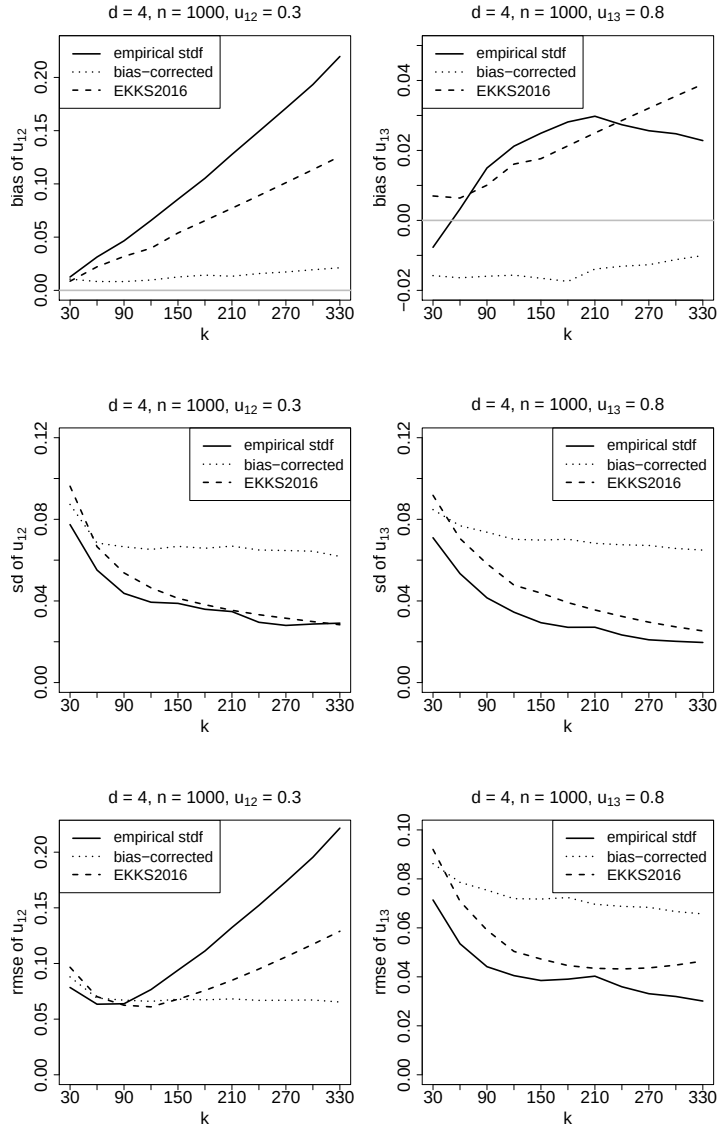


Figure 3.5: Bias, standard deviation and RMSE for estimators of $u_{12} = 0.3$ (left) and $u_{13} = 0.8$ (right) for the max-linear structural equation model based on a directed acyclic graph; 300 samples of size $n = 1000$.

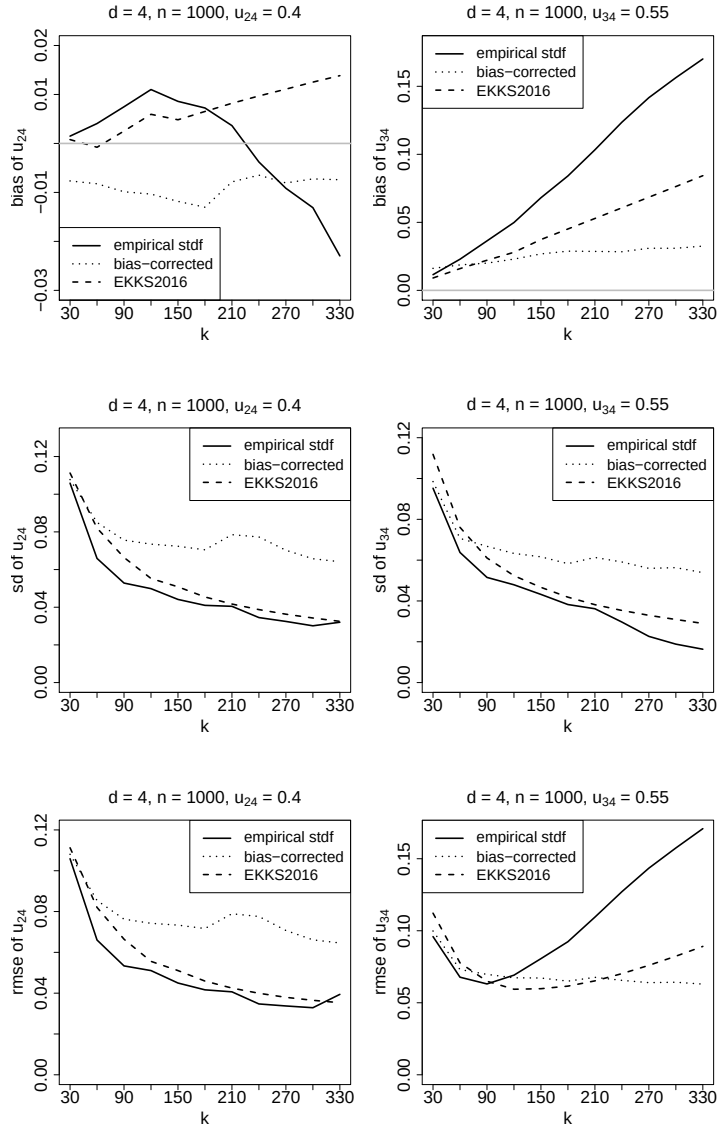


Figure 3.6: Bias, standard deviation and RMSE for estimators of $u_{24} = 0.4$ (left) and $u_{34} = 0.55$ (right) for the max-linear structural equation model based on a directed acyclic graph; 300 samples of size $n = 1000$.

rection is that some eigenvalues of $\Gamma(\boldsymbol{\theta})$ are (near) zero, which can in turn be due to the fact that for max-linear models such as this one, $\ell(\boldsymbol{c}_m; \boldsymbol{\theta})$ may hit its lower bound $\max(c_{m,1}, \ldots, c_{m,d})$ for some $m \in \{1, \ldots, q\}$.

3.3.4 Goodness-of-fit test

We compare the performance of the goodness-of-fit test presented in Corollary 3.2.5 to the three goodness-of-fit test statistics κ_n, ω_n^2, and A_n^2 proposed in Can et al. (2015, page 18). In the simulation study there, the observed rejection frequencies are reported at the 5% significance level under null and alternative hypotheses for two bivariate models for ℓ; a bivariate logistic model with $\theta \in (0, 1)$ and

$$\ell(x_1, x_2; \psi) = (1 - \psi)(x_1 + x_2) + \psi\sqrt{x_1^2 + x_2^2}, \qquad \psi \in (0, 1),$$

i.e., a mixture between a logistic model with parameter $1/2$ and tail independence. For both models, they generate 300 samples of size $n = 1500$ from a "null hypothesis" distribution function, for which the model is correct, and 100 samples of $n = 1500$ from an "alternative hypothesis" distribution function, for which the model is incorrect. These distribution functions are described in equations (32), (33), (35), and (36) of Can et al. (2015). We take $k = 200$ and $m = 1, \ldots, 4$ with $\boldsymbol{c}_m \in \{(1/2, 1/2), (1/2, 1), (1, 1/2), (1, 1)\}$.

Table 3.2 shows the observed fractions of Type I errors under the null hypotheses and the observed fraction of rejections under the alternative hypotheses. The results for κ_n, ω_n^2, and A_n^2 are taken from Can et al. (2015, Table 1). We see that our goodness-of-fit test performs comparably to the test statistics in Can et al. (2015).

	Null		Alternative	
	logistic	mixture	logistic	mixture
κ_n	19/300	9/300	92/100	97/100
ω_n^2	11/300	13/300	90/100	97/100
A_n^2	17/300	18/300	95/100	100/100
$kf_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k})$	16/300	14/300	100/100	82/100

Table 3.2: Observed rejection frequencies at the 5% significance level under null and alternative hypotheses.

It should be noted that the texts are of very different nature. The three test statistics in Can et al. (2015) are functionals of a transformed empirical process and are therefore of omnibus-type. The results in there are based on the full max-domain of attraction condition on F and the procedure is computationally complicated and therefore difficult to apply in dimensions (much) higher

than two. The present test only performs comparisons at q points and avoids integration. Therefore it is computationally much easier to apply in dimension $d > 2$.

3.4 Tail dependence in European stock markets

We analyze data from the EURO STOXX 50 Index, which represents the performance of the largest 50 companies among 19 different "supersectors" within the 12 main Eurozone countries. Since Germany (DE) and France (FR) together form 68% of the index, we will focus on these two countries only. Every company belongs to a supersector, of which there are 19 in total. We select two of them as an illustration: chemicals and insurance. We study the following five stocks: Bayer (DE, chemicals), BASF (DE, chemicals), Allianz (DE, insurance), Axa (FR, insurance), and Airliquide (FR, chemicals), and we take the weekly negative log-returns of the stock prices of these companies from `http://finance.yahoo.com/` for the period January 2002 to November 2015, leading to a sample of size $n = 711$.

We fit a structural equation model based on the directed acyclic graph given in Figure 3.7. Germany and France are represented by their national stock market indices, the DAX and the CAC40, respectively, while the supersectors chemicals and insurance are represented by the corresponding sub-indices of the EURO STOXX 50 Index. Note that this is a model for the tail dependence function only, i.e., we only assume that the joint distribution of the negative log-returns has tail dependence function ℓ as in (3.3.2) with coefficient matrix B given in Table 3.3. We have $d = 10$ and the parameter vector is given by $\boldsymbol{\theta} = (u_{12}, u_{13}, u_{14}, u_{15}, u_{26}, u_{46}, u_{27}, u_{47}, u_{38}, u_{48}, u_{39}, u_{59}, u_{2,10}, u_{5,10})$.

We perform the goodness-of-fit test described in Corollary 3.2.6, based on the $q = 1140$ points $\boldsymbol{c}_m$ in the grid $\{0, 1/2, 1\}^8$ having either two or three non-zero coordinates. We take $\Omega(\boldsymbol{\theta}) = I_q$, $k = 40$, and we choose s such that $\kappa_s > 0.1$, leading in this case to $s = 11$. The value of the test statistic is 5.28; the 95% quantile of a χ^2_{11} distribution is 19.68, so that the tail dependence model is not rejected.

The resulting parameter estimates are pictured at the edges of Figure 3.7, where the relative width of each edge is proportional to its parameter value. The standard errors are given in parentheses. We note that, except for Allianz, the influence of the stock market indices DAX and CAC40 is (much) stronger than the influence of the sector indices chemicals and insurance.

The univariate sample extremograms (Davis and Mikosch, 2009) suggest that there is some temporal dependence in most of our weekly stock return series. This dependence is limited to short-range dependence, i.e., two time points are independent provided they are far enough apart. In Bücher and Volgushev (2013), it is shown that the empirical copula is asymptotically normal under various weak dependence concepts; similar results could be deduced

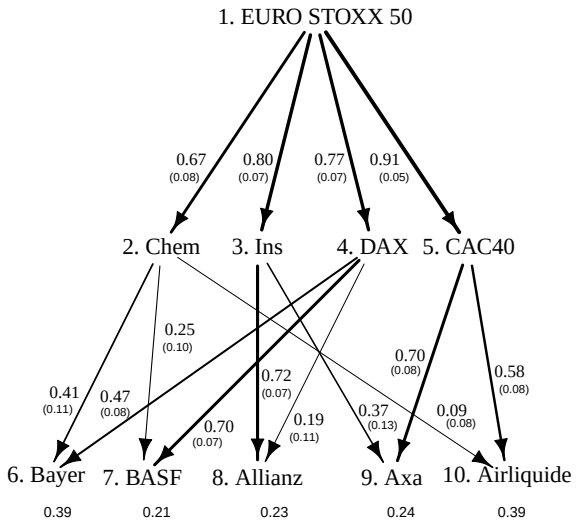


Figure 3.7: European stock market data: DAG with 14 parameters, whose estimates are shown near the corresponding edges. The relative width of each edge is proportional to its parameter value. The bottom row shows the estimated diagonal elements $u_6, \ldots, u_{10}$ of the matrix B in Table 3.3.

$$\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
u_{12} & u_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
u_{13} & 0 & u_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
u_{14} & 0 & 0 & u_4 & 0 & 0 & 0 & 0 & 0 & 0 \\
u_{15} & 0 & 0 & 0 & u_5 & 0 & 0 & 0 & 0 & 0 \\
u_{12}u_{26} \vee u_{14}u_{46} & u_2u_{26} & 0 & u_4u_{46} & 0 & u_6 & 0 & 0 & 0 & 0 \\
u_{12}u_{27} \vee u_{14}u_{47} & u_2u_{27} & 0 & u_4u_{47} & 0 & 0 & u_7 & 0 & 0 & 0 \\
u_{13}u_{38} \vee u_{14}u_{48} & 0 & u_3u_{38} & u_4u_{48} & 0 & 0 & 0 & u_8 & 0 & 0 \\
u_{13}u_{39} \vee u_{15}u_{59} & 0 & u_3u_{39} & 0 & u_5u_{59} & 0 & 0 & 0 & u_9 & 0 \\
u_{12}u_{2,10} \vee u_{15}u_{5,10} & u_2u_{2,10} & 0 & 0 & u_5u_{5,10} & 0 & 0 & 0 & 0 & u_{10}
\end{pmatrix}$$

Table 3.3: European stock market data: coefficient matrix B of the max-linear model stemming from the directed acyclic graph in Figure 3.7. The diagonal elements u_i, for $i = 2, \ldots, 10$ are such that the rows sum up to one.

for the empirical tail dependence function. Although the consistency of our estimator would still hold, the asymptotic variance matrix is no longer correct in the presence of serial dependence, so that standard errors and the optimal weight matrix are not entirely accurate.

Another concern is the (non)stationarity of our data, especially since our period of study contains the 2008/2009 financial crisis. To investigate whether the parameter estimates obtained in Figure 3.7 are stable over different time periods, we estimate our model parameters over a rolling window of five years starting from the beginning of every calendar year, i.e., for the time periods January 2002 - January 2007, January 2003 - January 2008, etc. The parameter estimates of the parameters u_{12}, u_{13}, u_{14}, u_{15} (characterizing the influence of the EURO STOXX index on the DAX, the CAC40 and on the sub-indices chemicals and insurance) changed only slightly, i.e., within the range of their standard errors as shown in Figure 3.7. The parameter estimates of the factors characterizing the influence of the DAX, CAC40 and the two sub-indices on the stocks show a bit more variation, especially for the parameters that are near zero such as u_{27}, u_{48}, and $u_{2,10}$; these are estimated anywhere between 0 and 0.5. In general, we can say that estimates of factors that have a large influence (i.e., estimates near one) fluctuate much less than estimates of factors that have a small influence (i.e., estimates near zero). However, these variations are still smaller than the ones stemming from the choice of k.

3.A Proofs

Proof of Theorem 3.2.1. This proof follows the same lines as the proof of Theorem 2.2.1 in Chapter 2; most differences are due to the continuous updating procedure. Let $\varepsilon_0 > 0$ be such that the closed ball $B_{\varepsilon_0}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \varepsilon_0\}$ is a subset of Θ; such an ε_0 exists since $\boldsymbol{\theta}_0$ is an interior point of Θ. Fix $\varepsilon > 0$ such that $0 < \varepsilon \leq \varepsilon_0$. Let, more precisely than in (3.2.8), $\widehat{\Theta}_{n,k}$ be the *set* of minimizers of the right-hand side of (3.2.8). We show first that

$$\mathbb{P}[\widehat{\Theta}_{n,k} \neq \varnothing \text{ and } \widehat{\Theta}_{n,k} \subset B_\varepsilon(\boldsymbol{\theta}_0)] \to 1, \qquad n \to \infty. \tag{3.A.1}$$

Because L is a homeomorphism, there exists $\delta > 0$ such that for $\boldsymbol{\theta} \in \Theta$, $\|L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_0)\| \leq \delta$ implies $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \varepsilon$. Equivalently, for every $\boldsymbol{\theta} \in \Theta$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \varepsilon$ we have $\|L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_0)\| > \delta$. Define the event

$$A_n = \left\{ \|L(\boldsymbol{\theta}_0) - \widehat{L}_{n.k}\| < \frac{\delta\sqrt{\lambda_1(\boldsymbol{\theta})}}{(1 + \sqrt{\lambda_1(\boldsymbol{\theta})})\max(1, \sqrt{\lambda_q(\boldsymbol{\theta}_0)})} \right\}.$$

If $\boldsymbol{\theta} \in \Theta$ is such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \varepsilon$, then on the event A_n, we have

$$\begin{aligned} \|D_{n,k}(\boldsymbol{\theta})\|_{\Omega(\boldsymbol{\theta})} &\geq \sqrt{\lambda_1(\boldsymbol{\theta})}\|D_{n,k}(\boldsymbol{\theta})\| \\ &\geq \sqrt{\lambda_1}\left\|L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta}) - \left(L(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{L}}_{n,k}\right)\right\| \end{aligned}$$

$$\geq \sqrt{\lambda_1}\left(\|L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta})\| - \|L(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{L}}_{n,k}\|\right)$$
$$> \sqrt{\lambda_1}\left(\delta - \frac{\delta\sqrt{\lambda_1}}{1+\sqrt{\lambda_1}}\right) = \frac{\delta\sqrt{\lambda_1}}{1+\sqrt{\lambda_1}}.$$

It follows that on A_n,

$$\inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\varepsilon} \|D_{n,k}(\boldsymbol{\theta})\|_{\Omega(\boldsymbol{\theta})} \geq \frac{\delta\sqrt{\lambda_1}}{1+\sqrt{\lambda_1}} > \sqrt{\lambda_q(\boldsymbol{\theta}_0)}\|L(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{L}}_{n,k}\|$$
$$\geq \|L(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{L}}_{n,k}\|_{\Omega(\boldsymbol{\theta}_0)}$$
$$\geq \inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\varepsilon} \|L(\boldsymbol{\theta}) - \widehat{\boldsymbol{L}}_{n,k}\|_{\Omega(\boldsymbol{\theta})}.$$

The infimum on the right-hand side is actually a minimum since L is continuous and $B_\varepsilon(\boldsymbol{\theta}_0)$ is compact. Hence on A_n the set $\widehat{\Theta}_{n,k}$ is non-empty and $\widehat{\Theta}_{n,k} \subset B_\varepsilon(\boldsymbol{\theta}_0)$. To show (3.A.1), it remains to prove that $\mathbb{P}[A_n] \to 1$ as $n \to \infty$, but this follows from (3.2.10).

Next we will prove that, with probability tending to one, $\widehat{\Theta}_{n,k}$ has exactly one element, i.e., the function $f_{n,k}$ has a unique minimizer. To do so, we will show that there exists $\varepsilon_1 \in (0, \varepsilon_0]$ such that, with probability tending to one, the Hessian of $f_{n,k}$ is positive definite on $B_{\varepsilon_1}(\boldsymbol{\theta}_0)$ and thus $f_{n,k}$ is strictly convex on $B_{\varepsilon_1}(\boldsymbol{\theta}_0)$. In combination with (3.A.1) for $\varepsilon \in (0, \varepsilon_1]$, this will yield the desired conclusion.

For $\boldsymbol{\theta} \in \Theta$, define the symmetric $p \times p$ matrix $\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$ by

$$\begin{aligned}
\big(\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)\big)_{i,j} := {} & 2\left(\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right)^T \Omega(\boldsymbol{\theta}) \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right) \\
& - 2\left(\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_i}\right)^T \Omega(\boldsymbol{\theta}) \big(L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta})\big) \\
& - 2\left(\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right)^T \frac{\partial \Omega(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} \big(L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta})\big) \\
& - 2\left(\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right)^T \frac{\partial \Omega(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} \big(L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta})\big) \\
& + \big(L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta})\big)^T \frac{\partial^2 \Omega(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_i} \big(L(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta})\big),
\end{aligned}$$

for $i, j \in \{1, \ldots, p\}$. The map $\boldsymbol{\theta} \mapsto \mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$ is continuous and

$$\mathcal{H}(\boldsymbol{\theta}_0) := \mathcal{H}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0) = 2\,\dot{L}(\boldsymbol{\theta}_0)^T\,\Omega(\boldsymbol{\theta}_0)\,\dot{L}(\boldsymbol{\theta}_0), \tag{3.A.2}$$

is a positive definite matrix. This $p \times p$ matrix is non-singular, since the $q \times q$ matrix $\Omega(\boldsymbol{\theta}_0)$ is non-singular and the $q \times p$ matrix $\dot{L}(\boldsymbol{\theta}_0)$ has rank p (recall $q \geq p$). Let $\|\cdot\|$ denote the spectral norm of a matrix. From Weyl's perturbation theorem (Jiang, 2010, page 145), there exists an $\eta > 0$ such that every

symmetric matrix $A \in \mathbb{R}^{p\times p}$ with $\|A - \mathcal{H}(\boldsymbol{\theta}_0)\| \leq \eta$ has positive eigenvalues and is therefore positive definite. Let $\varepsilon_1 \in (0, \varepsilon_0]$ be sufficiently small such that the second-order partial derivatives of L and Ω are continuous on $B_{\varepsilon_1}(\boldsymbol{\theta}_0)$ and such that $\|\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0) - \mathcal{H}(\boldsymbol{\theta}_0)\| \leq \eta/2$ for all $\boldsymbol{\theta} \in B_{\varepsilon_1}(\boldsymbol{\theta}_0)$.

Let $\mathcal{H}_{n,k,\Omega}(\boldsymbol{\theta}) \in \mathbb{R}^{p\times p}$ denote the Hessian matrix of $f_{n,k}$. Its (i,j)-th element is

$$\begin{aligned}
\big(\mathcal{H}_{n,k,\Omega}(\boldsymbol{\theta})\big)_{ij} &= \frac{\partial^2}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_i}\left[D_{n,k}(\boldsymbol{\theta})^T\,\Omega(\boldsymbol{\theta})\,D_{n,k}(\boldsymbol{\theta})\right] \\
&= \frac{\partial}{\partial\boldsymbol{\theta}_j}\left[-2D_{n,k}(\boldsymbol{\theta})^T\,\Omega(\boldsymbol{\theta})\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i} + D_{n,k}(\boldsymbol{\theta})^T\frac{\partial\Omega(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i}D_{n,k}(\boldsymbol{\theta})\right] \\
&= 2\left(\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j}\right)^T\Omega(\boldsymbol{\theta})\left(\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i}\right) - 2\left(\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_i}\right)^T\Omega(\boldsymbol{\theta})\,D_{n,k}(\boldsymbol{\theta}) \\
&\quad - 2\left(\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i}\right)^T\frac{\partial\Omega(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j}\,D_{n,k}(\boldsymbol{\theta}) \\
&\quad - 2\left(\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j}\right)^T\frac{\partial\Omega(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i}\,D_{n,k}(\boldsymbol{\theta}) \\
&\quad + D_{n,k}(\boldsymbol{\theta})^T\frac{\partial^2\Omega(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_i}\,D_{n,k}(\boldsymbol{\theta}).
\end{aligned}$$

Since $D_{n,k}(\boldsymbol{\theta}) = \widehat{\boldsymbol{L}}_{n,k} - L(\boldsymbol{\theta})$ and since $\widehat{\boldsymbol{L}}_{n,k}$ converges in probability to $L(\boldsymbol{\theta}_0)$, we obtain

$$\sup_{\boldsymbol{\theta}\in B_{\varepsilon_1}(\boldsymbol{\theta}_0)}\|\mathcal{H}_{n,k,\Omega}(\boldsymbol{\theta}) - \mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}_0)\| \xrightarrow{p} 0, \qquad n\to\infty. \tag{3.A.3}$$

By the triangle inequality, it follows that

$$\mathbb{P}\left[\sup_{\boldsymbol{\theta}\in B_{\varepsilon_1}(\boldsymbol{\theta}_0)}\|\mathcal{H}_{n,k,\Omega}(\boldsymbol{\theta}) - \mathcal{H}(\boldsymbol{\theta}_0)\| \leq \eta\right] \to 1, \qquad n\to\infty. \tag{3.A.4}$$

In view of our choice for η, this implies that, with probability tending to one, $\mathcal{H}_{n,k}(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta} \in B_{\varepsilon_1}(\boldsymbol{\theta}_0)$, as required. □

Proof of Theorem 3.2.2. Let $\nabla f_{n,k}(\boldsymbol{\theta})$, a $1\times q$ vector, be the gradient of $f_{n,k}$ at $\boldsymbol{\theta}$. By (3.2.11), we have

$$\begin{aligned}
\sqrt{k}\,\nabla f_{n,k}(\boldsymbol{\theta}_0) &= -2\sqrt{k}\,D_{n,k}(\boldsymbol{\theta}_0)^T\,\Omega(\boldsymbol{\theta}_0)\,\dot{L}(\boldsymbol{\theta}_0) \\
&\qquad + \sqrt{k}D_{n,k}(\boldsymbol{\theta}_0)^T\big(\nabla\Omega(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\big)D_{n,k}(\boldsymbol{\theta}_0) &\text{(3.A.5)} \\
&= -2\sqrt{k}\,D_{n,k}(\boldsymbol{\theta}_0)^T\,\Omega(\boldsymbol{\theta}_0)\,\dot{L}(\boldsymbol{\theta}_0) + o_P(1), &\text{(3.A.6)}
\end{aligned}$$

as $n\to\infty$. Since $\widehat{\boldsymbol{\theta}}_{n,k}$ is a minimizer of $f_{n,k}$, we have $\nabla f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) = 0$. An application of the mean value theorem to the function $t \mapsto \nabla f_{n,k}\big(\boldsymbol{\theta}_0 + t(\widehat{\boldsymbol{\theta}}_{n,k} -$

$\boldsymbol{\theta}_0)\big)$ at $t = 0$ and $t = 1$ yields

$$0 = \nabla f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k})^T = \nabla f_{n,k}(\boldsymbol{\theta}_0)^T + \mathcal{H}_{n,k,\Omega}(\widetilde{\boldsymbol{\theta}}_{n,k})\,(\widehat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_0), \tag{3.A.7}$$

where $\widetilde{\boldsymbol{\theta}}_{n,k}$ is a random vector on the segment connecting $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_{n,k}$ and $\mathcal{H}_{n,k,\Omega}$ is the Hessian matrix of $f_{n,k}$ as in the proof of Theorem 3.2.1. Since $\widehat{\boldsymbol{\theta}}_{n,k} \xrightarrow{p} \boldsymbol{\theta}_0$, we have $\widetilde{\boldsymbol{\theta}}_{n,k} \xrightarrow{p} \boldsymbol{\theta}_0$ as $n \to \infty$ too. By (3.A.3) and (3.A.2) and continuity of $\boldsymbol{\theta} \mapsto \mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$, it then follows that

$$\mathcal{H}_{n,k,\Omega}(\widetilde{\boldsymbol{\theta}}_{n,k}) \xrightarrow{p} \mathcal{H}(\boldsymbol{\theta}_0) = 2\dot{L}(\boldsymbol{\theta}_0)^T\,\Omega(\boldsymbol{\theta}_0)\,\dot{L}(\boldsymbol{\theta}_0), \qquad \text{as } n \to \infty. \tag{3.A.8}$$

Since $\mathcal{H}(\boldsymbol{\theta}_0)$ is non-singular, the matrix $\mathcal{H}_{n,k,\Omega}(\widetilde{\boldsymbol{\theta}}_{n,k})$ is non-singular with probability tending to one as well. Combine equations (3.A.6), (3.A.7) and (3.A.8) to see that

$$\begin{aligned}\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_0\big) &= -\mathcal{H}_{n,k,\Omega}(\widetilde{\boldsymbol{\theta}}_{n,k})^{-1}\,\sqrt{k}\,\nabla f_{n,k}(\boldsymbol{\theta}_0)^T + o_p(1)\\ &= \big(\dot{L}(\boldsymbol{\theta}_0)^T\Omega(\boldsymbol{\theta}_0)\dot{L}(\boldsymbol{\theta}_0)\big)^{-1}\,\dot{L}(\boldsymbol{\theta}_0)^T\Omega(\boldsymbol{\theta}_0)\,\sqrt{k}\,D_{n,k}(\boldsymbol{\theta}_0) + o_p(1),\end{aligned}$$

as $n \to \infty$. Convergence in distribution to the stated normal distribution follows from (3.2.11) and Slutsky's lemma. □

Proof of Corollary 3.2.4. Since $D_{n,k}(\boldsymbol{\theta}) = \widehat{\boldsymbol{L}}_{n,k} - L(\boldsymbol{\theta})$, we have

$$\sqrt{k}\,D_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) = \sqrt{k}\,D_{n,k}(\boldsymbol{\theta}_0) - \sqrt{k}\big(L(\widehat{\boldsymbol{\theta}}_{n,k}) - L(\boldsymbol{\theta}_0)\big).$$

By (3.2.12) and the delta method, we have

$$\begin{aligned}\sqrt{k}\big(L(\widehat{\boldsymbol{\theta}}_{n,k}) - L(\boldsymbol{\theta}_0)\big) &= \dot{L}\,\sqrt{k}(\widehat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_0) + o_p(1)\\ &= \dot{L}\,(\dot{L}^T\Omega\dot{L})^{-1}\dot{L}^T\Omega\,\sqrt{k}\,D_{n,k}(\boldsymbol{\theta}_0) + o_p(1)\\ &= P(\boldsymbol{\theta}_0)\,\sqrt{k}\,D_{n,k}(\boldsymbol{\theta}_0) + o_p(1), \qquad \text{as } n \to \infty,\end{aligned}$$

where $\dot{L}$ and Ω are evaluated at $\boldsymbol{\theta}_0$. Combination of the two previous displays yields

$$\sqrt{k}\,D_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) = (I_q - P(\boldsymbol{\theta}_0))\,\sqrt{k}\,D_{n,k}(\boldsymbol{\theta}_0) + o_p(1), \qquad \text{as } n \to \infty.$$

By (3.2.11) and Slutsky's lemma, we arrive at (3.2.16), as required.

The $q \times q$ matrix P has rank p since the $q \times p$ matrix $\dot{L}$ has rank p and the $q \times q$ matrix Ω is non-singular. Since $P^2 = P$, it follows that $\operatorname{rank}(I_q - P) = \operatorname{rank}(I_q) - \operatorname{rank}(P) = q - p$. □

Proof of Corollary 3.2.5. Equation (3.2.11) can be written as

$$\boldsymbol{Z}_{n,k} := \sqrt{k}\,D_{n,k}(\boldsymbol{\theta}_0) \xrightarrow{d} \boldsymbol{Z} \sim \mathcal{N}_q\big(\boldsymbol{0}, \Gamma(\boldsymbol{\theta}_0)\big), \qquad \text{as } n \to \infty.$$

In view of (3.2.16) and $\Omega(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}$, we find, by Slutsky's lemma and the continuous mapping theorem,

$$\begin{aligned} k\, f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) &= k\, D_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k})^T\, \Gamma(\widehat{\boldsymbol{\theta}}_{n,k})^{-1}\, D_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) \\ &= \boldsymbol{Z}_{n,k}^T\, (I_q - P(\boldsymbol{\theta}_0))^T\, \Gamma(\widehat{\boldsymbol{\theta}}_{n,k})^{-1}\, (I_q - P(\boldsymbol{\theta}_0))\, \boldsymbol{Z}_{n,k} + o_p(1) \\ &\xrightarrow{d} \boldsymbol{Z}^T\, (I_q - P(\boldsymbol{\theta}_0))^T\, \Gamma(\boldsymbol{\theta}_0)^{-1}\, (I_q - P(\boldsymbol{\theta}_0))\, \boldsymbol{Z}, \qquad \text{as } n \to \infty; \end{aligned}$$

here $P = \dot{L}\,(\dot{L}^T\Gamma^{-1}\dot{L})^{-1}\,\dot{L}^T\Gamma^{-1}$, with $\dot{L}$ and Γ evaluated at $\boldsymbol{\theta}_0$.

It remains to identify the distribution of the limit random variable. The random vector $\boldsymbol{Z}$ is equal in distribution to $\Gamma^{1/2}\boldsymbol{Y}$, where $\boldsymbol{Y} \sim \mathcal{N}_q(\boldsymbol{0}, I_q)$ and where $\Gamma^{1/2}$ is a symmetric square root of Γ. Straightforward calculation yields

$$\boldsymbol{Z}^T (I_q - P)^T\, \Gamma^{-1}\, (I_q - P)\, \boldsymbol{Z} \stackrel{\mathrm{d}}{=} \boldsymbol{Y}^T (I_q - B) \boldsymbol{Y}$$

where $B = \Gamma^{-1/2}\dot{L}\,(\dot{L}^T\Gamma^{-1}\dot{L})^{-1}\,\dot{L}^T\Gamma^{-1/2}$. It is easily checked that B is a projection matrix ($B = B^T = B^2$). Moreover, B has rank p. It follows that $I_q - B$ is a projection matrix too and that it has rank $q - p$. The distribution of the limit random variable now follows by standard properties of quadratic forms of normal random vectors. □

Proof of Corollary 3.2.6. Let $\boldsymbol{Z} \sim \mathcal{N}_q(\boldsymbol{0}, \Gamma(\boldsymbol{\theta}_0))$, which by (3.2.11) is the limit in distribution of $\sqrt{k}\, D_{n,k}(\boldsymbol{\theta}_0)$. By (3.2.16) and the continuous mapping theorem, we have, as $n \to \infty$,

$$\begin{aligned} k\, D_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k})^T\, A(\widehat{\boldsymbol{\theta}}_{n,k})\, D_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) \\ \xrightarrow{d} Z^T\, (I_q - P(\boldsymbol{\theta}_0))^T\, A(\boldsymbol{\theta}_0)\, (I_q - P(\boldsymbol{\theta}_0))\, Z. \quad \text{(3.A.9)} \end{aligned}$$

We can represent $(I_q - P)\boldsymbol{Z}$ as $V D^{1/2} Y$, with $\boldsymbol{Y} \sim \mathcal{N}_q(\boldsymbol{0}, I_q)$. The limiting random variable in (3.A.9) is then given by

$$\boldsymbol{Y}^T D^{1/2} V^T\, V_s D_s^{-1} V_s^T\, V D^{1/2} \boldsymbol{Y}.$$

Since V is an orthogonal matrix, this expression simplifies to $\sum_{j=1}^{s} Y_j^2$, which has the stated χ_s^2 distribution. □

Proof of Remark 3.2.1. Inspection of the proofs of Corollaries 3.2.5 and 3.2.6 shows that the difference between the two test statistics converges in distribution to the random variable $\boldsymbol{Z}^T\, R(\boldsymbol{\theta}_0)\, \boldsymbol{Z}$, where $\boldsymbol{Z}$ is a certain q-variate normal random vector and where

$$R(\boldsymbol{\theta}_0) = \big(I_q - P(\boldsymbol{\theta}_0)\big)^T\, \big(\Gamma(\boldsymbol{\theta}_0)^{-1} - A(\boldsymbol{\theta}_0)\big)\, \big(I_q - P(\boldsymbol{\theta}_0)\big).$$

The matrix $R(\boldsymbol{\theta}_0)$ can be shown to be equal to zero, proving the claim of the remark. To see why $R(\boldsymbol{\theta}_0)$ is zero, note first that, suppressing $\boldsymbol{\theta}_0$ and writing

$Q = I_q - P$, we have $Q^2 = Q$ and $\Gamma Q^T = Q\Gamma = Q\Gamma Q^T$. Recall the eigenvalue equation $Q\Gamma Q^T v_j = \kappa_j v_j$ for $j = 1, \ldots, q$. Note that $\kappa_j > 0$ if $j \leq s$ and $\kappa_j = 0$ if $j \geq s + 1$. The eigenvalue equation implies that $Qv_j = v_j$ for $j \leq s$ while $Q\Gamma v_j = 0$ for $j \geq s+1$. Since the vectors $v_1, \ldots, v_q$ are orthogonal, we find that the vectors $v_1, \ldots, v_s, \Gamma v_{s+1}, \ldots, \Gamma v_q$ are linearly independent. It then suffices to show that $Rv_j = 0$ for all $j \leq s$ and $R\Gamma v_j = 0$ for all $j \geq s + 1$. The first property follows from the fact that $\Gamma^{-1} v_j = \kappa_j^{-1} Q^T v_j$ and $Av_j = \kappa_j^{-1} v_j$ for $j \leq s$ (use the eigenvalue equation again), while the second property follows from $Q\Gamma v_j = 0$ for $j \geq s + 1$. □

Chapter 4

The R package tailDepFun

4.1 Introduction

We illustrate the use of the tailDepFun package, which provides functions for the estimation of tail dependence parameters for a variety of models. The estimators that are implemented are

- The pairwise M-estimator described in Chapter 2, which will be referred to as Mestimator in the R code.
- The continuous updating weighted least squares estimator described in Chapter 3, which will be called WLS in the R code.

The three main functions of this package are the following:

- EstimationGumbel: estimation of the parameter of a Gumbel model, also called a logistic model.
- EstimationMaxLinear: estimation of the parameters of a max-linear model, possibly defined on a directed acyclic graph (DAG).
- EstimationBR: estimation of the parameters of an (anisotropic) Brown–Resnick process.

All the above models are defined by means of their stable tail dependence function ℓ; see (1.3.4). Besides the main functions that are used for estimation, functions that compute the (bias-corrected) empirical stable tail dependence function are available as well. Finaly, the function SelectGrid aids to define a regular grid of indices in which the stable tail dependence function should be evaluated.

4.2 Choosing a grid

In the definition of the weighted least squares estimator in (3.2.8), we evaluate ℓ in the points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0, \infty)^d$, with $\boldsymbol{c}_m = (c_{m1}, \ldots, c_{md})$ for $m = 1, \ldots, q$ for $q \geq p$, where p denotes the dimension of the parameter vector $\boldsymbol{\theta}$ of ℓ. The function selectGrid aids in selecting these points. For instance, the grids used in the simulation study on the logistic model in Subsection 3.3.1 for dimension $d = 5$ and the weighted least squares estimator can be generated as follows

```
> selectGrid(cst = c(0,1), d = 5)
> selectGrid(cst = c(0,1), d = 5, nonzero = 3)
```

for pairs and triples respectively. For the simulation study on the Brown–Resnick process in Subsection 3.3.2, we first need to define the locations of the stations on a 3×4 unit distance grid

```
> loc <- cbind(rep(1:3, 4), rep(1:4, each = 3))
> selectGrid(cst = c(0,1), d = 12, locations = loc,
+            maxDistance = sqrt(2))
```

Finally, the grid used in the simulation study on a max-linear model in Subsection 3.3.3 for the weighted least squares estimator can be generated as follows

```
> selectGrid(cst = c(0,0.5,1), d = 4, nonzero = c(2:4))
```

Note that we always set `cst = c(0,1)` when we want to use an estimator based on extremal coefficients only. This function can also be used to select the pairs for the pairwise M-estimator. Then one has to set `cst = c(0,1)` and `nonzero = 2`, obtaining q rows, where the two ones in each row correspond to the pair that is selected.

4.3 Logistic model

The stable tail dependence function for the logistic model is defined in (1.3.1). We can generate observations from it using the copula package (Hofert et al., 2015), here in three dimensions with parameter value $\theta = 0.5$, and transform them to unit Pareto margins using ranks.

```
> library(copula)
> set.seed(1)
> n <- 1000
> data <- rCopula(n = n,
+          copula = gumbelCopula(param = 1/0.5, dim = 3))
> x <- apply(data, 2, function(i) n/(n + 0.5 - rank(i)))
```

Then, we can estimate θ using either the pairwise M-estimator

```
> indices <- selectGrid(c(0,1), d = 3)
> EstimationGumbel(x, indices, k = 50,
+                  method = "Mestimator")$theta
[1] 0.4123743
```

or using the weighted least squares estimator, based on the points $\boldsymbol{c}_m$ on the grid $\{0, 0.5, 1\}^3$ having at least two positive coordinates

```
> indices <- selectGrid(c(0,0.5,1), d = 3,
+                       nonzero = c(2,3))
> EstimationGumbel(x, indices, k = 50,
+                  method = "WLS")$theta
[1] 0.4177336
```

In Subsection 3.3.1, we simulate from a process in the max-domain of attraction of the Gumbel model, also known as the outer power Clayton copula (Hofert et al., 2015). The outer power Clayton copula is the Archimedean copula with generator $\psi(t) = \psi_\beta(t^\theta)$, where $\theta \in (0, 1]$ and

$$\psi_\beta(t) = \frac{1}{(1+\beta t)^{1/\beta}}, \qquad \beta > 0.$$

We can fix $\beta = 1$ and hence focus on the generator $\psi(t) = (1+t^\theta)^{-1}$. A sample from this copula for $\theta = 0.7$ and $d = 3$ can be obtained using the copula package as follows.

```
> opc <- opower(copClayton, thetabase = 1)
> copClayton <- onacopulaL(opc, list(theta = 1/0.7,
+                                    comp = c(1:3)))
> data <- rnacopula(n = n, copula = copClayton)
```

4.4 Brown–Resnick process

A full definition of the process and its stable tail dependence function can be found in Example 1.3.2 and in Subsections 2.3.1 and 3.3.2. Let $\boldsymbol{\theta} = (\alpha, \rho, \beta, c)$ denote the parameters of the anisotropic Brown–Resnick process, where $\alpha \in (0, 2]$, $\rho > 0$, $\beta \in [0, \pi/2)$ and $c > 0$.

We first illustrate the estimation of an isotropic Brown–Resnick process using the pairwise M-estimator. We define coordinates of four locations and we select all pairs of locations.

```
> locations <- rbind(c(1,1),c(2,1),c(1,2),c(2,2))
> indices <- selectGrid(cst = c(0,1), d = 4,
+                       locations = locations)
```

Then we generate data from the Brown–Resnick process using the SpatialExtremes package (Ribatet, 2015).

```
> set.seed(2)
> library(SpatialExtremes)
> x <- rmaxstab(n = 5000, coord = locations,
+               cov.mod = "brown",
+               range = 3, smooth = 1)
```

We calculate the estimator for $k = 300$. This could take a couple of minutes.

```
> EstimationBR(x, locations, indices, k = 300,
+              method = "Mestimator", iterate = TRUE,
+              isotropic = TRUE, Tol = 1e-04)
$theta
[1] 1.235120 2.689146

$theta_pilot
[1] 1.235124 2.689140

$covMatrix
             [,1]        [,2]
[1,]  0.009990198 -0.02362328
[2,] -0.023623279  0.11335579

$value
[1] 0.01474498
```

Standard errors are the square roots of the diagonal elements of `covMatrix`. Setting `iterate = TRUE` means that we use the two-step optimal weighting procedure as described in Corollary 2.2.3: `theta` returns the parameter estimates obtained using the optimal weight matrix, which is defined as the inverse of the matrix Σ evaluated in `theta_pilot`.

Next we estimate the parameters using the weighted least squares estimator. We use the bias-corrected stable tail dependence function (Beirlant et al., 2016). For `method = "WLS"` the option `iterate = TRUE` means that we use the continuous updating procedure as described in Corollary 3.2.3.

```
> EstimationBR(x, locations, indices, 300, method = "WLS",
+              isotropic = TRUE, biascorr = TRUE,
+              iterate = TRUE)
$theta
[1] 1.113500 3.087141

$theta_pilot
[1] 1.115788 3.092344

$covMatrix
            [,1]        [,2]
```

```
[1,]  0.02644506 -0.07761775
[2,] -0.07761775  0.31759008

$value
[1] 0.006486834
```

Note that since we set `iterate = TRUE`, we can use Corollary 3.2.5 to test the goodness-of-fit of the model. The test statistic is $300 \times 0.006486834 = 1.94605$, which we should compare to a high quantile of the χ^2_4 distribution.

If we want to mimic the two-step weighting procedure from Einmahl et al. (2016a) instead of continuous updated weighting, we could do so as follows

```
> result <- EstimationBR(x, locations, indices, k = 300,
+                        method = "WLS", isotropic = TRUE,
+                        biascorr = TRUE, covMat = FALSE)
> Sigma <- AsymVarBR(locations, indices, method = "WLS",
+                    par = result$theta_pilot)
> EstimationBR(x, locations, indices, 300, method= "WLS",
+              isotropic = TRUE, biascorr = TRUE,
+              Omega = solve(Sigma))$theta
[1] 1.115812 3.092290
```

If we want to estimate an isotropic Brown–Resnick process, we need to transform the coordinates of our locations, since we can only simulate isotropic Brown–Resnick processes. Hence, we multiply the coordinates of our locations with $V^{-1}(\beta, c)$; see Example 1.3.3. Here we take $\beta = 0.25$ and $c = 1$.

```
> Vmat <- matrix(c(cos(0.25), 1.5*sin(0.25),
+                  -sin(0.25), 1.5*cos(0.25)), nrow = 2)
> locationsAniso <- locations %*% t(solve(Vmat))
> EstimationBR(x, locationsAniso, indices, 300,
+              method = "WLS", biascorr = TRUE,
+              iterate = TRUE)$theta
[1] 1.1275481 3.1058154 0.4099169 1.5394794
```

Finally, some tips for the use of this function:

- If the number of locations d is small ($d < 8$ say), a sufficiently large sample size (eg $n > 2000$) is needed to obtain a satisfying result. However, if d is large, a sample size of $n = 500$ should suffice.
- The tolerance parameter is only used when calculating the three- and four-dimensional integrals in the asymptotic covariance matrix Σ; see Appendix 2.B in Chapter 2. A tolerance of 10^{-4} often suffices, although the default tolerance is a safer choice.
- For an anisotropic process, it is advised to try a couple of starting values if d is small, preferably a starting value with $c < 1$ and one with $c > 1$.

- Setting `iterate = TRUE` has a more significant effect when d is large.
- If the number of pairs q is large, then `method = "Mestimator"` will be rather slow. This is due to the calculation of the weight matrix Ω and the covariance matrix. Setting `iterate = FALSE` and `covMat = FALSE` will make estimation fast even for several hundreds of pairs of locations.
- `method = "WLS"`, it is not advised to change the values of `k1` or `tau`; the default values are chosen as advised in Beirlant et al. (2016).

Note that an extension two triples and more general grids (i.e., where `cst` is not necessarily `c(0,1)`) might be available in the future.

4.5 Max-linear model

The max-linear model is described in detail in Example 1.3.3 and Subsection 3.3.3. Its parameter matrix is a $d \times r$ matrix $B := (b_{jt})_{j,t}$, where r denotes the number of factors and d the dimension. The factor loadings b_{jt} are non-negative constants such that $\sum_{t=1}^{r} b_{jt} = 1$ for every $j \in \{1, \ldots, d\}$ and all column sums of B are positive. Note that B has $p = d \times (r-1)$ free elements. The parameter vector $\theta \in \mathbb{R}^p$ is defined by stacking the columns of B in decreasing order of their sums, leaving out the column with the lowest sum.

To illustrate estimation of a 2-factor model in dimension $d = 3$, we simulate data with parameter vector $\theta = c(b_{11}, b_{12}, b_{13}) = c(0.3, 0.5, 0.9)$.

```
> set.seed(1)
> n <- 1000
> fr <- matrix(-1/log(runif(2*n)), nrow = n, ncol = 2)
> data <- cbind(pmax(0.3*fr[,1], 0.7*fr[,2]),
+               pmax(0.5*fr[,1], 0.5*fr[,2]),
+               pmax(0.9*fr[,1], 0.1*fr[,2]))
```

Then we transform to unit Pareto, select a grid and estimate the parameters using the weighted least squares estimator. Note that the choice `cst = c(0,1)` will usually not lead to a valid estimator; see Subsection 3.3.3.

```
> x <- apply(data, 2, function(i) n/(n + 0.5 - rank(i)))
> indices <- selectGrid(cst = c(0,0.5,1), d = 3)
> EstimationMaxLinear(x, indices, k = 100, method = "WLS",
+                     iterate = TRUE, GoFtest = TRUE,
+                     startingValue = c(0.3,0.5,0.9))
$theta
[1] 0.3284527 0.5007583 0.9115711

$theta_pilot
[1] 0.3152544 0.4893125 0.8980283
```

```
$covMatrix
              [,1]          [,2]          [,3]
[1,] 0.0022120835 0.0021556005 0.0006626992
[2,] 0.0021556005 0.0026694102 0.0009185456
[3,] 0.0006626992 0.0009185456 0.0011039448

$value
[1] 0.2069704

$GoFresult
$GoFresult$value
[1] 1.961863

$GoFresult$s
[1] 2
```

The results of `GoFtest` permit us to do the test described in Corollary 3.2.6. We need to compare `GoFresult$value` to the 95% quantile of a χ^2_s distribution with $s = 2$ degrees of freedom, given by 5.99. For the two-step weighting procedure from Chapter 2 we would do

```
> result <- EstimationMaxLinear(x, indices, k = 100,
+                      method = "WLS",
+                      startingValue = c(0.3,0.5,0.9))
> Sigma <- AsymVarML(indices, par = result$theta_pilot)
> while(rcond(Sigma) < 1e-05){
+    Sigma <- Sigma + (1e-05)*diag(nrow(indices))
+ }
> EstimationMaxLinear(x, indices, 100, method = "WLS",
+           Omega = solve(Sigma),
+           startingValue = result$theta_pilot)$theta
[1] 0.3277286 0.4988443 0.9105570
```

The correction on `Sigma` is done because it is not invertible otherwise; see Remark 3.3.1.

In the above estimation, the function `EstimationMaxLinear` assumed a 2-factor model because we did not provide `Bmatrix` and `Ldot`. If we want to fit a max-linear model based on a directed acyclic graph, for instance the one in Gissibl and Klüppelberg (2015, Example 2.1) or Subsection 3.3.3, we need to define a `Bmatrix`, corresponding to the matrix of coefficients B, and a `Ldot`, corresponding to the total derivative of $L(\theta) = (\ell(c_m;\theta))_{m=1,\dots,q}$. For instance, B is given by

```
> Bmatrix <- function(th){
```

```
+  temp <- max(th[1]*th[3],th[2]*th[4])
+  B <- cbind(c(1, th[1], th[2], temp),
+             c(0, 1 - th[1], 0, (1 - th[1])*th[3]),
+             c(0, 0, 1 - th[2], (1 - th[2])*th[4]),
+             c(0, 0, 0, 1 - temp - (1 - th[1])*th[3] -
+                                   (1 - th[2])*th[4]))
+  return(B)
+ }
```

and then we generate data from the DAG as follows

```
> d <- r <- 4
> n <- 1000
> theta <- c(0.3, 0.8, 0.4, 0.55)
> B <- Bmatrix(theta)
> set.seed(1)
> fr <- matrix(-1/log(runif(r*n)), nrow = n, ncol = r)
> data <- cbind(B[1,1]*fr[,1],
+               pmax(B[2,1]*fr[,1], B[2,2]*fr[,2]),
+               pmax(B[3,1]*fr[,1], B[3,3]*fr[,3]),
+               pmax(B[4,1]*fr[,1], B[4,2]*fr[,2],
+                    B[4,3]*fr[,3], B[4,4]*fr[,4]))
> x <- apply(data, 2, function(i) n/(n + 0.5 - rank(i)))
```

We then estimate using a grid as in Subsection 3.3.3,

```
> indices <- selectGrid(cst = c(0,0.5,1), d = 4,
+                       nonzero = c(2:4))
> EstimationMaxLinear(x, indices, k = 100, method = "WLS",
+           Bmatrix = Bmatrix, covMat = FALSE,
+           startingValue = c(0.3,0.8,0.4,0.55))$theta
[1] 0.3502117 0.8692499 0.4124804 0.5909954
```

Note that in order to calculate `covMat`, we would also need to provide `Ldot`.

Chapter 5

Peaks-over-thresholds modelling with multivariate generalized Pareto distributions

Abstract

Statistical modelling using multivariate generalized Pareto distributions constitutes the multivariate analogue of peaks-over-thresholds modelling with the univariate generalized Pareto distribution. We recall three different representations of a multivariate generalized Pareto distribution described in Rootzén, Segers and Wadsworth (2016) and we propose a construction tool which allows to generate suitable parametric tail dependence models. Several concrete examples are proposed, and the densities necessary for censored likelihood estimation are derived. Finally, we present a new parametric model for data with structured components, and illustrate it with an application aimed at estimating the probability of a landslide in northern Sweden.

5.1 Introduction

Peaks-over-thresholds modelling of univariate time series has been common practice since it was proposed in Davison and Smith (1990), who advocated the use of the asymptotically motivated generalized Pareto distribution as a model for exceedances over high thresholds (see Subsection 1.2.2). However, many peaks-over-thresholds data are of a multivariate nature: imagine for instance a flooding, where the amount of damage depends on the number of dykes that have been breached. Another example is the modelling of landslides, where a multivariate dataset can be constructed from a univariate time series of precipitation measurements: here, a d-variate dataset is created by taking cumulative sums of up to d days of precipitation amounts. This type of construction is of

interest because a landslide might occur after either an extreme rainfall on one day or after moderate rainfall amounts on several consecutive days.

When generalizing peaks-over-thresholds modelling from the univariate to the multivariate setting, different definitions of what constitutes an exceedance might arise: for instance, one might either be interested in events where all components are large, or in events where at least one component is large; this corresponds to the distinction between the tail copula and the stable tail dependence function, presented in Subsection 1.3.1. Here we consider the latter, as we did in the previous chapters.

The *multivariate generalized Pareto* (MGP) distribution was introduced originally in Tajvidi (1996), Beirlant et al. (2004, Chapter 8) and Rootzén and Tajvidi (2006). There has been a growing body of probabilistic literature devoted to MGP distributions ever since; see for instance Falk and Guillou (2008), Falk and Michel (2009), Ferreira and de Haan (2014), or Rootzén et al. (2016). However, to our knowledge, statistical modelling using MGP distributions has thusfar received relatively little attention. Some examples include Michel (2009), where two likelihood-based estimation approaches based on the spectral density are presented, Huser et al. (2015), where the MGP likelihood is studied for a logistic model, and Thibaud et al. (2015), where the focus is on so-called elliptic Pareto processes.

One of the reasons for the lack of literature on statistical modelling of MGP distributions might be the existence of theoretically equivalent dependence modelling approaches, based on a point process, that have already been introduced in Coles and Tawn (1991). Nonetheless, the MGP distribution has some conceptual advantages over that of the point process representation, since it represents a proper multivariate distribution on an L-shaped region (see the left-hand side of Figure 1.5). Furthermore, the MGP distribution permits modelling of all data on this region, up to truncation from below, without the need to perform any marginal transformation, which is common in other extremal dependence modelling approaches.

In Chapters 2 and 3, we assumed a parametric model for the stable tail dependence function ℓ and we focused on semiparametric estimation for the parameters of ℓ using threshold exceedances. In this chapter, we will estimate the parameters of models for an MGP distribution using a likelihood approach, allowing us to model the marginal distributions jointly with the dependence structure if the dimension of the model is not too high. Note that, since ℓ is related to an MGP distribution through a simple formula, see (5.3.2) below, we could plug in nonparametric estimators of ℓ to obtain nonparametric estimators of an MGP distribution with standardized marginals.

After having fixed a high threshold, we select the episodes where at least one component exceeds the threshold, and we model the difference between the values of the components and the threshold. Components falling below the threshold are then censored at the threshold. This censored likelihood approach decreases bias (Huser et al., 2015) and avoids problems arising by

the fact that the lower bound of a MGP distribution depends on its parameter values. It was first introduced in Ledford and Tawn (1996) and Smith et al. (1997), then extended to the spatial framework in Wadsworth and Tawn (2014) and Thibaud et al. (2015). In high dimensions, sometimes a pairwise censored likelihood procedure is used (Thibaud et al., 2013; Huser and Davison, 2014).

In Rootzén et al. (2016), three representations for an MGP distribution are proposed, and linked to three corresponding point process models. From these representations a convenient tool for the construction of parametric models is deduced. This construction tool permits easy simulation from such models, thus enabling estimation of any quantity related to the extremes of a random vector. We present several models for potentially high-dimensional data, some of which are identifiable as well-known tail dependence models, whereas others are entirely new. A feature shared by all models is that their densities are analytically computable, allowing for fast estimation.

The remainder of this chapter is structured as follows. In Section 5.2 we recall some of the key results and properties of MGP distributions that will be useful for statistical modelling. Section 5.3 discusses a construction device for MGP distributions and the censored likelihood estimation procedure, whilst concrete parametric models and their simulation are presented in Section 5.4. Section 5.5 explains the links between the point process representations and the MGP distribution, providing an intuitive outline of the derivation; formal proofs can be found in Rootzén et al. (2016). Finally, an application to precipitation data, aimed at modelling landslides, is described in Section 5.6. We derive the formulas necessary for censored likelihood inference in Appendix 5.A.

5.2 Multivariate generalized Pareto distributions

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a random vector in $\mathbb{R}^d$ with cumulative distribution function F. Suppose that F is in the max-domain of attraction of a GEV distribution G (see Subsection 1.3.3) with $0 < G(\boldsymbol{0}) < 1$ and margins $G_1, \ldots, G_d$. Let $\boldsymbol{l} = (l_1, \ldots, l_d)$ denote the vector of marginal lower endpoints, i.e., l_j is the lower endpoint of G_j for $j \in \{1, \ldots, d\}$.

The multivariate generalized Pareto distribution arises as the only possible non-degenerate limit of $\boldsymbol{X}$, suitably normalized, conditioned upon at least one component of $\boldsymbol{X}$ being extreme. Specifically, if there exist scaling and threshold functions $\boldsymbol{a}_n \in (0, \infty)^d$ and $\boldsymbol{b}_n \in \mathbb{R}^d$ with $F_j(b_{n,j}) < 1$ for all $j \in \{1, \ldots, d\}$ and $F(\boldsymbol{b}_n) \to 1$ as $n \to \infty$, such that

$$\mathbb{P}\left[\max\left(\frac{\boldsymbol{X} - \boldsymbol{b}_n}{\boldsymbol{a}_n}, \boldsymbol{l}\right) \leq \cdot \,\middle|\, \boldsymbol{X} \not\leq \boldsymbol{b}_n\right] \xrightarrow{d} H(\,\cdot\,), \qquad \text{as } n \to \infty, \tag{5.2.1}$$

where H has non-degenerate margins $H_1, \ldots, H_d$ and $H_j(0) < 1$ for all $j \in \{1, \ldots, d\}$, then we say that H is a multivariate generalized Pareto distribution

(Rootzén and Tajvidi, 2006). We require $H_j(0) < 1$ so that all components have a positive probability of exceeding their threshold. We say that F belongs to the *threshold-domain of attraction* of H. Note that F is in the max-domain of attraction of a GEV distribution G with $0 < G(\mathbf{0}) < 1$ if and only if F is in the threshold-domain of attraction of an MGP distribution H. The link between H and G is given below in (5.2.3). The definition in (5.2.1) corresponds to the ones in Rootzén and Tajvidi (2006) and Beirlant et al. (2004) but is slightly different from the ones in Falk and Guillou (2008) or Ferreira and de Haan (2014).

Recall that the marginal distributions of G, denoted by $G_1, \ldots, G_d$, may be written as

$$G_j(x) = \exp\left\{-\left(1 + \xi_j \frac{x - \mu_j}{\sigma_j}\right)_+^{-1/\xi_j}\right\}, \qquad x \in \mathbb{R}. \tag{5.2.2}$$

Define $\boldsymbol{\eta} := \boldsymbol{\sigma} - \boldsymbol{\xi}\boldsymbol{\mu}$. We will assume throughout that $\boldsymbol{\eta} > \mathbf{0}$, which is equivalent to $H_j(0) < 1$ for $j = 1, \ldots, d$. Let $w_j \in (-\infty, \infty]$ denote the upper endpoints of G_j for $j = 1, \ldots, d$. Then the support of G_j, which we saw in Subsection 1.2, can be written in terms of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ as

$$(l_j, w_j) = \begin{cases} (-\eta_j/\xi_j, \infty) & \text{if } \xi_j > 0, \\ (-\infty, \infty) & \text{if } \xi_j = 0, \\ (-\infty, -\eta_j/\xi_j) & \text{if } \xi_j < 0. \end{cases}$$

An MGP distribution is then supported on

$$\{\boldsymbol{y} \in [\boldsymbol{l}, \boldsymbol{w}] : \boldsymbol{y} \not\leq \mathbf{0}\} = [\boldsymbol{l}, \boldsymbol{w}] \setminus [\boldsymbol{l}, \mathbf{0}].$$

Recall the point process defined in (1.3.12), whose intensity measure converges to the exponent measure μ, so that the limit of the expected number of points in a set $\mathcal{A}$ is equal to $\mu(\mathcal{A})$. Then, assuming for $\boldsymbol{x} > \boldsymbol{l}$,

$$\begin{aligned} H(\boldsymbol{x}) &= \lim_{n\to\infty} \frac{\mathbb{P}\left[\frac{\boldsymbol{X} - \boldsymbol{b}_n}{\boldsymbol{a}_n} \leq \boldsymbol{x}\right] - \mathbb{P}\left[\frac{\boldsymbol{X} - \boldsymbol{b}_n}{\boldsymbol{a}_n} \leq \min(\mathbf{0}, \boldsymbol{x})\right]}{\mathbb{P}\left[\frac{\boldsymbol{X} - \boldsymbol{b}_n}{\boldsymbol{a}_n} \not\leq \mathbf{0}\right]} \\ &= \frac{-\mu(\{\boldsymbol{y} : \boldsymbol{y} \not\leq \boldsymbol{x}\}) + \mu(\{\boldsymbol{y} : \boldsymbol{y} \not\leq \min(\boldsymbol{x}, \mathbf{0})\})}{\mu(\{\boldsymbol{y} : \boldsymbol{y} \not\leq \mathbf{0}\})}. \end{aligned}$$

Since the exponent measure satisfies $\mu(\{\boldsymbol{y} : \boldsymbol{y} \not\leq \boldsymbol{x}\}) = -\log G(\boldsymbol{x})$ we have

$$H(\boldsymbol{x}) = \begin{cases} \dfrac{1}{-\log G(\mathbf{0})} \log\left(\dfrac{G(\boldsymbol{x})}{G(\min(\boldsymbol{x}, \mathbf{0}))}\right), & \text{if } \boldsymbol{x} > \boldsymbol{l}, \\ 0 & \text{if } \exists\, j : x_j < l_j, \end{cases} \tag{5.2.3}$$

for any GEV G with $0 < G(\mathbf{0}) < 1$ (Rootzén and Tajvidi, 2006; Rootzén et al., 2016). For $\boldsymbol{x} > \mathbf{0}$, we get $H(\boldsymbol{x}) = 1 - \log G(\boldsymbol{x}) / \log G(\mathbf{0})$, which is exactly like the univariate case (1.2.4).

Let $J \subset \{1, \ldots, d\}$ and let H_J denote the $|J|$-variate marginal distribution of H. Let H_J^+ denote the margin H_J conditioned to have at least one positive component. In Rootzén et al. (2016, Proposition 2.1), it is shown that, if $H_J(\mathbf{0}) < 1$, then H_J^+ is an MGP distribution as well, associated to a GEV distribution G_J, i.e., the $|J|$-variate marginal distribution of G. Let $H_j := H_{\{j\}}$. For $x > 0$ and $j \in \{1, \ldots, d\}$,

$$H_j^+(x) = 1 - \frac{\log G_j(x)}{\log G_j(0)} = 1 - (1 + \xi_j x/\eta_j)_+^{-1/\xi_j}, \tag{5.2.4}$$

is the univariate GPD, with $\eta_j = \sigma_j - \xi_j \mu_j > 0$ and $\xi_j \in \mathbb{R}$ for $j \in \{1, \ldots, d\}$. The fact that H_J or H_j are typically not MGP distributions is rather intuitive, since the conditioning event involves all d random variables $X_1, \ldots, X_d$, whereas H_J concerns only the variables $(X_j)_{j \in J}$.

Following common practice in the statistical modelling of extremes, H may be used as a model for data which arise as multivariate threshold exceedances in the sense $\boldsymbol{X} \not\leq \boldsymbol{u}$. In particular, if $\boldsymbol{u} \in \mathbb{R}^d$ is a threshold that is sufficiently high in each margin, then (5.2.1) justifies the use of an MGPD as a model for exceedances over a high threshold, since

$$\mathbb{P}[\boldsymbol{X} - \boldsymbol{u} \leq \cdot \mid \boldsymbol{X} \not\leq \boldsymbol{u}] \approx H(\cdot / \boldsymbol{a}_n) = H_0(\cdot),$$

where H_0 is some other MGP distribution. Hence, the distribution of $\boldsymbol{X} - \boldsymbol{u} \mid \boldsymbol{X} \not\leq \boldsymbol{u}$ can be approximated by a member of the class of MGP distributions, with $\boldsymbol{\eta}$, $\boldsymbol{\xi}$ and the dependence structure to be estimated. From now on, we will do as in the univariate case and set $\boldsymbol{u} = \mathbf{0}$ without loss of generality: in what follows, we focus on threshold exceedances $\boldsymbol{Z} \mid \boldsymbol{Z} \not\leq \mathbf{0}$.

5.3 Model construction

Any vector $\boldsymbol{Z} = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ following an MGP distribution with distribution function H as in (5.2.3) can be written as

$$\boldsymbol{Z} \overset{\mathrm{d}}{=} \boldsymbol{\eta} \frac{e^{\boldsymbol{\xi} \boldsymbol{Z}^*} - \mathbf{1}}{\boldsymbol{\xi}}, \tag{5.3.1}$$

where $\boldsymbol{Z}^*$ is a "standard form" MGP random vector, that is, an MGP vector which is standardized to $\boldsymbol{\xi} = \mathbf{0}$ and $\boldsymbol{\eta} = \mathbf{1}$. This section discusses the construction of a vector $\boldsymbol{Z}^*$. For $\boldsymbol{\xi} = \mathbf{0}$, the right-hand side of equation (5.3.1) is simply $\boldsymbol{\eta} \boldsymbol{Z}^*$. The distribution function of $\boldsymbol{Z}^*$ will be denoted by H^*. It can be written in terms of the stable tail dependence function (1.3.4),

$$H^*(\boldsymbol{x}) = \frac{\ell\left(e^{\boldsymbol{\mu} - \min(\boldsymbol{x}, \mathbf{0})}\right) - \ell\left(e^{\boldsymbol{\mu} - \boldsymbol{x}}\right)}{\ell(e^{\boldsymbol{\mu}})}. \tag{5.3.2}$$

We focus on how to construct suitable densities for the random vector $\boldsymbol{Z}^*$ in equation (5.3.1), which will lead to densities for the MGP vector $\boldsymbol{Z}$ with general marginal forms. Let $T \sim \text{Exp}(1)$, and let $\boldsymbol{S}_0$ be a random vector that satisfies $\mathbb{P}[\max_{1\leq j\leq d} S_{0,j} = 0] = 1$ and is independent of T. Then the random vector $\boldsymbol{Z}^* = T + \boldsymbol{S}_0$ has the required properties to be a MGP vector on $\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} \not\leq \boldsymbol{0}\}$ with $\boldsymbol{\eta} = \boldsymbol{1}$ and $\boldsymbol{\xi} = \boldsymbol{0}$ (interpreted as the limit); see Rootzén et al. (2016). In terms of a general random vector $\widetilde{\boldsymbol{S}} \in [-\infty, \infty)^d \setminus \{-\boldsymbol{\infty}\}$, with $\mathbb{P}[S_j > -\infty] > 0$ for each $j \in 1, \ldots, d$, this can be written as

$$\boldsymbol{Z}^* = T + \widetilde{\boldsymbol{S}} - \max_{1\leq j\leq d} \widetilde{S}_j. \tag{5.3.3}$$

Let $f_{\widetilde{\boldsymbol{S}}}$ denote the density of $\widetilde{\boldsymbol{S}}$. If $\mathbb{P}[\widetilde{S}_j < \infty] = 1$ for each $j \in 1, \ldots, d$, then the density of $\boldsymbol{Z}^*$ defined in (5.3.3) is

$$h^*_{\widetilde{\boldsymbol{S}}}(\boldsymbol{z}) = \mathbb{1}(\max_{1\leq j\leq d} z_j > 0) e^{-\max_{1\leq j\leq d} z_j} \int_{-\infty}^{\infty} f_{\widetilde{\boldsymbol{S}}}(\boldsymbol{z} + t)\, \mathrm{d}t. \tag{5.3.4}$$

One way to construct models therefore is to assume different distributions for $\widetilde{\boldsymbol{S}}$, which provide flexible forms for h, and for which ideally the integral in (5.3.4) can be evaluated analytically. Some examples are provided in Section 5.4.

If the term $e^{-\max_{1\leq j\leq d} z_j}$ in (5.3.4) is inconvenient, an alternative approach to model construction that avoids this is proposed in Rootzén et al. (2016). Suppose that the density $f_{\widetilde{\boldsymbol{S}}}$ is a density formed by "tilting" another density $f_{\boldsymbol{S}}$ with some function $b : \mathbb{R}^d \to \mathbb{R}_+$, where $\mathbb{E}[b(\boldsymbol{S})] < \infty$, that is,

$$f_{\widetilde{\boldsymbol{S}}}(\boldsymbol{s}) = \frac{f_{\boldsymbol{S}}(\boldsymbol{s}) b(\boldsymbol{s})}{\mathbb{E}[b(\boldsymbol{S})]},$$

then the form of the density is

$$h^*_{\boldsymbol{S}}(\boldsymbol{z}) = \frac{\mathbb{1}(\max_{1\leq j\leq d} z_j > 0)}{\mathbb{E}[b(\boldsymbol{S})]} e^{-\max_{1\leq j\leq d} z_j} \int_{-\infty}^{\infty} f_{\boldsymbol{S}}(\boldsymbol{z} + t) b(\boldsymbol{z} + t)\, \mathrm{d}t.$$

An adequate choice of b, specifically $b(\boldsymbol{s}) = e^{\max_{1\leq j\leq d} s_j}$, means that the term $e^{-\max_{1\leq j\leq d} z_j}$ is eliminated, leaving the density

$$h^*_{\boldsymbol{S}}(\boldsymbol{z}) = \frac{\mathbb{1}(\max_{1\leq j\leq d} z_j > 0)}{\mathbb{E}\left[e^{\max_{1\leq j\leq d} S_j}\right]} \int_{-\infty}^{\infty} f_{\boldsymbol{S}}(\boldsymbol{z} + t) e^t\, \mathrm{d}t. \tag{5.3.5}$$

Because of the similarity of the integrals in (5.3.4) and (5.3.4), if one can be evaluated, then typically so can the other; the normalization constant in (5.3.4) is often more challenging to compute. In Section 5.4 we will assume a variety of probability distributions for either $\widetilde{\boldsymbol{S}}$ or $\boldsymbol{S}$, computing densities (5.3.4) or (5.3.5) respectively.

Using transformation (5.3.1), we have a general form for MGP densities as

$$h(\boldsymbol{z}) = h^*\left(\frac{\log(1+\boldsymbol{\xi}\boldsymbol{z}/\boldsymbol{\eta})}{\boldsymbol{\xi}}\right)\left(\prod_{j=1}^{d}\frac{1}{\eta_j+\xi_j z_j}\right). \tag{5.3.6}$$

Modelling the full densities is then done by plugging in (5.3.4) or (5.3.5) in (5.3.6).

We can obtain an alternative expression by defining the random variable $\boldsymbol{R} = (\boldsymbol{\eta}/\boldsymbol{\xi})\exp(\boldsymbol{\xi}\boldsymbol{S})$ for $\boldsymbol{\xi} \neq 0$, giving

$$h_{\boldsymbol{R}}(\boldsymbol{z}) = \mathbb{1}\left(\max_{1\leq j\leq d} z_j > 0\right)\frac{\int_0^\infty s^{\sum_{j=1}^d \xi_j} f_{\boldsymbol{R}}\left(s^{\boldsymbol{\xi}}\left(\boldsymbol{z}+\frac{\boldsymbol{\eta}}{\boldsymbol{\xi}}\right)\right)\mathrm{d}s}{\mathbb{E}\left[\max_{1\leq j\leq d}\left(\frac{\xi_j R_j}{\eta_j}\right)^{1/\xi_j}\right]}, \tag{5.3.7}$$

which is equal to (5.3.5) if $\boldsymbol{\xi} = \boldsymbol{0}$ and $\boldsymbol{\eta} = 1$. The motivation for the use of this variable $\boldsymbol{R}$ will be clear in Section 5.5, where we will illustrate how to obtain densities (5.3.4), (5.3.5) and (5.3.7) with the help of Poisson point processes.

5.4 Parametric models

Here we provide details of certain probability distributions for $\widetilde{\boldsymbol{S}}$, $\boldsymbol{S}$ and $\boldsymbol{R}$ that generate tractable MGP distributions. The use of random vectors to generate dependence structures for extremes is quite common, as we saw for instance in Subsection 1.3.4. In the application we will use a censored likelihood (see Subsection 5.6.1) and thus we should not just to be able to calculate densities, but also integrals of those densities. For each model we give the uncensored densities in the subsequent examples, whilst their censored versions are given in Appendix 5.A.

5.4.1 Independent components

Let $\widetilde{\boldsymbol{S}}$ and $\boldsymbol{S}$ in (5.3.4) and (5.3.5) respectively be vectors with independent and identically distributed components. The dependence structure of the associated MGP distribution is determined by the relative heaviness of the tails of the marginal distributions of $\widetilde{\boldsymbol{S}}$ or $\boldsymbol{S}$. The support for each density given in this subsection is $\{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{z} \not\leq \boldsymbol{0}\}$.

Example 5.4.1 (*Gumbel*)**.** Suppose $S_1, \ldots, S_d$ and $\widetilde{S}_1, \ldots, \widetilde{S}_d$ are iid Gumbel random variables with location parameters $\lambda_1, \ldots, \lambda_d \in \mathbb{R}$ and scale parameters $\alpha_1^{-1}, \ldots, \alpha_d^{-1} > 0$, i.e.,

$$f_{\boldsymbol{S}}(\boldsymbol{s}) = f_{\widetilde{\boldsymbol{S}}}(\boldsymbol{s}) = \prod_{j=1}^{d}\alpha_j \exp\{-\alpha_j(s_j-\lambda_j)\}\exp\{-\exp\{-\alpha_j(s_j-\lambda_j)\}\},$$

for $\alpha_j > 0$ and $\lambda_j > 0$. Then density (5.3.4) is

$$h^*_{\widetilde{\boldsymbol{S}}}(\boldsymbol{z}) = e^{\sum_{j=1}^d z_j - \max_{1\le j\le d} z_j} \times \int_0^\infty \prod_{j=1}^d \left(\frac{\alpha_j}{e^{\lambda_j}} \left(te^{z_j-\lambda_j}\right)^{-\alpha_j-1} e^{-(te^{z_j-\lambda_j})^{-\alpha_j}} \right) t^{d-1}\,\mathrm{d}t.$$

If $\alpha_1 = \cdots = \alpha_d = \alpha$ then this integral can be calculated explicitly:

$$h^*_{\widetilde{\boldsymbol{S}}}(\boldsymbol{z}) = e^{-\max_{1\le j\le d} z_j} \alpha^{d-1} \frac{\Gamma(d)\prod_{j=1}^d e^{-\alpha(z_j-\lambda_j)}}{\left(\sum_{j=1}^d (e^{z_j-\lambda_j})^{-\alpha}\right)^d}.$$

The marginal expectation of the exponentiated variable is

$$\mathbb{E}\left[e^{S_j}\right] = \begin{cases} e^{\lambda_j}\Gamma(1-1/\alpha_j), & \alpha_j > 1, \\ \infty, & \alpha_j \le 1. \end{cases} \tag{5.4.1}$$

Let $\min_{1\le j\le d} \alpha_j > 1$. Then density (5.3.5) is

$$h^*_{\boldsymbol{S}}(\boldsymbol{z}) = e^{\sum_{j=1}^d z_j} \frac{\int_0^\infty \prod_{j=1}^d \left(\frac{\alpha_j}{e^{\lambda_j}} \left(te^{z_j-\lambda_j}\right)^{-\alpha_j-1} e^{-(te^{z_j-\lambda_j})^{-\alpha_j}} \right) t^d\,\mathrm{d}t}{\int_0^\infty 1 - \prod_{j=1}^d e^{-(te^{-\lambda_j})^{-\alpha_j}}\,\mathrm{d}t}.$$

If $\alpha_1 = \cdots = \alpha_d = \alpha$ then this simplifies to:

$$h^*_{\boldsymbol{S}}(\boldsymbol{z}) = \frac{\alpha^{d-1}\Gamma(d-1/\alpha)\prod_{j=1}^d e^{-\alpha(z_j-\lambda_j)}}{\left(\sum_{j=1}^d (e^{z_j-\lambda_j})^{-\alpha}\right)^{d-1/\alpha}\Gamma(1-1/\alpha)\left(\sum_{j=1}^d e^{\alpha\lambda_j}\right)^{1/\alpha}}.$$

Observe that if in addition to $\alpha_1 = \cdots = \alpha_d = \alpha$, we have $\lambda_1 = \cdots = \lambda_d = 0$, then this is the MGP distribution associated to the well-known *logistic* model; see Example 1.3.1.

Example 5.4.2 (*Log-gamma*)**.** Suppose $e^{S_1}, \ldots, e^{S_d}$ and $e^{\widetilde{S}_1}, \ldots, e^{\widetilde{S}_d}$ are iid Gamma random variables with shape parameters $\alpha_1, \ldots, \alpha_d > 0$ and rate parameters all equal to one; i.e.,

$$f_{\boldsymbol{S}}(\boldsymbol{s}) = f_{\widetilde{\boldsymbol{S}}}(\boldsymbol{s}) = \prod_{j=1}^d \exp\{\alpha_j s_j\}\exp\{-\exp(s_j)\}/\Gamma(\alpha_j). \tag{5.4.2}$$

Recall that this is the density leading to the Dirichlet model, presented in Example 1.3.2. Density (5.3.4) is

$$\begin{aligned} h^*_{\widetilde{\boldsymbol{S}}}(\boldsymbol{z}) &= e^{-\max_{1\le j\le d} z_j} \prod_{j=1}^d \left(\frac{e^{\alpha_j z_j}}{\Gamma(\alpha_j)}\right) \int_0^\infty t^{\sum_{j=1}^d \alpha_j - 1} \exp\left\{-t\sum_{j=1}^d e^{z_j}\right\} \mathrm{d}t \\ &= \frac{\Gamma(\alpha_1+\cdots+\alpha_d)}{\prod_{j=1}^d \Gamma(\alpha_j)} \frac{e^{\sum_{j=1}^d \alpha_j z_j - \max_{1\le j\le d} z_j}}{(e^{z_1}+\cdots+e^{z_d})^{\alpha_1+\cdots+\alpha_d}}. \end{aligned}$$

For density (5.3.5) we first calculate

$$\begin{aligned}\mathbb{E}\left[e^{\max_j S_j}\right] &= \int_0^\infty 1 - \mathbb{P}[\exp S_1 \le t, \dots, \exp S_d \le t]\,\mathrm{d}t \\ &= \frac{\Gamma\left(\sum_{j=1}^d \alpha_j + 1\right)}{\prod_{j=1}^d \Gamma(\alpha_j)} \int_{\Delta_{d-1}} \max(u_1, \dots, u_d) \prod_{j=1}^d u_j^{\alpha_j - 1}\,\mathrm{d}\boldsymbol{u}_{1:d-1},\end{aligned}$$

where Δ_{d-1} is the unit simplex and $\boldsymbol{u}_{1:d-1} = \mathrm{d}u_1 \cdots \mathrm{d}u_{d-1}$. Define

$$C_d := \int_{\Delta_{d-1}} \max(u_1, \dots, u_d) \prod_{j=1}^d u_j^{\alpha_j - 1}\,\mathrm{d}u_1 \cdots \mathrm{d}u_{d-1}. \tag{5.4.3}$$

Then density (5.3.5) is

$$\begin{aligned}h^*_{\boldsymbol{S}}(\boldsymbol{z}) &= \frac{1}{\mathbb{E}\left[e^{\max_j S_j}\right]} \prod_{j=1}^d \left(\frac{e^{\alpha_j z_j}}{\Gamma(\alpha_j)}\right) \int_0^\infty t^{\sum_{j=1}^d \alpha_j} \exp\left\{-t \sum_{j=1}^d e^{z_j}\right\} \mathrm{d}t \\ &= \frac{1}{\mathbb{E}\left[e^{\max_j S_j}\right]} \prod_{j=1}^d \left(\frac{e^{\alpha_j z_j}}{\Gamma(\alpha_j)}\right) \frac{\Gamma\left(\sum_{j=1}^d \alpha_j + 1\right)}{(e^{z_1} + \cdots + e^{z_d})^{\alpha_1 + \cdots + \alpha_d + 1}} \\ &= \frac{\exp\{\sum_{j=1}^d \alpha_j z_j\}}{C_d\,(e^{z_1} + \cdots + e^{z_d})^{\alpha_1 + \cdots + \alpha_d + 1}}.\end{aligned}$$

Other examples are possible as well, for instance assuming that the components of $\widetilde{\boldsymbol{S}}$ or $\boldsymbol{S}$ are reverse Gumbel or reverse exponential variables. If we do not want to assume that the components of our vector are independent, we can for instance assume a Gaussian distribution, or the model presented in the next subsection.

5.4.2 Structured components

In Subsection 5.4.1 we considered several distributions for $\boldsymbol{S}$ and $\widetilde{\boldsymbol{S}}$, assuming they have independent components. Here we present a model for $\boldsymbol{R}$, based on partial sums of exponential random variables. We give the uncensored densities for both $\boldsymbol{\xi} = \boldsymbol{0}$ and $\boldsymbol{\xi} \neq \boldsymbol{0}$; see (5.3.7). This model will be of interest in Subsection 5.6, where we will focus on modelling landslides.

Case $\boldsymbol{\xi} = \boldsymbol{0}$ Recall from (5.3.7) that the densities $h_{\boldsymbol{R}}(\,\cdot\,)$ and $h^*_{\boldsymbol{S}}(\,\cdot\,)$ coincide if $\boldsymbol{\eta} = \boldsymbol{1}$. Let $\boldsymbol{R} \in \mathbb{R}^d$ be the random vector whose components are defined by

$$R_j = \log\left(\sum_{i=1}^j E_j\right), \qquad E_j \overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda_j), \qquad j = 1, \dots, d, \tag{5.4.4}$$

where $\lambda_1, \ldots, \lambda_d > 0$ are the rate parameters, i.e.,

$$\mathbb{P}[E_j > x_j] = \exp\left(-\lambda_j x_j\right), \qquad x_j \in (0, \infty),\ j = 1, \ldots, d.$$

The density $f_{\boldsymbol{R}}$ is given by

$$f_{\boldsymbol{R}}(\boldsymbol{r}) = \begin{cases} \left(\prod_{j=1}^{d} \lambda_j e^{-r_j}\right) \exp\left\{-\sum_{j=1}^{d} (\lambda_j - \lambda_{j+1}) e^{r_j}\right\}, & \text{if } r_1 < \cdots < r_d, \\ 0 & \text{otherwise.} \end{cases}$$

where we set $\lambda_{d+1} := 0$. Since by (5.3.3), $R_1 < \cdots < R_d$ implies $Z_1 < \cdots < Z_d$, the density of $\boldsymbol{Z}$ becomes, for $z_1 < \cdots < z_d$,

$$\begin{aligned} h_{\boldsymbol{R}}(\boldsymbol{z}) &= \frac{\mathbb{1}\left(z_d > 0\right)}{\mathbb{E}[e^{R_d}]} \left(\prod_{j=1}^{d} \lambda_j e^{z_j}\right) \int_0^\infty t^d \exp\left\{-t \left(\sum_{j=1}^{d} (\lambda_j - \lambda_{j+1}) e^{z_j}\right)\right\} \mathrm{d}t \\ &= \frac{\mathbb{1}(z_d > 0)\, d! \prod_{j=1}^{d} \lambda_j e^{z_j}}{\left(\sum_{j=1}^{d} \lambda_j^{-1}\right) \left(\sum_{j=1}^{d} (\lambda_j - \lambda_{j+1}) e^{z_j}\right)^{d+1}}. \end{aligned} \tag{5.4.5}$$

Note that for any constant $c > 0$ the parameters $(\lambda_1, \ldots, \lambda_d)$ and $(c\lambda_1, \ldots, c\lambda_d)$ lead to the same model - for identifiability we fix the value of λ_1.

Case $\boldsymbol{\xi} > \mathbf{0}$ Let $\boldsymbol{R} \in [0, \infty)^d$ be the random vector whose components are defined by

$$R_j = \sum_{i=1}^{j} E_j, \qquad E_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda_j), \qquad j = 1, \ldots, d,$$

where $\lambda_1, \ldots, \lambda_d > 0$ are the rate parameters. The density $f_{\boldsymbol{R}}$ is given by

$$f_{\boldsymbol{R}}(\boldsymbol{r}) = \begin{cases} \left(\prod_{j=1}^{d} \lambda_j\right) \exp\left\{-\sum_{j=1}^{d} (\lambda_j - \lambda_{j+1}) r_j\right\}, & \text{if } r_1 < \cdots < r_d, \\ 0 & \text{otherwise.} \end{cases}$$

where we set $\lambda_{d+1} = 0$. Suppose that $\eta = \eta_1 = \cdots = \eta_d$ and $\xi = \xi_1 = \cdots = \xi_d$. Then

$$\mathbb{E}\left[\max_{1 \le j \le d} \left(\frac{\xi R_j}{\eta}\right)^{1/\xi}\right] = \left(\frac{\xi}{\eta}\right)^{1/\xi} \mathbb{E}\left[R_d^{1/\xi}\right].$$

The distribution of R_d is called hypo-exponential if $\lambda_i \neq \lambda_j$ for all $i \neq j$, and we get

$$\begin{aligned}\left(\frac{\xi}{\eta}\right)^{1/\xi} \mathbb{E}\left[R_d^{1/\xi}\right] &= \left(\frac{\xi}{\eta}\right)^{1/\xi} \int_0^\infty r^{1/\xi} f_d(r)\,\mathrm{d}r \\ &= \left(\frac{\xi}{\eta}\right)^{1/\xi} \sum_{i=1}^d \lambda_i \left(\prod_{j=1, j\neq i}^d \frac{\lambda_j}{\lambda_j - \lambda_i}\right) \int_0^\infty r^{1/\xi} e^{-r\lambda_i}\,\mathrm{d}r \\ &= \left(\frac{\xi}{\eta}\right)^{1/\xi} \Gamma\left(\frac{1}{\xi}+1\right) \sum_{i=1}^d \lambda_i^{-1/\xi} \left(\prod_{j=1, j\neq i}^d \frac{\lambda_j}{\lambda_j - \lambda_i}\right).\end{aligned}$$

The density of $\boldsymbol{Z}$ becomes, for $z_1 > -\eta/\xi$ and $z_d > 0$,

$$\begin{aligned} h_{\boldsymbol{R}}(\boldsymbol{z}) &= \frac{\int_0^\infty t^{d\xi} \exp\left\{-t^\xi \sum_{j=1}^d (\lambda_j - \lambda_{j+1})(z_j + \eta/\xi)\right\}\,\mathrm{d}t}{\left(\prod_{j=1}^d \lambda_j^{-1}\right)\left(\frac{\xi}{\eta}\right)^{1/\xi} \mathbb{E}\left[R_d^{1/\xi}\right]} \qquad (5.4.6) \\ &= \frac{\left(\prod_{j=1}^d \lambda_j\right)\left(\frac{\xi}{\eta}\right)^{-1/\xi} \Gamma\left(d+\frac{1}{\xi}\right)/\Gamma\left(\frac{1}{\xi}\right)}{\left(\sum_{j=1}^d (\lambda_j - \lambda_{j+1}) z_j + (\eta/\xi)\lambda_1\right)^{d+1/\xi} \sum_{i=1}^d \lambda_i^{-1/\xi} \left(\prod_{j=1, j\neq i}^d \frac{\lambda_j}{\lambda_j - \lambda_i}\right)}. \end{aligned}$$

Again, we fix the value of λ_1 for identifiability.

5.4.3 Simulation

We focus on the simulation of a standardized MGP distributed vector $\boldsymbol{Z}^*$, since a non-standardized vector $\boldsymbol{Z}$ is easily obtained by expression (5.3.1). In Section 5.3, we saw that such a vector can be simulated either from density (5.3.4) or from density (5.3.5), i.e., starting from either a vector $\widetilde{\boldsymbol{S}}$ or a vector $\boldsymbol{S}$.

First, note that simulation from the $\widetilde{\boldsymbol{S}}$-density is immediate because of (5.3.3): we simulate an exponential variable $T \sim \mathrm{Exp}(1)$ and a vector $\widetilde{\boldsymbol{S}}$ from $f_{\widetilde{\boldsymbol{S}}}$ independently, and set $\boldsymbol{Z} = T + \widetilde{\boldsymbol{S}} - \max_{j=1,\ldots,d} \widetilde{S}_j$. Simulation from the density with $\boldsymbol{S}$ can be done by exploiting the formula

$$f_{\boldsymbol{S}}(\boldsymbol{s}) = \frac{e^{\max_{j=1,\ldots,d} s_j} f_{\widetilde{\boldsymbol{S}}}(\boldsymbol{s})}{\mathbb{E}[e^{\max_{j=1,\ldots,d} \widetilde{S}_j}]}. \qquad (5.4.7)$$

If $\widetilde{\boldsymbol{S}}$ is a vector we can sample from directly, then we can use rejection sampling to sample from the vector $\boldsymbol{S}$ as follows. Let $f_{\boldsymbol{Q}}$ be a density such that

$$\sup_{s\in\mathbb{R}} \frac{f_{\boldsymbol{S}}(\boldsymbol{s})}{f_{\boldsymbol{Q}}(\boldsymbol{s})} = K < \infty.$$

Then an observation from $\boldsymbol{S}$ is obtained by:

1. Sample a vector $\boldsymbol{Q}$ from $f_{\boldsymbol{Q}}$.
2. Sample a uniform random variable $U \sim U(0, K)$.
3. If $U < f_{\boldsymbol{S}}(\boldsymbol{Q})/f_{\boldsymbol{Q}}(\boldsymbol{Q})$, set $\boldsymbol{S} = \boldsymbol{Q}$; if not, restart at step 1.

Simulation of a density (5.3.7), based on a vector $\boldsymbol{R}$, can either be done by approximate simulation as described in Method 4 in Rootzén et al. (2016, Section 6), or by simulating a vector $\boldsymbol{S}$ and setting $\boldsymbol{R} = \exp(\boldsymbol{\xi S})(\boldsymbol{\sigma}/\boldsymbol{\xi})$.

Example 5.4.3. We show how to simulate from the log-gamma model, presented in Example 5.4.2. Let $f_{\widetilde{S}}$ be as in (5.4.2) and C_d as in (5.4.3). Then (5.4.7) becomes

$$f_{\widetilde{\boldsymbol{S}}}(\boldsymbol{s}) = \frac{e^{\max_{j=1,\dots,d} s_j + \sum_{j=1}^d (\alpha_j s_j - \exp(s_j))}}{C_d \Gamma(\sum_{j=1}^d \alpha_j + 1)}.$$

Let $\exp Q_j \sim \Gamma(\alpha_j, \beta)$ for $j = 1, \dots, d$ be iid random variables, where β is a common rate parameter such that $\beta < 1$. Then we find

$$\begin{aligned} K &= \frac{\prod_{j=1}^d (\beta^{-\alpha_j}\Gamma(\alpha_j))}{\Gamma(\sum_{j=1}^d \alpha_j + 1) C_d} \sup_{\boldsymbol{s} \in (-\infty, \infty)} e^{\max_{j=1,\dots,d} s_j + (\beta - 1)\sum_{j=1}^d e^{s_j}} \\ &= \frac{\prod_{j=1}^d (\beta^{-\alpha_j}\Gamma(\alpha_j))}{\Gamma(\sum_{j=1}^d \alpha_j + 1) C_d e (1-\beta)}, \end{aligned}$$

since the maximum is attained at $(0, \dots, 0, -\log(1-\beta), 0, \dots, 0)$. To minimize K we can choose $\beta = (\sum_{j=1}^d \alpha_j)/(\sum_{j=1}^d \alpha_j + 1)$. Simulation is very efficient: usually, $K \approx 2$.

5.5 Point process representations

A related construction of MGP distributions comes from Poisson point process representations. These representations will allow us to derive expressions for the GEV G, and thus for H or H^*. In the following, we let $(U_i)_{i \geq 1}$ denote the points of a Poisson process on $(0, \infty)$ with unit intensity.

Point process I Let $\boldsymbol{S}$ be any random vector such that $0 < \mathbb{E}[\exp S_j] < \infty$ for $j = 1, \dots, d$. Let $(\boldsymbol{S}_i)_{i \geq 1}$ be iid copies of $\boldsymbol{S}$. Consider the Poisson process $N_{\boldsymbol{S}} = \sum_{i \geq 1} \delta_{\boldsymbol{Z}_i}$ where

$$\boldsymbol{Z}_i = \begin{cases} \boldsymbol{\eta} \dfrac{(\exp(\boldsymbol{S}_i)/U_i)^{\boldsymbol{\xi}} - 1}{\boldsymbol{\xi}} = & \text{if } \boldsymbol{\xi} \neq \boldsymbol{0}, \boldsymbol{\eta} > \boldsymbol{0}, \\ \boldsymbol{S}_i - \log U_i & \text{if } \boldsymbol{\xi} = \boldsymbol{0}, \boldsymbol{\eta} = \boldsymbol{1}. \end{cases} \tag{5.5.1}$$

It is well-known (de Haan, 1984; Schlather, 2002) that $\max_{i\geq 1} \boldsymbol{Z}_i$ is max-stable (cf. expression (1.4.2)) so that on sets of the form $\mathcal{A} = [-\boldsymbol{\infty}, \boldsymbol{\infty}] \setminus [-\boldsymbol{\infty}, \boldsymbol{x}]$,

$$G(\boldsymbol{x}) = \mathbb{P}[\max_{i\geq 1} \boldsymbol{Z}_i \leq \boldsymbol{x}] = \mathbb{P}[N(\mathcal{A}) = 0] = \exp\left\{-\nu(\mathcal{A})\right\}.$$

The intensity measure of this Poisson process is then

$$\begin{aligned} \nu\left(\mathcal{A}\right) = \mathbb{E}\left[N(\mathcal{A})\right] &= \int_0^\infty \mathbb{P}\left[\eta \frac{(\exp(\boldsymbol{S})/u)^{\boldsymbol{\xi}} - 1}{\boldsymbol{\xi}} \not\leq \boldsymbol{x}\right] \mathrm{d}u \\ &= \int_0^\infty \mathbb{P}\left[\max_{j=1,\ldots,d} e^{\boldsymbol{S}} \left(1 + \frac{\xi_j x_j}{\eta_j}\right)^{-1/\xi_j} \geq u\right] \mathrm{d}u \\ &= \mathbb{E}\left[\max_{j=1,\ldots,d} e^{\boldsymbol{S}} \left(1 + \frac{\xi_j x_j}{\eta_j}\right)^{-1/\xi_j}\right]. \end{aligned} \tag{5.5.2}$$

We can identify the marginal parameters $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$ of G by comparing

$$G_j(x_j) = \exp\left\{-\mathbb{E}\left[e^{S_j}\right] \left(\frac{\xi_j x_j}{\eta_j} + 1\right)^{-1/\xi_j}\right\},$$

with (5.2.2); we find

$$\mu_j = (\eta_j/\xi_j)\left(\mathbb{E}[\exp S_j]^{\xi_j} - 1\right), \quad \sigma_j = \eta_j \mathbb{E}[\exp S_j]^{\xi_j}, \qquad j \in \{1, \ldots, d\}.$$

Then combining the expression for $\nu(\mathcal{A})$ with the expressions for the marginals $G_1, \ldots, G_d$ and recalling (1.3.10), we can express the stable tail dependence function as

$$\ell(\boldsymbol{x}) = \mathbb{E}\left[\max_{j=1,\ldots,d} x_j \frac{\exp S_j}{\mathbb{E}[\exp S_j]}\right], \qquad \boldsymbol{x} \in [0, \infty)^d. \tag{5.5.3}$$

Another expression for the intensity measure, and thus for G, is obtained from (5.5.2) as

$$G(\boldsymbol{x}) = \exp\left\{-\int_0^\infty 1 - F_{\boldsymbol{S}}\left(\log(t) + \boldsymbol{\xi}^{-1} \log\left(1 + \boldsymbol{\xi}\boldsymbol{x}/\boldsymbol{\eta}\right)\right) \mathrm{d}t\right\}.$$

For $\boldsymbol{\xi} = \boldsymbol{0}$ and $\boldsymbol{\eta} = \boldsymbol{1}$ we find that $\max_{i\geq 1} \boldsymbol{Z}_i$ follows a multivariate GEV distribution with $\boldsymbol{\sigma} = 1$ and $\boldsymbol{\mu} = \log \mathbb{E}[\exp \boldsymbol{S}]$ and we get the simple expression

$$G(\boldsymbol{x}) = \exp\left\{-\int_0^\infty 1 - F_{\boldsymbol{S}}\left(\log t + x\right) \mathrm{d}t\right\}.$$

Plugging this into (5.2.3) leads to the corresponding standardized MGP distribution function

$$H^*(\boldsymbol{z}) = \frac{\int_0^\infty F_{\boldsymbol{S}}(\boldsymbol{z} + \log t) - F_{\boldsymbol{S}}(\min(\boldsymbol{z}, 0) + \log t)\, \mathrm{d}t}{\int_0^\infty 1 - F_{\boldsymbol{S}}(\log t \boldsymbol{1})\, \mathrm{d}t}. \tag{5.5.4}$$

Differentiating under the integral sign leads to the density

$$h^*(\boldsymbol{z}) = \frac{\mathbb{1}(\max_{1\le j\le d} z_j > 0)}{\mathbb{E}[\exp \max_{1\le j\le d} S_j)]} \int_0^\infty f_{\boldsymbol{S}}(\boldsymbol{z} + \log t)\,\mathrm{d}t,$$

where we recognize (5.3.5). For a formal proof, see Rootzén et al. (2016).

Point process II. Let $\boldsymbol{S}_0$ be such that $\mathbb{P}[\max_{1\le j\le d} S_0 = 0] = 1$. Consider the Poisson process $N_{\boldsymbol{S}_0} = \sum_{i\ge 1} \delta_{\boldsymbol{Z}_i}$ where $\boldsymbol{Z}_i$ are as in (5.5.1) with $\boldsymbol{S}$ replaced by $\boldsymbol{S}_0 = \widetilde{\boldsymbol{S}} - \max_{1\le j\le d} \widetilde{S}_j$ for some vector $\widetilde{\boldsymbol{S}} \in \mathbb{R}^d$. In (5.5.4) the denominator disappears because $\mathbb{P}[\boldsymbol{S}_0 \not\le \log t\mathbf{1}] = 0$ for $t \ge 1$, so that

$$\begin{aligned}
H^*(\boldsymbol{z}) &= \int_0^\infty F_{\boldsymbol{S}_0}(\boldsymbol{z} + \log t) - F_{\boldsymbol{S}_0}(\min(\boldsymbol{z}, 0) + \log t)\,\mathrm{d}t \\
&= \int_0^\infty \int_{\boldsymbol{s}} \mathbb{1}\left(\min(\boldsymbol{z}, \mathbf{0}) + \log t \le \boldsymbol{s} - \max_{1\le j\le d} s_j \le \boldsymbol{z} + \log(t)\right) f_{\widetilde{\boldsymbol{S}}}(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}\,\mathrm{d}t \\
&= \int_0^\infty \int_{\min(\boldsymbol{z},\mathbf{0})}^{\boldsymbol{z}} v^{-1} f_{\widetilde{\boldsymbol{S}}}(\boldsymbol{y} + \log v) e^{-\max_{1\le j\le d} y_j}\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}v,
\end{aligned}$$

where for the last step we have set $v = te^{\max_{1\le j\le d} s_j}$ and $\boldsymbol{s} = \boldsymbol{y} + \log v$. The density is then

$$h^*(\boldsymbol{z}) = \frac{\mathbb{1}(\max_{1\le j\le d} z_j > 0)}{e^{-\max_{1\le j\le d} s_j}} \int_0^\infty t^{-1} f_{\widetilde{\boldsymbol{S}}}(\boldsymbol{z} + \log t)\,\mathrm{d}t,$$

which corresponds to (5.3.4).

Point process III. Let $\boldsymbol{R}$ be a random vector such that, for all $j = 1, \dots, d$,

$$\begin{aligned}
\mathbb{E}\left[|R_j|^{1/\xi_j}\right] < \infty \text{ if } \xi_j \ne 0, \\
\mathbb{E}\left[\exp R_j\right] < \infty \text{ if } \xi_j = 0.
\end{aligned}$$

Let $(\boldsymbol{R}_i)_{i\ge 1}$ be iid copies of $\boldsymbol{R}$, and define

$$N_{\boldsymbol{R}} = \sum_{i\ge 1} \delta_{\boldsymbol{Z}_i} = \begin{cases} \displaystyle\sum_{i\ge 1} \delta_{\boldsymbol{R}_i / U_i^{\boldsymbol{\xi}}} & \text{if } \boldsymbol{\xi} \ne \mathbf{0}, \\ \displaystyle\sum_{i\ge 1} \delta_{\boldsymbol{R}_i - \log U_i} & \text{if } \boldsymbol{\xi} = \mathbf{0}. \end{cases}$$

Note that for $\boldsymbol{\xi} = \mathbf{0}$, this is equivalent to (5.5.1) above. Then the intensity measure of this Poisson process is, for $\mathcal{A} = [-\boldsymbol{\infty}, \boldsymbol{\infty}] \setminus [-\boldsymbol{\infty}, \boldsymbol{x}]$,

$$\nu(\mathcal{A}) = \begin{cases} \displaystyle\int_0^\infty \mathbb{P}\left[\frac{\boldsymbol{R}_i}{u^{\boldsymbol{\xi}}} \not\le \boldsymbol{x}\right] \mathrm{d}u = \int_0^\infty 1 - F_{\boldsymbol{R}}(u^{\boldsymbol{\xi}})\,\mathrm{d}u & \text{if } \boldsymbol{\xi} \ne \mathbf{0}, \\ \displaystyle\int_0^\infty 1 - F_{\boldsymbol{R}}\left(\log t + x\right)\,\mathrm{d}t & \text{if } \boldsymbol{\xi} = \mathbf{0}, \end{cases}$$

and the corresponding GEV distribution G is

$$G(\boldsymbol{x}) = \begin{cases} \exp\left\{-\int_0^\infty 1 - F_{\boldsymbol{R}}(t^{\boldsymbol{\xi}}\boldsymbol{x})\,\mathrm{d}t\right\} & \text{if } \boldsymbol{\xi} \neq \boldsymbol{0}, \\ \exp\left\{-\int_0^\infty 1 - F_{\boldsymbol{R}}(\log t + \boldsymbol{x})\,\mathrm{d}t\right\} & \text{if } \boldsymbol{\xi} = \boldsymbol{0}. \end{cases}$$

Recall that for $\boldsymbol{\xi} \neq \boldsymbol{0}$, the lower limits of the univariate GEV distributions $G_1, \ldots, G_d$ are $l_j = -\eta_j/\xi_j$. Here we consider the conditional distribution of $\boldsymbol{X} - \boldsymbol{\eta}/\boldsymbol{\xi} \mid \boldsymbol{X} \not\leq \boldsymbol{\eta}/\boldsymbol{\xi}$, so setting $H_{\boldsymbol{R}}(\boldsymbol{x}) = H(\boldsymbol{x} + \boldsymbol{\eta}/\boldsymbol{\xi})$ in (5.2.3) we get,

$$H_{\boldsymbol{R}}(\boldsymbol{z}) = \frac{\int_0^\infty F_{\boldsymbol{R}}\left(t^{\boldsymbol{\xi}}\left(\boldsymbol{z} + \frac{\boldsymbol{\eta}}{\boldsymbol{\xi}}\right)\right) - F_{\boldsymbol{R}}\left(t^{\boldsymbol{\xi}}\left(\min(\boldsymbol{z}, \boldsymbol{0}) + \frac{\boldsymbol{\eta}}{\boldsymbol{\xi}}\right)\right)\,\mathrm{d}t}{\int_0^\infty 1 - F_{\boldsymbol{R}}\left(t^{\boldsymbol{\xi}}\frac{\boldsymbol{\eta}}{\boldsymbol{\xi}}\right)\,\mathrm{d}t}. \tag{5.5.5}$$

Finally, differentiating under the integral sign leads to the density

$$h(\boldsymbol{z}) = \mathbb{1}\left(\max_{1\leq j\leq d} z_j > 0\right) \frac{\int_0^\infty t^{\sum_{j=1}^d \xi_j} f_{\boldsymbol{R}}\left(t^{\boldsymbol{\xi}}\left(\boldsymbol{z} + \frac{\boldsymbol{\eta}}{\boldsymbol{\xi}}\right)\right)\,\mathrm{d}t}{\int_0^\infty 1 - F_{\boldsymbol{R}}\left(t^{\boldsymbol{\xi}}\frac{\boldsymbol{\eta}}{\boldsymbol{\xi}}\right)\,\mathrm{d}t},$$

for $\boldsymbol{\xi} \neq \boldsymbol{0}$, where we recognize (5.3.7). If $\boldsymbol{\xi} = \boldsymbol{0}$ we get expression (5.5.4).

Remark 5.5.1. We chose to standardize to $\boldsymbol{\xi} = \boldsymbol{0}$ because of the simplicity of the formulas, but standardization to $\boldsymbol{\xi} = \boldsymbol{1}$ (or $\boldsymbol{\xi} = -\boldsymbol{1}$) is possible as well: starting from the point process (5.5.1) with $\boldsymbol{W}_i = \exp(\boldsymbol{S}_i)$, we get the MGP distribution function

$$H_0(\boldsymbol{z}) = \frac{\int_0^\infty \left(F_{\boldsymbol{W}}(r(\boldsymbol{1} + \boldsymbol{\xi}\boldsymbol{z}/\boldsymbol{\eta})^{1/\boldsymbol{\xi}}) - F_{\boldsymbol{W}}(r(\boldsymbol{1} + \boldsymbol{\xi}\min(\boldsymbol{z}, \boldsymbol{0})/\boldsymbol{\eta})^{1/\boldsymbol{\xi}}\right)\,\mathrm{d}r}{\int_0^\infty 1 - F_{\boldsymbol{W}}(r\boldsymbol{1})\,\mathrm{d}r}.$$

Then letting $\boldsymbol{X} \sim H_0$ with $\boldsymbol{\eta} = \boldsymbol{\xi} = \boldsymbol{1}$ and setting $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{1}$, we get a standardized MGP distribution function

$$H_0^*(\boldsymbol{z}) = \frac{\int_0^\infty F_{\boldsymbol{W}}(r\boldsymbol{z}) - F_{\boldsymbol{W}}(r\min(\boldsymbol{z}, \boldsymbol{1}))\,\mathrm{d}r}{\int_0^\infty 1 - F_{\boldsymbol{W}}(r\boldsymbol{1})\,\mathrm{d}r}.$$

5.6 Applications

5.6.1 Censored likelihood inference

We will use the density (5.3.6) as a contribution to the likelihood only when all components of the observed vector $\boldsymbol{X}$ are "large", in the sense of exceeding the threshold $\boldsymbol{u}$. The reasoning for this is twofold.

1. For $\xi_j > 0$, the lower endpoint of a MGP distribution is $-\eta_j/\xi_j$. Using a censored likelihood means that for small values of a component, we only need to assume they are between $-\eta_j/\xi_j$ and u_j.

2. It has become well established that without censoring, bias is induced in the estimation of dependence parameters; see, for instance, Huser et al. (2015).

Let $D := \{1, \ldots, d\}$ and $C \subset D$ be the subset of indices denoting which components of $\boldsymbol{X}$ fall below the corresponding component of $\boldsymbol{u}$, i.e., $X_j \leq u_j$ for $j \in C$, and $X_j > u_j$ for $j \in D \setminus C$, with at least one j such that $X_j > 0$. Set $\boldsymbol{x}_C := \{x_j : j \in C\}$. For each realization of $\boldsymbol{X}$, we use the likelihood contribution

$$h_C(\boldsymbol{x}_{D\setminus C}, \boldsymbol{u}_C) = \int_{\prod_{j\in C}(-\infty, u_j]} h(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}_C, \tag{5.6.1}$$

which is equal to (5.3.6) if C is empty. Appendix 5.A contains forms of censored likelihood contributions for the models included in Section 5.4. Thus, for n independent repeated observations of $\boldsymbol{X}$, $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$, the form of the censored likelihood function to be maximized is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} h_{C_i}(\boldsymbol{x}_i; \boldsymbol{\theta}),$$

where C_i denotes the censoring subset for $\boldsymbol{x}_i$, which may be empty, and $\boldsymbol{\theta}$ is the parameter vector containing $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ and the dependence parameters stemming from the construction vector $\widetilde{\boldsymbol{S}}$, $\boldsymbol{S}$ or $\boldsymbol{R}$.

Working within a likelihood-based framework for inference offers clear benefits: for instance, comparison of nested models is straightforward by likelihood ratio tests. This is important as the number of parameters can quickly grow large if margins and dependence are fitted simultaneously. Moreover, incorporating covariate effects, such as a linear trend in one of the parameters, is straightforward. Such ideas were introduced in the univariate framework by Davison and Smith (1990), but nonstationarity in the dependence structure has received comparatively little attention; an exception is Huser and Genton (2016).

5.6.2 Case study: landslides

Rainfall might cause water pressure to build up into the ground and is therefore a trigger of landslides. Landslides may be caused either by extreme rainfall on a certain day, or by moderate rainfall amounts on consecutive days. It is possible to establish a threshold relation between the intensity (I) of the rainfall, i.e., the amount of precipitation accumulated in a period, and the duration (D) of the rainfall. Guzzetti et al. (2007) consolidate many previous studies on landslides, obtaining 853 landslide events between 1841 and 2002. They find that the majority of rainfall events that cause landslides have a duration between one hour and three days (Guzzetti et al., 2007, Figure 6A),

and propose the following threshold relation for highland climates in Europe:

$$I = 7.56 \times D^{-0.48}, \tag{5.6.2}$$

where the intensity is in mm/hour and the duration is expressed in hours. The amount of rainfall necessary to cause a landslide during a 3-day period is then 69.9 mm, during a 2-day period 56.59 mm and during a 1-day period 39.47.

The total cost of landslides in Sweden is typically around SEK 200 million/year. In Abisko (northern Sweden), there have been several landslides in the past century, for instance in October 1959, August 1998, and in July 2004 (Rapp and Strömquist, 1976; Jonasson and Nyberg, 1999; Beylich and Sandberg, 2005). We have daily accumulated precipitation data (in mm) from Abisko between January 1st 1913 to December 31st 2014, leading to a sample size of 37 255. The previously mentioned landslides are visible in our data, with 24.5 mm rain on the 5th of October 1959, 21.0 mm rain on the 24th of August 1998 and 61.9 mm rain on the 21st of July 2004, which is the largest value of our entire dataset. Figure 5.1 shows the daily precipitation amounts $P_1, \ldots, P_n$ (left) and a mean residual life plot of $P_1, \ldots, P_n$ (right). Based on the mean residual life plot, we choose the threshold $u = 12$, which corresponds roughly to the 99% quantile, leading to 335 threshold exceedances. The parameter stability plot in Figure 5.2 confirms the choice of this threshold.

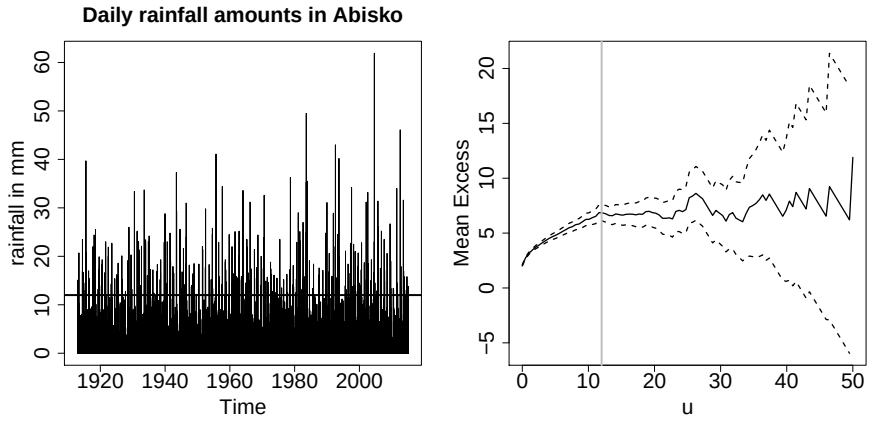


Figure 5.1: Daily precipitation $P_1, \ldots, P_n$ in Abisko (left) and the mean residual life plot with the threshold $u = 12$ (right).

We wish to construct a dataset $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N \in \mathbb{R}^3$, whose components represent daily, two-day, and three-day extreme rainfall amounts respectively. We limit ourselves to $d = 3$ because of the findings in Guzzetti et al. (2007). Let $P_{(1)}$ denote the first value of $P_1, \ldots, P_n$ which exceeds the threshold u, or the

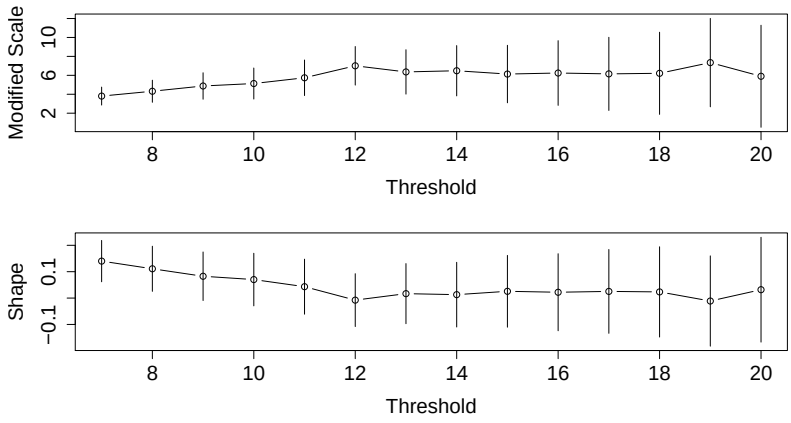


Figure 5.2: Parameter stability plots for the daily precipitation amounts $P_1, \ldots, P_n$.

maximum of two consecutive non-zero datapoints whose sum is larger than u, or the maximum of three consecutive non-zero datapoints whose sum is larger than u. Let the first cluster consist of $P_{(1)}$ plus the five values preceding it and the five values following it. Then let X_{11} be the largest value in the first cluster, X_{12} the largest sum of two consecutive non-zero values in the first cluster, and X_{13} the largest sum of three consecutive non-zero values in the first cluster. Find the second cluster and compute $\boldsymbol{X}_2 = (X_{21}, X_{22}, X_{23})$ in the same way, starting with the first value after the first cluster. If the second cluster overlaps with the first one, then we let the second cluster only consist of the values which are not already in the first cluster. Continuing this way, we obtain a dataset $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, with $N = 580$. Recall that $\boldsymbol{X}_i - u \mid \boldsymbol{X}_i \not\leq u$ can be approximated by an MGPD $\boldsymbol{X}$ on $\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} \not\leq 0\}$, whose margins, conditionally on being positive, are univariate GP distributions; see (5.2.4).

Time trend

A similar dataset has previously been analysed in Rudvik (2012), where a univariate GEV model with a linear trend in the location parameter was fitted to annual maxima, concluding there is no significant trend. We investigate the question whether there is a linear trend in the daily, two-day or three-day rainfall amounts by fitting a univariate GP distribution with $\log \eta(t) = a_2 + b_2 t$ for $t \in (0, 1]$ to the marginal distributions of $(\boldsymbol{X}_i)_{i=1,\ldots,N}$, where the time t corresponds to the time of the threshold exceedances. To this end, we need to select marginal threshold above which we fit the univariate GP distributions. For the first component, we take $u = 12$ as found previously; for the second

and third component, we take $u = 13.5$ and $u = 14$ respectively, based on inspection of parameter stability plots.

In Table 5.1, we report the parameter estimates for the univariate GP fit above these thresholds. The last line shows the deviance, i.e., -2 times the difference in log-likelihood with respect to a model with $\eta(t) \equiv \eta$. We compare to the 95% quantile of a χ^2_1 distribution, given by 3.84. Likelihood ratio tests show that a linear trend in the logarithm of the scale parameter is rejected for all margins.

We do not adopt any trend and Table 5.2 shows the result of fitting univariate GP distributions to the margins conditional on exceeding the previously mentioned thresholds. Observing that the estimated shape parameters are all around zero, we would like to test if the simpler model $\xi_1 = \xi_2 = \xi_3 = 0$ would suffice and we find that likelihood ratio tests can not be rejected. The assumption of an equal η for all margins can not be rejected either (for $\eta = 8.6$), providing some evidence for an even simpler model where $\eta_1 = \eta_2 = \eta_3$. In the following analysis, we will set $\boldsymbol{\eta} = \eta\mathbf{1}$ and $\boldsymbol{\xi} = \xi\mathbf{1}$, and we will fit both a model with the restriction $\xi = 0$ and one without.

	$\boldsymbol{X}_{i1}$	$\boldsymbol{X}_{i2}$	$\boldsymbol{X}_{i3}$
$\widehat{\xi}$	-0.09 (0.06)	-0.05 (0.06)	-0.03 (0.06)
$\widehat{a}_2$	2.01 (0.12)	2.13 (0.11)	2.20 (0.11)
$\widehat{b}_2$	0.31 (0.27)	0.32 (0.25)	0.27 (0.22)
deviance	1.27	1.62	1.52

Table 5.1: Estimates of the parameters of a GP model with rate $\log \eta = a_2 + b_2 t$ and shape ξ for thresholds $u = 12$, $u = 13.5$ and $u = 14$ respectively; standard errors in parentheses.

	$\boldsymbol{X}_{i1}$	$\boldsymbol{X}_{i2}$	$\boldsymbol{X}_{i3}$
$\widehat{\xi}$	-0.06 (0.05)	-0.02 (0.06)	-0.01 (0.05)
$\widehat{\eta}$	8.26 (0.69)	9.34 (0.74)	9.96 (0.74)

Table 5.2: Estimates of the parameters of a GP model with rate η and shape ξ for thresholds $u = 12$, $u = 13.5$ and $u = 14$ respectively; standard errors in parentheses.

Dependence structure

We fit the MGP density corresponding to the structured components model from Section 5.4.2 to the data $(\boldsymbol{X}_i)_{i=1,\dots,N}$. Note that the components of our data are strictly increasing, so that the model from Section 5.4.2 is applicable.

We consider a model with the restriction $\xi = 0$ and a more general model with $\xi \in \mathbb{R}$. Recall that we need to impose a restriction on the parameters for identifiability; we set $\lambda_1 = 7.79$ for $\xi = 0$, obtained by fitting an exponential distribution to the first margin, and $\lambda_1 = 8.26$ for $\xi \in \mathbb{R}$ (see Table 5.2). We estimate the parameters $(\lambda_2, \lambda_3, \eta)$ in the first case and $(\lambda_2, \lambda_3, \eta, \xi)$ in the second case. We choose $u = 24$ and we continue with the 142 datapoints whose third component exceeds $u = 24$. Table 5.3 shows the parameter estimates obtained using censored likelihood. Again, the hypothesis of $\xi = 0$ can not be rejected. We see that the estimates of η are somewhat higher than we saw in the marginal analysis.

$u = 24$	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\eta}$	$\widehat{\xi}$	nllh
restricted	7.78	6.57	8.40	10.17	–	870.0
	–	(0.99)	(1.43)	(0.80)	–	
unrestricted	8.26	6.84	8.79	9.14	0.11	868.9
	–	(1.02)	(1.50)	(0.99)	(0.08)	

Table 5.3: Parameters for the ordered components model with $u = 24$; standard errors in parentheses.

Let $\boldsymbol{X} = (X_1, X_2, X_3)$ denote a MGP vector whose density is the one of the three-dimensional structured components model. We wish to estimate the probability of a future landslide using formula (5.6.2), i.e., we wish to calculate $\mathbb{P}[\boldsymbol{X} \not\leq \boldsymbol{x}]$ where $\boldsymbol{x} = (39.5, 56.6, 69.9)$, and $\boldsymbol{x} > \boldsymbol{u}$. We can write

$$\begin{aligned} \mathbb{P}[\boldsymbol{X} \not\leq \boldsymbol{x}] &= \mathbb{P}[\boldsymbol{X} - \boldsymbol{u} \not\leq \boldsymbol{x} - \boldsymbol{u} \mid \boldsymbol{X} \not\leq \boldsymbol{u}]\, \mathbb{P}[\boldsymbol{X} \not\leq \boldsymbol{u}] \\ &= \Big(1 - H(\boldsymbol{x} - \boldsymbol{u})\Big)\, \mathbb{P}[X_3 > u]. \end{aligned} \tag{5.6.3}$$

The first term of (5.6.3) can be written as

$$1 - H(\boldsymbol{x} - \boldsymbol{u}) = \ell\Big(1 - H_1(x_1 - u), \ldots, 1 - H_d(x_d - u)\Big),$$

and $1 - H_j(x_j - u)$ can be approximated using (5.2.4) for $j \in \{1, 2, 3\}$. Recall that the rainfall data we used to obtain the estimates in Table 5.3 comprises 102 years. If $n_{j,u}$ denotes the number of observed exceedances of component j over the threshold u, then the number of yearly exceedances follows a Poisson distribution with parameter $n_{j,u}/102$ and the probability of at least one exceedance in a given year is $1 - \exp(n_{j,u}/102)$, which we use to estimate $\mathbb{P}[X_j > u]$. Plugging in the parameter estimates $(\widehat{\lambda}_1, \widehat{\lambda}_2, \widehat{\lambda}_3, \widehat{\eta})$ from the top row of Table 5.3 gives

$$\mathbb{P}[X_1 > 39.5 \text{ or } X_2 > 56.6 \text{ or } X_3 > 69.9] \approx 0.0563. \tag{5.6.4}$$

We find that the probability of a landslide in any given year is 0.0563, which is very similar to the result in Rudvik (2012).

Goodness-of-fit For a visual test we consider QQ-plots for each of the univariate GP distributions given by $(X_{ij} - u) \mid X_{ij} > u$, where $u = 24$ as before. Figure 5.3 shows the QQ-plots for the restricted model with $\xi = 0$ (upper panels) and for the unrestricted model (lower panels). Both fits are less good for the first component, due to the restriction $\boldsymbol{\eta} = \eta\mathbf{1}$.

Let χ_{12}, χ_{13} and χ_{23} denote the pairwise tail dependence coefficients and χ_{123} the full tail dependence coefficient; see (1.3.6). We would like to compare empirical estimates of the tail dependence coefficient with model-based ones. The pairwise empirical estimates are given by expression (1.3.7); the full empirical estimate is computed in a similar way.

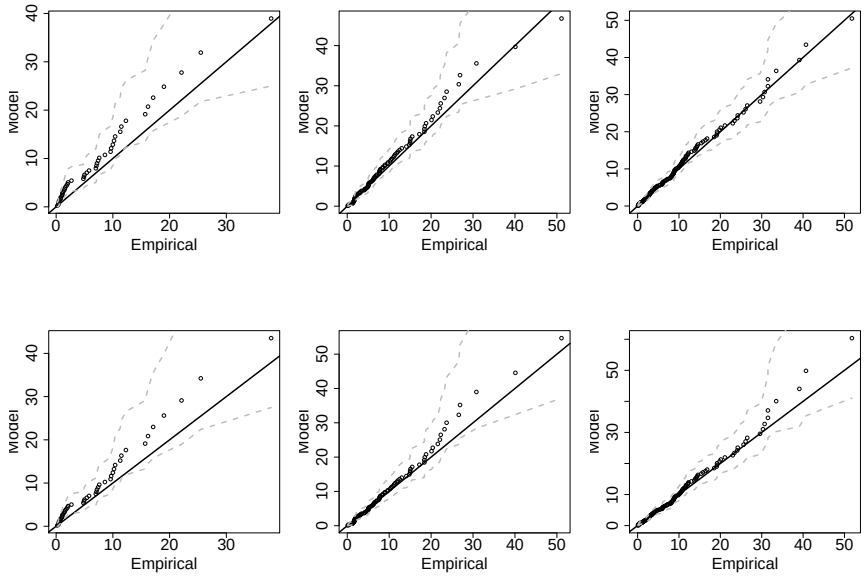


Figure 5.3: QQ-plots for the univariate GP distributions with parameters implied by Table 5.3, for the restricted model (upper panels) and the unrestricted model (lower panels).

Figure 5.4 shows the results, where the horizontal lines represent the model-based pairwise tail dependence coefficients for $u = 24$, i.e., we plugged in $\widehat{\lambda}_1$, $\widehat{\lambda}_2$ and $\widehat{\lambda}_3$ from the top row in Table 5.3 in

$$\chi_{12} = 1 - \frac{\lambda_1}{2(\lambda_1 + \lambda_2)},$$
$$\chi_{13} = 1 - \frac{\lambda_1(\lambda_2 + \lambda_3)^3}{(\lambda_3 + 2\lambda_2)(\lambda_2 + 2\lambda_3)(\lambda_2\lambda_3 + \lambda_1\lambda_3 + \lambda_1\lambda_2)},$$

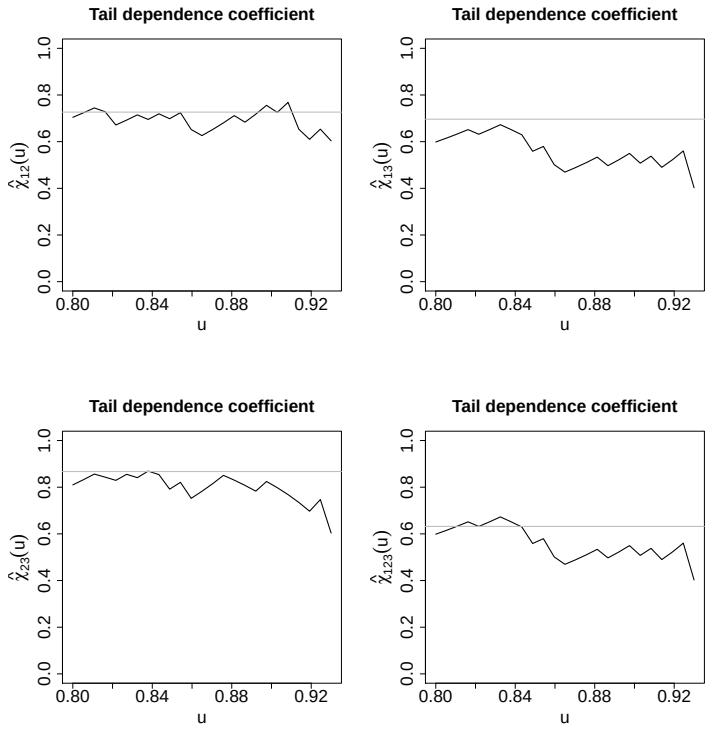


Figure 5.4: Pairwise and full empirical tail dependence coefficients for a range of thresholds. The horizontal lines are the model-based tail dependence coefficients for $u = 24$, with parameters implied by Table 5.3 for $\boldsymbol{\xi} = \mathbf{0}$.

$$\chi_{23} = 1 - \frac{\lambda_1\lambda_2(\lambda_1+\lambda_2)^2}{(\lambda_1+2\lambda_2)(\lambda_2+2\lambda_1)(\lambda_2\lambda_3+\lambda_1\lambda_3+\lambda_1\lambda_2)},$$
$$\chi_{123} = 1 - \frac{\lambda_1}{2(\lambda_1+\lambda_2)} - \frac{\lambda_1\lambda_2(4\lambda_1\lambda_2+\lambda_1\lambda_3+3\lambda_2^2+\lambda_2\lambda_3)}{3(2\lambda_1+\lambda_2)(2\lambda_2+\lambda_3)(\lambda_1\lambda_2+\lambda_1\lambda_3+\lambda_2\lambda_3)}.$$

Some properties of the structured components model can be inferred from the simple expression for χ_{12}; when $\lambda_1 = \lambda_2$, then $\chi_{12} = 0.75$ regardless of the value of λ_1. If $\lambda_1 \gg \lambda_2$, then $\chi_{12} \to 0.5$; if $\lambda_2 \gg \lambda_1$, then $\chi_{12} \to 1$. It is natural that the model cannot approach asymptotic independence, since it is based on cumulative sums.

Finally, we apply the goodness-of-fit test presented in Corollary 3.2.5. To this end, we need to define a list of points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0,\infty)^3$ for $q > 2$. We let $\boldsymbol{c}_m \in \{(1,1,0),(1,0,1),(0,1,1),(1,1,1)\}$ for $m \in \{1,2,3,4\}$. The test statistic from Corollary 3.2.5 then converges to a chi-square distribution with

$q - p = 2$ degrees of freedom; its 95% quantile is equal to 5.99. Computing the test statistic for $k \in \{50, 75, 100, 125, 150\}$, where we set again $\lambda_1 = 7.79$, we find the values 1.08, 4.48, 1.17, 5.42, and 0.99 respectively, so that we can not reject the structured components model for any value of k.

Using the parameter estimates obtained when calculating the test statistic for the goodness-of-fit test for $k = 142$ we obtain a yearly landslide probability of

$$\mathbb{P}[X_1 > 39.5 \text{ or } X_2 > 56.6 \text{ or } X_3 > 69.9] \approx 0.0564,$$

which is practically equal the result in (5.6.4).

Discussion

We fitted a three-dimensional structured components model to cumulative precipitation data. The fit of the marginals of this model could be further improved by removing the restriction $\boldsymbol{\eta} = \eta\mathbf{1}$. Unfortunately, it is not obvious how to do so: although Table 5.2 suggests that $\eta_1 < \eta_2 < \eta_3$, we can not fit such a model since it will destroy the ordering of the data components. However, we saw that even a model with the restrictions $\boldsymbol{\eta} = \eta\mathbf{1}$ and $\boldsymbol{\xi} = \mathbf{0}$ provides an adequate fit in terms of the dependence structure.

5.A Censored likelihoods

We describe the expressions needed to perform censored likelihood estimation for the models detailed in Section 5.4. For simplicity they are presented in standardized ($\boldsymbol{\eta} = \mathbf{1}$, $\boldsymbol{\xi} = \mathbf{0}$) form, i.e.,

$$h_C^*(\boldsymbol{z}_{D\setminus C}, \tilde{\boldsymbol{v}}_C) = \int_{\prod_{j\in C}(-\infty, \tilde{v}_j]} h^*(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}_C, \tag{5.A.1}$$

for $\boldsymbol{v} \leq \mathbf{0}$. The generalized form of a censored likelihood is then easily obtained from (5.A.1) as

$$h_C(\boldsymbol{x}_{D\setminus C}, \tilde{\boldsymbol{v}}_C) = \prod_{j\in C} \frac{1}{(\eta_j + \xi_j x_j)} h^* \left(\frac{\log(1 + \boldsymbol{\xi}\boldsymbol{x}_{D\setminus C}/\boldsymbol{\eta})}{\boldsymbol{\xi}}, \frac{\log(1 + \boldsymbol{\xi}\tilde{\boldsymbol{v}}_C/\boldsymbol{\eta})}{\boldsymbol{\xi}} \right).$$

The support for each density presented here is $\{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{z} \not\leq 0\}$.

Gumbel model For $\widetilde{\boldsymbol{S}}$, we obtain

$$h_C^*(\boldsymbol{z}_{D\setminus C}, \tilde{\boldsymbol{v}}_C) = \int_0^\infty r^{d-|C|-1} \prod_{j\in C} e^{-(re^{\tilde{v}_j - \lambda_j})^{-\alpha_j}} \prod_{j\in D\setminus C} \frac{\alpha_j}{e^{\lambda_j}} \left(\frac{re^{z_j}}{e^{\lambda_j}} \right)^{-\alpha_j - 1} \frac{e^{z_j}}{e^{(re^{z_j - \lambda_j})^{-\alpha_j}}} \, \mathrm{d}r.$$

If all α_j are equal to α:

$$h_C^*(\boldsymbol{z}_{D\setminus C}, \tilde{\boldsymbol{v}}_C) = \frac{\alpha^{d-|C|-1} e^{-\max_j z_j} \Gamma(d-|C|) \prod_{j\in D\setminus C} e^{-\alpha(z_j-\lambda_j)}}{\left(\sum_{j\in C}(e^{\tilde{v}_j-\lambda_j})^{-\alpha} + \sum_{j\in D\setminus C}(e^{z_j-\lambda_j})^{-\alpha}\right)^{d-|C|}}.$$

For $\boldsymbol{S}$, we obtain

$$h_C^*(\boldsymbol{z}_{D\setminus C}, \tilde{\boldsymbol{v}}_C) = \int_0^\infty r^{d-|C|} \prod_{j\in C} e^{-(re^{\tilde{v}_j-\lambda_j})^{-\alpha_j}} \prod_{j\in D\setminus C} \frac{\alpha_j}{e^{\lambda_j}} \left(\frac{re^{z_j}}{e^{\lambda_j}}\right)^{-\alpha_j-1} \frac{e^{z_j}}{e^{(re^{z_j-\lambda_j})^{-\alpha_j}}} \,\mathrm{d}r.$$

If all α_j are equal to α:

$$h_C^*(\boldsymbol{z}_{D\setminus C}, \tilde{\boldsymbol{v}}_C) = \frac{\alpha^{d-|C|-1}\Gamma(d-|C|-1/\alpha) \prod_{j\in D\setminus C} e^{-\alpha(z_j-\lambda_j)}}{\left(\sum_{j\in C}(e^{\tilde{v}_j-\lambda_j})^{-\alpha} + \sum_{j\in D\setminus C}(e^{z_j-\lambda_j})^{-\alpha}\right)^{d-|C|-1/\alpha}}.$$

Log-gamma model Let F_j, $j \in C$, denote the cumulative distribution function of a Gamma$(\alpha_j, 1)$ random variable. For $\widetilde{\boldsymbol{S}}$,

$$h_C^*(\boldsymbol{z}_{D\setminus C}, \tilde{\boldsymbol{v}}_C) = e^{-\max_{1\le j\le d} z_j} \prod_{j\in D\setminus C} \frac{e^{\alpha_j z_j}}{\Gamma(\alpha_j)} \times \prod_{j\in D\setminus C} \int_0^\infty r^{-1} \left(\prod_{j\in D\setminus C} r^{\alpha_j} e^{-re^{z_j}}\right) \left(\prod_{j\in C} F_j(re^{\tilde{v}_j})\right) \,\mathrm{d}r.$$

If C_d is as defined in (5.4.3), then for $\boldsymbol{S}$,

$$h_C(\boldsymbol{z}_{D\setminus C}, \tilde{\boldsymbol{v}}_C) = \frac{C_d^{-1}}{\Gamma\left(\sum_{j=1}^d \alpha_j + 1\right)} \prod_{j\in D\setminus C} e^{\alpha_j z_j} \times \prod_{j\in C} \Gamma(\alpha_j) \int_0^\infty \left(\prod_{j\in D\setminus C} r^{\alpha_j} e^{-re^{z_j}}\right) \left(\prod_{j\in C} F_j(re^{\tilde{v}_j})\right) \,\mathrm{d}r.$$

Structured components model Recall that since this is a model on $\boldsymbol{R}$, we need to differentiate between $\boldsymbol{\xi} = \boldsymbol{0}$ and $\boldsymbol{\xi} > \boldsymbol{0}$.

Case $\boldsymbol{\xi} = \boldsymbol{0}$. The censored likelihood has an analytical expression but is tedious to write down. We show the result for $\tilde{\boldsymbol{v}} = \tilde{v}\boldsymbol{1}$. Note that, since the density $h(\boldsymbol{z}; 0, 1)$ is non-zero only for $z_1 < \cdots < z_d$, we censor in $|C| = k$ components if

$$z_1 < \cdots < z_k < \tilde{v} < z_{k+1} < \cdots < z_d.$$

We show the result for $k = 1$; expressions for $k > 1$ follow naturally by repeated integration of this result. Then for $\mathbb{1}(\tilde{v} < z_2 < \cdots < x_d)$ and $\mathbb{1}(z_d > 0)$,

$$\begin{aligned} h_C(z_2, \ldots, z_d, \tilde{v}) &= \frac{d! \prod_{j=1}^d \lambda_j}{\sum_{j=1}^d \lambda_j^{-1}} \int_{-\infty}^{\tilde{v}} \frac{\prod_{j=1}^d e^{z_j}}{\left(\sum_{j=1}^d (\lambda_j - \lambda_{j+1}) e^{z_j}\right)^{d+1}} \, \mathrm{d}z_k \\ &= \frac{(d-1)! \lambda_1 \prod_{j=2}^d \lambda_j e^{z_j}}{(\lambda_1 - \lambda_2) \sum_{j=1}^d \lambda_j^{-1}} \left\{ \left(\sum_{j=2}^d (\lambda_j - \lambda_{j+1}) e^{z_j} \right)^{-d} \right. \\ &\qquad \left. - \left(\sum_{j=2}^d (\lambda_j - \lambda_{j+1}) e^{z_j} + (\lambda_1 - \lambda_2) e^{\tilde{v}} \right)^{-d} \right\} \end{aligned}$$

Case $\boldsymbol{\xi} > \mathbf{0}$. Similar to the above.

Chapter 6

Max-factor individual risk models with application to credit portfolios

Abstract

Individual risk models need to capture possible correlations as failing to do so typically results in an underestimation of extreme quantiles of the aggregate loss. Such dependence modelling is particularly important for managing credit risk, for instance, where joint defaults are a major cause of concern. Often, the dependence between the individual loss occurrence indicators is driven by a small number of unobservable factors. Conditional loss probabilities are then expressed as monotone functions of linear combinations of these hidden factors. However, combining the factors in a linear way allows for some compensation between them. Such diversification effects are not always desirable and this is why we propose a new model replacing linear combinations with maxima. These max-factor models give more insight into which of the factors is dominant. This chapter is based on Denuit, Kiriliouk and Segers (2015).

6.1 Introduction and motivation

Individual risks are often exposed to the same environment and this induces some dependence that leads to bias in calculations of stop-loss premiums and other risk measures. There are many situations in practice where dependence affects occurrences of losses. Typical cases arise for policies covering natural disasters (hurricane, tornado, flood, etc.). We refer the reader to the book of Denuit et al. (2005) for an introduction to the modelling of dependence and to the review paper of Anastasiadis and Chukova (2012) for an overview of the various multivariate insurance models suggested in the literature.

In this chapter, we model the occurrence of losses at the individual level. Recall that portfolios of risks are generally described by means of either a bottom-up approach or a top-down approach. In insurance, these two approaches are referred to as the individual and the collective models of risk theory. The bottom-up approach is also known as a name-per-name approach in the credit risk literature. It starts from a description of the individual risks from which the distribution of the aggregate loss is derived. The bottom-up approach has some clear advantages over the top-down approach, such as the possibility to easily account for heterogeneity.

In credit risk models, default indicators can in general not be considered as being mutually independent. Dependence between the defaults of different firms can be caused by direct links between them (e.g., one firm is the other's largest customer) or by more indirect links. In the latter category, we find industrial firms using the same resources, and thus exposed to the same price shocks, or selling on the same markets, and thus tributary of the same demand and subject to the same regulation.

A number of macroeconomic factors may influence many default indicators at once; examples include business cycles, level of unemployment, or shifts in monetary policy. To account for these situations, vectors of default indicators are often modelled via common mixture models. The idea is that there exists a limited number of systematic factors such that the default indicators are conditionally independent when the factors are controlled. Unconditionally, however, the default indicators are dependent because they are subject to the same unobservable macroeconomic factors. These factor models are among the few models that can replicate a realistic correlated default behavior while dramatically reducing the numerical complexity when computing the distribution of the aggregate portfolio loss.

In general, conditional default probabilities are functions of linear combinations of the hidden factors, with weights reflecting the relative sensitivity to the risk factor. This is the case for the majority of industry models, including the CreditRisk$^+$ and KMV models. We refer the reader to Bluhm et al. (2002) for a general introduction. The hidden factors are typically associated to different levels of the economy in a hierarchical way, accounting for global effects and sector-specific ones. Replacing linear combinations of hidden factors with maxima is attractive in some applications. The max-decomposition better accounts for shocks specific to a given category of risks, whereas linear combinations of factors tend to dilute the shock within the contributions of each factor to the sum.

The remainder of this chapter is organized as follows. In Section 6.2, we describe the proposed max-factor specification to induce dependence between loss indicators. Section 6.3 is devoted to calibration techniques. First, we describe the general set-up and introduce some parametric factor models. Second, we propose efficient numerical procedures to obtain the maximum likelihood estimates for max-factor models. In Section 6.4, new nonparametric estimators

are proposed that can be used as a benchmark to evaluate the goodness-of-fit of parametric risk models. A simulation study assessing the performances of the estimators is given in Appendix B, whereas formal proofs of the results proposed in Sections 6.3 and 6.4 can be found in Appendix A. In Section 6.5, we work out a detailed numerical illustration performed on a classical credit risk data set provided in Standard and Poor's (2001). Finally, Section 6.6 briefly discusses the results obtained in this chapter and concludes.

6.2 Max-factor risk model

Consider a portfolio of m risks split into k categories observed over a given reference period. Each category, r, contains m_r individual risks, $r = 1, \ldots, k$. The indicator $Y_{r,i}$ is equal to 1 if risk i from category r brings some financial loss and to 0 otherwise. The random variables $Y_{r,i}$ may be associated to a borrower's default in credit risk, to a policyholder's death in life insurance, or to the occurrence of a claim in general insurance, for instance. Henceforth, we refer to $Y_{r,i}$ as the loss (occurrence) indicator.

As individual contracts are subject to a common environment, loss indicators are impacted by a number of identical risk factors. The max-factor decomposition accounts for this positive correlation by means of a global risk factor Ψ_0 affecting all the m contracts and category-specific factors $\Psi_1, \ldots, \Psi_k$ whose influence is restricted to the contracts in the same class. The random variables $\Psi_0, \Psi_1, \ldots, \Psi_k$ are assumed to be independent with common distribution function F_Ψ. All the contracts in the same risk class r share the common random effect Ψ_r but are also subject to a competing global effect Ψ_0 affecting the entire block of business. In homeowners insurance, this global effect may be related to storms or earthquakes. In life insurance, it typically accounts for the sudden increase in death probabilities due to the occurrence of pandemics.

Write $\boldsymbol{\Psi} = (\Psi_0, \Psi_1, \ldots, \Psi_k)$. Whereas the majority of factor models are based on linear combinations of the hidden risk factors, here we specify a latent-shock or competing-risk mechanism. Specifically, the conditional loss probability $\mathbb{P}[Y_{r,i} = 1 \mid \boldsymbol{\Psi}]$ is expressed as an increasing function of the latent factor

$$\max\{\nu_r + \sigma_r \Psi_r, \mu_r + \sigma_r \Psi_0\}, \tag{6.2.1}$$

where the class-specific parameters satisfy $\nu_r, \mu_r \in \mathbb{R}$ and $\sigma_r \geq 0$. Then, the effect in (6.2.1) is mapped to the unit interval with the help of the distribution function F_Ψ, i.e.,

$$\mathbb{P}[Y_{r,i} = 1 \mid \boldsymbol{\Psi}] = F_\Psi\big(\max\{\nu_r + \sigma_r \Psi_r, \mu_r + \sigma_r \Psi_0\}\big). \tag{6.2.2}$$

There is thus a competition between the class-specific effect, $\nu_r + \sigma_r \Psi_r$, and the global effect, $\mu_r + \sigma_r \Psi_0$. Only the larger of the two has an impact on the occurrences of losses. The parameters ν_r and μ_r represent the sensitivity of the conditional loss probability to the class-specific factor Ψ_r and to the global risk

factor Ψ_0, respectively: the smaller μ_r, the less sensitive the loss indicators in category r to Ψ_0.

Natural candidates for F_Ψ are to be found among the max-stable distributions. Max-stability ensures that the distribution of the maximum in (6.2.1) stays in the same family. In this paper, we consider the Gumbel distribution, but a similar analysis can be carried out with any other max-stable family of distributions. Recall that the (standardized) distribution function of the Gumbel distribution is

$$F_\Psi(x) = \exp\big(-\exp(-x)\big), \qquad x \in \mathbb{R}.$$

The choice of the Gumbel distribution explains why we have chosen the latent shock to be of the form (6.2.1): the maximum in (6.2.1) is again a Gumbel distributed random variable, due to the fact that the multiplicative coefficient σ_r is equal for every element of $\mathbf{\Psi}$. The smaller the constants ν_r and μ_r, the less sensitive the contract is to the corresponding factor.

The model we propose is related to similar constructions suggested in the literature, but applied to different levels. For instance, in Denuit et al. (2002, Example 2.7) it is suggested, following Cossette et al. (2002), to represent the loss indicator $Y_{r,i}$ in terms of independent Bernoulli random variables $J_0, J_1, \ldots, J_k$ as

$$Y_{r,i} = \min(J_r + J_0, 1) = \max(J_0, J_r), \qquad i = 1, \ldots, n;$$

see also Valdez (2014). In credit risk modelling, time-to-defaults are sometimes assumed to be subject to a competing-risk mechanism (Giesecke, 2003). Default indicators are then of the form

$$Y_{r,i} = \mathbb{1}\big\{\min(E_r, E_0) \le 1\big\} = \max\big(\mathbb{1}\{E_r \le 1\}, \mathbb{1}\{E_0 \le 1\}\big),$$

where $E_0, E_1, \ldots, E_k$ are independent, positive random variables. The factor E_0 impacting all obligors accounts for a systematic shock threatening the solvency of the entire portfolio. In the model we propose, the max-factor decomposition affects the conditional loss probability and not the loss indicators directly. Contrarily to the two models described above, where the occurrence of the common shock (J_0 in the first case, or $\{E_0 \le 1\}$ in the second case) leads to the simultaneous occurrence of losses, the factors $\Psi_0, \ldots, \Psi_k$ only impact the conditional loss probabilities in (6.2.2). As long as this conditional probability stays below unity, there is still room for distinct individual default experiences. In this sense, the max-factor model appears to be more flexible.

The max-factor model can also be seen as a regime-switching construction, where the maximum drives the switch from standard to severe conditions. Think for instance of life insurance. The indicator $Y_{r,i}$ is now equal to 1 if individual i from risk class r dies during the year. Modern actuarial calculations recognize the uncertainty surrounding one-year death probabilities. The max-factor model can account for the occurrence of pandemics increasing the mortality of the population: Ψ_0 is related to the severity of the pandemics

and the parameters μ_r and σ_r modulate its consequences for the different risk categories (typically, flu pandemics can have different consequences depending on age category). There is thus a switch in the mortality regime, from standard to high.

Compared to the classical linear specification, the maximum in (6.2.1) prohibits any compensation between the global factor, Ψ_0, and the category-specific factors, $\Psi_1, \ldots, \Psi_k$. Indeed, the linear combination $\mu_r + \tau_r \Psi_r + \sigma_r \Psi_0$, where $\mu_r \in \mathbb{R}$, $\tau_r \geq 0$, $\sigma_r \geq 0$, allows for diversification between Ψ_0 and Ψ_r: a large realization for Ψ_0 can be compensated by a small realization for Ψ_r, leaving the corresponding linear combination unchanged. Assume for instance that the global economy is booming, so that Ψ_0 is small (default probabilities being increasing in the linear combination of risk factors). However, firms in some category r may experience severe problems because of new regulations, embargo, emerging new technologies, etc., so that Ψ_r may be large. The linear combination somewhat compensates the difficulties specific to category r with the excellent global conditions. In contrast, the max-factor specification (6.2.1) focuses on the worst factor, which is Ψ_r in our example, and recognizes the particular problems faced by the firms in category r. Another possible interpretation of the max-factor specification is that it measures the degree of "tolerance" an individual risk has in the presence of a common global shock. Depending on the kind of application, linear or max-factor decompositions may be considered to represent the correlation structure of the individual loss indicators.

6.3 Calibration of max-factor models

6.3.1 General setup

Assume that a portfolio of risks has been observed for n calendar years. Define the indicator variables $Y_{r,j,i}$, $r \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n\}$, $i \in \{1, \ldots, m_{r,j}\}$, where $Y_{r,j,i} = 1$ corresponds to the occurrence of losses for individual i in category r during calendar year j, while $m_{r,j}$ denotes the number of risks in category r and calendar year j. In the credit risk data that we will study in Section 6.5, the categories will correspond to the rating classes.

For fixed r and j, the number of risks producing losses is $M_{r,j} = \sum_{i=1}^{m_{r,j}} Y_{r,j,i}$. We assume that, within a category r, individual risks are exchangeable. More specifically, let $\boldsymbol{Q}_j = (Q_{1,j}, \ldots, Q_{k,j})$ be the conditional loss probabilities for calendar year j. Assume that $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ are independent and identically distributed. Given $\boldsymbol{Q}_j$, the random variables $Y_{r,j,i}$ are independent Bernoulli random variables with respective means $Q_{r,j}$, so that the conditional distribution of $M_{r,j}$ is given by

$$\mathbb{P}[M_{r,j} = l \mid \boldsymbol{Q}_j] = \binom{m_{r,j}}{l} Q_{r,j}^l \left(1 - Q_{r,j}\right)^{m_{r,j}-l}, \qquad l \in \{0, \ldots, m_{r,j}\}.$$

Conditionally on $\boldsymbol{Q}_j$, the numbers $M_{1,j}, \ldots, M_{k,j}$ of risks producing losses are independent and binomially distributed.

6.3.2 Quantities of interest

We are interested in the estimation of the following quantities:

Marginal loss probabilities. The probability that risk i in category r produces a loss during year j is given by

$$\pi_r = \mathbb{P}[Y_{r,j,i} = 1] = \mathbb{E}[Y_{r,j,i}] = \mathbb{E}[Q_{r,j}]. \tag{6.3.1}$$

Joint loss probabilities. The probability that two different risks i_1 and i_2 in the same or different categories r and s produce losses during the same year j is given by

$$\pi_{rs} = \mathbb{P}[Y_{r,j,i_1} = 1,\, Y_{s,j,i_2} = 1] = \mathbb{E}[Y_{r,j,i_1} Y_{s,j,i_2}] = \mathbb{E}[Q_{r,j} Q_{s,j}]. \tag{6.3.2}$$

Intra-class higher-order loss probabilities. The probability that $l \geq 1$ risks within the same category r produce losses during the same year j is equal to

$$\pi_r^{(l)} = \mathbb{P}[Y_{r,j,1} = \ldots = Y_{r,j,l} = 1] = \mathbb{E}[Q_{r,j}^l].$$

Clearly, $\pi_r^{(1)} = \pi_r$ and $\pi_r^{(2)} = \pi_{rr}$.

Inter-class higher-order joint loss probabilities. The probability that $l_1 \geq 1$ risks in category r and $l_2 \geq 1$ risks in category s, where $r \neq s$, produce losses during the same year j is given by

$$\begin{aligned}\pi_{rs}^{(l_1,l_2)} &= \mathbb{P}[Y_{r,j,1} = \ldots = Y_{r,j,l_1} = 1, Y_{s,j,1} = \ldots = Y_{s,j,l_2} = 1] \\ &= \mathbb{E}[Q_{r,j}^{l_1} Q_{s,j}^{l_2}].\end{aligned}$$

Clearly, $\pi_{rs}^{(1,1)} = \pi_{rs}$.

The higher-order (joint) loss probabilities are not of primary interest; they will appear in Section 6.4 where we will define nonparametric estimators for π_r and π_{rs}.

Dependence measures are easily expressed in terms of the probabilities defined above. For instance, the relative risk measure, or risk ratio, used in Valdez (2014) in motor insurance, can be written as

$$\frac{\mathbb{P}[Y_{r,j,i_1} = 1 | Y_{s,j,i_2} = 1]}{\mathbb{P}[Y_{r,j,i_1} = 1 | Y_{s,j,i_2} = 0]} = \frac{\pi_{rs}(1 - \pi_s)}{(\pi_r - \pi_{rs})\pi_s}.$$

Borrowed from medical studies, this quantity measures the tendency of one risk to induce another risk to produce losses. As pointed out in Valdez (2014), the linear correlation coefficient is less suitable as a measure of association between binary random variables. For more details, see e.g. Denuit and Lambert (2005).

6.3.3 Factor models

We assume a parametric model for the conditional default probabilities $\boldsymbol{Q}_j$ by setting $Q_{r,j} = Q_r(\boldsymbol{\Psi}_j; \boldsymbol{\theta})$, where $\boldsymbol{\Psi}_j = (\Psi_{1,j}, \ldots, \Psi_{p,j})$ with $p < m_{r,j}$ for $j \in \{1, \ldots, n\}$ are independent and identically distributed latent factors with some known distribution, $\boldsymbol{\theta}$ is the parameter vector, and $Q_r(\,\cdot\,; \boldsymbol{\theta})$ are functions from $\mathbb{R}^p$ to $[0, 1]$. To simplify the notation, we will usually omit the dependence on $\boldsymbol{\theta}$.

Formally, for fixed r and j, given p-dimensional vectors $\boldsymbol{\Psi}_j$ with $p < m_{r,j}$, $\boldsymbol{Y}_{r,j}$ follows a Bernoulli mixture model with factor vector $\boldsymbol{\Psi}_j$ if there exist functions $Q_r : \mathbb{R}^p \to [0, 1]$, $r \in \{1, \ldots, k\}$, such that given $\boldsymbol{\Psi}_j = \boldsymbol{\psi}_j$, $\boldsymbol{Y}_{r,j}$ is a vector of independent Bernoulli variables with $\mathbb{P}[Y_{r,j,i} = 1 \mid \boldsymbol{\Psi}_j = \boldsymbol{\psi}_j] = Q_r(\boldsymbol{\psi}_j)$ for $i \in \{1, \ldots, m_{r,j}\}$, where $\boldsymbol{\psi}_j = (\psi_{j,1}, \ldots, \psi_{j,p})$. Dependence between loss indicators is essentially dependence of conditional loss probabilities on a set of factors.

As described in Section 6.2, our focus is on a Gumbel max-factor model. For comparison, we consider factor models based on the normal distribution and a Gumbel one-factor model as well.

Model (1a) The one-factor Probit-normal specification assumes that for a year $j \in \{1, \ldots, n\}$

$$Q_r\left(\Psi_j\right) = \Phi(\mu_r + \sigma_r \Psi_j), \qquad \sigma_r > 0,\, r \in \{1, \ldots, k\},$$

where Φ denotes the cumulative distribution function of a standard normal random variable and the factors $\Psi_1, \ldots, \Psi_k$ are independent with common distribution function Φ for $j \in \{1, \ldots, n\}$. The model parameters are $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k)$. This classical model has been applied in Frey and McNeil (2003) to the same dataset appearing in Section 6.5.

Model (2a) A direct extension of model (1a) is

$$Q_r\left(\boldsymbol{\Psi}_j\right) = \Phi\left(\mu_r + \tau_r \Psi_{r,j} + \sigma_r \Psi_{0,j}\right), \;\; \sigma_r, \tau_r > 0,\, r \in \{1, \ldots, k\},$$

where the $k + 1$ components $\Psi_{0,j}, \Psi_{1,j}, \ldots, \Psi_{k,j}$ of $\boldsymbol{\Psi}_j$ are independent with common distribution function Φ. The model parameters are $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_k, \tau_1, \ldots, \tau_k, \sigma_1, \ldots, \sigma_k)$. If $\tau_r \to 0$ for every r, we retrieve model (1a).

Model (1b) The Gumbel one-factor can be defined as

$$Q_r\left(\Psi_j\right) = F_\Psi(\mu_r + \sigma_r \Psi_j), \qquad \sigma_r > 0,\, r \in \{1, \ldots, k\},$$

where the factors $\Psi_1, \ldots, \Psi_k$ are independent random variables with common distribution function $F_\Psi(x) = \exp\left(-\exp(-x)\right)$. The vector of model parameters is $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k)$.

Model (2b) For the Gumbel max-factor model, we take

$$Q_r\left(\boldsymbol{\Psi}_j\right) = F_\Psi\left(\max\left\{\nu_r + \sigma_r\Psi_{r,j}, \mu_r + \sigma_r\Psi_{0,j}\right\}\right), \qquad \sigma_r > 0,$$

where $\Psi_{0,j}, \Psi_{1,j}, \ldots, \Psi_{k,j}$ are independent and have a common distribution function F_Ψ. The model parameters are $\boldsymbol{\theta} = (\nu_1, \ldots, \nu_k, \mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k)$. If $\nu_r \to -\infty$ for every r, then we are back at model (1b).

Models (1a)-(1b) involve a single factor but differ in the right tails of the conditional loss probabilities $Q_r\left(\boldsymbol{\Psi}_j\right)$: the probability that these conditional probabilities exceed high thresholds is typically larger under the Gumbel specification compared to the Gaussian one. Considering models (2a)-(2b), a global effect $\Psi_{0,j}$ is now combined with category-specific effects $\Psi_{r,j}$. This gives more flexibility as conditional loss probabilities now become dependent, sharing the common random effect $\Psi_{0,j}$. It is worth mentioning that the interpretation of the parameters is different under models (2a) and (2b). In model (2a), the coefficients τ_r and σ_r multiplying the random effects measure the sensitivity of the individual risks in category r to $\Psi_{r,j}$ and $\Psi_{0,j}$, respectively, whereas these sensitivities are measured by the additive parameters ν_r and μ_r in model (2b).

As described in Section 6.3.2, we focus on the marginal loss probabilities π_r and the joint loss probabilities π_{rs}. If $F_{\boldsymbol{\Psi}}$ is the distribution function of a generic risk factor $\boldsymbol{\Psi}$, then these loss probabilities are obtained directly from (6.3.1) and (6.3.2) by

$$\pi_r = \mathbb{E}[Q_r(\boldsymbol{\Psi})] = \int Q_r(\boldsymbol{\psi})\,\mathrm{d}F_{\boldsymbol{\Psi}}(\boldsymbol{\psi}), \tag{6.3.3}$$

$$\pi_{rs} = \mathbb{E}[Q_r(\boldsymbol{\Psi})\,Q_s(\boldsymbol{\Psi})] = \int Q_r(\boldsymbol{\psi})\;Q_s(\boldsymbol{\psi})\,\mathrm{d}F_{\boldsymbol{\Psi}}(\boldsymbol{\psi}). \tag{6.3.4}$$

6.3.4 Likelihood

Let $F_{\boldsymbol{\Psi}}$ denote again the distribution function of a generic risk factor $\boldsymbol{\Psi}$. For category r and year j, the unconditional distribution of the number of risks producing losses is given by

$$\mathbb{P}[M_{r,j} = l_{r,j}] = \binom{m_{r,j}}{l_{r,j}} \int Q_r\left(\boldsymbol{\psi}_j\right)^{l_{r,j}} \left(1 - Q_r(\boldsymbol{\psi}_j)\right)^{m_{r,j}-l_{r,j}}\,\mathrm{d}F_{\boldsymbol{\Psi}}(\boldsymbol{\psi}_j).$$

We write $\boldsymbol{M}_j = (M_{1,j}, \ldots, M_{k,j})$ and $\boldsymbol{l}_j = (l_{1,j}, \ldots, l_{k,j})$ for $j = 1, \ldots, n$. Notice that since the loss indicators are independent given the vectors $\boldsymbol{\Psi}_j$, we can write

$$\mathbb{P}[\boldsymbol{M}_j = \boldsymbol{l}_j \mid \boldsymbol{\Psi}_j = \boldsymbol{\psi}_j] = \prod_{r=1}^{k} \binom{m_{r,j}}{l_{r,j}} Q_r(\boldsymbol{\psi}_j)^{l_{r,j}} \big(1 - Q_r(\boldsymbol{\psi}_j)\big)^{m_{r,j}-l_{r,j}}. \tag{6.3.5}$$

For every year we have expression (6.3.5) and the log-likelihood takes the form

$$L_n(\boldsymbol{\theta}; \boldsymbol{M}_1, \ldots, \boldsymbol{M}_n) = \sum_{j=1}^{n} \sum_{r=1}^{k} \log \binom{m_{r,j}}{M_{r,j}} + \sum_{j=1}^{n} \log I_j,$$

where

$$I_j = \int \prod_{r=1}^{k} Q_r(\boldsymbol{\psi}_j)^{M_{r,j}} \left(1 - Q_r(\boldsymbol{\psi}_j)\right)^{m_{r,j} - M_{r,j}} \mathrm{d}F_{\boldsymbol{\Psi}}(\boldsymbol{\psi}_j).$$

For the one-factor models (1a) and (1b), we find it convenient to make the substitution $q = F_\Psi(\psi_j)$ and to evaluate I_j as

$$I_j = \int_0^1 \exp\left(\sum_{r=1}^{k} M_{r,j} \log\left\{ Q_r(F_\Psi^{-1}(q)) \right\} + (m_{r,j} - M_{r,j}) \log\left\{ 1 - Q_r(F_\Psi^{-1}(q)) \right\} \right) \mathrm{d}q.$$

For models (2a) and (2b), we can make the substitutions $q_l = F_\Psi(\psi_{j,l})$ for $l \in \{0, \ldots, k\}$ since $\boldsymbol{\psi}_j = (\psi_{j,0}, \ldots, \psi_{j,k})$. Then, we can write the likelihood as

$$I_j = \int_{[0,1]^{k+1}} \left(\prod_{r=1}^{k} f_{r,j}(q_r, q_0) \right) \mathrm{d}q_0 \cdots \mathrm{d}q_k, \tag{6.3.6}$$

where

$$f_{r,j}(q_r, q_0) = \left\{ Q_r\left(F_\Psi^{-1}(q_r), F_\Psi^{-1}(q_0)\right) \right\}^{M_{r,j}} \times \left\{ 1 - Q_r\left(F_\Psi^{-1}(q_r), F_\Psi^{-1}(q_0)\right) \right\}^{m_{r,j} - M_{r,j}}. \tag{6.3.7}$$

Each likelihood term involves high-dimensional numerical integration over a complicated function. Especially for model (2b), when the integrand is a product of maxima, a nondifferentiable function, this is a computational burden. Fortunately, we can simplify the likelihood to a sum of lower-dimensional integrals over smoother functions thanks to the following result.

Lemma 6.3.1. *Define I_j and $f_{r,j}$ for $r = 1, \ldots, k$ and $j = 1, \ldots, n$ as in* (6.3.6) *and* (6.3.7)*, where Q_r is the function corresponding to the Gumbel max-factor model, model (2b). Define*

$$g_r(q) = \exp\left\{ \log(q) \exp\left(\frac{\nu_r - \mu_r}{\sigma_r} \right) \right\},$$

$$h_{r,j}(q; \mu_r) = F_\Psi\left(\mu_r - \sigma_r \log\left\{-\log(q)\right\}\right)^{M_{r,j}} \times \left[1 - F_\Psi(\mu_r - \sigma_r \log\left\{-\log(q)\right\})\right]^{m_{r,j} - M_{r,j}}.$$

Let $R = \{1, \ldots, k\}$ and let $\mathcal{P}(R)$ denote the power set of R. Then

$$I_j = \sum_{I \in \mathcal{P}(R)} \int_0^1 \left(\prod_{r \in R \setminus I} g_r(q_0)\, h_{r,j}(q_0; \mu_r) \right) \left(\prod_{r \in I} \int_{g_r(q_0)}^1 h_{r,j}(q_r; \nu_r)\, \mathrm{d}q_r \right) \mathrm{d}q_0.$$

The proof of this result is provided in Appendix 6.A.

The parameter vector $\boldsymbol{\theta}$ is estimated by maximizing the log-likelihood $L_n(\boldsymbol{\theta})$. After estimating $\boldsymbol{\theta}$, the implied marginal and joint loss probabilities, (6.3.1) and (6.3.2), are obtained by plugging in the estimator of $\boldsymbol{\theta}$ in expressions (6.3.3) and (6.3.4), yielding $\widehat{\pi}_r$ and $\widehat{\pi}_{rs}$, respectively.

6.4 Nonparametric estimation

Nonparametric estimators can be useful as a benchmark for model-based estimators, especially in the case of model uncertainty. Usually, the nonparametric estimators presented in Section 6.4.1 are used, see for example Frey and McNeil (2003). However, since the numbers $m_{r,j}$ may vary strongly over the years, more accurate nonparametric estimators are obtained by assigning more weight to those years for which there is more information (Section 6.4.2).

6.4.1 Preliminary estimators

Define the observed proportions of risks producing losses as

$$\widehat{Q}_{r,j} = M_{r,j}/m_{r,j}, \qquad \text{for } r \in \{1, \ldots, k\},\ j \in \{1, \ldots, n\}.$$

For $r \neq s$, define the estimators

$$\widehat{\pi}_r^{(l)} = \frac{1}{n} \sum_{j=1}^{n} \frac{M_{r,j}(M_{r,j}-1)\cdots(M_{r,j}-l+1)}{m_{r,j}(m_{r,j}-1)\cdots(m_{r,j}-l+1)}, \tag{6.4.1}$$

$$\widehat{\pi}_{rs}^{(l_1,l_2)} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{M_{r,j}(M_{r,j}-1)\cdots(M_{r,j}-l_1+1)}{m_{r,j}(m_{r,j}-1)\cdots(m_{r,j}-l_1+1)} \right. \tag{6.4.2}$$

$$\left. \times \frac{M_{s,j}(M_{s,j}-1)\cdots(M_{s,j}-l_2+1)}{m_{s,j}(m_{s,j}-1)\cdots(m_{s,j}-l_2+1)} \right), \tag{6.4.3}$$

for $l < m_{r,j}$, $l_1 < m_{r,j}$ and $l_2 < m_{s,j}$. To see that (6.4.1) and (6.4.3) are unbiased estimators, recall that if the random variable M is binomially distributed with n trials and success probability p, then we have for $l \in \{1, \ldots, n\}$ that

$$\mathbb{E}[M(M-1)\cdots(M-l+1)] = n(n-1)\cdots(n-l+1)\, p^l. \tag{6.4.4}$$

Conditionally on $\boldsymbol{Q}_j$, the random variable $M_{r,j}$ follows the binomial distribution with $m_{r,j}$ trials and success probability $Q_{r,j}$. Hence, for $l < m_{r,j}$,

$$\begin{aligned}\mathbb{E}[\widehat{\pi}_r^{(l)}] &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\frac{M_{r,j}(M_{r,j}-1)\cdots(M_{r,j}-l+1)}{m_{r,j}(m_{r,j}-1)\cdots(m_{r,j}-l+1)}\,\middle|\,\boldsymbol{Q}_j\right]\right] \\ &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[Q_{r,j}^l] = \pi_r^{(l)},\end{aligned}$$

and similarly, $\mathbb{E}[\widehat{\pi}_{rs}^{(l_1,l_2)}] = \pi_{rs}^{(l_1,l_2)}$.

6.4.2 Weighted estimators

For the marginal loss probabilities π_r, consider estimators of the form

$$\widetilde{\pi}_r(\boldsymbol{w}_r) = \sum_{j=1}^{n} w_{r,j}\widehat{Q}_{r,j}, \qquad r \in \{1,\dots,k\},$$

where $\widehat{Q}_{r,1},\dots,\widehat{Q}_{r,n}$ have a common expectation $\mathbb{E}[\widehat{Q}_{r,j}] = \pi_r$ and possibly different variances $\mathrm{Var}[\widehat{Q}_{r,j}] = \sigma_{r,j}^2$. The weight vector $\boldsymbol{w}_r = (w_{r,1},\dots,w_{r,n})$ has non-negative entries. We seek optimal weights, in the sense that we minimize the mean squared error of $\widetilde{\pi}_r(\boldsymbol{w}_r)$ as a function of $\boldsymbol{w}_r$, leading to weights $w_{r,1,\mathrm{opt}},\dots,w_{r,n,\mathrm{opt}}$. The same recipe can be followed for the joint loss probabilities π_{rr} and π_{rs}.

Theorem 6.4.1. *The estimators* $(\pi_{r,\mathrm{opt}}, \pi_{rr,\mathrm{opt}}, \pi_{rs,\mathrm{opt}})$ *for* $r, s, \in \{1,\dots,k\}$ *and* $r \neq s$ *that minimize the mean squared error of* $(\widetilde{\pi}_r(\boldsymbol{w}_r), \widetilde{\pi}_{rr}(\boldsymbol{w}_r), \widetilde{\pi}_{rs}(\boldsymbol{w}_r))$ *are*

1. $$\pi_{r,\mathrm{opt}} = \sum_{j=1}^{n} w_{r,j,\mathrm{opt}}\frac{M_{r,j}}{m_{r,j}}, \quad w_{r,j,\mathrm{opt}} = \frac{\sigma_{r,j}^{-2}}{\pi_r^{-2} + \sum_{t=1}^{n}\sigma_{r,t}^{-2}},$$
$$\sigma_{r,j}^2 = \frac{\pi_r}{m_{r,j}} + \left(1 - \frac{1}{m_{r,j}}\right)\pi_{rr} - \pi_r^2;$$

2. $$\pi_{rr,\mathrm{opt}} = \sum_{j=1}^{n} w_{rr,j,\mathrm{opt}}\frac{M_{r,j}(M_{r,j}-1)}{m_{r,j}(m_{r,j}-1)}, \quad w_{rr,j,\mathrm{opt}} = \frac{\sigma_{rr,j}^{-2}}{\pi_r^{-2} + \sum_{t=1}^{n}\sigma_{rr,t}^{-2}},$$
$$\begin{aligned}\sigma_{rr,j}^2 = m_{r,j}(m_{r,j}-1)\Big\{&(2 - m_{r,j}(m_{r,j}-1)\pi_{rr})\pi_{rr} \\ &+ 4(m_{r,j}-2)\pi_r^{(3)} + (m_{r,j}-2)(m_{r,j}-3)\pi_r^{(4)}\Big\};\end{aligned}$$

3. $$\pi_{rs,\text{opt}} = \sum_{j=1}^{n} w_{rs,j,\text{opt}} \frac{M_{r,j}M_{s,j}}{m_{r,j}m_{s,j}}, \quad w_{rs,j,\text{opt}} = \frac{\sigma_{rs,j}^{-2}}{\pi_r^{-2} + \sum_{t=1}^{n} \sigma_{rs,t}^{-2}},$$
$$\begin{aligned}\sigma_{rs,j}^2 = m_{r,j}^{-1}m_{s,j}^{-1}\Big\{&(1 - m_{r,j}m_{s,j}\pi_{rs})\pi_{rs} \\ &+ (m_{s,j} - 1)\pi_{rs}^{(1,2)} + (m_{r,j} - 1)\pi_{rs}^{(2,1)} \\ &+ (1 - m_{s,j} - m_{r,j} + m_{r,j}m_{s,j}\pi_{rs}^{(2,2)})\Big\}.\end{aligned}$$

The variances of these estimators are given by

$$\begin{aligned}\text{Var}[\pi_{r,\text{opt}}] &= \frac{1}{\sum_{j=1}^{n} \sigma_{r,j}^{-2}}, \\ \text{Var}[\pi_{rr,\text{opt}}] &= \frac{1}{\sum_{j=1}^{n} \sigma_{rr,j}^{-2}}, \quad \text{Var}[\pi_{rs,\text{opt}}] = \frac{1}{\sum_{j=1}^{n} \sigma_{rs,j}^{-2}}.\end{aligned}$$

The proof of this result is provided in Appendix 6.A. As the quantities $\sigma_{r,j}$, $\sigma_{rr,j}$, and $\sigma_{rs,j}$ depend on the unknown quantities $\pi_r^{(l)}$ and $\pi_{rs}^{(l_1,l_2)}$, we replace these with their preliminary estimators $\widehat{\pi}_r^{(l)}$ and $\widehat{\pi}_{rs}^{(l_1,l_2)}$ from (6.4.1) and (6.4.3) respectively.

Definition 6.4.1. Let $\widehat{\pi}_r^{(l)}$ and $\widehat{\pi}_{rs}^{(l_1,l_2)}$ be defined as in (6.4.1) and (6.4.3). The estimators $(\widehat{\pi}_{r,\text{opt}}, \widehat{\pi}_{rr,\text{opt}}, \widehat{\pi}_{rs,\text{opt}})$ of $(\pi_r, \pi_{rr}, \pi_{rs})$ for $r, s, \in \{1, \ldots, k\}$ and $r \neq s$ are defined as

1. $$\widehat{\pi}_{r,\text{opt}} = \sum_{j=1}^{n} \widehat{w}_{r,j,\text{opt}} \frac{M_{r,j}}{m_{r,j}}, \quad \widehat{w}_{r,j,\text{opt}} = \frac{\widehat{\sigma}_{r,j}^{-2}}{\widehat{\pi}_r^{-2} + \sum_{t=1}^{n} \widehat{\sigma}_{r,t}^{-2}},$$
$$\widehat{\sigma}_{r,j}^2 = \frac{\widehat{\pi}_r}{m_{r,j}} + \left(1 - \frac{1}{m_{r,j}}\right)\widehat{\pi}_{rr} - \widehat{\pi}_r^2;$$

2. $$\widehat{\pi}_{rr,\text{opt}} = \sum_{j=1}^{n} \widehat{w}_{rr,j,\text{opt}} \frac{M_{r,j}(M_{r,j} - 1)}{m_{r,j}(m_{r,j} - 1)}, \quad \widehat{w}_{rr,j,\text{opt}} = \frac{\widehat{\sigma}_{rr,j}^{-2}}{\widehat{\pi}_r^{-2} + \sum_{t=1}^{n} \widehat{\sigma}_{rr,t}^{-2}},$$
$$\begin{aligned}\widehat{\sigma}_{rr,j}^2 = m_{r,j}(m_{r,j} - 1)\Big\{&(2 - m_{r,j}(m_{r,j} - 1)\widehat{\pi}_{rr})\widehat{\pi}_{rr} \\ &+ 4(m_{r,j} - 2)\widehat{\pi}_r^{(3)} + (m_{r,j} - 2)(m_{r,j} - 3)\widehat{\pi}_r^{(4)}\Big\};\end{aligned}$$

3. $$\widehat{\pi}_{rs,\text{opt}} = \sum_{j=1}^{n} \widehat{w}_{rs,j,\text{opt}} \frac{M_{r,j}M_{s,j}}{m_{r,j}m_{s,j}}, \quad \widehat{w}_{rs,j,\text{opt}} = \frac{\widehat{\sigma}_{rs,j}^{-2}}{\widehat{\pi}_r^{-2} + \sum_{t=1}^{n} \widehat{\sigma}_{rs,t}^{-2}},$$
$$\begin{aligned}\widehat{\sigma}_{rs,j}^2 = m_{r,j}^{-1}m_{s,j}^{-1}\Big\{&(1 - m_{r,j}m_{s,j}\widehat{\pi}_{rs})\widehat{\pi}_{rs} \\ &+ (m_{s,j} - 1)\widehat{\pi}_{rs}^{(1,2)} + (m_{r,j} - 1)\widehat{\pi}_{rs}^{(2,1)} \\ &+ (1 - m_{s,j} - m_{r,j} + m_{r,j}m_{s,j}\widehat{\pi}_{rs}^{(2,2)})\Big\}.\end{aligned}$$

Approximate standard errors of these estimators can be obtained by plugging in the estimators of $\sigma^2_{r,j}$, $\sigma^2_{rr,j}$ and $\sigma^2_{rs,j}$ into the variances obtaind in Theorem 6.4.1.

A simulation study illustrating the performance of these estimators is provided in Appendix 6.B. We found that the relative root mean squared error (RRMSE) of the estimators of π_r and π_{rs} is significantly lower for the weighted estimators than for the preliminary estimators. Moreover, the weighted nonparametric estimators have low RRMSE in comparison with the maximum likelihood estimators of Section 6.3.4, especially when the parametric model is misspecified, that is, when we maximize the likelihood of the parameters of a factor model that is different from the true underlying model.

6.5 Application to credit risk

6.5.1 Credit risk data

We study one-year default rates for groups of obligors formed into static pools (cohorts). The default rates are taken from Table 13 in Standard and Poor's (2001), where the period of study is 1981–2000. The total data comprises around 9200 obligors rated as of January 1st, 1981, or first rated between that date and December 31st, 1999. A company is considered defaulted on the date when it is unable to fulfill a payment or any other financial obligation for the first time. Companies are given credit ratings ranging from AAA to CCC. We consider here the ratings BB, B, and CCC which form the group "speculative grade", since for higher-rated classes the data contain a too small amount of defaults to do meaningful inference. The starting year, 1981, does not include companies that defaulted in that year. Since it contains zero defaults by construction, we removed that year from our study.

Few obligors default early in their rating history. If default rates are obtained by dividing the number of defaults by all outstanding ratings, then consequently the default rates will be comparatively low during periods of high rating activity. To avoid any misleading results, the data are presented for cohorts called static pools. A static pool is formed on the first day of each year, and includes all companies in the study. The pools are called static because their membership remains constant over time. The obligors are followed from year to year within each pool. The ratings of the first and last days of each year are compared. Companies that default (D) or whose ratings have been withdrawn (N.R., not rated) are excluded from subsequent pools. For instance, we start with all companies that had outstanding non-defaulted ratings on January 1st, 1981. The 1982 static pool consisted of all companies that survived 1981 plus all companies that were first rated in 1981. In the scope of our time period, 9169 first-time rated organizations were added to the static pools, 746 companies defaulted and 3118 companies were excluded due to a N.R. rating. A company usually obtains a N.R. rating due to paid-off debt, a result of mer-

gers and acquisitions, or a lack of cooperation with the rating agency. Figure 6.1 shows the total number of firms per rating class, the number of defaulted firms per rating class and the proportion of defaults per rating class.

6.5.2 Nonparametric estimation

Table 6.1 displays the estimators $\widehat{\pi}_{r,\text{opt}}$ and $\widehat{\pi}_{rs,\text{opt}}$ from Definition 6.4.1, where $r, s \in \{\text{BB}, \text{B}, \text{CCC}\}$. The standard errors are in parentheses. These values serve as benchmarks to evaluate the accuracy of the parametric factor models fitted in the next section.

	BB	B	CCC
$\widehat{\pi}_{r,\text{opt}}$	0.0107 (0.0024)	0.0511 (0.0064)	0.2069 (0.0225)
$\widehat{\pi}_{rs,\text{opt}} \times 1000$	BB	B	CCC
BB	0.151 (0.081)	0.649 (0.206)	2.438 (0.682)
B	0.649 (0.206)	3.075 (0.935)	11.64 (2.438)
CCC	2.438 (0.692)	11.64 (2.438)	49.02 (8.887)

Table 6.1: Standard and Poor's (2001) data: nonparametric estimators of the marginal default probabilities π_r and the joint default probabilities π_{rs}.

6.5.3 Parametric factor models

We estimated the parameters of the parametric factor models (1a), (2a), (1b), and (2b). Table 6.2 shows the AIC and BIC for these models. Note that the number of parameters for models (2a) and (2b) is reduced:

- for model (2a), we find $\tau_B \to 0$, i.e. class B does not require a specific random effect and is influenced by the global effect only;
- for model (2b), we find $\nu_B, \nu_{CCC} \to -\infty$, i.e. classes B and CCC do not require specific random effects and are influenced by the global effect only.

In terms of both AIC and BIC, the one-factor models (1a) and (1b) perform better than a multi-factor normal model (model 2a). However, the Gumbel max-factor model (2b) appears to be the best alternative. Between models (1b) and (2b) we can also perform a likelihood ratio test, since model (1b) is a submodel of model (2b) with $\nu_{BB} \to -\infty$. The value of the likelihood ratio test statistic is equal to 2.76. The hypothesis for $\nu_{BB} = -\infty$ concerns a value at the boundary of the parameter space. The asymptotic null distribution of the likelihood ratio test statistic $2\log(L_n)$ is a mixture of two chi-squared distributions; see Self and Liang (1987). We reject the null hypothesis of the

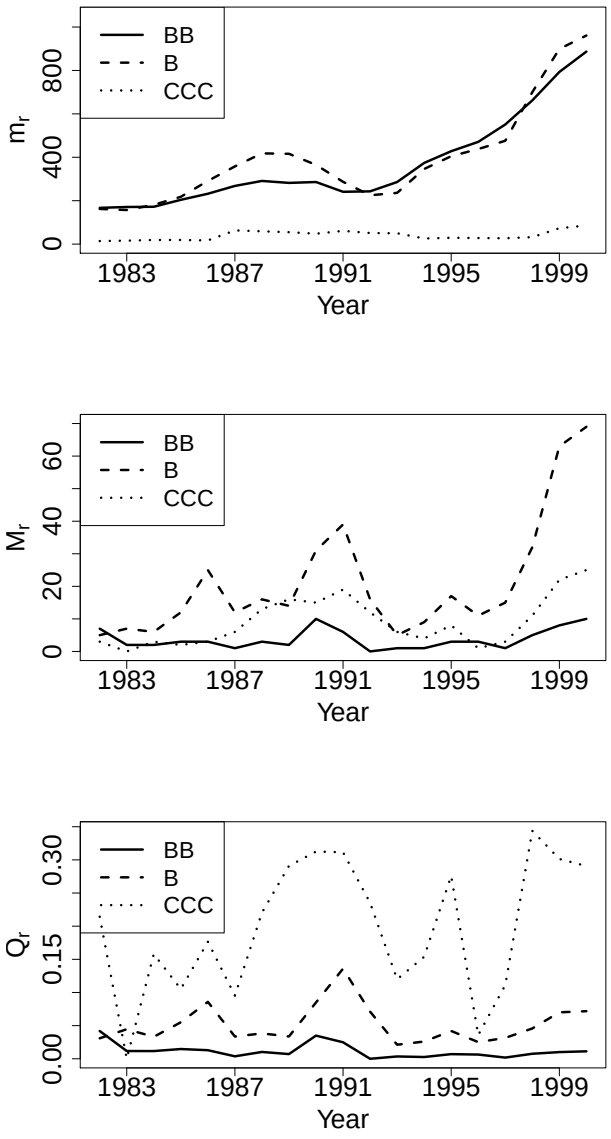


Figure 6.1: Standard and Poor's (2001) data: the total number of firms per rating class m_r (top), the number of defaulted firms per rating class M_r (middle) and the proportion of defaults per rating class $\widehat{Q}_r = M_r/m_r$ for the years 1982–2000, where $r \in \{\text{BB}, \text{B}, \text{CCC}\}$.

Model	# parameters	$-\log L_n$	AIC	BIC
(1a)	6	154.707	321.41	316.05
(2a)	8	154.445	324.89	317.74
(1b)	6	154.517	321.03	315.67
(2b)	7	153.138	320.28	314.02

Table 6.2: Standard and Poor's (2001) data: overview of the number of parameters, the negative log-likelihood, AIC, and BIC for the four parametric models presented in Subsection 6.3.3

.

one-factor model at a significance level of $\alpha = 0.05$, corresponding to a critical value of 1.92.

Table 6.3 shows estimates and standard errors for the parameters of model (2b), together with implied estimates of the marginal default probabilites π_r and the joint default probabilities π_{rs}, obtained using expressions (6.3.3) and (6.3.4). Both $\widehat{\pi}_r$ and $\widehat{\pi}_{rs}$ match the nonparametric ones reasonably well.

A visual test is presented in the form of prediction intervals for the numbers of defaults, $M_{r,j}$. We simulate 5000 realizations of Q_{BB}, Q_B, Q_{CCC}, accounting for the correlation structure, i.e., we simulate 5000 realizations of model (2b) using the parameter values that we obtained for the corresponding model. Using Q_{BB}, Q_B, Q_{CCC} we simulate 5000 realizations of $M_{r,j}$, where $m_{r,j}$ is given by the credit risk data. Finally, we calculate the prediction intervals, obtained by isolating the 4500 central realizations. The results are presented in Figure 6.2. The observed number of defaults generally stays within the prediction intervals; departures are in line with the 90% confidence level. These intervals provide the risk manager with useful ranges for the number of defaults.

It is also interesting to compare the distribution function of the conditional default probabilities $Q_r(\mathbf{\Psi})$ for models (1a)–(1b) to models (2a)–(2b). Figure 6.3 shows the survival functions of $Q_r(\mathbf{\Psi})$ for $r \in \{\text{BB}, \text{B}, \text{CCC}\}$. The fatness of the right tail of $Q_r(\mathbf{\Psi})$ greatly distinguishes the Gumbel models from the normal ones, even if all distribution functions agree to a large extent around the mean value. The impact of replacing the traditional normally distributed latent factor with a Gumbel one is clearly visible for high quantiles.

6.5.4 Value-at-Risk and Expected Shortfall

The estimated models in the previous section can be used to obtain estimates for risk measures such as Value-at-Risk or Expected Shortfall. Similarly to Frey and McNeil (2003), we proceed as follows. Consider a portfolio of 1000 obligors where the numbers of BB, B, and CCC-rated firms are 450, 500 and 50 respectively; these proportions correspond roughly to the numbers in the Standard and Poor's data. Let $m = 1000$ denote the total number of obligors

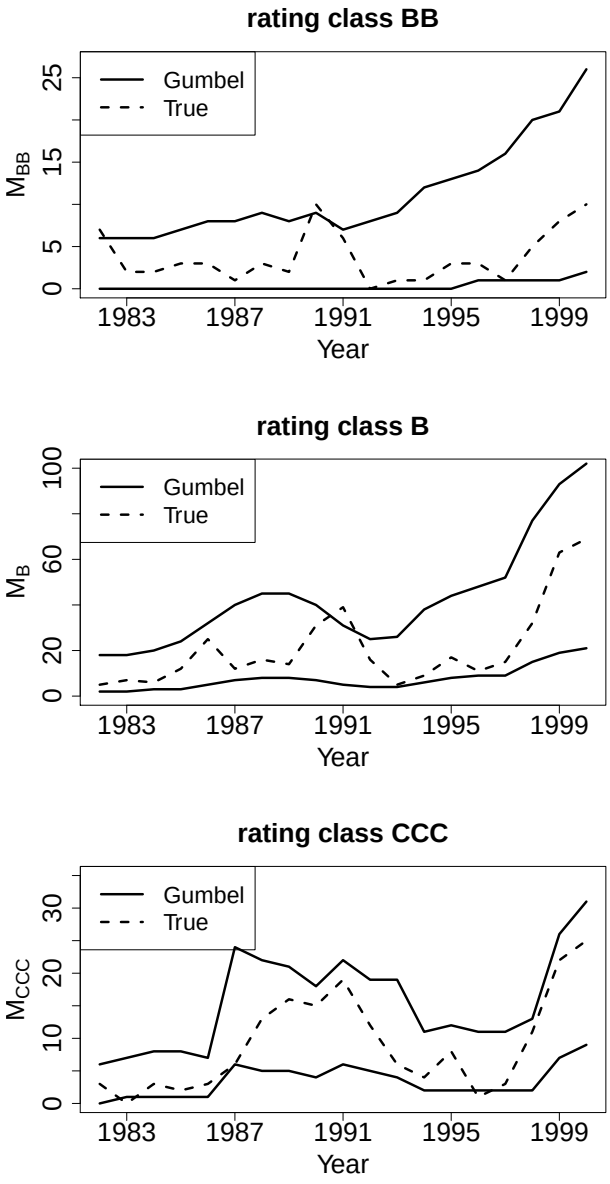


Figure 6.2: Standard and Poor's (2001) data: prediction intervals for the number of defaults obtained by simulating 5000 default matrices $M_{r,j}$ from Q_{BB}, Q_B, Q_{CCC} and isolating the 4500 central observations for the Gumbel max-factor model. The dashed lines show the actual number of defaults.

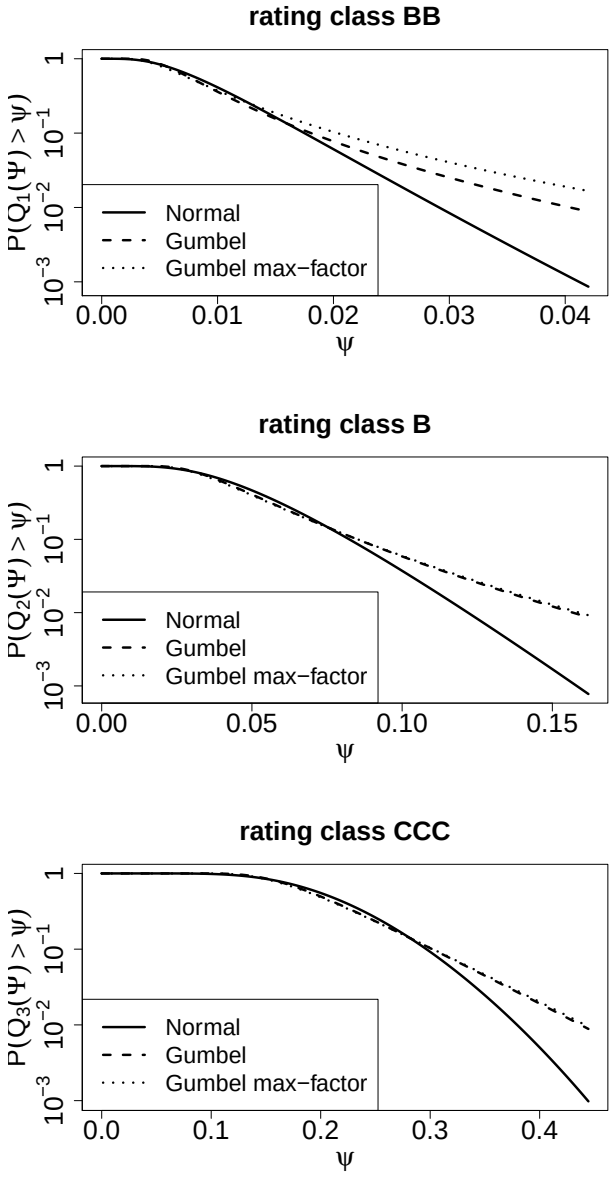


Figure 6.3: Standard and Poor's (2001) data: excess probabilities for conditional default probabilities $Q_r(\mathbf{\Psi})$ for $r \in \{\mathrm{BB}, \mathrm{B}, \mathrm{CCC}\}$.

	BB	B	CCC
μ_r	-1.66 (0.07)	-1.18 (0.04)	-0.54 (0.07)
ν_r	-1.73 (0.11)	—	—
σ_r	0.112 (0.033)	0.124 (0.029)	0.162 (0.053)
$\widehat{\pi}_r$	0.0109	0.0520	0.2120
$\widehat{\pi}_{rs} \times 1000$	BB	B	CCC
BB	0.215	0.781	2.795
B	0.781	3.512	12.96
CCC	2.795	12.96	49.73

Table 6.3: Standard and Poor's (2001) data: maximum likelihood parameter estimates and standard errors for the Gumbel max-factor model (2b), together with the implied estimates of default probabilities.

and let $m_r \in \{450, 500, 50\}$ denote the number of obligors per rating class. Let $Y_{r,i} = 1$ correspond to the default of obligor i in rating class r. We are interested in computing the distribution of the overall loss L,

$$L = \sum_{r=1}^{k} \sum_{i=1}^{m_r} e_{r,i} \Delta_{r,i} Y_{r,i},$$

where $e_{r,i}$ denotes the overall exposure of company i in rating class r and $\Delta_{r,i}$ is the random proportion of the exposure which is lost in case of default. We will assume $e_{r,i} = \Delta_{r,i} = 1$ for all i and r, leading to

$$L = \sum_{r=1}^{k} \sum_{i=1}^{m_r} Y_{r,i} = \sum_{r=1}^{k} M_r = M,$$

where M_r represents the number of defaults in rating class r and M represents the total number of defaults. For $k = 3$ we have

$$\begin{aligned}
\mathbb{P}[M = y] &= \mathbb{E}\left[\mathbb{P}\left[M_1 + M_2 + M_3 = y \mid \mathbf{\Psi}\right]\right] \\
&= \sum_{x_1=l_1}^{u_1} \sum_{x_2=l_2}^{u_2} \Big\{ \mathbb{E}[\mathbb{P}[M_1 = x_1 \mid \mathbf{\Psi}]\, \mathbb{P}[M_2 = x_2 \mid \mathbf{\Psi}] \\
&\qquad \times \mathbb{P}[M_3 = y - x_1 - x_2 \mid \mathbf{\Psi}]] \Big\},
\end{aligned}$$

where $l_1 = \max\{0, y - m_2 - m_3\}$, $u_1 = \min\{y, m_1\}$, $l_2 = \max\{0, y - x_1 - m_3\}$ and $u_2 = \min\{y - x_1, m_2\}$. Moreover,

$$\mathbb{P}[M_r = x_r \mid \mathbf{\Psi}] = \binom{m_r}{x_r} Q_r(\mathbf{\Psi})^{x_r} (1 - Q_r(\mathbf{\Psi}))^{m_r - x_r}. \tag{6.5.1}$$

The expected value is then computed by integrating over the distribution of $\mathbf{\Psi}$. We calculate the Value-at-Risk and the Expected Shortfall, i.e.,

$$\begin{aligned}\mathrm{VaR}_\alpha(M) &= \min\{x : \mathbb{P}[M \leq x] \geq \alpha\},\\ \mathrm{ES}_\alpha(M) &= \frac{1}{\mathbb{P}[M > \mathrm{VaR}_\alpha]} \sum_{x=\mathrm{VaR}_\alpha+1}^{m} x\,\mathbb{P}[M = x],\end{aligned}$$

for $\alpha = 0.99$. This is done by plugging in the parameter estimates obtained in Section 5.3 into (6.5.1), for the one-factor normal model and the max-factor Gumbel model. We obtain $\mathrm{VaR}_\alpha(M) = 96$ and $\mathrm{ES}_\alpha(M) = 109$ for the normal model and $\mathrm{VaR}_\alpha(M) = 124$ and $\mathrm{ES}_\alpha(M) = 155$ for the Gumbel model. The financial impact of the max-factor model thus appears to be considerable, as measured by the increase in VaR and ES when moving from the traditional normal setting to the max-Gumbel one.

6.6 Discussion

In this paper, we have proposed max-factor models to account for dependencies between individual loss occurrence indicators. Compared to the more traditional approach where the correlation is induced by linear combinations of random effects, the max-factor specification prohibits diversification or compensation between hidden factors as only the largest effect controls individual risk levels. The max-factor specification appears to be particularly appealing to model the occurrence of shocks affecting policies in the portfolio, as well as the effect of common economic conditions. Compared to previous literature, these shocks increase the conditional loss probability without systematically inducing losses on all the contracts.

This new model produces a good fit on the Standard and Poor's (2001) credit risk data set. Besides classical goodness-of-fit measures based on the log-likelihood (such as AIC and BIC), we have also proposed novel nonparametric estimators, minimizing the mean squared error, that can be used as a benchmark to evaluate the relative merits of the different models.

Future work might include a dynamic component or extend the methodology to take into account both latent and observable factors. However, the Standard and Poor's data studied in this paper are not adequate to tackle such issues, since they do not contain any information on the companies included nor on their transitions between rating classes.

The max-factor decomposition may also be interesting for credibility models decomposing the individual unobservable risk proneness in a hierarchical way. Again, this approach is desirable in situations where no compensation is possible between the random effects associated to the different levels but the worst case drives the individual risk proneness. We leave this topic for a future investigation.

6.A Proofs

In order to establish the validity of Theorem 6.4.1, we will need expressions for some moments of the random variables $M_{r,j}$.

Lemma 6.A.1. *For $r \in \{1, \ldots, k\}$, we have*

1. $\mathbb{E}[M_{r,j}] = m_{r,j}\pi_r$;
2. $\mathbb{E}[M_{r,j}(M_{r,j}-1)] = m_{r,j}(m_{r,j}-1)\pi_{rr}$;
3. $\mathrm{Var}[M_{r,j}] = m_{r,j}(\pi_r + (m_{r,j}-1)\pi_{rr} - m_{r,j}^2\pi_r^2)$;
4. $$\begin{aligned}\mathrm{Var}[M_{r,j}(M_{r,j}-1)] = m_{r,j}(m_{r,j}-1)\big[&(2 - m_{r,j}(m_{r,j}-1)\pi_{rr})\pi_{rr} \\ &+ 4(m_{r,j}-2)\pi_r^{(3)} \\ &+ (m_{r,j}-2)(m_{r,j}-3)\pi_r^{(4)}\big];\end{aligned}$$

and for $r, s \in \{1, \ldots, k\}$, $r \neq s$,

5. $\mathbb{E}[M_{r,j}M_{s,j}] = m_{r,j}m_{s,j}\pi_{rr}$;
6. $$\begin{aligned}\mathrm{Var}[M_{r,j}M_{s,j}] = m_{r,j}m_{s,j}\big[&(1 - m_{r,j}m_{s,j}\pi_{rs})\pi_{rs} \\ &+ (m_{s,j}-1)\pi_{rs}^{(1,2)} + (m_{r,j}-1)\pi_{rs}^{(2,1)} \\ &+ (1 - m_{s,j} - m_{r,j} + m_{r,j}m_{s,j}\pi_{rs}^{(2,2)})\big].\end{aligned}$$

Proof of Lemma 6.A.1.

1. $$\begin{aligned}EE[M_{r,j}] &= \mathbb{E}[\mathbb{E}[M_{r,j} \mid \boldsymbol{Q}_j]] \\ &= \mathbb{E}[m_{r,j}Q_{r,j}] = m_{r,j}\pi_r.\end{aligned}$$

2. $$\begin{aligned}\mathbb{E}[M_{r,j}(M_{r,j}-1)] &= \mathbb{E}[\mathbb{E}[M_{r,j}(M_{r,j}-1) \mid \boldsymbol{Q}_j]] \\ &= \mathbb{E}[m_{r,j}(m_{r,j}-1)Q_{r,j}^2] \\ &= m_{r,j}(m_{r,j}-1)\pi_{rr},\end{aligned}$$
   where the second step follows from equation (6.4.4).

3. $$\begin{aligned}\mathrm{Var}[M_{r,j}] &= \mathbb{E}[M_{r,j}^2] - \mathbb{E}[M_{r,j}]^2 \\ &= \mathbb{E}[M_{r,j}(M_{r,j}-1)] + \mathbb{E}[M_{r,j}] - \mathbb{E}[M_{r,j}]^2 \\ &= m_{r,j}(m_{r,j}-1)\pi_{rr} + m_{r,j}\pi_r - m_{r,j}^2\pi_r^2.\end{aligned}$$

4. We first note that
   $$\begin{aligned}\mathbb{E}[M_{r,j}^3] &= \mathbb{E}[M_{r,j}(M_{r,j}-1)(M_{r,j}-2)] + 3\mathbb{E}[M_{r,j}(M_{r,j}-1)] + \mathbb{E}[M_{r,j}], \\ \mathbb{E}[M_{r,j}^4] &= \mathbb{E}[M_{r,j}(M_{r,j}-1)(M_{r,j}-2)(M_{r,j}-3)] + 6\mathbb{E}[M_{r,j}^3] \\ &\quad + 7\mathbb{E}[M_{r,j}(M_{r,j}-1)] + \mathbb{E}[M_{r,j}].\end{aligned}$$

Then, using again equation (6.4.4),

$$\begin{aligned}\mathrm{Var}[M_{r,j}(M_{r,j}-1)] &= \mathbb{E}[M_{r,j}^4] - 2\mathbb{E}[M_{r,j}^3] + \mathbb{E}[M_{r,j}^2] - \mathbb{E}[M_{r,j}(M_{r,j}-1)]^2 \\ &= \mathbb{E}[M_{r,j}(M_{r,j}-1)(M_{r,j}-2)(M_{r,j}-3)] \\ &\quad + 2\mathbb{E}[M_{r,j}(M_{r,j}-1)] \\ &\quad + 4\mathbb{E}[M_{r,j}(M_{r,j}-1)(M_{r,j}-2)] - \mathbb{E}[M_{r,j}(M_{r,j}-1)]^2 \\ &= m_{r,j}(m_{r,j}-1)\big[(2-m_{r,j}(m_{r,j}-1)\pi_{rr})\pi_{rr} \\ &\quad + 4(m_{r,j}-2)\pi_r^{(3)} + (m_{r,j}-2)(m_{r,j}-3)\pi_r^{(4)}\big].\end{aligned}$$

5. $$\begin{aligned}\mathbb{E}[M_{r,j}M_{s,j}] &= \mathbb{E}[\mathbb{E}[M_{r,j}M_{s,j} \mid \boldsymbol{Q}_j]] \\ &= \mathbb{E}[m_{r,j}m_{s,j}Q_{r,j}Q_{s,j}] = m_{r,j}m_{s,j}\pi_{rs}.\end{aligned}$$

6. $$\begin{aligned}\mathrm{Var}[M_{r,j}M_{s,j}] &= \mathbb{E}[\mathbb{E}[M_{r,j}^2 \mid \boldsymbol{Q}_j]\,\mathbb{E}[M_{s,j}^2 \mid \boldsymbol{Q}_j]] - \mathbb{E}[M_{r,j}M_{s,j}]^2 \\ &= m_{r,j}m_{s,j}\big[(1-m_{r,j}m_{s,j}\pi_{rs})\pi_{rs} \\ &\quad + (m_{s,j}-1)\pi_{rs}^{(1,2)} + (m_{r,j}-1)\pi_{rs}^{(2,1)} \\ &\quad + (1-m_{s,j}-m_{r,j}+m_{r,j}m_{s,j}\pi_{rs}^{(2,2)})\big].\end{aligned}$$

□

We are now ready to proceed to the proof of the announced result.

Proof of Theorem 6.4.1. Write the estimators of π_r as

$$\widetilde{\pi}_r(\boldsymbol{w}_r) = \sum_{j=1}^{n} w_{r,j}\widehat{Q}_{r,j}, \qquad r \in \{1,\ldots,k\},$$

where $\widehat{Q}_{r,1},\ldots,\widehat{Q}_{r,n}$ have common expectation $\mathbb{E}[\widehat{Q}_{r,j}] = \pi_r$ (Lemma 6.A.1, item 1) and possibly different variances $\mathrm{Var}[\widehat{Q}_{r,j}] = \sigma_{r,j}^2$, where

$$\sigma_{r,j}^2 = \frac{1}{m_{r,j}^2}\mathrm{Var}[M_{r,j}] = \frac{\pi_r}{m_{r,j}} + \left(1-\frac{1}{m_{r,j}}\right)\pi_{rr} - \pi_r^2,$$

by Lemma 6.A.1 (item 3) and where the weight vector $\boldsymbol{w}_r = (w_{r,1},\ldots,w_{r,n})$ has nonnegative entries. We wish to minimize the mean squared error (MSE) of $\widetilde{\pi}_r(\boldsymbol{w}_r)$ as a function of $\boldsymbol{w}_r$,

$$\begin{aligned}\mathrm{MSE}[\widetilde{\pi}_r(\boldsymbol{w}_r)] &= \mathrm{Var}[\widetilde{\pi}_r(\boldsymbol{w}_r)] + \left(\mathbb{E}[\widetilde{\pi}_r(\boldsymbol{w}_r) - \pi_r]\right)^2 \\ &= \sum_{j=1}^{n} w_{r,j}^2\sigma_{r,j}^2 + \left(\sum_{j=1}^{n} w_{r,j} - 1\right)^2\mu_r^2.\end{aligned}$$

Setting the partial derivatives with respect to $w_{r,1}, \ldots, w_{r,n}$ equal to zero gives the solution

$$w_{r,j,\text{opt}} = \frac{\sigma_{r,j}^{-2}}{\pi_r^{-2} + \sum_{t=1}^{n} \sigma_{r,t}^{-2}}, \qquad j = 1, \ldots, n,$$
$$\text{MSE}[\widetilde{\pi}_r(\boldsymbol{w}_{r,\text{opt}})] = \frac{1}{\pi_r^{-2} + \sum_{j=1}^{n} \sigma_{r,j}^{-2}},$$

where $\boldsymbol{w}_{r,\text{opt}} = (w_{r,1,\text{opt}}, \ldots, w_{r,n,\text{opt}})$. For the second-order probabilities π_{rr} and π_{rs}, the same recipe can be followed. Their estimators will use the quantities $\text{Var}[M_{r,j}(M_{r,j}-1)]$ and $\text{Var}[M_{r,j}M_{s,j}]$. These are calculated in Lemma 6.A.1 (items 4 and 6). □

Remark 6.A.1. In practice, the estimators $\widehat{\sigma}_{r,j}^2$ and $\widehat{\sigma}_{rr,j}^2$ from Definition 6.4.1 can be negative. When this happens, a simple solution is to replace the preliminary estimator $\widehat{\pi}_r^{(l)}$ by the estimator

$$\check{\pi}_r^{(l)} = \frac{1}{n} \sum_{j=1}^{n} \frac{M_{r,j}^l}{m_{r,j}^l}.$$

Note that $\check{\pi}_r^{(l)} > \widehat{\pi}_r^{(l)}$ for $l > 1$. Using $\check{\pi}_r^{(l)}$ instead of $\widehat{\pi}_r^{(l)}$ as a preliminary estimator in Definition 6.4.1 ensures that both $\widehat{\sigma}_{r,j}^2$ and $\widehat{\sigma}_{rr,j}^2$ are positive. Asymptotically as $m_{r,j} \to \infty$, this method is equivalent to the one described in Definition 6.4.1.

Proof of Lemma 6.3.1. To prove Lemma 6.3.1 for every $k \in \mathbb{N}$, first observe that

$$\nu_r - \sigma_r \log\left(-\log(q_r)\right) > \mu_r - \sigma_r \log\left(-\log(q_0)\right) \quad \Longleftrightarrow \quad q_r > g_r(q_0).$$

Suppose $k = 1$. Then $R = \{1\}$ and

$$\begin{aligned} I_j &= \int_0^1 \int_0^1 f_{1,j}(q_0, q_1)\, \mathrm{d}q_1\, \mathrm{d}q_0 \\ &= \int_0^1 \int_0^1 \mathbb{1}\left(q_1 \leq g_1(q_0)\right) h_{1,j}(q_0; \mu_1) + \mathbb{1}\left(q_1 > g_1(q_0)\right) h_{1,j}(q_1; \nu_1)\, \mathrm{d}q_1\, \mathrm{d}q_0 \\ &= \int_0^1 g_1(q_0) h_{1,j}(q_0; \mu_1)\, \mathrm{d}q_0 + \int_0^1 \int_{g_1(q_0)}^1 h_{1,j}(q_1; \nu_1)\, \mathrm{d}q_1\, \mathrm{d}q_0, \end{aligned}$$

which is equal to the result of Lemma 6.3.1 since $\mathcal{P}(R) = \{\varnothing, \{1\}\}$. Next, define $R_k = \{1, \ldots, k\}$ and assume that the final expression in Lemma 6.3.1 is

valid. Let $\mathrm{d}\boldsymbol{q}_{1:k}$ denote $\mathrm{d}q_1 \cdots \mathrm{d}q_k$. Then for $R_{k+1} = \{1, \ldots, k+1\}$,

$$
\begin{aligned}
I_j &= \int_{[0,1]^2} f_{k+1,j}(q_0, q_{k+1}) \left(\int_{[0,1]^k} \left(\prod_{r=1}^{k} f_{r,j}(q_0, q_r) \right) \mathrm{d}\boldsymbol{q}_{1:k} \right) \mathrm{d}q_{k+1}\, \mathrm{d}q_0 \\
&= \int_0^1 g_{k+1}(q_0) h_{k+1,j}(q_0; \mu_{k+1}) \left(\int_{[0,1]^k} \left(\prod_{r=1}^{k} f_{r,j}(q_0, q_r) \right) \mathrm{d}\boldsymbol{q}_{1:k} \right) \mathrm{d}q_0 \\
&\quad + \int_0^1 \int_{g_{k+1}(q_0)}^1 h_{k+1,j}(q_{k+1}; \tau_{k+1}) \left(\int_{[0,1]^k} \left(\prod_{r=1}^{k} f_{r,j}(q_0, q_r) \right) \mathrm{d}\boldsymbol{q}_{1:k} \right) \mathrm{d}q_{k+1}\, \mathrm{d}q_0 \\
&= \sum_{I \in \mathcal{P}(R_k)} \int_0^1 \left(\prod_{r \in \{R_k \setminus I\} \cup \{k+1\}} g_r(q_0) h_{r,j}(q_0; \mu_r) \right) \\
&\qquad\qquad \times \left(\prod_{r \in I} \int_{g_r(q_0)}^1 h_{r,j}(q_r; \nu_r)\, \mathrm{d}q_r \right) \mathrm{d}q_0 \\
&\quad + \sum_{I \in \mathcal{P}(R_k)} \int_0^1 \left(\prod_{r \in R_k \setminus I} g_r(q_0) h_{r,j}(q_0; \mu_r) \right) \\
&\qquad\qquad \times \left(\prod_{r \in I \cup \{k+1\}} \int_{g_r(q_0)}^1 h_{r,j}(q_r; \nu_r)\, \mathrm{d}q_r \right) \mathrm{d}q_{k+1}\, \mathrm{d}q_0 \\
&= \sum_{I \subset \mathcal{P}(R_{k+1})} \int_0^1 \left(\prod_{r \in R_{k+1} \setminus I} g_r(q_0) h_{r,j}(q; \mu_r) \right) \\
&\qquad\qquad \times \left(\prod_{r \in I} \int_{g_r(q_0)}^1 h_{r,j}(q_r; \nu_r)\, \mathrm{d}q_r \right) \mathrm{d}q_{k+1}\, \mathrm{d}q_0.
\end{aligned}
$$

For the last step, note that if $I \in \mathcal{P}(R_{k+1})$, then either $I \in \mathcal{P}(R_k)$ so that $\{R_{k+1} \setminus I\} = \{R_k \setminus I\} \cup \{k+1\}$ and we get the first term on the penultimate line, or $I \notin \mathcal{P}(R_k)$ and we get the second term on the penultimate line. □

6.B Simulation study

In Section 6.3.4, the parameter vector $\boldsymbol{\theta}$ of a factor model is estimated using maximum likelihood estimation, after which the implied marginal and joint default probabilities π_r and π_{rs} are obtained by plugging in the estimate of $\boldsymbol{\theta}$ in expressions (6.3.3) and (6.3.4). In Section 6.4, two nonparametric estimators of π_r and π_{rs} are introduced. Suppose that the conditional default probabilities $Q_{r,j}$ are generated from one of the multi-factor models in Section 6.3.3, i.e., model (2a) or (2b). We wish to answer the following questions.

- Do the weighted nonparametric estimators of (π_r, π_{rs}) defined in Section 6.4.2 perform better than the unweighted nonparametric estimators from Section 6.4.1?
- Does nonparametric estimation (weighted or unweighted) lead to better or worse estimates of (π_r, π_{rs}) than via maximum likelihood estimation of the parameters of the true model?
- Does maximum likelihood estimation of the parameters of a one-factor submodel provide us with worse estimates of (π_r, π_{rs}) than maximum likelihood estimation of the parameters of the true model?
- Does maximum likelihood estimation of the parameters of another multi-factor model lead to worse estimators of (π_r, π_{rs}) than maximum likelihood estimation of the parameters of the true model, or is the estimation quality of the (joint) default probabilities independent of the underlying data-generating process?

More specifically, we proceed as follows. The number of risks, $m_{r,j}$, in risk category $r \in \{1, \ldots, k\}$ and time period $j \in \{1, \ldots, n\}$ is generated randomly using a beta-binomial model, for $k = 2$ and $n = 19$. The conditional default probabilities $Q_{r,j}$ are then generated using models (2a) and (2b), where the parameter values are chosen in such a way that π_r and π_{rs} very roughly resemble the default probabilities of the S&P rating classes B and CCC; see Table 6.4. The quantities π_r and π_{rs} are then estimated by the two nonparametric estimators and by maximizing the likelihood under the assumption of one of the parametric models (1a), (2a), (1b), and (2b). We repeat this 1000 times and we compare the results using the relative root mean squared error (RRMSE), i.e., the root mean squared error divided by the true parameter value. Note that if the weighted nonparametric estimator leads to a negative value of $\widehat{\sigma}_{r,j}$ or $\widehat{\sigma}_{rr,j}$, we use the unweighted nonparametric estimator; see Remark 6.A.1. This happens less than 1% of the time.

The results are presented in Tables 6.5 and 6.6. Table 6.5 shows the RRMSE of the estimators of the marginal default probabilities, π_r. The methods are compared in terms of the decrease, Δ, of RRMSE in percent with respect to the best method. Consequently, the best method has $\Delta = 0$. When calculating Δ, we take the sum of the values of the two categories. In terms of RRMSE, the weighted nonparametric estimator performs slightly better than the non-weighted one. Maximum likelihood estimation beats nonparametric estimation when the data are generated from model (2b), but it is the other way around when data are generated from model (2a). Misspecification of the model, for example, estimating the parameters of model (2a) although data are generated from model (2b), has no negative effect when data are generated from model (2b).

Table 6.6 shows the RRMSE of the estimators of the joint default probabilities π_{rs}. Again, the weighted nonparametric estimator performs much better

Gumbel			normal		
	$r=1$	$r=2$		$r=1$	$r=2$
μ_r	-1.15	-0.55	μ_r	-1.60	-0.85
ν_r	-1.30	-1.00	τ_r	0.13	0.16
σ_r	0.11	0.15	σ_r	0.18	0.28
π_r	0.0585	0.2091	π_r	0.0591	0.2093
$\pi_{rs} \times 100$	0.397	1.365	$\pi_{rs} \times 100$	0.394	1.401
	1.365	4.759		1.401	4.980

Table 6.4: Parameter values for the max-factor Gumbel model (left) and the sum-factor normal model (right) used in the simulation study.

	Gumbel				normal		
	$r=1$	$r=2$	Δ		$r=1$	$r=2$	Δ
NP	0.116	0.095	4	NP	0.110	0.119	1
NP weighted	0.112	0.092	1	NP weighted	0.109	0.119	0
Model (1a)	0.110	0.094	0	Model (1a)	0.112	0.119	2
Model (2a)	0.110	0.094	0	Model (2a)	0.111	0.119	1
Model (1b)	0.109	0.093	0	Model (1b)	0.120	0.121	6
Model (2b)	0.110	0.094	0	Model (2b)	0.121	0.121	6

Table 6.5: Relative root mean squared error (RRMSE) of estimators of π_r for data generated from a Gumbel max-factor model (left) and a normal sum-factor model (right) with parameter values as in Table 6.4. The methods are compared using Δ, the increase of RRMSE in percent with respect to the best method, which has $\Delta = 0$.

than the non-weighted estimator. Model misspecification has again less effect when data are generated from model (2b) than when data are generated from model (2a). Compared with the results in Table 6.5, for the joint default probabilities we see a larger increase in RRMSE if we estimated the parameters of a one-factor submodel instead of a multi-factor model.

A final thing worth noticing is that when estimating the parameters of models based on the normal distribution we obtain better estimators of low default probabilities (here $r = 1$) than when using models based on the Gumbel distribution. For higher default probabilities (here $r = 2$), the quality of estimation differs less. The higher RRMSEs for the joint default probabilities based on the parameters of the Gumbel models are entirely caused by a rather high bias; while all estimators stemming from the Gumbel models exhibit (high) positive bias, the estimators stemming from the normal models are all negatively

biased. Thus, although using factor models based on the normal distribution leads to estimators with lower RRMSE, it is an important drawback of models (1a) and (2a) that they are underestimating the true default probabilities.

	Gumbel				normal		
NP	$r=1$	$r=2$	Δ	NP	$r=1$	$r=2$	Δ
$r=1$	0.316	0.222	13	$r=1$	0.253	0.207	7
$r=2$	0.222	0.211		$r=2$	0.207	0.255	
NP weighted	$r=1$	$r=2$	Δ	NP weighted	$r=1$	$r=2$	Δ
$r=1$	0.266	0.202	0	$r=1$	0.230	0.202	0
$r=2$	0.202	0.197		$r=2$	0.202	0.238	
Model (1a)	$r=1$	$r=2$	Δ	Model (1a)	$r=1$	$r=2$	Δ
$r=1$	0.263	0.204	1	$r=1$	0.263	0.216	8
$r=2$	0.204	0.202		$r=2$	0.216	0.245	
Model (2a)	$r=1$	$r=2$	Δ	Model (2a)	$r=1$	$r=2$	Δ
$r=1$	0.258	0.204	0	$r=1$	0.248	0.211	5
$r=2$	0.204	0.202		$r=2$	0.211	0.243	
Model (1b)	$r=1$	$r=2$	Δ	Model (1b)	$r=1$	$r=2$	Δ
$r=1$	0.289	0.210	6	$r=1$	0.397	0.274	41
$r=2$	0.210	0.203		$r=2$	0.274	0.271	
Model (2b)	$r=1$	$r=2$	Δ	Model (2b)	$r=1$	$r=2$	Δ
$r=1$	0.271	0.209	3	$r=1$	0.358	0.259	32
$r=2$	0.209	0.203		$r=2$	0.259	0.270	

Table 6.6: Relative root mean squared error (RRMSE) of estimators of π_{rs} for data generated from a Gumbel max-factor model (left) and a normal sum-factor model (right) with parameter values as in Table 6.4. The methods are compared using Δ, the increase of RRMSE in percent with respect to the best method, which has $\Delta = 0$.

Chapter 7

Conclusion

One of the aims of this thesis was to develop new estimation methods for parametric models for extremal dependence of high-dimensional data. Using Einmahl et al. (2012) as a starting point, in Chapter 2 we proposed the pairwise M-estimator, which is particularly adapted to high-dimensional (spatial) models since only pairs of random variables are used in the estimation procedure. Although for dimensions d in the order of a hundred, using all $d(d-1)/2$ pairs of variables is feasible, this is no longer a possibility when one focuses on a spatial problem with thousands of measuring stations. Since we saw that the quality of estimation might actually improve when limiting the number of pairs to pairs of neighbouring stations only, this is not necessarily a bad thing. In Chapter 2 we saw that numerically this method might be cumbersome when using the optimal weight matrix, since it requires four-dimensional numerical integration. Moreover, estimation without an optimal weight matrix was fast because the models we considered had analytical expressions for the integrated bivariate stable tail dependence function, but this does not need to be the case for all parametric models.

A direct method of reducing the number of integrals and therefore reducing the computation time is by estimating the *Pickands dependence function* $A : \Delta_{d-1} \to [1/d, 1]$ instead of the stable tail dependence function ℓ. These two functions are connected through

$$\ell(\boldsymbol{x}) = \left(\sum_{j=1}^{d} x_j\right) A\left(\frac{x_1}{\sum_{j=1}^{d} x_j}, \ldots, \frac{x_{d-1}}{\sum_{j=1}^{d} x_j}\right), \qquad \boldsymbol{x} \in [0,\infty)^d.$$

Let $\widehat{A}_{n,k}$ denote any initial estimator of A. For locations $\boldsymbol{s}_u, \boldsymbol{s}_v \in \mathbb{R}^2$, the bivariate margins of A and $\widehat{A}_{n,k}$ are

$$A_{uv}(t;\theta) = \ell_{uv}(t, 1-t;\theta), \qquad \widehat{A}_{n,k,uv}(t;\theta) = \widehat{\ell}_{n,k,uv}(t, 1-t).$$

and a new estimator is obtained by following the same approach as in Chapter 2; the non-weighted estimator has the form

$$\breve{\boldsymbol{\theta}}_n = \operatorname*{arg\,min}_{\theta \in \Theta} \sum_{(u,v)} \left(\int_0^1 \widehat{A}_{n,k,uv}(t) - A_{uv}(t) \, \mathrm{d}t \right)^2 .$$

Only two-dimensional numerical integration is necessary if we add an optimal weight matrix. Moreover, there exist a lot of other initial estimators of A. For instance, $\widehat{A}_n(t)$ does not satisfy the lower bound $\max(t, 1-t)$ of A nor is it convex. A better nonparametric estimator is obtained by taking the greatest convex minorant of $\max(\widehat{A}_n(t), t, 1-t)$ (Genest and Segers, 2009; Bücher et al., 2011). Other initial estimators of A could be used as well (Pickands III, 1981; Deheuvels, 1991; Capéraà et al., 1997; Marcon et al., 2016; Vettori et al., 2016). More complicated spatial models like the extremal t-process, of which the Brown–Resnick process is a special case, could be estimated using an estimator of this form.

In Chapter 3 we presented the continuous updating weighted least squares estimator, which is numerically easier than the pairwise M-estimator and is thus not limited to pairwise inference. We focused on the estimation of a non-differentiable model, the max-linear model, although the estimator performs satisfactory for all types of models. Similarly to a Pickands dependence function estimator described above, the possibilities for an initial estimator are not limited to the (bias-corrected) empirical tail dependence function: we could use block maxima or threshold exceedances, parametric or non-parametric estimators, the possibilities are endless.

In Chapters 2 and 3, we have always treated the margins by computing their ranks and using the integral probability transform. However, we could also model the margins parametrically above a high threshold by a generalized Pareto distribution; in this case, an estimator of the stable tail dependence function could be obtained as

$$\breve{\ell}_{n,k}(\boldsymbol{x}) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1} \left\{ \left(1 + \widehat{\boldsymbol{\xi}} \frac{\boldsymbol{X}_i - \boldsymbol{u}}{\widehat{\boldsymbol{\eta}}} \right)_+^{1/\widehat{\boldsymbol{\xi}}} \not\leq \frac{1}{\boldsymbol{x}} \right\},$$

where $\widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\xi}}$ are vectors of estimators of the parameters of the generalized Pareto distribution above the threshold $\boldsymbol{u} = (u_1, \ldots, u_d)$ and u_j is the $(k+1)$th largest observation of $X_{1j}, \ldots, X_{nj}$ for $j \in \{1, \ldots, d\}$.

An open problem related to Chapters 2 and 3 is hypothesis testing at the boundary of the parameter space. When estimating a Brown–Resnick process, a shape parameter close to 2 might indicate that the smooth Smith model would suffice. For the max-linear model, a factor loading which is estimated close to zero suggests that we do not need to include this factor in our model. To test for these hypotheses we need a result similar to Corollary 2.2.5, but which is also valid at the boundary of Θ.

Chapter 4 is aimed at guiding the reader through the programs that were written for Chapters 2 and 3. On the long term, the aim of this R package is to contain a complete set of tools for distance-based estimation methods for both multivariate and spatial extreme value analysis; more models or methods might be added in the future.

Chapter 5 is a careful first attempt to the statistical modelling of multivariate extremes using multivariate generalized Pareto distributions. The construction device that is proposed in Rootzén et al. (2016) allows for new parametric models, which is important in the context of high-dimensional extremes because not so many models exist whose numerical implementation is doable. In this chapter, the focus is on densities rather than on stable tail dependence functions — models that do not have an analytical expression for the stable tail dependence function might still have a relatively easy density: the ordered components model proposed to model landslides is such an example.

In Chapter 5 we focused on censored likelihood estimation only. It follows from expression (5.3.2) that one could use the nonparametric estimator of the stable tail dependence function to obtain a nonparametric estimator of the (standardized) generalized Pareto distribution H^*. This could in turn be a stepping stone for semi-parametric estimation of H^*.

The subject of Chapter 6 deviates from the title of this thesis, focusing on the modelling of dependent default in credit risk. Although the proposed max-factor model seems to capture the shocks stemming from both global factors and rating class-specific factors quite well, the computational burden is significantly larger than for a linear model based on the Gaussian distribution. This forms the main obstacle to the application of this model on a larger scale. A possible extension of the model would make use of a richer dataset, taking into account transition probabilities between rating classes or specific observable factors.

This thesis only casually mentions the issues of serial dependence and stationarity, since these subjects could both constitute a book on its own. Whereas we usually assumed that the univariate time series we use are independent and identically distributed, in practice one often encounters either temporal dependence (e.g., even a series of weekly negative log-returns on a stock can exhibit short-term temporal dependence), seasonal variability (e.g., rainfall series that follow another distribution in summer than in winter), or long-term trends (e.g., an increase in temperature over the last decades due to climate change). Temporal dependence in a stationary time series does not have an influence on the consistency result of our proposed estimators, but the asymptotic variance matrix, and thus the optimal weight matrix, would need to be adjusted. When using censored likelihood estimation for the multivariate generalized Pareto distribution, our threshold exceedances are not necessarily independent when our series exhibit temporal dependence, and some declustering might be needed. However, a cluster is already hard to define in the univariate setting, let alone

in the multivariate setting, so we have chosen not to focus on this issue.

Seasonal variability can be dealt with by splitting up our data in blocks that are stationary: for instance, in Chapter 2, we studied only the summer wind speed maxima. Including a long-term trend, for instance in the mean value of our process of interest, although straightforward for likelihood methods, is less obvious for our estimators.

Another topic this thesis has not touched upon is asymptotic independence. When the data we consider are asymptotically independent, the stable tail dependence function is equal to its upper bound, i.e., $\ell(\boldsymbol{x}) = x_1 + \cdots + x_d$. In that case, the theory described in this thesis is of no use and alternative methods to model the joint tail need to be used. In the multivariate setting, this is a subject which has been developed since several decades (Ledford and Tawn, 1996, 1997; Heffernan and Tawn, 2004; Ramos and Ledford, 2009; Wadsworth et al., 2013, 2016). In the spatial setting, models incorporating asymptotic independence have started to be developed as well (Wadsworth and Tawn, 2012; Bacro et al., 2016; Opitz, 2016).

Bibliography

Abadir, K. M. and Magnus, J. R. (2005). *Matrix Algebra*, volume 1. Cambridge University Press.

Anastasiadis, S. and Chukova, S. (2012). Multivariate insurance models: an overview. *Insurance: Mathematics and Economics*, 51(1):222–227.

Aulbach, S., Falk, M., and Zott, M. (2015). The space of D-norms revisited. *Extremes*, 18(1):85–97.

Bacro, J.-N. and Gaetan, C. (2014). Estimation of spatial max-stable models using threshold exceedances. *Statistics and Computing*, 24(4):651–662.

Bacro, J.-N., Gaetan, C., and Toulemonde, G. (2016). A flexible dependence model for spatial extremes. *Journal of Statistical Planning and Inference*, 172:36–52.

Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804.

Balkema, A. A. and Resnick, S. I. (1977). Max-infinite divisibility. *Journal of Applied Probability*, 14(2):309–319.

Ballani, F. and Schlather, M. (2011). A construction principle for multivariate extreme value distributions. *Biometrika*, 98(3):633–645.

Beirlant, J., Escobar-Bach, M., Goegebeur, Y., and Guillou, A. (2016). Bias-corrected estimation of stable tail dependence function. *Journal of Multivariate Analysis*, 143(1):453–466.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley.

Beylich, A. A. and Sandberg, O. (2005). Geomorphic effects of the extreme rainfall event of 20–21 july, 2004 in the Latnjavagge catchment, northern Swedish Lapland. *Geografiska Annaler: Series A, Physical Geography*, 87(3):409–419.

Bienvenüe, A. and Robert, C. Y. (2014). Likelihood based inference for high-dimensional extreme value distributions. Available at `http://arxiv.org/abs/1403.0065`.

Blanchet, J. and Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics*, 5(3):1699–1724.

Bluhm, C., Overbeck, L., and Wagner, C. (2002). *An introduction to credit risk modeling*. CRC Press.

Brown, B. M. and Resnick, S. I. (1977). Extreme values of independent stochastic processes. *Journal of Applied Probability*, 14(4):732–739.

Bücher, A., Dette, H., and Volgushev, S. (2011). New estimators of the Pickands dependence function and a test for extreme-value dependence. *The Annals of Statistics*, 39(4):1963–2006.

Bücher, A. and Segers, J. (2014). Extreme value copula estimation based on block maxima of a multivariate stationary time series. *Extremes*, 17(3):495–528.

Bücher, A. and Segers, J. (2016). On the maximum likelihood estimator for the Generalized Extreme-Value distribution. Available at `http://arxiv.org/abs/1601.05702`.

Bücher, A., Segers, J., and Volgushev, S. (2014). When uniform weak convergence fails: empirical processes for dependence functions and residuals via epi- and hypographs. *The Annals of Statistics*, 42(4):1598–1634.

Bücher, A. and Volgushev, S. (2013). Empirical and sequential empirical copula processes under serial dependence. *Journal of Multivariate Analysis*, 119:61–70.

Buishand, T., de Haan, L., and Zhou, C. (2008). On spatial extremes: with application to a rainfall problem. *Annals of Applied Statistics*, 2(2):624–642.

Can, S. U., Einmahl, J. H., Khmaladze, E. V., and Laeven, R. J. (2015). Asymptotically distribution-free goodness-of-fit testing for tail copulas. *The Annals of Statistics*, 43(2):878–902.

Capéraà, P., Fougères, A.-L., and Genest, C. (1997). A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, 84(3):567–577.

Castruccio, S., Huser, R., and Genton, M. G. (2016). High-order composite likelihood inference for max-stable distributions and processes. *Journal of Computational and Graphical Statistics*. To appear.

Ceppi, P., Della-Marta, P., and Appenzeller, C. (2008). Extreme value analysis of wind speed observations over Switzerland. *Arbeitsberichte der MeteoSchweiz*, 219.

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53(2):377–392.

Cooley, D., Cisewski, J., Erhardt, R. J., Jeon, S., Mannshardt, E., Omolo, B. O., and Sun, Y. (2012a). A survey for spatial extremes: measuring spatial dependence and modelling spatial effects. *REVSTAT*, 10(1):135–165.

Cooley, D., Davis, R. A., and Naveau, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117.

Cooley, D., Davison, A. C., and Ribatet, M. (2012b). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22(2):813–845.

Cossette, H., Gaillardetz, P., Marceau, É., and Rioux, J. (2002). On two dependent individual risk models. *Insurance: Mathematics and Economics*, 30(2):153–166.

Davis, R. A., Klüppelberg, C., and Steinkohl, C. (2013). Statistical inference for max-stable processes in space and time. *Journal of the Royal Statistical Society: Series B (Statistical Methodolgy)*, 75(5):791–819.

Davis, R. A. and Mikosch, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli*, 15(4):977–1009.

Davison, A. C. and Gholamrezaee, M. M. (2012). Geostatistics of extremes. *Proceedings of the Royal Society A*, 468(2138):581–608.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science*, 27(2):161–186.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with comments). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 52(3):393–442.

de Haan, L. (1984). A spectral representation for max-stable processes. *The Annals of Probability*, 12(4):1194–1204.

de Haan, L. and de Ronde, J. (1998). Sea and wind: Multivariate extremes at work. *Extremes*, 1(1):7–45.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: an Introduction*. Springer-Verlag Inc.

de Haan, L. and Lin, T. (2001). On convergence toward an extreme value distribution in C[0,1]. *The Annals of Probability*, 29(1):467–483.

de Haan, L. and Pereira, T. T. (2006). Spatial extremes: Models for the stationary case. *The Annals of Statistics*, 34(1):146–168.

de Haan, L. and Resnick, S. I. (1977). Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 40(4):317–337.

Deheuvels, P. (1991). On the limiting behavior of the pickands estimator for bivariate extreme-value distributions. *Statistics & Probability Letters*, 12(5):429–439.

Denuit, M., Dhaene, J., Goovaerts, M., and Kaas, R. (2005). *Actuarial theory for dependent risks: measures, orders and models*. John Wiley & Sons.

Denuit, M., Kiriliouk, A., and Segers, J. (2015). Max-factor individual risk models with application to credit portfolios. *Insurance: Mathematics and Economics*, 62(1):162–172.

Denuit, M. and Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93(1):40–57.

Denuit, M., Lefèvre, C., and Utev, S. (2002). Measuring the impact of dependence between claims occurrences. *Insurance: Mathematics and Economics*, 30(1):1–19.

Draisma, G., Drees, H., Ferreira, A., and De Haan, L. (2004). Bivariate tail estimation: dependence in asymptotic independence. *Bernoulli*, 10(2):251–280.

Drees, H. and Huang, X. (1998). Best attainable rates of convergence for estimators of the stable tail dependence function. *Journal of Multivariate Analysis*, 64(1):25–47.

Einmahl, J. H., Kiriliouk, A., Krajina, A., and Segers, J. (2016a). An M-estimator of spatial tail dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):275–298.

Einmahl, J. H., Kiriliouk, A., and Segers, J. (2016b). A continuous updating weighted least squares estimator of tail dependence in high dimensions. Available at `http://arxiv.org/abs/1601.04826`.

Einmahl, J. H. J., de Haan, L., and Li, D. (2006). Weighted approximations of tail copula processes with application to testing the bivariate extreme value condition. *The Annals of Statistics*, 34(4):1987–2014.

Einmahl, J. H. J., Krajina, A., and Segers, J. (2008). A method of moments estimator of tail dependence. *Bernoulli*, 14(4):1003–1026.

Einmahl, J. H. J., Krajina, A., and Segers, J. (2012). An M-estimator for tail dependence in arbitrary dimensions. *The Annals of Statistics*, 40(3):1764–1793.

Engelke, S., Malinowski, A., Kabluchko, Z., and Schlather, M. (2015). Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):239–265.

Falk, M. and Guillou, A. (2008). Peaks-over-threshold stability of multivariate generalized Pareto distributions. *Journal of Multivariate Analysis*, 99(4):715–734.

Falk, M., Hofmann, M., and Zott, M. (2015). On generalized max-linear models and their statistical interpolation. *Journal of Applied Probability*, 52(3):736–751.

Falk, M., Hüsler, J., and Reiss, R.-D. (2010). *Laws of small numbers: extremes and rare events*. Springer Science & Business Media.

Falk, M. and Michel, R. (2009). Testing for a multivariate generalized Pareto distribution. *Extremes*, 12(1):33–51.

Ferreira, A. and de Haan, L. (2014). The generalized Pareto process; with a view towards application and simulation. *Bernoulli*, 20(4):1717–1737.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge Univ Press.

Fougères, A.-L., de Haan, L., and Mercadier, C. (2015). Bias correction in multivariate extremes. *The Annals of Statistics*, 43(2):903–934.

Fougères, A.-L., Nolan, J. P., and Rootzén, H. (2009). Models for dependent extremes using stable mixtures. *Scandinavian Journal of Statistics*, 36(1):42–59.

Frey, R. and McNeil, A. J. (2003). Dependent defaults in models of portfolio credit risk. *Journal of Risk*, 6:59–92.

Genest, C. and Segers, J. (2009). Rank-based inference for bivariate extreme-value copulas. *The Annals of Statistics*, 37(5B):2990–3022.

Genton, M. G., Ma, Y., and Sang, H. (2011). On the likelihood function of Gaussian max-stable processes. *Biometrika*, 98(2):481–488.

Genton, M. G., Padoan, S. A., and Sang, H. (2015). Multivariate max-stable spatial processes. *Biometrika*, 102(1):215–230.

Giesecke, K. (2003). A simple exponential model for dependent defaults. *The Journal of Fixed Income*, 13(3):74–83.

Gissibl, N. and Klüppelberg, C. (2015). Max-linear models on directed acyclic graphs. Available at `http://arxiv.org/abs/1512.07522`.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, 22(3):423–453.

Gudendorf, G. and Segers, J. (2011). Nonparametric estimation of an extreme-value copula in arbitrary dimensions. *Journal of multivariate analysis*, 102(1):37–47.

Gudendorf, G. and Segers, J. (2012). Nonparametric estimation of multivariate extreme-value copulas. *Journal of Statistical Planning and Inference*, 142(12):3073–3085.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707.

Guzzetti, F., Peruccacci, S., Rossi, M., and Stark, C. P. (2007). Rainfall thresholds for the initiation of landslides in central and southern europe. *Meteorology and atmospheric physics*, 98(3-4):239–267.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3):262–280.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2015). *copula: multivariate dependence with copulas*. R package version 0.999-13.

Hosking, J., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261.

Hosking, J. R. M. and Wallis, J. R. (2005). *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press.

Huang, X. (1992). *Statistics of bivariate extreme values*. PhD thesis, Tinbergen Institute Research Series.

Huser, R. and Davison, A. (2013). Composite likelihood estimation for the Brown–Resnick process. *Biometrika*, 100(2):511–518.

Huser, R. and Davison, A. (2014). Space–time modelling of extreme events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):439–461.

Huser, R., Davison, A. C., and Genton, M. G. (2015). Likelihood estimators for multivariate extremes. *Extremes*, 19(1):1–25.

Huser, R. and Genton, M. G. (2016). Non-stationary dependence structures for spatial extremes. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 1–22.

Hüsler, J. and Li, D. (2009). Testing asymptotic independence in bivariate extremes. *Journal of Statistical Planning and Inference*, 139(3):990–998.

Jiang, J. (2010). *Large sample techniques for statistics*. Springer.

Jonasson, C. and Nyberg, R. (1999). The rainstorm of August 1998 in the Abisko area, northern Sweden: preliminary report on observations of erosion and sediment transport. *Geografiska Annaler: Series A, Physical Geography*, 81(3):387–390.

Kabluchko, Z., Schlather, M., and de Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. *Annals of Probability*, 37(5):2042–2065.

Karpa, O. and Naess, A. (2013). Extreme value statistics of wind speed data by the ACER method. *Journal of Wind Engineering and Industrial Aerodynamics*, 112:1–10.

Kiriliouk, A. (2016). *tailDepFun: Minimum Distance Estimation of Tail Dependence Models*. R package version 1.0.0.

Kiriliouk, A., Segers, J., and Warchoł, M. (2016). Nonparametric estimation of extremal dependence. In *Extreme Value Modeling and Risk Analysis: Methods and Applications*. CRC Press.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Ledford, A. W. and Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59:475–499.

Li, D. X. (2001). On default correlation: a copula function approach. *Journal of Fixed Income*, 9(4):43–53.

Marcon, G., Padoan, S., Naveau, P., Muliere, P., and Segers, J. (2016). Multivariate nonparametric estimation of the pickands dependence function using Bernstein polynomials. Available at `http://arxiv.org/abs/1405.5228`.

Michel, R. (2009). Parametric estimation procedures in multivariate generalized Pareto models. *Scandinavian Journal of Statistics*, 36(1):60–75.

Molchanov, I. (2008). Convex geometry of max-stable distributions. *Extremes*, 11(3):235–259.

Nikoloulopoulos, A. K., Joe, H., and Li, H. (2009). Extreme value properties of multivariate t copulas. *Extremes*, 12(2):129–148.

Nolan, J., Fougères, A.-L., and Mercadier, C. (2015). Estimation for multivariate extreme value distributions using max projections. Presentation available at `http://sites.lsa.umich.edu/eva2015/program/`.

Oesting, M., Schlather, M., and Friedrichs, P. (2015). Statistical post-processing of forecasts for extremes using bivariate Brown–Resnick processes with an application to wind gusts. Available at `http://arxiv.org/abs/1312.4584`.

Opitz, T. (2013). Extremal t processes: Elliptical domain of attraction and a spectral representation. *Journal of Multivariate Analysis*, 122(1):409–413.

Opitz, T. (2016). Modeling asymptotically independent spatial extremes based on laplace random fields. *Spatial Statistics*, 16:1–18.

Padoan, S., Ribatet, M., and Sisson, S. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association (Theory and Methods)*, 105(489):263–277.

Palutikof, J., Brabson, B., Lister, D., and Adcock, S. (1999). A review of methods to calculate extreme wind speeds. *Meteorological Applications*, 6(2):199–132.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131.

Pickands III, J. (1981). Multivariate extreme value distributions (STMA V25 119). *Bulletin of the International Statistical Institute*, 49:859–878.

Prescott, P. and Walden, A. (1980). Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika*, 67(3):723–724.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramos, A. and Ledford, A. (2009). A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):219–241.

Rapp, A. and Strömquist, L. (1976). Slope erosion due to extreme rainfall in the Scandinavian mountains. *Geografiska Annaler. Series A. Physical Geography*, 58(3):193–200.

Reich, B. J. and Shaby, B. A. (2012). A hierarchical max-stable spatial model for extreme precipitation. *The Annals of Applied Statistics*, 6(4):1430–1451.

Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer, New York.

Ressel, P. (2013). Homogeneous distributions—and a spectral representation of classical mean values and stable tail dependence functions. *Journal of Multivariate Analysis*, 117(1):246–256.

Ribatet, M. (2013). Spatial extremes: max-stable processes at work. *Journal de la Société Française de Statistique*, 154(2):1–22.

Ribatet, M. (2015). *SpatialExtremes: Modelling Spatial Extremes*. R package version 2.0-2.

Rootzén, H., Segers, J., and Wadsworth, J. (2016). Multivariate peaks over threshold models. Available at `http://arxiv.org/abs/1603.06619`.

Rootzén, H. and Tajvidi, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli*, 12(5):917–930.

Rudvik, A. (2012). *Dependence structures in stable mixture models with an application to extreme precipitation*. PhD thesis, Chalmers University of Technology.

Salmon, F. (2009). Recipe for Disaster: The Formula That Killed Wall Street.

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1):33–60.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, 5(1):33–44.

Schlather, M. and Tawn, J. (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, 90(1):139–156.

Schmidt, R. and Stadtmüller, U. (2006). Non-parametric estimation of tail dependence. *Scandinavian Journal of Statistics*, 33(2):307–335.

Segers, J. (2012). Max-stable models for multivariate extremes. *REVSTAT — Statistical Journal*, 10(1):61–92.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Sibuya, M. (1960). Bivariate extreme statistics, I. *Annals of the Institute of Statistical Mathematics*, 11(2):195–210.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. Technical report, Publications de l'Institut de statistique de l'Université de Paris 8.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90.

Smith, R. L. (1990). Max-stable processes and spatial extremes. Unpublished manuscript.

Smith, R. L., Tawn, J. A., and Coles, S. G. (1997). Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268.

Standard and Poor's (2001). Ratings performance 2000: Default, transition, recovery, and spreads.

Stephenson, A. and Tawn, J. (2005). Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika*, 92(1):213–227.

Strokorb, K., Schlather, M., et al. (2015). An exceptional max-stable process fully parameterized by its extremal coefficients. *Bernoulli*, 21(1):276–302.

Tajvidi, N. (1996). *Characterisation and Some Statistical Aspects of Univariate and Multivariate Generalized Pareto Distributions*. PhD thesis, Department of Mathemtics, Chalmers, Göteborg.

Tawn, J. A. (1990). Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245–253.

Thibaud, E., Aalto, J., Cooley, D. S., Davison, A. C., and Heikkinen, J. (2015). Bayesian inference for the Brown–Resnick process, with an application to extreme low temperatures. Available at `http://arxiv.org/abs/1506.07836`.

Thibaud, E., Mutzner, R., and Davison, A. (2013). Threshold modeling of extreme spatial rainfall. *Water resources research*, 49(8):4633–4644.

Thibaud, E. and Opitz, T. (2015). Efficient inference and simulation for elliptical Pareto processes. *Biometrika*, 102(4):855–870.

Valdez, E. A. (2014). Empirical investigation of insurance claim dependencies using mixture models. *European Actuarial Journal*, 4(1):1–25.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Vettori, S., Hüser, R., and Genton, M. G. (2016). A comparison of non-parametric and parametric estimators of the dependence function in multivariate extremes. Submitted.

Wadsworth, J., Tawn, J., Davison, A., and Elton, D. (2016). Modelling across extremal dependence classes. To be published in the Journal of the Royal Statistical Society: Series B (Statistical Methodology).

Wadsworth, J., Tawn, J., et al. (2013). A new representation for multivariate tail probabilities. *Bernoulli*, 19(5):2689–2714.

Wadsworth, J. L. and Tawn, J. A. (2012). Dependence modelling for spatial extremes. *Biometrika*, 99(2):253–272.

Wadsworth, J. L. and Tawn, J. A. (2014). Efficient inference for spatial extreme-value processes associated to log-Gaussian random functions. *Biometrika*, 101(1):1–15.

Wang, Y. and Stoev, S. (2011). Conditional sampling for spectrally discrete max-stable random fields. *Advances in Applied Probability*, 43(2):461–483.

Yuen, R. A. and Stoev, S. (2014). CPRS M-estimation for max-stable models. *Extremes*, 17(3):387–410.