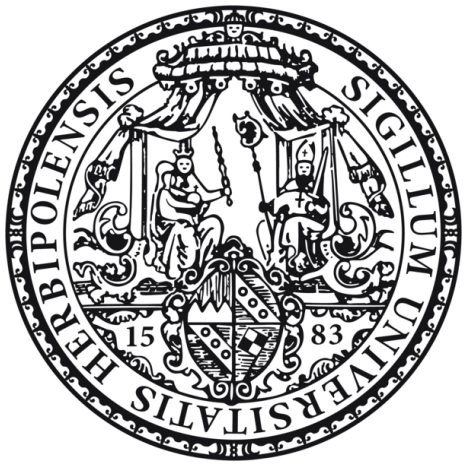
Contributions to the Multivariate Max-Domain of Attraction



PhD Thesis
Timo Fuller
February 2020

under the guidance of
Prof. Dr. Michael Falk
Chair of Mathematics VIII
Würzburg University

Contents

Chapter 1

Preliminaries

In Section 1.1 we start with some questions that motivate extreme value theory and introduce the bare necessities we need to know about univariate extreme value theory to study multivariate extreme value theory. We introduce the multivariate max-domain of attraction and the class of simple max-stable random vectors.

Section 1.2 then introduces many different, but equivalent characterizations of when a random vector $\mathbf{X}$ is in the max-domain of attraction of a simple max-stable vector $\mathbf{Y}$. Because of how often this theorem appears in the rest of the thesis it got the name Rosetta Stone theorem.

Section 1.3 then procedes to prove the Rosetta Stone theorem with a great deal of measure theory.

Section 1.4 investigates what happens if a sequence of D-norms converges pointwise from a multivariate extreme value view.

Section 1.5 goes into detail how and when we should use the Rosetta Stone theorem in practical applications. It gives an interesting characterization for when a random vector has the right marginal distributions, but the wrong dependence structure for the Rosetta Stone theorem.

1.1 Tail-behavior and the max-domain of attraction

Many questions that motivate (univariate) extreme value theory fall in one of the three categories below. In all of them X is a random quantity with a cumulative distribution function F defined by $F(t) = P(X \leq t)$.

1. **(Probability problem)** Given a high threshold t what is the probability
$$P(X > t) = 1 - F(t)?$$
2. **(Quantile problem)** Given a low probability p what is a threshold t such that
$$P(X > t) = p$$
holds?
3. **(Behavior above a threshold problem)** Given a high threshold t what are the properties of the X under the condition $X > t$?

Those are abstractions of questions from real life, of which there are some examples below.

Example 1 (From a dike-building perspective)**.** *X is the maximum water level during a year.*

'We have built a dike of height t. How likely is it that the water rises higher than that in a year?' is a probability problem.

'We want to build a dike such that the water level exceeds its height on average once per 100 years. How how high do we have to build it?' is a quantile problem with $p = 1/100$.

'We have build a dike of height t. How high is the average flood that goes above t?' is a question about the behavior above a threshold.

Example 2 (From finance perspective)**.** *X is the loss from an investment.*

'How likely will our investment have a loss of at least t?' is a probability problem.

'What is the Value at risk of this investment at level 99%?' is a quantile problem with $p = 1/100$.

'What is the Expected Shortfall of our investment at level 99%?' is a question about the behavior above the threshold t, that is the value at risk at level $p = 1/100$.

In this thesis we will treat the term 'tail-behavior of X' not as a mathematical object, but as mathematical properties of X that help us answer those kinds of questions.

The most common approach to those questions requires the assumption that there are scaling constants $a_n > 0$, and shifting constants $b_n \in \mathbb{R}$, such that we have the limit
$$a_n \cdot \max_{i=1,\ldots,n} X^{(i)} + b_n \xrightarrow{\mathcal{D}} Y,$$

as $n \to \infty$ and where $X^{(i)}$ are iid copies of X, and Y is a non-degenerate limit random variable. $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. By non-degenerate we mean that Y is not almost surely a constant. We say X is in the max-domain of attraction of Y and Y has a property called max-stability. Max-stability means there exist constants $c_n > 0, d_n \in \mathbb{R}$ such that

$$c_n \cdot \max_{i=1,\dots,n} Y^{(i)} + d_n \stackrel{\mathcal{D}}{=} Y$$

for iid copies $Y^{(i)}$ of Y. $\stackrel{\mathcal{D}}{=}$ denotes that the left hand side and the right hand side follow the same distribution. By setting $a_n := c_n$ and $b_n := d_n$ we see that a max-stable distribution is always in its own max-domain of attraction. One example of a max-stable distribution is the standard Fréchet one with

$$P(Y \leq 1/y) = \exp(-y)$$

for all $y > 0$. One can easily check that for all n we have

$$\frac{1}{n} \cdot \max_{i=1,\dots,n} Y^{(i)} \stackrel{\mathcal{D}}{=} Y,$$

where $Y^{(i)}$ are iid copies of Y.

Because this thesis is not about univariate extreme value theory, we will not go into detail about the following facts:

- The family of non-degenerate max-stable value distributions is a parametric family. One parameter is for the shape, while the other two parameters are about additive shift und multiplicative scale. In the literature this is known as the Fisher–Tippett–Gnedenko theorem (see Theorem 1.1.3 in the book by de Haan and Ferreira (2006))
- If we ignore shift and scale, then a random variable is in the max-domain of attraction of at most one non-degenerate max-stable distribution.
- Not every random variable is in the max-domain of attraction of a max-stable distribution (see Example 2.6.1 in the book by Galambos (1978)).

Also we refer to the case studies in the book by de Haan and Ferreira (2006) on how the max-domain of attraction affects the questions in the beginning of this section.

In multivariate EVT we are interested in the simultaneous appearance of extremes in more than one component of a random vector $\mathbf{X} = (X_1, \dots, X_d)^\intercal$. There is a natural extension of the univariate max-domain of attraction, which we will call the multivariate max-domain of attraction.

For this it is necessary that there are positive scaling constants a_{jn} and shifting constants b_{jn} such that

$$a_{jn} \cdot \max_{i=1,\dots,n} X_j^{(i)} + b_{jn} \xrightarrow{\mathcal{D}} Y_j$$

as $n \to \infty$ for all $j = 1, \ldots, d$, where Y_j are max-stable univariate random variables and $X_j^{(i)}$ are iid copies of X_j. If it turns out that $Y_1, \ldots, Y_d$ are the components of a random vector $\mathbf{Y}$ and we have the convergence

$$\left[a_{jn} \cdot \max_{i=1,\ldots,n} X_j^{(i)} + b_{jn} \right]_{j=1}^{d} \xrightarrow{\mathcal{D}} \mathbf{Y},$$

as $n \to \infty$, where the $\mathbf{X}^{(i)}$ are iid copies of $\mathbf{X}$, then we say $\mathbf{X}$ is in the multivariate max-domain of attraction of $\mathbf{Y}$.

One example of a multivariate limit distribution might be that the limit $\mathbf{Y}$ has independent components $Y_1, \ldots, Y_d$. Then the components $X_1, \ldots, X_d$ are called tail-independent.

Another example of a limit distribution might be a limit $\mathbf{Y}$ wich $Y_1 = Y_2 = \cdots = Y_d$ almost surely. Then the components $X_1, \ldots, X_d$ have the strongest tail-dependence there is.

If we transform the margins of $\mathbf{Y}$ to be standard Fréchet distributed and call the result $\mathbf{Y}^*$ and we transform the margins of $\mathbf{X}$ to be standard Pareto and call the result $\mathbf{X}^*$, then we get

$$\frac{1}{n} \cdot \max_{i=1,\ldots,n} (\mathbf{X}^*)^{(i)} \xrightarrow{\mathcal{D}} \mathbf{Y}^*,$$

as $n \to \infty$, where $(\mathbf{X}^*)^{(i)}$ are iid copies of $\mathbf{X}^*$. We have the max-stability

$$\frac{1}{n} \cdot \max_{i=1,\ldots,n} (\mathbf{Y}^*)^{(i)} \overset{\mathcal{D}}{=} \mathbf{Y}^*, \tag{1.1}$$

for all $n \in \mathbb{N}$, where $(\mathbf{Y}^*)^{(i)}$ are iid copies of $\mathbf{Y}^*$. The details of these results can be looked up in the book by Resnick (1987). That's why there is nothing lost by only investigating max-stable distributions with standard Fréchet margins.

Definition 1. *We will call a max-stable distribution with standard Fréchet margins a simple max-stable distribution.*

Every simple max-stable distribution can be characterized by a special type of norm. Because this norm entails a dependence structure we refer to it as a D-norm. Further details and a proof of Theorem 1 below can be found in the book by Falk (2019).

Definition 2. *A d-dimensional random vector $\boldsymbol{Z}$ with $\mathbb{E}(Z_j) = 1$ and $Z_j \geq 0$ almost surely for all $j = 1, \ldots, d$ is called a D-norm generator and the corresponding D-norm $\|\cdot\|_D$ is defined by*

$$\|\boldsymbol{x}\|_D := \mathbb{E}\left(\max_{j=1,\ldots,d} |x_j| Z_j \right)$$

for all $\boldsymbol{x} \in \mathbb{R}^d$. If there are two D-norm generators $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ with

$$\mathbb{E}\left(\max_{j=1,\ldots,d} |x_j| Z_j^{(1)} \right) = \mathbb{E}\left(\max_{j=1,\ldots,d} |x_j| Z_j^{(2)} \right) \tag{1.2}$$

for all $\boldsymbol{x} \in \mathbb{R}^d$ *then we say* $\boldsymbol{Z}^{(1)}$ *and* $\boldsymbol{Z}^{(2)}$ *generate the same D-norm.*

The most basic instances of different D-norm generators that generate the same D-norm are given in the following example:

Example 3. *If* $\boldsymbol{Z}$ *generates a D-norm and* X *is a random variable independent of* $\boldsymbol{Z}$*, with* $X \geq 0$ *almost surely and with expectation* $\mathbb{E}(X) = 1$*, then* $\boldsymbol{Z}$ *and* $X \cdot \boldsymbol{Z}$ *generate the same D-norm. This is shown immediately by the following equation:*

$$\begin{aligned}\mathbb{E}\left(\max_{j=1,\dots,d} |x_j|(X \cdot Z_j)\right) &= \mathbb{E}\left(X \cdot \max_{j=1,\dots,d} |x_j| \cdot Z_j\right) \\ &= \underbrace{\mathbb{E}(X)}_{=1} \cdot \mathbb{E}\left(\max_{j=1,\dots,d} |x_j| \cdot Z_j\right),\end{aligned}$$

which holds for all $\boldsymbol{x} \in \mathbb{R}^d$*. The second step comes from the fact that* X *is independent of* $\max_{j=1,\dots,d} |x_j| \cdot Z_j$ *and that the expected value of products of independent random variables is the product of the individual expected values.*

While one does not need to know anything about extreme value theory to understand D-norms, they have a central place in the theory of multivariate max-stability as we will see in the following theorem:

Theorem 1. *A random vector* $\boldsymbol{Y}$ *is simple max-stable if and only if there exists a D-norm* $\|\cdot\|_D$ *such that*

$$P\left(\boldsymbol{Y} \leq \frac{1}{\boldsymbol{y}}\right) = \exp(-\|\boldsymbol{y}\|_D) \tag{1.3}$$

for all $\boldsymbol{y} > \boldsymbol{0}$*.*

Note that while Equation (1.3) only covers the multivariate cumulative distribution function $F(\mathbf{x}) = P(\mathbf{Y} \leq \mathbf{x})$ for positive vectors $\mathbf{x}$, we also have $F(\mathbf{x}) = 0$ for all other vectors $\mathbf{x}$ for monotonicity reasons. This is used several times implicitely throughout the thesis.

So far we have defined the multivariate max-domain of attraction, but we have not seen how the max-stable attractor $\mathbf{Y}$ gives us informations about a random vector $\mathbf{X}$ in its max-domain of attraction. The next section will change that.

1.2 The Rosetta Stone theorem

In Ptolemaic Egypt a royal decree was carved into a stone. Part of the stone was discovered in Rosetta (Rashid, Egypt). The same content was written in three languages and for this reason the Rosetta Stone was chosen as the name-sake for the following theorem, which gives several equivalent definitions of the same multivariate max-domain attraction. Not only is this a compact translation guide for multivariate extreme value theory, but also a toolbox, because depending on the context one characterization is more useful or intuitive than another.

For one characterization we need the following definition:

Definition 3. *A function $h : [0,\infty)^d \to [0,\infty)$ is called homogeneous of order 1 if it fulfills*

$$h(\lambda \cdot \boldsymbol{x}) = \lambda \cdot h(\boldsymbol{x})$$

for all $\lambda \geq 0$ and all $\boldsymbol{x} \in [0,\infty)^d$.

Example 4 (Examples of homogeneous functions)**.** *The following functions are homogeneous of order 1:*

- $h(x_1,\ldots,x_d) = \max_{j=1,\ldots,d} x_j$.
- $h(x_1,\ldots,x_d) = \min_{j=1,\ldots,d} x_j$.
- $h(x_1,\ldots,x_d) = \max_{j=1,\ldots,d} x_j \cdot |y_j|$, *where $\boldsymbol{y}$ is a fixed vector.*
- $h(x_1,\ldots,x_d) = \min_{j=1,\ldots,d} x_j \cdot |y_j|$, *where $\boldsymbol{y}$ is a fixed vector.*
- $h(x_1,\ldots,x_d) = x_j$, *where j is a fixed index from the set $\{1,\ldots,d\}$.*
- $h(x_1,\ldots,x_d) = \sum_{j=1}^{d} x_j$.
- $h(x_1,\ldots,x_d) = \max(x_i, x_j)$, *where i and j are fixed indices from the set $\{1,\ldots,d\}$.*
- $h(x_1,\ldots,x_d) = \sqrt{x_i x_j}$, *where i and j are fixed indices from the set $\{1,\ldots,d\}$.*
- $h(x_1,x_2) := x_1 \cdot 1_{x_1 > x_2} = \begin{cases} x_1 & \text{if } x_1 > x_2 \\ 0 & \text{else.} \end{cases}$

The last example shows that homogeneous functions need not be continuous.

Example 4 illustrates how flexible the class of homogeneous functions is. But an individual function of this class always turns a random vector $\mathbf{X} \geq \mathbf{0}$ into a random variable $h(\mathbf{X}) \geq 0$. The Rosetta Stone theorem explains how the multivariate tail-behavior of $\mathbf{X}$ affects the univariate tail-behavior of $h(\mathbf{X})$ and vice versa.

Theorem 2 (Rosetta Stone)**.** *Let $\boldsymbol{Z}$ be the generator of a D-norm $\|\cdot\|_D$ and let $\boldsymbol{X}$ be a random vector on $[0,\infty)^d$ with joint cumulative distribution function F. Then there exists a uniquely determined measure ν on the set $[0,\infty)^d \setminus \{\boldsymbol{0}\}$ that fulfills*

$$\nu\left(\left[\boldsymbol{0}, \frac{1}{\boldsymbol{y}}\right]^{\complement}\right) = \mathbb{E}\left(\max_{j=1,\dots,d} y_j Z_j\right) = \|\boldsymbol{y}\|_D$$

for all $\boldsymbol{y} > \boldsymbol{0}$ and the following five statements are equivalent:

(i)

$$\lim_{t\to\infty} t \cdot P(h(\boldsymbol{X}) > t) = \mathbb{E}(h(\boldsymbol{Z}))$$

for all non-negative, continuous functions h, that are homogeneous of order 1.

(ii)

$$\lim_{t\to\infty} t \cdot P\left(\max_{j=1,\dots,d} y_j X_j > t\right) = \mathbb{E}\left(\max_{j=1,\dots,d} y_j Z_j\right) = \|\boldsymbol{y}\|_D$$

for all $\boldsymbol{y} \geq \boldsymbol{0}$.

(iii)

$$\lim_{n\to\infty} F^n\left(n \cdot \frac{1}{\boldsymbol{y}}\right) = \exp\left(-\mathbb{E}\left(\max_{j=1,\dots,n} |y_j| Z_j\right)\right) = \exp(-\|\boldsymbol{y}\|_D)$$

for all $\boldsymbol{y} > \boldsymbol{0}$.

(iv)

$$\frac{1}{n} \cdot \max_{i=1,\dots,n} \boldsymbol{X}^{(i)} \xrightarrow{\mathcal{D}} \boldsymbol{Y},$$

as $n \to \infty$, where the maximum is meant component-wise, $\boldsymbol{X}^{(i)}$ are iid copies of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is a max-stable random vector with standard Fréchet margins and

$$P\left(\boldsymbol{Y} \leq \frac{1}{\boldsymbol{y}}\right) = \exp\left(-\mathbb{E}\left(\max_{j=1,\dots,n} y_j Z_j\right)\right) = \exp(-\|\boldsymbol{y}\|_D)$$

for all $\boldsymbol{y} > \boldsymbol{0}$.

(v)

$$\lim_{t\to\infty} t \cdot P(\boldsymbol{X} \in t \cdot M) = \nu(M) \tag{1.4}$$

for all measurable $M \subset [0,\infty)^d \setminus \{\boldsymbol{0}\}$ with $\nu(\partial M) = 0$, where ∂M is the topological boundary of M.

Also if $\|.\|$ is an arbitrary norm, such that $\|\boldsymbol{Z}\| = c$ holds almost surely, then the five statements further imply:

(vi) We have the weak limit

$$\lim_{t\to\infty} P\left(c\cdot\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}\in M\middle|\,\|\boldsymbol{X}\|>t\right)=P(\boldsymbol{Z}\in M) \tag{1.5}$$

for all measurable sets M with $P(\boldsymbol{Z}\in\partial M)=0$.

The Rosetta Stone theorem shows that there are several equivalent characterizations for the same multivariate max-domain of attraction. The equivalences (ii) $\Leftrightarrow$(iii) $\Leftrightarrow$ (iv) $\Leftrightarrow$ (v) are well known, see e.g. Resnick (1987).

The author has found no prior appearance of characterization (i), but this limit was at least observed for some homogeneous functions (see e.g. Jessen and Mikosch (2006) and Segers (2012)).

In the following we will call Statements (i)-(v) the equivalent statements of the Rosetta Stone theorem. Statement (vi) is slightly weaker than the rest as we can see in the following example:

Example 5. *Let $\boldsymbol{X}=(X_1,X_2)$ be a random vector that fulfills $X_1=X_2$ almost surely and $P(X_1>t)=\frac{1}{\sqrt{t}}$ for all $t\geq 1$. Further let $\boldsymbol{Z}$ be a D-norm generator with $\boldsymbol{Z}=(1,1)^\intercal$ almost surely.*

First we will show that Statement (vi) of Theorem 2 holds:

For an arbitrary norm $\|\cdot\|$ we have $\|\boldsymbol{Z}\|=\|(1,1)^\intercal\|=:c$ almost surely and consequently

$$c\cdot\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}=c\cdot\frac{X_1\cdot(1,1)^\intercal}{\|X_1\cdot(1,1)^\intercal\|}=(1,1)^\intercal$$

almost surely. For all measurable sets M and all thresholds t this implies

$$P\left(c\cdot\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}\in M\middle|\,\|\boldsymbol{X}\|>t\right)=\begin{cases}1 \text{ if } (1,1)^\intercal\in M\\ 0 \text{ else,}\end{cases}$$

which is exactly the same as

$$P(\boldsymbol{Z}\in M)=\begin{cases}1 \text{ if } (1,1)^\intercal\in M\\ 0 \text{ else.}\end{cases}$$

As a trivial consequence the limit in Equation (1.5) *holds and so Statement* (vi) *of the Rosetta Stone theorem is fulfilled. However Statement (i) is violated, as we have*

$$\lim_{t\to\infty} t\cdot P(X_1>t)=\lim_{t\to\infty}\sqrt{t}=\infty\neq\mathbb{E}(Z_1).$$

Nonetheless Statement (vi) is an important part of the Rosetta Stone theorem because it opens the way for the peaks-over-threshold approach to infer the extremal dependence structure (see Section 3.1).

The proof of the Rosetta Stone theorem can be found in Section 1.3.

An immediate consequence of the Rosetta Stone theorem is the following:

Corollary 1 (Additivity in the Tail)**.** *Let $\boldsymbol{X}$ be a random vector that fulfills one of the equivalent statements of the Rosetta Stone theorem. If $h_1, \ldots, h_n$ are non-negative, continuous functions that are homogeneous of order 1, then the following equation holds:*

$$\lim_{t\to\infty} t \cdot P\left(\sum_{i=1}^{n} h_i(\boldsymbol{X}) > t\right) = \sum_{i=1}^{n} \lim_{t\to\infty} t \cdot P(h_i(\boldsymbol{X}) > t)$$

As a special case we get that

$$\lim_{t\to\infty} t \cdot P\left(\sum_{j=1}^{d} X_j > t\right) = d$$

regardless of the underlying dependence structure.

Proof. $\mathbf{x} \mapsto \sum_{i=1}^{n} h_i(\mathbf{x})$ defines a function that obviously is non-negative, continuous and homogeneous of order 1, so we can apply the Rosetta Stone theorem to it. If $\mathbf{Z}$ is a generator of the underlying D-norm, we get:

$$\begin{aligned}\lim_{t\to\infty} t \cdot P\left(\sum_{i=1}^{n} h_i(\mathbf{X}) > t\right) =& \mathbb{E}\left(\sum_{i=1}^{n} h_i(\mathbf{Z})\right) \\ &= \sum_{i=1}^{n} \mathbb{E}(h_i(\mathbf{Z})) = \sum_{i=1}^{n} \lim_{t\to\infty} t \cdot P(h_i(\mathbf{X}) > t).\end{aligned}$$

By choosing $n = d$ and $h_i(\mathbf{x}) = x_i$, we get the special case:

$$\lim_{t\to\infty} t \cdot P\left(\sum_{j=1}^{d} h_j(\mathbf{X}) > t\right) = \sum_{j=1}^{d} \lim_{t\to\infty} P(h_j(\mathbf{X}) > t) = \sum_{j=1}^{d} \mathbb{E}(h_j(\mathbf{Z})) = \sum_{j=1}^{d} \underbrace{\mathbb{E}(Z_j)}_{=1},$$

as by definition a D-norm generator fulfills $\mathbb{E}(Z_j) = 1$ for all $j = 1, \ldots, d$. □

Even though in the setting of the Rosetta Stone theorem the functions h have the trivial minumum 0 and supremum ∞ (as long as h is not the constant 0), it is quite useful to introduce the following notation:

Definition 4. *Let h be a non-negative function on $[0, \infty)^d$ that is homogeneous of order 1. Then we set*

$$h_{\max} := \max\left\{h(\boldsymbol{x}) : \boldsymbol{x} \geq \boldsymbol{0}, \sum_{j=1}^{d} x_j = d\right\} \text{ and}$$

$$h_{\min} := \min\left\{h(\boldsymbol{x}) : \boldsymbol{x} \geq \boldsymbol{0}, \sum_{j=1}^{d} x_j = d\right\}.$$

Effectively we maximize and minimize h on a compact subset, which seems arbitrary in nature, but there is a thought behind that.

Corollary 2. *We have*

$$\lim_{t\to\infty} t \cdot P(h(\boldsymbol{X}) > t) \in [h_{\min}, h_{\max}]$$

for all random vectors $\boldsymbol{X}$ *that fulfills one of the equivalent statements of the Rosetta Stone theorem and all* h *that are non-negative, continuous and homogeneous of order* 1*.*

Proof. According to Corollary 4 below there exists a generator $\mathbf{Z}$ which fulfills $Z_1 + \cdots + Z_d = d$ almost surely and which generates the underlying D-norm. For this generator the random variable $h(\mathbf{Z})$ almost surely falls into the interval $[h_{\min}, h_{\max}]$, which implies

$$\lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t) = \mathbb{E}(h(\mathbf{Z})) \in [h_{\min}, h_{\max}].$$

□

The notation $h_{\max}$ and $h_{\min}$ will turn out to be useful for investigating the performance of estimators for the quantity $\mathbb{E}(h(\mathbf{Z}))$ in Section 3.1.

The following theorem is also very useful:

Corollary 3. *Let* $\boldsymbol{Z}$ *be a D-norm generator and let* U *be uniformly distributed on the interval* $(0, 1)$ *and independent of* $\boldsymbol{Z}$*. Then the random vector*

$$\boldsymbol{X} := \frac{1}{U} \cdot \boldsymbol{Z} = (Z_1/U, Z_2/U, \ldots, Z_d/U)^\intercal$$

fulfills the equivalent statements of the Rosetta Stone theorem with D-norm generators $\boldsymbol{Z}$*.*

Proof. According to the Rosetta Stone theorem it is sufficient to show the limit

$$\lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t) = \mathbb{E}(h(\mathbf{Z}))$$

for all non-negative, continuous h that are homogeneous of order 1. Using the homogeneity of h and Fubini's theorem we get:

$$\begin{aligned} t \cdot P(h(\mathbf{X}) > t) = t \cdot P(h(\mathbf{Z}) > t \cdot U) &= t \cdot \int_0^1 P(h(\mathbf{Z}) > t \cdot u)\,\mathrm{d}u \\ &= \int_0^t P(h(\mathbf{Z}) > s)\,\mathrm{d}s, \end{aligned}$$

where the last step comes from the substitution $u \mapsto s = t \cdot u$. However, $\mathbb{E}(h(\mathbf{Z})) = \int_0^\infty P(h(\mathbf{Z}) > s)\,\mathrm{d}s$ is the natural limit of this term, as $t \to \infty$. □

Corollary 3 implies that as long we can simulate the D-norm generator, we can also simulate a random vector in the corresponding max-domain of attraction. In Section 3.6 we will use this for a simulation study and in Section 3.4 we will use this for some helpful examples.

In the next section we will prove the Rosetta Stone theorem and in Section 1.5 we will talk about when and how the Rosetta Stone theorem is actually applicable in practice. The whole of Chapter 2 is about applications of the Rosetta Stone theorem.

1.3 Proof of the Rosetta Stone theorem

Before we can prove the Rosetta Stone theorem we need some preparatory results.

Theorem 3 below can be motivated by Example 3. In the example we realized that if $\mathbf{Z}$ generates a D-norm, then $\mathbf{Z}' = X \cdot \mathbf{Z}$ generates the same D-norm, as long as $X \geq 0$ is independent of $\mathbf{Z}$ and has expectation $E(X) = 1$.

In that case we have for every non-negative, measurable function h that is homogeneous of order 1 the following equation:

$$E(h(\mathbf{Z}')) = \mathbb{E}(X \cdot h(\mathbf{Z})) = \mathbb{E}(X) \cdot \mathbb{E}(h(\mathbf{Z})) = \mathbb{E}(h(\mathbf{Z})).$$

This is a special case of a more general result:

Theorem 3. *Let $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}$ be two D-norm generators. Then the following two statements are equivalent:*

(i) $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ generate the same D-norm.

(ii) For every non-negative, measurable function h that is homogeneous of order 1 we have

$$\mathbb{E}(h(\boldsymbol{Z}^{(1)})) = \mathbb{E}(h(\boldsymbol{Z}^{(2)})).$$

This result implies the values $\mathbb{E}(h(\mathbf{Z}))$ are uniquely determined by the D-norm $\|\cdot\|_D$, which itself is uniquely determined by a max-stable random vector, which itself is uniquely determined by an arbitrary random vector $\mathbf{X}$ in its max-domain of attraction. This suggests that there is a link between the tail-behavior of the random vector $\mathbf{X}$ and the constants $\mathbb{E}(h(\mathbf{Z}))$. This is indeed the case, as can be seen in Statement (i) the Rosetta Stone theorem under the additional assumption that h is continuous.

To proof this theorem we need a preparatory result:

Lemma 1. *Let $\boldsymbol{Z}$ be a D-norm generator. Then there exists a measure ν on the set $[0,\infty)^d \setminus \{\boldsymbol{0}\}$ with the property*

$$\mathbb{E}(h(\boldsymbol{Z})) = \nu(\{\boldsymbol{x} : h(\boldsymbol{x}) > 1\}) = \nu(\{\boldsymbol{x} : h(\boldsymbol{x}) \geq 1\})$$

for every non-negative, measurable function h that is homogeneous of order 1.

Further the measure ν has the property

$$\nu(c \cdot M) = \frac{1}{c} \cdot \nu(M) \tag{1.6}$$

for $c > 0$ and every measurable set M.

Proof. Let $T : (0,\infty) \times [0,\infty)^d \to [0,\infty)^d \setminus \{\mathbf{0}\}$ be the continuous transformation

$$T(s, \mathbf{z}) = \frac{1}{s} \cdot \mathbf{z}.$$

Further let λ be the Lebesgue measure on $(0, \infty)$ and let P be the probability measure of $\mathbf{Z}$ on $[0, \infty)^d$. Then we can introduce the measure ν by

$$\nu(M) := (\lambda \times P)(T^{-1}(M)),$$

where $T^{-1}(M)$ is the pre-image of M under the transformation T and $(\lambda \times P)$ denotes the product measure.

Now let h be a non-negative, measurable function that is homogeneous of order 1. By using indicator functions and Fubini's theorem we get

$$\begin{aligned}
\nu(\{\mathbf{x} : h(\mathbf{x} > 1\}) &= \int \int_0^\infty 1(h \circ T(s, \mathbf{z}) > 1) \, \mathrm{d}s \, \mathrm{d}P(\mathbf{Z} = \mathbf{z}) \\
&= \int \int_0^\infty 1(h(\mathbf{z}) > s) \, \mathrm{d}s \, \mathrm{d}P(\mathbf{Z} = \mathbf{z}) \qquad (1.7) \\
&= \int \int_0^{h(\mathbf{z})} 1 \, \mathrm{d}s \, \mathrm{d}P(\mathbf{Z} = \mathbf{z}) = \mathbb{E}(h(\mathbf{Z})).
\end{aligned}$$

Note that in Equation (1.7) we can replace all $>$ signs by $\geq$ signs without changing the resulting value $\mathbb{E}(h(\mathbf{Z}))$.

Equation (1.6) is a consequence of the following:

$$\begin{aligned}
\nu(c \cdot M) &= (\lambda \times P) \left(\left\{ (s, \mathbf{x}) : \frac{1}{s} \cdot \mathbf{x} \in c \cdot M \right\} \right) \\
&= (\lambda \times P) \left(\left\{ (s, \mathbf{x}) : \frac{1}{c \cdot s} \cdot \mathbf{x} \in M \right\} \right) \\
&= (\lambda \times P) \left(\left\{ (s/c, \mathbf{x}) : \frac{1}{s} \cdot \mathbf{x} \in M \right\} \right) \\
&= (\lambda/c \times P) \left(\left\{ (s, \mathbf{x}) : \frac{1}{s} \cdot \mathbf{x} \in M \right\} \right) \\
&= \frac{1}{c} \cdot (\lambda \times P) \left(\left\{ (s, \mathbf{x}) : \frac{1}{s} \cdot \mathbf{x} \in M \right\} \right) = \frac{1}{c} \cdot \nu(M).
\end{aligned}$$

□

With that result we can proceed to prove Theorem 3.

Proof of Theorem 3. The direction (ii) ⇒ (i) is obvious. By definition generating the same D-norm means

$$\mathbb{E}\left(\max_{j=1,\ldots,d} |x_j| Z_j^{(1)} \right) = \mathbb{E}\left(\max_{j=1,\ldots,d} |x_j| Z_j^{(2)} \right)$$

for every vector $\mathbf{x} \in \mathbb{R}^d$, which is implied by (ii) with the specific choices $h_{\mathbf{x}}(\mathbf{z}) := \max_{j=1,\ldots,d} |x_j| z_j$.

For the direction (i)⇒(ii) let $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ be two generators that generate the same D-norm. From Lemma 1 we already know of the existence of two measures ν_1, ν_2 with the properties

$$\mathbb{E}(h(\mathbf{Z}^{(1)})) = \nu_1(\{\mathbf{x} : h(\mathbf{x}) \geq 1\})$$
$$\mathbb{E}(h(\mathbf{Z}^{(2)})) = \nu_2(\{\mathbf{x} : h(\mathbf{x}) \geq 1\})$$

for every non-negative, measurable function h, that is homogeneous of order 1. So to show (ii) it is sufficient to prove that the two measure ν_1 and ν_2 coincide.

As $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ generate the same D-norm $\|\cdot\|_D$ we get:

$$\begin{aligned}\nu_1\left([\mathbf{0},\mathbf{x})^\complement\right) &= \nu_1\left(\left\{\mathbf{y} \geq \mathbf{0} : \max_{j=1,\dots,d} 1/x_j \cdot y_j \geq 1\right\}\right) \\ &= \mathbb{E}\left(\max_{j=1,\dots,d} 1/x_j \cdot Z_j^{(1)}\right) = \left\|\frac{1}{\mathbf{x}}\right\|_D = \mathbb{E}\left(\max_{j=1,\dots,d} 1/x_j \cdot Z_j^{(2)}\right) \\ &= \nu_2\left(\{\mathbf{y} \geq \mathbf{0} : \max_{j=1,\dots,d} 1/x_j \cdot y_j \geq 1\}\right) = \nu_2\left([\mathbf{0},\mathbf{x})^\complement\right)\end{aligned}$$

for all $\mathbf{x} > \mathbf{0}$. The proof of Lemma 21 below shows that this result implies $\nu_1([\mathbf{y},\mathbf{z})) = \nu_2([\mathbf{y},\mathbf{z}))$ for every $\mathbf{y} \geq \mathbf{0}, \mathbf{y} \neq \mathbf{0}$ and $\mathbf{z} \geq \mathbf{y}$. Both measures ν_1, ν_2 are continuous from below, so we get the following:

$$\begin{aligned}\nu_1([\mathbf{y},\infty)) &= \lim_{n\to\infty} \nu_1([\mathbf{y}, \mathbf{y} + n \cdot \mathbf{1}]) \\ &= \lim_{n\to\infty} \nu_2([\mathbf{y}, \mathbf{y} + n \cdot \mathbf{1}]) = \nu_2([\mathbf{y},\infty))\end{aligned}$$

for all $\mathbf{y} \geq \mathbf{0}, \mathbf{y} \neq \mathbf{0}$.

At this point it should be obvious that $\nu_1 = \nu_2$ the same way two random vectors follow the same probability distribution if they have the same multivariate cumulative distribution function. Thus

$$\mathbb{E}(h(\mathbf{Z}^{(1)})) = \nu_1(\{\mathbf{x} : h(\mathbf{x}) \geq 1\}) = \nu_2(\{\mathbf{x} : h(\mathbf{x}) \geq 1\}) = \mathbb{E}(h(\mathbf{Z}^{(2)}))$$

holds for all non-negative measurable functions h that are homogeneous of order 1. □

For a norm $\|\cdot\|$ let us call it the generator-finding-problem to find a D-norm generator $\mathbf{Z} \geq \mathbf{0}$, such that

$$\|\mathbf{x}\| = \mathbb{E}\left(\max_{j=1,\dots,d} |x_j| \cdot Z_j\right)$$

holds for all $\mathbf{x} \in \mathbb{R}^d$ is a mathematical problem. By definitions the D-norms are exactly the norms, where the generator-finding-problem is solvable. The solution is not unique in the sense that there are always at least two solutions $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ with $\mathbf{Z}^{(1)} \overset{\mathcal{D}}{\neq} \mathbf{Z}^{(2)}$, see Example 3. The following result shows us that we can impose a constraint on the generators $\mathbf{Z}$ in the generator-finding-problem and still get a solution, which is now unique (in distribution).

Lemma 2. *Let $\|\cdot\|_D$ be a D-norm and let $\|\cdot\|$ be an arbitrary norm. Then this D-norm has a generator $\boldsymbol{Z}_B$ with $\|\boldsymbol{Z}_B\| = const$ almost surely. Both the constant and the distribution of the normed generator $\boldsymbol{Z}$ are uniquely determined by the D-norm.*

Also we have

$$P(Z_B \in A) = \nu((1,\infty)\cdot A)) \tag{1.8}$$

for every measurable subset

$$A \subset B := \{\boldsymbol{x} : \boldsymbol{x} \in [0,\infty)^d, \|\boldsymbol{x}\| = c\}$$

and where ν is the measure from Lemma 1.

Proof. The main bulk of this lemma is the so-called normed-generators theorem (1.7.1) in the book by Falk (2019). What remains is the proof of Equation (1.8). Let P_B be the probability distribution of $\mathbf{Z}_B$. According to the definition of ν in the proof of Lemma 1 we have

$$\begin{aligned}
\nu((1,\infty)\cdot A) &= (\lambda \times P_B)\left(\left\{(s,\mathbf{x}) : \frac{1}{s}\cdot\mathbf{x} \in (1,\infty)\cdot A\right\}\right) \\
&= (\lambda \times P_B)\left(\left\{(s,\mathbf{x}) : \frac{1}{s}\cdot\mathbf{x} \in (1,\infty)\cdot A, \|\mathbf{x}\| = c\right\}\right) \\
&\quad + \underbrace{(\lambda \times P_B)\left(\left\{(s,\mathbf{x}) : \frac{1}{s}\cdot\mathbf{x} \in (1,\infty)\cdot A, \|\mathbf{x}\| \neq c\right\}\right)}_{=0} \\
&= (\lambda \times P)((0,1)\cdot A) = P_B(A) = P(\mathbf{Z}_B \in A)
\end{aligned}$$

□

As straightforward consequence of this lemma is the following result:

Corollary 4. *Let $\|\cdot\|_D$ be a D-norm. Then it has a generator $\boldsymbol{Z}$ with the property $\sum_{j=1}^d Z_j = d$ almost surely.*

Proof. $\|\mathbf{x}\|_1 = \sum_{j=1}^d |x_j|$ defines a norm (the so-called Manhattan-norm), so we can use Lemma 2 to find a generator $\mathbf{Z}$ that fulfills $\|\mathbf{Z}\|_1 = c$ for some positive constant c. Because D-norm generators fulfill $\mathbb{E}(Z_j) = 1$ for all $j = 1,\ldots,d$, the constant c is equal to the number of dimensions d. □

The following topological result is crucial for multivariate peaks-over-threshold methods and is necessary to show (vi) in the Rosetta Stone theorem.

Lemma 3. *Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^d$ and let A be Borel subset of the normed surface $B = \{\boldsymbol{x} : \boldsymbol{x} \in [0,\infty)^d, \|\boldsymbol{x}\| = 1\}$. For the set*

$$M_A := (1,\infty)\cdot A = \left\{\boldsymbol{x} : \boldsymbol{x} \in [0,\infty)^d, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \in A, \|\boldsymbol{x}\| > 1, \right\}$$

we have

$$\partial M_A \subset B \cup (1,\infty) \cdot \left[\overline{A} \cap \overline{B \setminus A}\right],$$

where $\overline{M}$ *denotes the topological closure of a set* M *and* ∂M *denotes the topological boundary of the set* M*.*

Proof. Let $\mathbf{x} \in \partial M_A$. If $\|\mathbf{x}\| = 1$, then $\mathbf{x} \in B$. So assume $\|\mathbf{x}\| > 1$. Then there exist two sequences $(\mathbf{x}_n)_{n\in\mathbb{N}}$, $(\mathbf{y}_n)_{n\in\mathbb{N}}$ with

$$\mathbf{x} = \lim_{n\to\infty} \underbrace{\mathbf{x}_n}_{\in M_A} = \lim_{n\to\infty} \underbrace{\mathbf{y}_n}_{\notin M_A}$$

Because norms on finite dimensional vectorspaces are continuous functions we can assume without loss of generality $\|\mathbf{x}_n\| > 1, \|\mathbf{y}_n\| > 1$ for all $n \in \mathbb{N}$ and thus we have

$$\frac{\mathbf{x}}{\|\mathbf{x}\|} = \lim_{n\to\infty} \underbrace{\frac{\mathbf{x}_n}{\|\mathbf{x}_n\|}}_{\in A} = \lim_{n\to\infty} \underbrace{\frac{\mathbf{y}_n}{\|\mathbf{y}_n\|}}_{\in B\setminus A}.$$

This representation clearly shows $\mathbf{x} \in (1,\infty) \cdot [\overline{A} \cap \overline{B \setminus A}]$, whenever $\mathbf{x} \in \partial M_A \setminus B$. □

The last prepatory result has nothing to do with measure theory, but is useful nonetheless.

Lemma 4. *Let* $(x_n)_{n\in\mathbb{N}}$ *be a sequence of real numbers with* $\lim_{n\to\infty} x_n = x \in \mathbb{R}$*. Then we have the limit*

$$\lim_{n\to\infty} \left(1 + \frac{x_n}{n}\right)^n = \exp(x). \tag{1.9}$$

Proof. For every $y \in \mathbb{R}$ we have the limit

$$\begin{aligned}
&\lim_{n\to\infty} \left(1 + \frac{y}{n}\right)^n \\
&= \lim_{n\to\infty} \exp\left(n \cdot \log\left(1 + \frac{y}{n}\right)\right) \\
&= \exp\left(\lim_{n\to\infty} n \cdot \left(\frac{y}{n} + o\left(\frac{y}{n}\right)\right)\right) = \exp(y),
\end{aligned}$$

where we used the continuity of exponential function, the Taylor expansion of the logarithm around 1 and Landau notation. Now for every $\epsilon > 0$ we have for monotonicity reasons:

$$\limsup_{n\to\infty} \left(1 + \frac{x_n}{n}\right)^n \leq \limsup_{n\to\infty} \left(1 + \frac{x+\epsilon}{n}\right)^n = \exp(x+\epsilon).$$

With the same reasoning we have $\liminf_{n\to\infty} \left(1 + \frac{x_n}{n}\right)^n \geq \exp(x-\epsilon)$. Because this holds for every $\epsilon > 0$ Equation (1.9) is true. □

Now all pieces are in place to prove the Rosette Stone theorem.

Proof of the Rosetta Stone theorem. The implication (i) $\Rightarrow$ (ii) is obvious.

Next we prove the implication (ii) $\Rightarrow$ (iii). We have

$$\begin{aligned}
&\lim_{n\to\infty} F^n\left(n \cdot \frac{1}{\mathbf{y}}\right) \\
&= \lim_{n\to\infty} \left(1 - \frac{1}{n} \cdot n \cdot P\left(\max_{j=1,\ldots,d} |y_j| X_j > n\right)\right)^n = \exp\left(-E\left(\max_{j=1,\ldots,d} |y_j| Z_j\right)\right),
\end{aligned}$$

where the last step comes Lemma 4.

The conclusion (iii) $\Rightarrow$ (iv) is obvious.

The implication (iv) $\Rightarrow$ (v) is given by Proposition 5.17 in the book by Resnick (2008). In its proof the book leaves it to the reader to confirm that if Equation (1.4) holds for rectangular sets, then we already have vague convergence. The technical details will not be too different from the Lemmata 19 to 24 in thesis if we replace the random variables $\nu_n(M)$ in those Lemmata by the deterministic values $n \cdot P(\mathbf{X} \in n \cdot M)$ and realize that the convergence in probability then turns into convergence of real numbers.

For (v) $\Rightarrow$ (i) let h be a non-negative, continuous function that is homogeneous order 1 and set $M = \{\mathbf{x} : h(\mathbf{x}) > 1\}$. If we can show that $\nu(\partial M) = 0$, then we have

$$\lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t) = \lim_{t\to\infty} t \cdot P(\mathbf{X} \in t \cdot M) = \nu(M) = \mathbb{E}(h(\mathbf{Z})).$$

The last step of this equation comes from two previous results: Lemma 1 shows that there exists a measure ν^* with the property $\nu^*(M) = \mathbb{E}(h(\mathbf{Z}))$ and the proof of Theorem 3 implies that the measure this ν^* is the same as the measure ν of the Rosetta Stone theorem. The continuity of h further implies

$$\begin{aligned}
\partial M = \overline{M} \cap \overline{M^{\complement}} &= \{\mathbf{x} : h(\mathbf{x}) \geq 1\} \cap \{\mathbf{x} : h(\mathbf{x}) \leq 1\} \\
&= \{\mathbf{x} : h(\mathbf{x}) = 1\}.
\end{aligned}$$

Lemma 1 then reveals that

$$\nu(\partial M) = \nu(\{\mathbf{x} : h(\mathbf{x}) \geq 1\}) - \nu(\{\mathbf{x} : h(\mathbf{x}) > 1\}) = 0.$$

This shows that M is in fact a continuity set of ν.

In order to prove (v) $\Rightarrow$ (vi) let us define be the normed surface

$$B = \{\mathbf{x} : \mathbf{x} \in [0, \infty)^d, \|\mathbf{x}\| = c\}.$$

Let M be a Borel set with $P(\mathbf{Z}_B \in \partial M) = 0$. Put $A := M \cap B$. Define $M_A := (1, \infty) \cdot A$. Then according to Lemma 3, applied to the norm $\|\cdot\|/c$, we have

$$\partial M_A \subset A \cup (1,\infty) \cdot \overline{A} \cap \overline{B \setminus A},$$
$$\nu(\partial M_A) \leq \nu(A) + \nu((1,\infty) \cdot \overline{A} \cap \overline{B \setminus A}).$$

Let us show that the right hand side of this inequality is 0. We have

$$\nu(A) \leq \nu(B) = \nu(\{\mathbf{x} : \mathbf{x} \geq \mathbf{0}, \|\mathbf{x}\| / c \geq 1\}) - \nu(\{\mathbf{x} : \mathbf{x} \geq \mathbf{0}, \|\mathbf{x}\| / c > 1\}) = 0,$$

where again the last step is a result of Lemma 1. Thus,

$$\nu((1,\infty) \cdot \overline{A} \cap \overline{B \setminus A}) = P(\mathbf{Z}_B \in \overline{A} \cap \overline{B \setminus A})) \leq P(\mathbf{Z}_B \in \partial M) = 0.$$

As a consequence M_A is a continuity set of ν, and statement (v) becomes applicable to the limit

$$\begin{aligned}
&\lim_{t\to\infty} P\left(c \cdot \frac{\mathbf{X}}{\|\mathbf{X}\|} \in M \middle| \|\mathbf{X}\| > t \right) \\
&= \lim_{t\to\infty} P\left(c \cdot \frac{\mathbf{X}}{\|\mathbf{X}\|} \in M \middle| \|\mathbf{X}\| > ct \right) \\
&= \lim_{t\to\infty} P\left(c \cdot \frac{\mathbf{X}}{\|\mathbf{X}\|} \in A \middle| \|\mathbf{X}\| > ct \right) \\
&= \lim_{t\to\infty} \frac{P\left(\mathbf{X} \in t \cdot M_A\right)}{P(\|\mathbf{X}\| > ct)} \\
&= \lim_{t\to\infty} \frac{t \cdot P\left(\mathbf{X} \in t \cdot M_A\right)}{t \cdot P(\|\mathbf{X}\| > ct)} = \frac{\nu(M_A)}{1} = P(\mathbf{Z}_B \in A).
\end{aligned}$$

□

1.4 The topology of D-norms

This section is concerned with the point-wise convergence of D-norms. The Rosetta Stone theorem says that all information given by the D-norm $\|\cdot\|_D$ can also be found within the simple max-stable vector $\mathbf{Y}$ or within normed D-norm generators $\mathbf{Z}$.

In practice we can only work with surrogates $\widehat{\|\cdot\|}_D, \widehat{\mathbf{Y}}, \widehat{\mathbf{Z}}$ inferred from the data. The Rosetta Stone theorem only says, that a perfect D-norm fit $\widehat{\|\cdot\|}_D = \|\cdot\|_D$ is equivalent to a perfect max-stable fit $\widehat{\mathbf{Y}} \overset{\mathcal{D}}{=} \mathbf{Y}$, which itself is equivalent to a perfect normed generator fit $\widehat{\mathbf{Z}} \overset{\mathcal{D}}{=} \mathbf{Z}$. But in practice there are no perfect fits, so Theorem 4 is a natural extension of the Rosetta Stone theorem.

Theorem 4. *Let $(\|\cdot\|_{D_n})_{n\in\mathbb{N}}$ be a sequence of D-norms and let $\|\cdot\|_D$ be a D-norm as well.*

Then the following are equivalent:

(i) We have the pointwise limit $\|\boldsymbol{x}\|_{D_n} \to \|\boldsymbol{x}\|_D$ as $n \to \infty$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

(ii) We have the weak convergence $\boldsymbol{Y}^{(n)} \overset{\mathcal{D}}{\to} \boldsymbol{Y}$ as $n \to \infty$, where $\boldsymbol{Y}^{(n)}, n \in \mathbb{N}$ and $\boldsymbol{Y}$ are simple max-stable random vectors with

$$P\left(\boldsymbol{Y}^{(n)} \leq \frac{1}{\boldsymbol{y}}\right) = \exp(-\|\boldsymbol{y}\|_{D_n}) \text{ for all } n \in \mathbb{N} \text{ and}$$
$$P\left(\boldsymbol{Y} \leq \frac{1}{\boldsymbol{y}}\right) = \exp(-\|\boldsymbol{y}\|_D)$$

for all $\boldsymbol{y} > \boldsymbol{0}$.

(iii) We have the weak convergence $\boldsymbol{Z}^{(n)} \overset{\mathcal{D}}{\to} \boldsymbol{Z}$ as $n \to \infty$, where $\boldsymbol{Z}^{(n)}, n \in \mathbb{N}$ and $\boldsymbol{Z}$ are generators of the corresponding D-norms that realize almost surely on the set $B := \{\boldsymbol{x} : \boldsymbol{x} \geq \boldsymbol{0}, \sum_{j=1}^{d} x_j = d\}$.

(iv) For every h that is non-negative, continuous and homogeneous of order 1 we have the limit

$$\lim_{t\to\infty} t \cdot P\left(h\left(\boldsymbol{X}^{(n)}\right) > t\right) \to \lim_{t\to\infty} t \cdot P(h(\boldsymbol{X}) > t) \tag{1.10}$$

as $n \to \infty$, where $\boldsymbol{X}^{(n)}, n \in \mathbb{N}$ and $\boldsymbol{X}$ are random vectors on $[0,\infty)^d$ that fulfills one of the equivalent statements of the Rosetta Stone theorem with underlying D-norms $\|\cdot\|_{D_n}, n \in \mathbb{N}$ and $\|\cdot\|_D$.

Note that Theorem 1 guarantees the existence of the max-stable random vectors in Statement (ii). Also Corollary 4 implies the existence of the generators in Statement (iii).

For the following technicals details we have to recall the notion of tightness, when it comes to probability measures:

Definition 5. *A sequence of probability measures $(P_n)_{n\in\mathbb{N}}$ on $\mathbb{R}^d$ is called tight if for every $\epsilon > 0$ there exists a d-dimensional compact rectangle Q with $P_n(Q) \geq 1-\epsilon$ for all $n \in \mathbb{N}$. We will call a sequence of random vectors $(\boldsymbol{X}^{(n)})_{n\in\mathbb{N}}$ tight if the sequence of the corresponding probability measures is tight.*

Lemma 5. *For a tight sequence of random vectors $(\boldsymbol{X}^{(n)})_{n\in\mathbb{N}}$ there exists a random vector $\boldsymbol{X}$ and a subsequence $(\boldsymbol{X}^{(n_i)})_{i\in\mathbb{N}}$ that converges in distribution to $\boldsymbol{X}$.*

If there is a topologically closed set B with $\boldsymbol{X}^{(n)} \in B$ almost surely for all $n \in \mathbb{N}$, then this $\boldsymbol{X}$ also fulfills $\boldsymbol{X} \in B$ almost surely.

Proof. See Theorem 29.3 in the book by Billingsley (1979) for why for every tight sequence of probability measures there exists a random vector $\mathbf{X}$ and a subsequence $(\mathbf{X}^{(n_i)})_{i\in\mathbb{N}}$ that converges in distribution to $\mathbf{X}$.

The second assertion can be proven with Skorohod's theorem (see Theorem 29.6 in the book by Billingsley (1979)). It says we can assume without loss of generality that $(\mathbf{X}^{(n_i)})_{i\in\mathbb{N}}$ and $\mathbf{X}$ are from the same probability space and we have $P(\lim_{i\to\infty} \mathbf{X}^{(n_i)} = \mathbf{X}) = 1$. Now let B be a closed set. Then we have

$$P(\mathbf{X} \notin B) \leq P\left(\mathbf{X} \neq \lim_{i\to\infty} \mathbf{X}^{(n_i)}\right) + \sum_{i=1}^{\infty} P\left(\mathbf{X}^{(i)} \notin B\right),$$

because if a realization of $\mathbf{X}$ was the limit of elements of B it would be in B itself. If all the summands are 0, then we have $P(\mathbf{X} \notin B) = 0$. □

The following extension of Lemma 5 becomes important more than once later.

Lemma 6. *Let $\boldsymbol{X}$ be a random vector. If a tight sequence of random vectors $(\boldsymbol{X}^{(n)})_{n\in\mathbb{N}}$ does not converge to $\boldsymbol{X}$, then there exists a subsequence $(\boldsymbol{X}^{(n_i)})_{i\in\mathbb{N}}$ that converges to a random vector $\boldsymbol{X}^* \overset{\mathcal{D}}{\neq} \boldsymbol{X}$.*

Proof. Because $(\mathbf{X}^{(n)})_{n\in\mathbb{N}}$ does not converge to $\mathbf{X}$ there exists a continuous, bounded function $f : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}(f(\mathbf{X}^{(n)})) \not\to \mathbb{E}(f(\mathbf{X}))$. This means there exists an $\epsilon > 0$ and a subsequence $(\mathbf{X}^{(n_i')})_{i\in\mathbb{N}}$ with

$$|E(f(\mathbf{X}^{(n_i')})) - E(f(\mathbf{X}))| > \epsilon$$

for all $i \in \mathbb{N}$. This subsequence is tight itself, so there exists a $\mathbf{X}^*$ that is the limit of a sub-subsequence $(\mathbf{X}^{(n_i)})_{i\in\mathbb{N}}$.

Also we have

$$|E(f(\mathbf{X}^*)) - E(f(\mathbf{X}))| = \lim_{i\to\infty} |E(f(\mathbf{X}^{(n_i)})) - \mathbb{E}(f(\mathbf{X}))| \geq \epsilon,$$

which shows that $\mathbf{X}^*$ has a different distribution than $\mathbf{X}$. □

The following Lemma is interesting on its own.

Lemma 7. *The set of D-norms is sequentially compact, in the sense that for every sequence $(\|\cdot\|_{D_n})_{n\in\mathbb{N}}$ of D-norms there exists a D-norm $\|\cdot\|_D$ and a subsequence $(\|\cdot\|_{D_{n(i)}})_{i\in\mathbb{N}}$ with $\|\boldsymbol{x}\|_{D_{n(i)}} \to \|\boldsymbol{x}\|_D$ as $i \to \infty$ for all $\boldsymbol{x} \in \mathbb{R}^d$.*

Proof. According to Corollary 4 there exists a sequence of D-norm generators $(\mathbf{Z}^{(n)})_{n\in\mathbb{N}}$, each of which generates the corresponding element of the D-norm sequence, but which all fulfill $\mathbf{Z}^{(n)} \in B := \{\mathbf{x} : \mathbf{x} \geq \mathbf{0}, \sum_{j=1}^d x_j = d\}$ almost surely.

Because B is a bounded set it is very easy to confirm that $(\mathbf{Z}^{(n)})_{n\in\mathbb{N}}$ is a tight sequence. According to Lemma 5 there exists a vector $\mathbf{Z}$ and a subsequence $(\mathbf{Z}^{(n_i)})_{i\in\mathbb{N}}$ with $\mathbf{Z}^{(n_i)} \xrightarrow{\mathcal{D}} \mathbf{Z}$ as $i \to \infty$. Also according the this lemma we have $\mathbf{Z} \in B$ almost surely.

Now according to the Portmanteau lemma the convergence in distribution implies $\mathbb{E}(f(\mathbf{Z}^{(n_i)})) \to \mathbb{E}(f(\mathbf{Z}))$ for all bounded, continuous functions f. But it also holds for arbitrary continuous functions f, as the restriction of a continuous function on the compact set B is automatically bounded.

For the choice $f(\mathbf{z}) := z_j, j = 1, \ldots, d$ this implies $\mathbb{E}(Z_j) = \lim_{i\to\infty} \mathbb{E}(Z_j^{(n_i)}) = 1$ for all $j = 1, \ldots, d$, so $\mathbf{Z}$ generates a D-norm $\|\cdot\|_D$. For the choide $f(\mathbf{z}) = \max_{j=1,\ldots,d} |x_j| z_j$ for an arbitrary fixed vector $\mathbf{x}$ this implies

$$\|\mathbf{x}\|_D = \mathbb{E}\left(\max_{j=1,\ldots,d} |x_j| Z_j\right) = \lim_{i\to\infty} \mathbb{E}\left(\max_{j=1,\ldots,d} |x_j| Z_j^{(n_i)}\right) = \lim_{i\to\infty} \|\mathbf{x}\|_{D_{n_i}}.$$

We have shown that the set of D-norms is sequentially compact when it comes to point-wise convergence. □

We now have everything we need to prove the equivalences of Theorem 4.

Proof of Theorem 4. The equivalence (i) ⇔ (ii) is obvious.

To prove the implication (iii) ⇒ (iv) let h be an arbitrary, non-negative, continuous function that is homogeneous of order 1. Restricted on the compact set B it is also bounded and the Portmanteau lemma then implies convergence of the moments

$$\mathbb{E}\left(h\left(\mathbf{Z}^{(n)}\right)\right) \to \mathbb{E}(h(\mathbf{Z})),$$

which by Statement (i) of the Rosetta Stone theorem already gives Equation (1.10).

For the implication (iv) ⇒ (i) we have to set $h(\mathbf{z}) := \max_{j=1,\ldots,d} |x_j| \cdot z_j$ for a constant vector $\mathbf{x}$ and use the Rosetta Stone theorem to get

$$\begin{aligned}
\|\mathbf{x}\|_D = \mathbb{E}(h(\mathbf{Z})) &= \lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t) \\
&\overset{(iv)}{=} \lim_{n\to\infty} \lim_{t\to\infty} t \cdot P\left(h\left(\mathbf{X}^{(n)} > t\right)\right) \\
&= \lim_{n\to\infty} \mathbb{E}\left(h\left(\mathbf{Z}^{(n)}\right)\right) = \lim_{n\to\infty} \|\mathbf{x}\|_{D_n}.
\end{aligned}$$

Because this construction can be done for every $\mathbf{x} \in \mathbb{R}^d$ we have pointwise convergence.

The implication (i) $\Rightarrow$ (iii) remains. Assume that (i) holds, but that $\mathbf{Z}^{(n)} \xrightarrow{\mathcal{D}} \not\to \mathbf{Z}$. We will show that those assumptions lead to a contradiction. According to Lemma 6 there exists a vector $\mathbf{Z}^* \overset{\mathcal{D}}{\neq} \mathbf{Z}$ and a subsequence $(\mathbf{Z}^{(n_i)})_{i\in\mathbb{N}}$ that converges to $\mathbf{Z}^*$ in distribution. With the same arguments as in the proof of Lemma 5 $\mathbf{Z}^*$ is a D-norm generator and for every vector $\mathbf{x} \in \mathbb{R}^d$ we have

$$\mathbb{E}\left(\max_{j=1,\dots,d} |x_j| \cdot Z_j^*\right) = \lim_{i\to\infty} \mathbb{E}\left(\max_{j=1,\dots,n} Z_j^{n_i}\right) = \lim_{i\to\infty} \|\mathbf{x}\|_{D_{n_i}} = \|\mathbf{x}\|_D .$$

So both $\mathbf{Z}$ and $\mathbf{Z}^*$ realize almost surely on the same normed surface B, generate the same D-norm, but have different distributions. This violates Lemma 2. We have a contradiction and therefore there can never be a case, where (i) is fulfilled but not (iii). □

So we are one step forward: Now we know that getting arbitrarily close to the true D-norm is equivalent to getting arbitrary close to the true max-stable attractor, which itself is equivalent to getting arbitrarily close to the true values $\lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t)$ for the attracted random vectors.

Still a proper metric on the set of D-norms would be good and Falk (2019) suggested using the Wasserstein metric between the normed generators (which we introduce below), because this results in

$$\left|\|\mathbf{x}\|_{D_1} - \|\mathbf{x}\|_{D_2}\right| \leq \max_{j=1,\dots,d} |x_j| \cdot d_W\left(\|\cdot\|_{D_1}, \|\cdot\|_{D_2}\right), \tag{1.11}$$

for all $\mathbf{x} \in \mathbb{R}^d$ see Lemma 1.8.4 in the mentioned reference.

Definition 6. *The Wasserstein metric between two D-norms is defined by*

$$d_W\left(\|\cdot\|_{D_1}, \|\cdot\|_{D_2}\right) = \inf\left\{\left\|\boldsymbol{Z}^{(1)} - \boldsymbol{Z}^{(2)}\right\|_1 : \boldsymbol{Z}^{(i)} \text{ generates } \|\cdot\|_{D_i} \text{ and realizes almost surely in } B, i = 1, 2\right\},$$

where $\|\boldsymbol{x}\|_1 = \sum_{j=1}^d |x_j|$ *is the Manhattan norm.*

Note that in the set in this definition is not empty, because we can always choose $\mathbf{Z}^{(1)}$ to be independent of $\mathbf{Z}^{(2)}$. If the D-norms are the same we can chose $\mathbf{Z}^{(1)} = \mathbf{Z}^{(2)}$ to see that the Wasserstein metric then becomes 0.

Equation (1.11) is a special case of a more general result:

Corollary 5. *Let* $\boldsymbol{X}^{(1)}$ *and* $\boldsymbol{X}^{(2)}$ *be two random vectors each of which fulfills the equivalent statements of the Rosetta Stone theorem with underlying D-norms* $\|\cdot\|_{D_1}$ *and* $\|\cdot\|_{D_2}$*. If* h *is a non-negative function, that is homogeneous of order* 1 *and which is Lipschitz bounded with a constant* L *such that* $|h(\boldsymbol{x}) - h(\boldsymbol{y})| \leq L \cdot \|\boldsymbol{x} - \boldsymbol{y}\|_1$ *holds for all* $\boldsymbol{x}, \boldsymbol{y} \in [0,\infty)^d$*. Then we have*

$$\left|\lim_{t\to\infty} t \cdot P\left(h\left(\boldsymbol{X}^{(1)}\right) > t\right) - \lim_{t\to\infty} t \cdot P\left(h\left(\boldsymbol{X}^{(1)}\right) > t\right)\right| \leq L \cdot d_W\left(\|\cdot\|_{D_1}, \|\cdot\|_{D_2}\right).$$

Proof. Whenever $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ are D-norm generators on the same Probability space we have

$$\begin{aligned}
&\left|\lim_{t\to\infty} t\cdot P\left(h\left(\mathbf{X}^{(1)}\right)>t\right)-\lim_{t\to\infty} t\cdot P\left(h\left(\mathbf{X}^{(1)}\right)>t\right)\right| \\
=&\left|\mathbb{E}\left(h\left(\mathbf{Z}^{(1)}\right)\right)-\mathbb{E}\left(h\left(\mathbf{Z}^{(2)}\right)\right)\right| \\
\leq& E\left(\left|h\left(\mathbf{Z}^{(1)}\right)-h\left(\mathbf{Z}^{(2)}\right)\right|\right)\leq L\cdot\mathbb{E}\left(\left\|\mathbf{Z}^{(1)}-\mathbf{Z}^{(2)}\right\|\right).
\end{aligned}$$

We now can take the infimum on the right hand side to end with the Wasserstein metric. □

This way the Wasserstein metric gives a reasonable bound for the loss of 'information' in the tail when we replace a D-norm $\|\cdot\|_{D_1}$ by another D-norm $\|\cdot\|_{D_2}$. We will do something very similar in Section 2.6, where we reduce the dimensionality of multivariate extreme events and need a measure for the cost of doing that.

1.5 How and when to use the Rosetta Stone theorem.

We cannot assume all real life data to originate from a random vector $\mathbf{X}$ that fulfills one of the equivalent statements of the Rosetta Stone theorem. Such a naive assumption would lead to strange consequences like the following example:

Example 6. *If the joint behavior financial loss X_j for different investments $j = 1, \ldots, d$ could be described by a random vector $\boldsymbol{X}$ fulfilling one of the equivalent statements of the Rosetta Stone theorem, then we would have for every portfolio composition $(w_1, \ldots, w_d)$ with weights $w_j \geq 0$ that sum up to 1 the following limit*

$$\lim_{t\to\infty} t \cdot P\left(\sum_{j=1} w_j X_j > t\right) = \mathbb{E}\left(\sum_{j=1} w_j Z_j\right) = \sum_{j=1} w_j = 1.$$

Ultimately the probability of extreme losses would not depend on the composition of the portfolio at all. This is a strange consequence.

However the Rosetta Stone framework is not too restricting after some transformation as the following theorem will show:

Theorem 5. *Let $\boldsymbol{X}$ be a random vector with continuous marginal distribution functions $F_j(x) = P(X_j \leq d)$. By setting*

$$X'_j := \frac{1}{1 - F_j(X_j)}$$

we get a random vector $\boldsymbol{X}'$ and for every $\boldsymbol{y} \in \mathbb{R}^d$ the function

$$f_{\boldsymbol{y}}(t) := t \cdot P\left(\max_{j=1,\ldots,d} |y_j| X'_j > t\right)$$

maps values $t \geq \max_{j=1,\ldots,d} |y_j|$ to values in the compact interval

$$\left[\max_{j=1,\ldots,d} |y_j|, \sum_{j=1} |y_j|\right]. \tag{1.12}$$

$\boldsymbol{X}'$ fulfills the equivalent statements of the Rosetta Stone theorem if and only if the bounded functions $f_{\boldsymbol{y}}$ also converge as $t \to \infty$.

Proof. For $\mathbf{y} = \mathbf{0}$ the bounds are trivial. So without loss of generality we can assume $\mathbf{y} \neq \mathbf{0}$. Then there exists an index j_0 with $|y_{j_0}| = \max_{j=1,\ldots,d} |y_j| > 0$.

Then we have

$$\begin{aligned} t \cdot P\left(\max_{j=1,\dots,d} |y_j| X_j' > t\right) &\geq t \cdot P(|y_{j_0}| \cdot X_{j_0}' > t) \\ &= t \cdot P\left(X_{j_0}' > \frac{t}{|y_{j_0}|}\right) \\ &= t \cdot P\left(F_j(X_j) > 1 - \frac{|y_{j_0}|}{t}\right) \\ &= t \cdot \frac{|y_{j_0}|}{t} = \max_{j=1,\dots,d} |y_j| \end{aligned}$$

for every $t \geq |y_{j_0}|$, where we used that $F_j(X_j)$ follows the uniform distribution on $(0,1)$. For the same reasons we get

$$\begin{aligned} t \cdot P\left(\max_{j=1,\dots,d} |y_j| X_j' > t\right) &\leq t \cdot \sum_{j=1}^{d} P(|y_j| \cdot X_j' > t) \\ &= t \cdot \sum_{j=1}^{d} \frac{|y_j|}{t} = \sum_{j=1}^{d} |y_j|. \end{aligned}$$

Now to prove the equivalence: If $\mathbf{X}'$ fulfills one of the equivalent statements of the Rosetta Stone theorem, then we have

$$\lim_{t \to \infty} f_{\mathbf{y}}(t) = \|\mathbf{y}\|_D$$

for every $\mathbf{y} \in \mathbb{R}^d$ where $\|\cdot\|_D$ is the underlying D-norm. This means every $f_{\mathbf{y}}$ has a limit.

To show the opposite direction of the equivalence we now assume that every $f_{\mathbf{y}}$ has a limit and we denote it by

$$g(\mathbf{y}) := \lim_{t \to \infty} t \cdot P\left(\max_{j=1,\dots,d} |y_j| X_j' > t\right).$$

A careful inspection of Statement (ii) of the Rosetta Stone theorem reveals that if we can show g to be a D-norm, then we are done with the proof. To do that we will find a simple max-stable random vector $\mathbf{Y}$ with

$$P\left(\mathbf{Y} \leq \frac{1}{\mathbf{y}}\right) = \exp(-g(\mathbf{y}))$$

for all $\mathbf{y} > 0$. Theorem 1 tells us that the existence of such a $\mathbf{Y}$ is both necessary and sufficient for g to be a D-norm.

$\mathbf{Y}$ will be the weak limit of the random variables

$$M_n := \frac{1}{n} \cdot \max_{i=1,\dots,n} (\mathbf{X}')^{(i)},$$

where $(\mathbf{X}')^{(i)}$ are iid copies of $\mathbf{X}'$ and where the maximum is meant componentwise.

For every $\mathbf{y} > \mathbf{0}$ we have

$$\begin{aligned} P\left(M_n \leq \frac{1}{\mathbf{y}}\right) &= P\left((\mathbf{X}')^{(i)} \leq n \cdot \frac{1}{\mathbf{y}} \text{ for all } i = 1, \ldots, n\right) \\ &= P\left(\mathbf{X}' \leq n \cdot \frac{1}{\mathbf{y}}\right)^n \\ &= \left(1 - \frac{1}{n} \cdot f_{\mathbf{y}}(n)\right)^n \\ &\to \exp\left(-\lim_{t\to\infty} f_{\mathbf{y}}(t)\right) = \exp(-g(\mathbf{y})). \end{aligned}$$

For the second to last step we used Lemma 4.

We will now prove that the sequence $(M_n)_{n\in\mathbb{N}}$ is tight. Let ϵ be an arbitrary positive constant. Then there exists a value t' such that

$$\exp\left(-g\left(\frac{1}{t'} \cdot \mathbf{1}\right)\right) \geq 1 - \epsilon/2,$$

where $\mathbf{1} = (1, \ldots, 1)^\intercal$ is the d-dimensional vector of ones. This is possible as $g(\mathbf{y})$ always falls into the interval in Equation (1.12). Also there exists an $N \in \mathbb{N}$ such that

$$\left|P\left(M_n \leq t' \cdot \mathbf{1}\right) - \exp\left(-g\left(\frac{1}{t'} \cdot \mathbf{1}\right)\right)\right| \leq \frac{\epsilon}{2} \text{ for all } n \geq N.$$

Consequently we have

$$P\left(M_n \in [\mathbf{0}, t' \cdot \mathbf{1}]\right) \geq 1 - \epsilon \text{ for all } n \geq N.$$

Also there exist positive constants $t_1, \ldots, t_{N-1}$ such that

$$P(M_n \in [\mathbf{0}, t_n \cdot \mathbf{1}]) \geq 1 - \epsilon \text{ for all } n = 1, \ldots, N-1.$$

Now we can confirm that $t := \max\{t_1, \ldots, t_{N-1}, t'\}$ has the property

$$P(M_n \in [\mathbf{0}, t \cdot \mathbf{1}]) \geq 1 - \epsilon \text{ for all } n \in \mathbb{N}.$$

As we can do this construction for every $\epsilon > 0$ the sequence of random vectors is tight. Lemma 5 now states that there is a sub-sequence of $(M_{n_i})_{i\in\mathbb{N}}$ and a limiting random vector $\mathbf{Y}$ such that

$$M_{n_i} \xrightarrow{\mathcal{D}} \mathbf{Y} \text{ as } i \to \infty,$$

where the convergence is meant in distribution. We would then have

$$P\left(\mathbf{Y} \leq \frac{1}{\mathbf{y}}\right) = \lim_{i\to\infty} P\left(M_{n_i} \leq \frac{1}{\mathbf{y}}\right) = \exp(-g(\mathbf{y}))$$

for all $\mathbf{y} > 0$. We will even show that the whole sequence (M_n) converges to $\mathbf{Y}$. We will prove this by contradiction. If this was not the case, then according to Lemma 6 there would be a random vector $\mathbf{Y}^* \overset{\mathcal{D}}{\neq} \mathbf{Y}$ and a subsequence $(M_{n_i'})_{i\in\mathbb{N}}$ that converges to $\mathbf{Y}^*$.

This would result in

$$P\left(\mathbf{Y}^* \leq \frac{1}{\mathbf{y}}\right) = \lim_{j\to\infty} P\left(M_{n_{i_j}} \leq \frac{1}{\mathbf{y}}\right) = \exp(-g(\mathbf{y})) \\ = P\left(\mathbf{Y} \leq \frac{1}{\mathbf{y}}\right),$$

for all $\mathbf{y} > \mathbf{0}$ and this contradicts $\mathbf{Y}^*$ having a different distribution than $\mathbf{Y}$.

By contradiction we have shown that the complete series (M_n) converges to $\mathbf{Y}$. Standard arguments from extreme value theory can be used to show that $\mathbf{Y}$ is simple max-stable and Theorem 1 then implies that g was a D-norm to begin with.

Ultimately we have shown that Statement (ii) of the Rosetta Stone theorem holds. □

Very loosely speaking Theorem 5 implies that if $\mathbf{X}$ has standard Pareto margins and does not fulfill one of the equivalent statements the Rosetta Stone theorem, then there exists a constant vector $\mathbf{y}$ such that the rescaled probabilities

$$t \cdot P\left(\max_{j=1,\ldots,d} X_j \cdot |y_j| > t\right)$$

'oscillate' between a limes superior and a limes inferior, two different finite real numbers.

Immediately we see that the Rosetta Stone theorem depends on the exact equality of lim sup and lim inf. A potential direction for future research might be the following: Under what bounds for the differences between lim sup and lim inf can weaker version of the statements in the Rosetta Stone be saved? One possible candidate might be inequalities of the type

$$\left|\limsup_{t\to\infty} t \cdot P(h(\mathbf{X}) > t) - \liminf_{t\to\infty} t \cdot P(h(\mathbf{X}) > t)\right| \leq \ldots,$$

where h is a continuous non-negative function that is homogeneous of order 1.

But these thoughts go beyond the scope of this thesis. Our aim for now is to use the Rosetta Stone theorem to the fullest. In Chapter 2 we will use it to prove an extreme value version of the central limit theorem from classical probability theory, battle the curse of dimensionality and do exploratory statistics in multivariate extremes. Chapter 3 is all about estimation and inference of the extremal dependence structure.

Chapter 2

Managing extremes in many dimensions

If $\mathbf{X} = (X_1, \ldots, X_d)^\intercal$ is a random vector, that fulfills one of the equivalent statements of the Rosetta Stone theorem, then a lot of information about the joint appearance of extremes in the different components $X_1, \ldots, X_d$ is contained in the D-norm.

A d-dimensional D-norm is a is a real function on the space $\mathbb{R}^d$. Approximating the complete D-norm might therefore not be feasible as the dimension d grows.

One option that remains feasable is to store one value for every pair of indices $(i, j), 1 \leq i < j \leq d$, because the size of the resulting object grows quadratically in d. This value can be something like a correlation coefficient, except for extremal events. This chapter investigates the merits of this approach: Section 2.1 introduces the co-extremality c_{ij}, which measures the extremal dependence between two components X_i and X_j.

Also useful for dealing with many dimensions is the Hüsler–Reiss model, which is the finite-dimensional version of what is known as the Brown–Resnick process in the literature. Section 2.2 introduces this parametric family of D-norms. The Hüsler–Reiss D-norms are generated by multivariate lognormal generators and are parametrized by their co-extremal matrices, which means the number of parameters grows quadratically in dimension. To prove this we need a result we will call the 'geometric mean characterization of D-norms', which is proven in Section 2.3.

If a model is easy to handle with mathematics, that does not automatically mean it is a good model for many practical scenarios. However for the Hüsler–Reiss distribution there is a central limit theorem which says that a global medium-tailed phenomenon perturbed by many independent phenomena has a tail-behavior that can be approximated well with a Hüsler–Reiss distribution. We introduce this theorem in Section 2.4 and in Section 2.5 we procede to prove it.

Section 2.6 approaches the problem of high dimensions with D-norm generators - a vector $\mathbf{X}$ is replaced by a lower-dimensional $\Phi(\mathbf{X})$ and the underlying D-norm generator $\mathbf{Z}$ is replaced by $\Phi(\mathbf{Z})$.

Regardless of how we ultimately model the multivariate tail-dependence, the co-extremal matrix can be used for exploratory statistics. Section 2.7 features an exploratory approach to find clusters of components, such that extreme events tend to not hit more than cluster at one time.

Section 2.8 aims to compare the joint behavior of extremal events with the joint behavior of non-extremal events in an exploratory fashion.

In Section 2.9 we apply those tools to German weather data for illustrative purposes.

2.1 Co-extremality

Co-extremality is a measure of pairwise tail-dependence. It is not the first of its kind as the following example shows:

Example 7. *For a d-dimensional D-norm $\|\cdot\|_D$, the value $\theta = \|(1,\dots,1)^\intercal\|_D$ is called the extremal coefficient (see e.g. the work by Schlather and Tawn (2002) and the references in there). This can immediately turned into a measure of pairwise tail-dependence by setting $e_{ij} = \|\boldsymbol{e}_i + \boldsymbol{e}_j\|_D, i \neq j$. These values are well understood, especially as they coincide with the stable tail dependence function evaluated at the positions $\boldsymbol{e}_i + \boldsymbol{e}_j$.*

If there already exists a measure of pairwise tail dependence, there has to be a good reason to introduce a new one. For the co-extremality those reasons will be the connection to the geometric mean characterization of D-norms (see Theorem 7 below), the applications for dimension reduction (see Section 2.6) and for exploratory statistics in multivariate extremes (see Section 2.7).

So let's introduce the co-extremality and investigate its properties.

Theorem 6. *Let $\boldsymbol{Y}$ be a max-stable random vector with*

$$P((Y_1,\dots,Y_d) \leq (1/y_1,\dots,1/y_d)) = \exp\left(-\|\boldsymbol{y}\|_D\right)$$

for all $\boldsymbol{y} > \boldsymbol{0}$, where $\|\cdot\|_D$ is a D-norm with generator $\boldsymbol{Z}$. Let $\boldsymbol{X}$ be a random vector on $[0,\infty)^d$ with

$$\frac{1}{n} \cdot \max_{i=1,\dots,n} \boldsymbol{X}^{(i)} \xrightarrow{\mathcal{D}} \boldsymbol{Y},$$

as $n \to \infty$, where $\boldsymbol{X}^{(i)}$ are iid copies of $\boldsymbol{X}$. Then we have the limit

$$\lim_{t\to\infty} t \cdot P(X_i X_j > t^2) = \mathbb{E}(\sqrt{Z_i Z_j}) \tag{2.1}$$

for all $1 \leq i,j \leq d$.

The fact that there is a limit in Equation (2.1) has already been established by Jessen and Mikosch (2006).

Proof. This is a consequence of Rosetta Stone theorem applied to the function $\mathbf{x} \mapsto \sqrt{x_i x_j}$, which is non-negative, continuous and homogeneous of order 1. □

Definition 7. *Let us call the constant $c_{ij} := \mathbb{E}(\sqrt{Z_i Z_j})$ in Theorem 6 the co-extremality of X_i and X_j. The $d \times d$ matrix $C := (c_{ij})_{1\leq i,j\leq d}$ we will call the co-extremal matrix.*

The name co-extremality was chosen because its properties are similar to the properties of correlation, as in the Pearson correlation coefficient

$$\rho_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)} \cdot \sqrt{\mathrm{Var}(X_j)}},$$

as will be stated in the next result:

Corollary 6. *With the same assumptions and notations as in Theorem 6 the number c_{ij} lies in the interval $[0,1]$ and we have $c_{ij} = 1$ if an only if $Y_i = Y_j$ almost surely. Also there is the alternative representation*

$$c_{ij} = \lim_{t\to\infty} \frac{P(X_i X_j > t^2)}{\sqrt{P(X_i > t)} \cdot \sqrt{P(X_j > t)}}. \tag{2.2}$$

The matrix C is symmetric and has 1's on its diagonal.

Proof. By the representation $c_{ij} = \mathbb{E}(\sqrt{Z_i Z_j})$ we obviously have $c_{ij} \geq 0$ and according to the Cauchy–Schwarz inequality we have $c_{ij} \leq \mathbb{E}(\sqrt{Z_i}^2) \cdot E(\sqrt{Z_j}^2) = 1$ with equality if and only if $Z_i = Z_j$ almost surely. In that case

$$\begin{aligned} P(Y_i \leq 1/y_i, Y_j \leq 1/y_j) &= \exp(-\mathbb{E}(\max\{y_i Z_i, y_j Z_j\})) \\ &= \exp(-\mathbb{E}(\max\{y_i, y_j\} Z_i) \\ &= \exp(-\max\{y_i, y_j\}) = P(Y_i \leq 1/y_i, Y_i \leq 1/y_j) \end{aligned}$$

for all $y_i, y_j > 0$, which shows that (Y_i, Y_j) follows the same distribution as (Y_i, Y_i) and thus $Y_i = Y_j$ almost surely.

To get the alternative representation we use the Rosetta Stone theorem to see $\lim_{t\to\infty} t \cdot P(X_i > t) = \mathbb{E}(Z_i) = 1$ and $\lim_{t\to\infty} t \cdot P(X_j > t) = 1$ as well. Then we obtain

$$c_{ij} = \frac{c_{ij}}{\sqrt{1} \cdot \sqrt{1}} = \lim_{t\to\infty} \frac{t \cdot P(X_i X_j > t^2)}{\sqrt{t \cdot P(X_i > t)} \cdot \sqrt{t \cdot P(X_j > j)}},$$

where we can cancel the factor t to get Equation (2.2). The symmetry of C is obvious and $c_{jj} = \mathbb{E}(\sqrt{Z_j Z_j}) = \mathbb{E}(Z_j) = 1$ for all $j = 1, \ldots, d$. □

The co-extremal matrix will turn out to parametrize an important class of multivariate max-stable distributions, the Hüsler–Reiss model, which we will introduce and investigate in the following sections.

2.2 The Hüsler–Reiss model

D-norms are inherently connected to their estimators. So any parametric family of D-norm generators

$$\{\mathbf{Z}_\theta : \theta \in \Theta\}$$

can be turned into a parametric family of D-norms

$$\{\|\cdot\|_D : \|\cdot\|_D \text{ is generated by } \mathbf{Z}_\theta, \theta \in \Theta\}.$$

A well understood parametric class of finite-dimensional distributions are the multivariate normal distributions. However a multivariate normal random vector $\mathcal{N} = (\mathcal{N}_1, \ldots, \mathcal{N}_d)^\intercal$ in general is not a D-norm generator, as its components realize in the negative number with positive probability. One solution to this problem is to use

$$\mathbf{Z}' := (\exp(\mathcal{N}_1, \ldots, \exp(\mathcal{N}_d))^\intercal$$

instead. This vector fulfills $\mathbf{Z}' \geq \mathbf{0}$ almost surely and its components have finite expected values. So by rescaling the components of $\mathbf{Z}'$ we can turn it into a proper D-norm generator.

Definition 8. *Let $\mathcal{N} = (\mathcal{N}_1, \ldots, \mathcal{N}_d)^\intercal$ follow a multivariate normal distribution with mean vector $\boldsymbol{0}$ and covariance matrix $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq d}$. The D-norm generated by*

$$(Z_1, \ldots, Z_d) := (\exp(\mathcal{N}_1 - \sigma_{11}/2), \ldots, \exp(\mathcal{N}_d - \sigma_{dd}/2))$$

is called a Hüsler–Reiss D-norm and the simple max-stable random vector $\boldsymbol{Y}$ with

$$P\left(\boldsymbol{Y} \leq \frac{1}{\boldsymbol{y}}\right) = \exp\left[-\mathbb{E}\left(\max_{j=1,\ldots,d} y_j \exp(\mathcal{N}_j - \sigma_{jj}/2)\right)\right]$$

for all $\boldsymbol{y} > \boldsymbol{0}$ is said to follow a Hüsler–Reiss distribution.

Lemma 8 will show that this $\mathbf{Z}$ is indeed a D-norm generator. The Hüsler–Reiss distribution is the finite dimensional version of what is called the Brown–Resnick process.

Definition 9. *Let S be an arbitrary index set and let $\boldsymbol{Y} = (Y_s)_{s \in S}$ be a random process with standard Fréchet margins. We call $\boldsymbol{Y}$ a Brown–Resnick process, if there exists a centered Gaussian process $\mathcal{N} = (\mathcal{N}_s)_{s \in S}$ with covariances $(\sigma_{st})_{s,t \in S}$ such that for every finite collection of indices $s_1, \ldots, s_d \in S$ we have*

$$P\left((Y_{s_1}, \ldots, Y_{s_d}) \leq (1/y_1, \ldots, 1/y_d)\right) = \exp\left[-\mathbb{E}\left(\max_{j=1,\ldots,d} y_j \exp(\mathcal{N}_{s_j} - \sigma_{s_j s_j}/2)\right)\right] \tag{2.3}$$

for all $y_1, \ldots, y_d > 0$.

When the index set S is finite, then the Brown–Resnick process follows a Hüsler–Reiss distribution. If the index set S is infinite, then at least all finite dimensional margins of the Brown–Resnick process follow Hüsler–Reiss distributions.

The Gaussian process $\mathcal{N}$ in Definition 9 is not unique: For every Brown–Resnick process there are many different centered Gaussian processes that satisfy Equation (2.3). Kabluchko et al. (2009) showed that two centered Gaussian processes produce the same Brown–Resnick process if and only if they have the same variogram $(\mathrm{Var}(\mathcal{N}_{s_1} - \mathcal{N}_{s_2}))_{s_1,s_2 \in S}$. In finite dimension this means the Hüsler–Reiss distribution only depends on the square matrix $(\mathrm{Var}(\mathcal{N}_i - \mathcal{N}_j))_{1 \le i,j \le d}$. We state this with D-norm notation in Theorem 8 below. But before we can get to that we need some preparatory results.

Lemma 8. *Let $\mathcal{N} = (\mathcal{N}_1, \ldots, \mathcal{N}_d)^\intercal$ be a multivariate normal distributed random vector with mean vector $\boldsymbol{0}$ and covariance matrix $\Sigma = (\sigma_{ij})_{1 \le i,j \le d}$. Then $\boldsymbol{Z}$ defined by*

$$(Z_1, \ldots, Z_d) := (\exp(\mathcal{N}_1 - \sigma_{11}/2), \ldots, \exp(\mathcal{N}_d - \sigma_{dd}/2)) \tag{2.4}$$

generates a D-norm and also we have

$$\mathbb{E}\left(\prod_{j=1}^{d} Z_j^{\lambda_j}\right) = \exp\left(-\frac{1}{4} \cdot \sum_{1 \le i,j \le d} \lambda_i \lambda_j \mathrm{Var}(\mathcal{N}_i - \mathcal{N}_j)\right) < \infty, \tag{2.5}$$

for all real numbers $\lambda_1, \ldots, \lambda_d$ that add up to 1.

Note that because $Z_j > 0$ holds almost surely for all $j = 1, \ldots, d$, negative exponents $\lambda_j < 0$ don't pose a problem at all in Equation (2.5).

Proof. For all $j = 1, \ldots, d$ the random variable $\exp(\mathcal{N}_j - \sigma_{jj}/2)$ is log-normal distributed with Parameters $-\sigma_{jj}/2$ and σ_{jj}. Thus, $Z_j \ge 0$ almost surely and $\mathbb{E}(Z_j) = 1$. This is already sufficient for $\mathbf{Z}$ to be a D-norm generator.

Let $\lambda_1, \ldots, \lambda_d$ be arbitrary real numbers adding up to 1. We will prove Equation (2.5). For that we will define a random variable $\widetilde{\mathcal{N}}$ by

$$\widetilde{\mathcal{N}} := \sum_{j=1}^{d} \lambda_j (\mathcal{N}_j - \sigma_{jj}/2) = \log\left(\prod_{j=1}^{d} Z_j^{\lambda_j}\right).$$

Then $\widetilde{\mathcal{N}}$ is a normal distributed random variable with mean

$$\mathbb{E}\left(\widetilde{\mathcal{N}}\right) = -\frac{1}{2} \cdot \sum_{j=1}^{d} \lambda_j \sigma_{jj}$$

and variance

$$\begin{aligned}
\mathrm{Var}\left(\widetilde{\mathcal{N}}\right) &= \sum_{1\le i,j\le d} \lambda_i\lambda_j\sigma_{ij} \\
&= \sum_{1\le i,j\le d} \lambda_i\lambda_j(\sigma_{ii}+\sigma_{jj}-\mathrm{Var}(\mathcal{N}_i-\mathcal{N}_j))/2 \\
&= \frac{1}{2}\cdot\left[\sum_{i=1}^d \lambda_i\sigma_{ii}\underbrace{\sum_{j=1}^d \lambda_j}_{=1} + \sum_{j=1}^d \lambda_j\sigma_{jj}\underbrace{\sum_{i=1}^d \lambda_i}_{=1} - \sum_{1\le i,j\le d}\lambda_i\lambda_j\mathrm{Var}(\mathcal{N}_i-\mathcal{N}_j)\right] \\
&= \sum_{i=1}^d \lambda_i\sigma_{ii} - \frac{1}{2}\cdot\sum_{1\le i,j\le d}\lambda_i\lambda_j\mathrm{Var}(\mathcal{N}_i-\mathcal{N}_j),
\end{aligned}$$

where we used that $\mathrm{Var}(\mathcal{N}_i-\mathcal{N}_j)=\sigma_{ii}+\sigma_{jj}-2\sigma_{ij}$ for all pairs of indices i,j. With this information we can calculate the expected value of the random variable $\exp\left(\widetilde{\mathcal{N}}\right)$. It follows a lognormal distribution and, thus,

$$\begin{aligned}
\mathbb{E}\left(\prod_{j=1}^d Z_j^{\lambda_j}\right) = \mathbb{E}\left(\exp\left(\widetilde{\mathcal{N}}\right)\right) &= \exp\left(\mathbb{E}\left(\widetilde{\mathcal{N}}\right)+\frac{1}{2}\cdot\mathrm{Var}\left(\widetilde{\mathcal{N}}\right)\right) \\
&= \exp\left(-\frac{1}{4}\cdot\sum_{1\le i,j\le d}\lambda_i\lambda_j\mathrm{Var}(\mathcal{N}_i-\mathcal{N}_j)\right),
\end{aligned}$$

which is what we wanted to prove. □

If $\mathbf{Z}$ is a D-norm generator with $Z_j>0$ almost surely for $j=1,\dots,d$, then there is no obvious connection between a D-norm $\|\mathbf{y}\|_D=\mathbb{E}(\max_{j=1,\dots,d}|y_j|Z_j)$ and the values $\mathbb{E}\left(\prod_{j=1}^d Z_j^{\lambda_j}\right)$, where λ_j are real numbers adding up to 1. But with a combination of probabilistic tools - namely the Cramér–Wold device and the method of moments - one can show that the D-norm $\|\cdot\|_D$ is uniquely determined by the values $\mathbb{E}\left(\prod_{j=1}^d Z_j^{\lambda_i}\right)$, if they are finite.

This is what we call the geometric mean characterization of D-Norms.

Theorem 7 (Geometric mean characterization of D-norms)**.** *Let $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ be two random vectors on $(0,\infty)^d$ that generate the D-norms $\|\cdot\|_{D_1}$ and $\|\cdot\|_{D_2}$. If we have*

$$\mathbb{E}\left(\prod_{j=1}^d (Z_j^{(1)})^{\lambda_j}\right) = \mathbb{E}\left(\prod_{j=1}^d (Z_j^{(2)})^{\lambda_j}\right) < \infty \tag{2.6}$$

for all real numbers λ_j adding up to 1, then $\|\cdot\|_{D_1}=\|\cdot\|_{D_2}$.

Note that once again $Z_j^{(1)}, Z_j^{(2)}>0$ almost surely by requirements of the theorem and, therefore, negative exponents $\lambda_j<0$ don't pose a problem. The proof of Theorem 7 is in Section 2.3.

With these preparatory results the proof of the following result becomes very easy to prove.

Theorem 8. *Let $\mathcal{N}^{(1)}, \mathcal{N}^{(2)}$ be two d-dimensional multivariate normal distributed random vectors with $\mathrm{Var}(\mathcal{N}_i^{(1)} - \mathcal{N}_j^{(1)}) = \mathrm{Var}(\mathcal{N}_i^{(2)} - \mathcal{N}_j^{(2)})$ for all $1 \leq i, j \leq d$. If $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ are defined as in Equation (2.4), then $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ generate the same D-norm.*

Proof. We have $\min_{j=1,\ldots,d} Z_j^{(1)} > 0$ and $\min_{j=1,\ldots,d} Z_j^{(2)} > 0$ almost surely. According to Lemma 8 we have $\mathbb{E}\left(\prod_{j=1}^d (Z_j^{(1)})^{\lambda_j}\right) = \mathbb{E}\left(\prod_{j=1}^d (Z_j^{(2)})^{\lambda_j}\right) < \infty$, whenever $\lambda_1, \ldots, \lambda_d$ are real numbers adding up to 1. Thus, Theorem 7 becomes applicable and $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ generate the same D-norm. □

Corollary 7. *A Hüsler–Reiss D-norm is uniquely determined by its co-extremal matrix.*

The multivariate tail-dependence of a random vector $\boldsymbol{X}$, that fulfills the equivalent statements of the Rosetta Stone theorem with an underlying Hüsler–Reiss D-norm, is determined by the joint extremal behavior of the bivariate random vectors $(X_i, X_j), 1 \leq i, j \leq d$.

Proof. If for two indices $i \neq j$ we set $\lambda_h := 1/2$ if $h \in \{i, j\}$ and $\lambda_h := 0$ else, then Equation (2.5) reduces to

$$c_{ij} = \mathbb{E}((Z_i \cdot Z_j)^{1/2}) = \exp\left(-\frac{1}{8} \cdot \mathrm{Var}(\mathcal{N}_i - \mathcal{N}_j))\right),$$

which is equivalent to

$$\mathrm{Var}(\mathcal{N}_i - \mathcal{N}_j) = -8 \cdot \log(c_{ij}), \tag{2.7}$$

so by Theorem 8 the Hüsler–Reiss D-norm is uniquely determined.

If we introduce a vector $\mathbf{X}$ that fulfills the equivalent statements of the Rosetta Stone theorem, then Equation (2.7) turns into

$$\mathrm{Var}(\mathcal{N}_i - \mathcal{N}_j) = -8 \cdot \log\left(\lim_{t\to\infty} t \cdot P(X_i \cdot X_j > t^2)\right),$$

where the right hand side only depends on the pair-wise tail dependence. □

To capture the complete dependence structure of the Hüsler–Reiss model Corollary 7 says it is sufficient to know the co-extremal matrix C, which only contains d^2 entries. We can reduce this even further, because we already know that C is symmetric and has 1's on the diagonal, so there are only $\frac{d\cdot(d-1)}{2}$ free parameters.

This is a nice mathematical property, when we deal with high dimensions. But when is it reasonable to assume a vector $\mathbf{X}$ to be in the max-domain of a Hüsler–Reiss distribution? Section 2.4 gives an answer to that. But before that we will prove the geometric mean characterization of D-norms.

2.3 Proof of the geometric mean characterization of D-norms

The following result is crucial in showing that the distribution of a Brown–Resnick process only depends on the variogram of the underlying Gaussian distribution. It was first proven by Kabluchko et al. (2009), but here is a proof using D-norm terminology.

Proof of the geometric mean characterization. Let $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ be two D-norm generators with

$$\min_{j=1,\dots,d} Z_j^{(i)} > 0$$

almost surely for $i = 1, 2$. We will show that if Equation (2.6) holds for all real numbers λ_j adding up to 1, then $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ generate the same D-norm.

Let us apply Lemma 1 to the generators $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}$ to get to measures ν_1, ν_2 on the set $[0,\infty)^d \setminus \{\mathbf{0}\}$ with the property

$$\nu_i(\{\mathbf{x} : h(\mathbf{x}) \geq 1\}) = \mathbb{E}(h(\mathbf{Z}^{(i)})) \text{ for } i = 1, 2$$

for every non-negative, measurable function h that is homogeneous of order 1. Define the sets

$$\begin{aligned} M_n &:= \left\{\mathbf{x} : \mathbf{x} \in [0,\infty)^d, n \cdot (x_1 + \cdots + x_d) \cdot 1_{\min_{j=1,\dots,d} x_j = 0} \geq 1\right\} \\ &\quad \text{for every } n \in \mathbb{N} \text{ and} \\ M_\infty &:= \bigcup_{n\in\mathbb{N}} M_n = \left\{\mathbf{x} : \mathbf{x} \in [0,\infty)^d, \mathbf{x} \neq \mathbf{0}, \min_{j=1,\dots,d} x_j = 0\right\}. \end{aligned}$$

The function $\mathbf{x} \mapsto n \cdot (x_1 + \cdots + x_d) \cdot 1_{\min_{j=1,\dots,d} x_j = 0}$ is non-negative, measurable and homogeneous of order 1. By choice of ν_1, ν_2 we have

$$\nu_i(M_n) = \mathbb{E}(n \cdot (Z_1 + \cdots + Z_d) \cdot \underbrace{1_{\min_{j=1,\dots,d} Z_j = 0}}_{=0 \text{ almost surely}}) = 0, \quad i = 1, 2 \quad n = 1, 2, 3, \dots$$

and consequently

$$\nu_1(M_\infty) = 0 = \nu_2(M_\infty). \tag{2.8}$$

Define the surface B by

$$B = \{\mathbf{x} : \mathbf{x} \in (0,\infty)^d, x_1 = 1\}.$$

Using Equation (2.8) and by the choice of ν_1 we have

$$\begin{aligned} \nu_1((1,\infty) \cdot B) &= \nu_1(\{\mathbf{x} : \mathbf{x} \in (0,\infty)^d, x_1 \geq 1\}) \\ &\overset{(2.8)}{=} \nu_1(\{\mathbf{x} : \mathbf{x} \in [0,\infty)^d, x_1 \geq 1\}) = \mathbb{E}(Z_1^{(1)}) = 1. \end{aligned}$$

The function

$$A \mapsto \nu_1([1,\infty) \cdot A) =: P_B^{(1)}(A)$$

for all measurable subsets $A \subset B$ obviously defines a probability measure $P_B^{(1)}$. Just like in the proof of Lemma 1 we set

$$\nu_1'(M) := (\lambda \times P_B^{(1)})(T^{-1}(M))$$

for all measurable M, where λ is the Lebesgue measure on $(0,\infty)$, $\times$ denotes the product measure and T is the transformation $(s, \mathbf{x}) \mapsto \frac{1}{s} \cdot \mathbf{x}$.

Let $\mathbf{Z}_B^{(1)}$ be a random vector that follows the probability distribution $P_B^{(1)}$. We will prove $\nu_1 = \nu_1'$, which implies that $\mathbf{Z}^{(1)}$ and $\mathbf{Z}_B^{(1)}$ generate the same D-norm. For an arbitrary vector $\mathbf{y} \in (0,\infty)^d$ define

$$A_{\mathbf{y}} := \{\mathbf{x} : x_1 = 1, \mathbf{x} \geq \mathbf{y}/y_1\} \subset B,$$

which gives us the convenient equality $[y_1,\infty) \cdot A_{\mathbf{y}} = [\mathbf{y},\infty)$. This can be used in the following:

$$\begin{aligned}
\nu_1([\mathbf{y},\infty)) &= \nu_1([y_1,\infty) \cdot A_{\mathbf{y}}) \\
&= \frac{1}{y_1} \cdot \nu_1([1,\infty) \cdot A_{\mathbf{y}}) \\
&= \frac{1}{y_1} \cdot P_B^{(1)}(A_{\mathbf{y}}) = \nu_1'([\mathbf{y},\infty)).
\end{aligned}$$

Further we have $\nu_1(M_\infty) = 0$ from Equation (2.8) and $\nu_1'(M_\infty) = 0$ because

$$T^{-1}(M_\infty) \cap (0,\infty) \cdot B = \emptyset$$

and the product measure $(\lambda \times P_B^{(1)})$ puts zero mass outside of the set $(0,\infty) \cdot B$. Combining this with $\nu_1([\mathbf{y},\infty)) = \nu_1'([\mathbf{y},\infty))$ for all $\mathbf{y} > \mathbf{0}$ we get $\nu_1 = \nu_1'$.

So there exists a D-norm generator $\mathbf{Z}_B^{(1)}$ of $\|\cdot\|_{D_1}$ with $P(\mathbf{Z}_B^{(1)} \in B) = 1$ and with exactly the same reasoning there exists a D-norm generator $\mathbf{Z}_B^{(2)}$ of $\|\cdot\|_{D_2}$ with $P(\mathbf{Z}_B^{(2)} \in B) = 1$ as well. We will proceed to show that the distributions of $\mathbf{Z}_B^{(1)}$ and $\mathbf{Z}_B^{(2)}$ coincide. It is sufficient to show that

$$(\log(Z_{B2}^{(1)}), \ldots, \log(Z_{Bd}^{(1)})) \stackrel{\mathcal{D}}{=} (\log(Z_{B2}^{(2)}), \cdots, \log(Z_{Bd}^{(2)}).$$

We prove this using the Cramér–Wold device and the method of moments as described by Billingsley (1979).

1. Two random vectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ follow the same distribution if and only if the linear combinations $\mathbf{y}^\intercal \mathbf{X}^{(1)}$ follows the same univariate distribution as $\mathbf{y}^\intercal \mathbf{X}^{(2)}$ for all constant vectors $\mathbf{y} \in \mathbb{R}^d$. This is the Cramér–Wold device.

2. If two random variables have the same moment generating function that is finite on an open interval around 0, then they follow the same distribution. This is called the method of moments.

Let $\mathbf{y} = (y_2, \ldots, y_d)^\intercal$ be an arbitrary vector in $\mathbb{R}^{d-1}$ and let t be an arbitrary real number. Set $\lambda_j := t \cdot y_j$ for $j = 2, \ldots, d$ and $\lambda_1 := 1 - \sum_{j=2}^{d} \lambda_j$. Then we have

$$\begin{aligned}
\mathbb{E}\left(\exp\left(t \cdot \sum_{j=2}^{d} y_j \cdot \log(Z_j^{(1)})\right)\right) &= \mathbb{E}\left(\prod_{j=2}^{d} (Z_{Bj}^{(1)})^{\lambda_j}\right) \\
&\overset{(a)}{=} \mathbb{E}\left(\prod_{j=1}^{d} (Z_{Bj}^{(1)})^{\lambda_j}\right) \\
&\overset{(b)}{=} \mathbb{E}\left(\prod_{j=1}^{d} (Z_{j}^{(1)})^{\lambda_j}\right) \\
&\overset{(c)}{=} \mathbb{E}\left(\prod_{j=1}^{d} (Z_{j}^{(2)})^{\lambda_j}\right) \\
&\overset{(d)}{=} \mathbb{E}\left(\prod_{j=1}^{d} (Z_{Bj}^{(2)})^{\lambda_j}\right) \\
&\overset{(e)}{=} \mathbb{E}\left(\prod_{j=2}^{d} (Z_{Bj}^{(2)})^{\lambda_j}\right) \\
&= \mathbb{E}\left(\exp\left(t \cdot \sum_{j=2}^{d} y_j \cdot \log(Z_{Bj}^{(2)})\right)\right),
\end{aligned}$$

where in (a) and (e) we used $Z_{B1}^{(1)} = 1$ and $Z_{B1}^{(2)} = 1$ almost surely. (c) is the exactly Equation (2.6), the main condition of the geometric mean characterization theorem. As for the steps (b) and (d), we have to introduce h_λ by

$$h_\lambda(\mathbf{x}) := \begin{cases} \prod_{j=1}^{d} x_j^{\lambda_j} & \text{for } \min_{j=1,\ldots,d} x_j > 0 \\ 0 & \text{else.} \end{cases}$$

It is non-negative, measurable and homogeneous of order 1. The second case of its definition is triggered only with probability 0 when we apply h_λ to the random vectors $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \mathbf{Z}_B^{(1)}, \mathbf{Z}_B^{(2)}$, because all of them are $> \mathbf{0}$ almost surely. Consequently we can use Theorem 3 to get

$$\mathbb{E}\left(\prod_{j=1}^{d}(Z_{Bj}^{(1)})^{\lambda_j}\right) = \mathbb{E}(h_\lambda(\mathbf{Z}_B^{(1)})) = \mathbb{E}(h_\lambda(\mathbf{Z}^{(1)})) = \mathbb{E}\left(\prod_{j=1}^{d}(Z_j^{(1)})^{\lambda_j}\right),$$

which is exactly (b) and with the same reasoning we get (d).

Thus, for every $\mathbf{y} \in \mathbb{R}^{d-1}$, the linear combination $\sum_{j=2}^{d} y_j \cdot \log(Z_j^{(1)})$ has the same finite moment generating function as $\sum_{j=2}^{d} y_j \cdot \log(Z_j^{(2)})$. This implies that $\mathbf{Z}_B^{(1)}$ follows the same distribution as $\mathbf{Z}_B^{(2)}$ and, thus, $\|\cdot\|_{D_1} = \|\cdot\|_{D_2}$.

□

2.4 A central limit theorem for the Hüsler–Reiss distribution

Often the reason for modelling real-life phenomena with normal distributions is the central limit theorem. The additive effect of many independent small perturbations results in distributions increasingly similar to a normal distribution.

There is something similar for the Hüsler–Reiss models. A lot of independent light-tailed perturbations of a single medium-tailed phenomenon develop a tail-behavior increasingly similar to that of a Hüsler–Reiss distribution. This will be verified in Theorem 10, but for proving the theorem we need some auxiliary results.

The first of those auxiliary results is strongly connected to what is called D-norm multiplication in a paper by Falk (2013).

Theorem 9. *Let $\boldsymbol{X}$ be a random vector on $[0,\infty)^d$ such that*

$$\frac{1}{n} \cdot \max_{i=1,\dots,n} \boldsymbol{X}^{(i)} \xrightarrow{\mathcal{D}} \boldsymbol{Y},$$

as $n \to \infty$, where $\boldsymbol{X}^{(i)}$ are iid copies of $\boldsymbol{X}$, and $\boldsymbol{Y}$ is the simple max-stable random vector with

$$P\left(\boldsymbol{Y} \leq \frac{1}{\boldsymbol{y}}\right) = \exp\left(-\mathbb{E}\left(\max_{j=1,\dots,d} y_j Z_j^{(1)}\right)\right)$$

for all $\boldsymbol{y} > \boldsymbol{0}$, where $\boldsymbol{Z}^{(1)}$ is a D-norm generator. Let $\boldsymbol{Z}^{(2)}$ be another D-norm generator independent of $\boldsymbol{X}$ and independent of $\boldsymbol{Z}^{(1)}$. Put $\boldsymbol{X}^* = \boldsymbol{X} \cdot \boldsymbol{Z}^{(2)}$ and $\boldsymbol{Z}^* = \boldsymbol{Z}^{(1)} \cdot \boldsymbol{Z}^{(2)}$ where in both cases the product is meant component-wise. Then we have

$$\frac{1}{n} \cdot \max_{i=1,\dots,n} (\boldsymbol{X}^*)^{(i)} \xrightarrow{\mathcal{D}} \boldsymbol{Y}^*$$

as $n \to \infty$, where $(\boldsymbol{X}^*)^{(i)}$ are iid copies of $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$ is the simple max-stable random vector with

$$P\left(\boldsymbol{Y}^* \leq \frac{1}{\boldsymbol{y}}\right) = \exp\left(-\mathbb{E}\left(\max_{j=1,\dots,d} y_j Z_j^*\right)\right)$$

for all $\boldsymbol{y} > \boldsymbol{0}$.

Note that $\mathbf{Z}^*$ is a D-norm generator itself, as it is a non-negative random vector with $\mathbb{E}(Z_j^*) = \mathbb{E}(Z_j^{(1)} \cdot Z_j^{(2)}) = \mathbb{E}(Z_j^{(1)}) \cdot \mathbb{E}(Z_j^{(2)}) = 1$ for all $j = 1, \dots, d$.

The interesting point of the theorem is that the operation that turns $\mathbf{X}$ into $\mathbf{X}^*$ is the same operation that turns the D-norm generator $\mathbf{Z}^{(1)}$ into the D-norm generator $\mathbf{Z}^*$.

Proof of Theorem 9. According to the Rosetta Stone theorem we have

$$\lim_{t\to\infty} t \cdot P\left(\max_{j=1,\dots,d} y_j z_j X_j > t\right) = \mathbb{E}\left(\max_{j=1,\dots,d} y_j z_j Z_j^{(1)}\right)$$

for all combination of constant vectors $\mathbf{y}, \mathbf{z} \geq \mathbf{0}$. What we have to show is the limit

$$\lim_{t\to\infty} t \cdot P\left(\max_{j=1,\dots,d} y_j Z_j^{(2)} X_j > t\right) = \mathbb{E}\left(\max_{j=1,\dots,d} y_j Z_j^{(2)} Z_j^{(1)}\right), \tag{2.9}$$

where only $\mathbf{y} \geq \mathbf{0}$ is constant and $\mathbf{z}$ was replaced by the random vector $\mathbf{Z}^{(2)}$. Because $\mathbf{Z}^{(2)}$ is independent of both $\mathbf{X}$ and $\mathbf{Z}^{(1)}$ we can use Fubini's theorem on both sides of Equation (2.9). The left hand side turns out to be:

$$\begin{aligned}
&\lim_{t\to\infty} t \cdot P\left(\max_{j=1,\dots,d} y_j Z_j^{(2)} X_j > t\right) \\
=&\lim_{t\to\infty} t \cdot \int P\left(\max_{j=1,\dots,d} y_j z_j X_j > t\right) \mathrm{d}P(\mathbf{Z}^{(2)} = \mathbf{z}) \\
=&\lim_{t\to\infty} \int t \cdot P\left(\max_{j=1,\dots,d} y_j z_j X_j > t\right) \mathrm{d}P(\mathbf{Z}^{(2)} = \mathbf{z}).
\end{aligned}$$

At the same time the right hand side of Equation (2.9) is the following:

$$\begin{aligned}
&\mathbb{E}\left(\max_{j=1,\dots,d} y_j Z_j^{(2)} Z_j^{(1)}\right) \\
=&\int \mathbb{E}\left(\max_{j=1,\dots,d} y_j z_j Z_j^{(1)}\right) \mathrm{d}P(\mathbf{Z}^{(2)} = \mathbf{z}) \\
=&\int \lim_{t\to\infty} t \cdot P\left(\max_{j=1,\dots,d} y_j z_j X_j > t\right) \mathrm{d}P(\mathbf{Z}^{(2)} = \mathbf{z}).
\end{aligned}$$

To prove Equation (2.9) it is sufficient to show that the pointwise convergence of $t \cdot P\left(\max_{j=1,\dots,d} y_j z_j X_j > t\right)$ is dominated by an integrable function in the probability space of $\mathbf{Z}^{(2)}$.

The function $f(t) = t \cdot P(\max_{j=1,\dots,d} X_j > t)$ is always lower than the function $t \mapsto t$ and converges to a constant as $t \to \infty$. It is elementary to show that these properties imply a constant upper bound M for f. If $\|\cdot\|_\infty$ is the supremum norm, that is $\|\mathbf{x}\|_\infty = \max_{j=1,\dots,d} |x_j|$, then for all pairs $\mathbf{y}, \mathbf{z} \geq \mathbf{0}$ we get the inequality

$$\begin{aligned}
&t \cdot P\left(\max_{j=1,\dots,d} y_j z_j X_j > t\right) \\
\leq& t \cdot P\left(\max_{j=1,\dots,d} X_j \cdot \|\mathbf{y}\|_\infty \|\mathbf{z}\|_\infty > t\right) \\
=& \|\mathbf{y}\|_\infty \|\mathbf{z}\|_\infty \cdot \frac{t}{\|\mathbf{y}\|_\infty \|\mathbf{z}\|_\infty} P\left(\max_{j=1,\dots,d} X_j > \frac{t}{\|\mathbf{y}\|_\infty \|\mathbf{z}\|_\infty}\right) \\
=& \|\mathbf{y}\|_\infty \|\mathbf{z}\|_\infty f\left(\frac{t}{\|\mathbf{y}\|_\infty \|\mathbf{z}\|_\infty}\right) \leq \|\mathbf{y}\|_\infty \|\mathbf{z}\|_\infty \cdot M.
\end{aligned}$$

The upper bound $\mathbf{z} \mapsto \|\mathbf{y}\|_\infty \|\mathbf{z}\|_\infty \cdot M$ is an integrable function with respect to the probability measure of $\mathbf{Z}^{(2)}$, as

$$\begin{aligned}\mathbb{E}\left(\|\mathbf{y}\|_\infty \cdot \left\|\mathbf{Z}^{(2)}\right\|_\infty \cdot M\right) &= \|\mathbf{y}\|_\infty \cdot M \cdot \mathbb{E}\left(\left\|\mathbf{Z}^{(2)}\right\|_\infty\right) \\ &\leq \|\mathbf{y}\|_\infty \cdot M \cdot \mathbb{E}\left(\sum_{j=1}^{d} Z_j^{(2)}\right) = \|\mathbf{y}\|_\infty \cdot M \cdot d,\end{aligned}$$

as we required $\mathbf{Z}^{(2)}$ to be a D-norm generator.

In the probability space of $\mathbf{Z}^{(2)}$ the pointwise convergence

$$\lim_{t\to\infty} t \cdot P\left(\max_{j=1,\ldots,d} y_j z_j X_j > t\right) \to \mathbb{E}(\max_{j=1,\ldots,d} y_j z_j X_j)$$

for all $\mathbf{z}$ is dominated by the integrable function $\mathbf{z} \mapsto \|\mathbf{y}\|_\infty \|\mathbf{z}\|_\infty \cdot M$ and therefore the well-known dominated convergence theorem is applicable and Equation (2.9) holds. □

By setting $\mathbf{Z}^{(1)} = (1, \ldots, 1)^\intercal$ almost surely in the preceding theorem we obtain the following two corollaries:

Corollary 8. *Let $X \geq 0$ be a random variable with $\lim_{t\to\infty} t \cdot P(X > t) = 1$ and let $\mathcal{N}$ be a multivariate normal distributed random vector independent of X. By setting*

$$\boldsymbol{X} := (X \cdot \exp(\mathcal{N}_1), \ldots, X \cdot \exp(\mathcal{N}_d))^\intercal,$$

we get the weak limit

$$\frac{1}{n} \cdot \boldsymbol{D} \cdot \max_{i=1,\ldots,n} \boldsymbol{X}^{(i)} \xrightarrow{\mathcal{D}} \boldsymbol{Y}$$

as $n \to \infty$, where $\boldsymbol{X}^{(i)}$ are iid copies of $\boldsymbol{X}$, $\boldsymbol{D}$ is the diagonal matrix containing the values $\frac{1}{\mathbb{E}(\exp(\mathcal{N}_j))}$ and $\boldsymbol{Y}$ follows a Hüsler–Reiss distribution with D-norm generator

$$\boldsymbol{Z} = \left(\frac{\exp(\mathcal{N}_1)}{\mathbb{E}(\exp(\mathcal{N}_1))}, \ldots, \frac{\exp(\mathcal{N}_d)}{\mathbb{E}(\exp(\mathcal{N}_d))}\right)^\intercal.$$

We can easily turn this multiplicative model into an additive model by taking the logarithm in every component.

Corollary 9. *Let X be a random variable with $\lim_{t\to\infty} t \cdot P(X > \log(t)) = 1$ and let $\mathcal{N}$ be a multivariate normal distributed random vector independent of X. By setting*

$$\boldsymbol{X} := (X + \mathcal{N}_1, \ldots, X + \mathcal{N}_d)^\intercal,$$

we get the weak limit

$$\max_{i=1,\dots,n} \boldsymbol{X}^{(i)} - \boldsymbol{b} - \log(\boldsymbol{n}) \xrightarrow{\mathcal{D}} \log(\boldsymbol{Y})$$

as $n \to \infty$, where $\boldsymbol{X}^{(i)}$ are iid copies of $\boldsymbol{X}$, $\boldsymbol{b}$ is the vector containing the values $\log(\mathbb{E}(\exp(\mathcal{N}_j)))$ and where we used the notation $\log(\boldsymbol{n}) := (\log(n), \dots, \log(n))^\intercal$ and $\boldsymbol{Y}$ follows the same Hüsler–Reiss distribution as in Corollary 8.

Results like Corollary 8 and 9 have already been established in Krupskii et al. (2018). This has some importance in real world applications. A system whose random behavior is the composite of a Gaussian process and some global effect, which acts on all components at once and behaves like a standard exponential distribution in its tail, has Hüsler–Reiss tail-behavior and the parameters of the Hüsler–Reiss distribution are directly determined by the covariance structure of the Gaussian process.

One problem remains: In practice we will not encounter pure normal distributions, but rather sums of many small perturbations, which we will denote by $\mathbf{V}$ in the following. The rest of this section is about showing that the assertion in Corollary 9 is robust in the sense that we can replace the $\mathcal{N}$'s by sums of small perturbations and still get a limit distribution close to a Hüsler–Reiss distribution.

Theorem 10 (Central limit theorem for Hüsler–Reiss distributions)**.** *Let $\boldsymbol{V} = (V_1, \dots, V_d)^\intercal$ be a random vector with mean vector $\boldsymbol{0} = (0, \dots, 0)^\intercal \in \mathbb{R}^d$ and covariance matrix $A \in \mathbb{R}^{d \times d}$ such that for every $j = 1, \dots, d$ the function $s \mapsto \mathbb{E}(\exp(s \cdot V_j))$ is finite on some open interval containing 0.*

In regular statistics we have a central limit theorem:

$$\boldsymbol{W}(m) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \boldsymbol{V}^{(i)} \xrightarrow{\mathcal{D}} \mathcal{N},$$

as $m \to \infty$ where $\mathcal{N}$ follows a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix A.

If we further introduce a random variable $X \geq 0$ that is both independent of all $\boldsymbol{V}^{(i)}, i \in \mathbb{N}$ and of $\mathcal{N}$ and which further fulfills $\lim_{t\to\infty} t \cdot P(X > \log(t)) = 1$, then we also have the limit

$$\begin{aligned} \boldsymbol{X}(m) &:= (X + W_1(m), \dots, X + W_d(m))^\intercal \\ &\xrightarrow{\mathcal{D}} \boldsymbol{X} := (X + \mathcal{N}_1, \dots, X + \mathcal{N}_d)^\intercal \end{aligned}$$

as $m \to \infty$.

This limit also extends to the tail-behavior. First there exists a number $M \in \mathbb{N}$ such that

$$b_j(m) := \log(\mathbb{E}(\exp(W_j(m))))$$

is well defined for all $m \geq M$ and all $j = 1, \ldots, d$. For every $m \geq M$ there exists a simple max-stable random variable $\boldsymbol{Y}(m)$ that fulfills

$$\max_{i=1,\ldots,n} \boldsymbol{X}^{(i)}(m) - \boldsymbol{b}(m) - \log(\boldsymbol{n}) \xrightarrow{\mathcal{D}} \log(\boldsymbol{Y}(m)),$$

where $\boldsymbol{X}^{(i)}(m)$ are iid copies of $\boldsymbol{X}(m)$, where both the maximum and the logarithm on the right hand side of the limit are meant component-wise and $\log(\boldsymbol{n}) := (\log(n), \ldots, \log(n))^\intercal$.

Also there exists simple max-stable $\boldsymbol{Y}$ following a Hüsler–Reiss distribution that fulfills

$$\max_{i=1,\ldots,n} \boldsymbol{X}^{(i)} - \boldsymbol{b} - \log(\boldsymbol{n}) \xrightarrow{\mathcal{D}} \log(\boldsymbol{Y}),$$

where $\boldsymbol{X}^{(i)}$ are iid copies of $\boldsymbol{X}$ and $\boldsymbol{b}$ is the constant vector containing the values $b_j := \log(\mathbb{E}(\exp(\mathcal{N}_j)))$.

The connection between the tail-behavior of $\boldsymbol{X}(m), m \geq M$ and the tail-behavior of $\boldsymbol{X}$ is

$$\boldsymbol{b}(m) \to \boldsymbol{b} \text{ and } \boldsymbol{Y}(m) \xrightarrow{\mathcal{D}} \boldsymbol{Y} \text{ as } m \to \infty.$$

This central limit theorem confirms that for m high enough we can model the tail-behavior of $\mathbf{X}(m)$ with a Hüsler–Reiss distribution. We have convergence of the shift constants $\mathbf{b}(m) \to \mathbf{b}$ and of the max-stable limit distributions $\mathbf{Y}(m) \xrightarrow{\mathcal{D}} \mathbf{Y}$ (the importance of that can be looked up in Theorem 4).

A good way to remember Theorem 10 is that

$$\text{attractor}\left(\lim_{m\to\infty} \mathbf{X}(m)\right) = \lim_{m\to\infty} \text{attractor}(\mathbf{X}(m))$$

holds, where 'attractor' stands for the max-stable attractor. Mapping a random vector to its attractor is not continuous in general (see the 'pitfalls' in Section 2.8 and also Section 1.4).

We will prove Theorem 10 in the next section.

2.5 Proof of the central limit theorem for the Hüsler–Reiss distribution

A multivariate central limit theorem with the weak limit

$$\mathbf{W}(m) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \mathbf{V}^{(i)} \xrightarrow{\mathcal{D}} \mathcal{N}$$

as $m \to \infty$ is not sufficient for our purposes. What we actually need is

$$\mathbb{E}\left(\max_{j=1,\ldots,d} |y_j| \exp\left(W_j(m)\right) \right) \to \mathbb{E}\left(\max_{j=1,\ldots,d} |y_j| \exp(\mathcal{N}_j) \right)$$

for all $\mathbf{y} \in \mathbb{R}^d$ as $m \to \infty$. In the proof of Theorem 10 we will see why this is so crucial.

Recall the definition of uniform integrability.

Definition 10. *A sequence of integrable random variables $(X_m)_{m\in\mathbb{N}}$ is called* uniformly integrable *if for every $\epsilon > 0$ there is an $\alpha > 0$ such that*

$$\mathbb{E}(|X_m| \cdot 1_{|X_m|>\alpha}) \leq \epsilon$$

for all $m \in \mathbb{N}$. By $1_{|X_m|>\alpha}$ we denote the indicator function which is 1, whenever $|X_m| > \alpha$, and 0 otherwise.

Compare this to the notion of tightness in Definition 5. Tightness means that within the sequence the probability outside of a compact interval has an upper bound, while uniform integrability means that the 'weighted probabilities' outside of a compact interval have an upper bound.

The following auxiliary result is obvious.

Lemma 9. *Let $((X_m, Y_m))_{m\in\mathbb{N}}$ be a sequence of bivariate random vectors with $0 \leq |Y_m| \leq |X_m|$ almost surely for all $m \in \mathbb{N}$. If the sequence $(X_m)_{m\in\mathbb{N}}$ is uniformly integrable, then so is the sequence $(Y_m)_{m\in\mathbb{N}}$.*

Our next auxiliary result is Theorem 5.4 in the book by Billingsley (1968).

Lemma 10. *Let X be an integrable random variable and $(X_m)_{m\in\mathbb{N}}$ be a uniformly integrable sequence of random variables such that the limit $X_m \xrightarrow{\mathcal{D}} X$ holds for $m \to \infty$. Then $\lim_{m\to\infty} \mathbb{E}(X_m) = \mathbb{E}(X)$. It is also true, that the convergence $X_n \xrightarrow{\mathcal{D}} X$ together with $\lim_{m\to\infty} \mathbb{E}(X_m) = \mathbb{E}(X)$ implies the uniform integrability the sequence $(X_m)_{m\in\mathbb{N}}$.*

Recall that the moment generating function M_V of a random variable V is defined as $M_V(s) := \mathbb{E}(\exp(s \cdot V))$. This value is not necessarily finite, but for our central limit theorem we need finiteness on a neighborhood around 0.

Lemma 11. *If V is a random variable with $\mathbb{E}(V) = 0$ and $\mathbb{E}(V^2) = \sigma^2$ and there is an $\epsilon > 0$ such that the moment generating function M_V is finite on $[-\epsilon, \epsilon]$, then we have the expansion*

$$M_V(s) = 1 + \frac{\sigma^2}{2} \cdot s^2 + o(s^2)$$

for $s \to 0$.

Proof. Denote $X_n(s) := \sum_{i=0}^{n} \frac{(sV)^i}{i!}$. Note that these random variables converge pointiwse to the random variable $\exp(s \cdot V)$. The following inequality shows that the dominated convergence theorem is applicable:

$$|X_n(s)| \leq \exp(|sV|) \leq \underbrace{\exp(\epsilon V) + \exp(-\epsilon V)}_{\text{integrable}}$$

for all $s \in [-\epsilon, \epsilon]$ and all $n \in \mathbb{N}$. This shows that $X_n(s)$ is integrable for all $s \in [-\epsilon, \epsilon]$ and all $n \in \mathbb{N}$ and also

$$V^n = \frac{i!}{s^i} \cdot (X_n - X_{n-1})$$

is integrable as well for all $n \in \mathbb{N}$.

We then get

$$M_V(s) = \mathbb{E}(\exp(sV)) = \mathbb{E}\left(\lim_{n\to\infty} X_n(s)\right) = \lim_{n\to\infty} \mathbb{E}(X_n(s)) = \sum_{i=0}^{\infty} \frac{s^i \mathbb{E}(V^i)}{i!}$$

for all $s \in [-\epsilon, \epsilon]$. We have found the Taylor expansion of the moment-generating function M_V around the center 0. This already implies in Landau notation

$$\sum_{i=0}^{\infty} \frac{s^i \mathbb{E}(V^i)}{i!} = 1 + \mathbb{E}(V)s + \frac{\mathbb{E}(V^2)}{2} s^2 + o(s^2) = 1 + \frac{\sigma^2}{2} \cdot s^2 + o(s^2),$$

where we used $\mathbb{E}(V) = 0$ and $E(V^2) = E(V^2) - E(V)^2 = \sigma^2$. □

The Taylor expansion of the moment-generating function is necessary in the proof of the following result:

Lemma 12. *Let $\boldsymbol{V}$ be a d-dimensional random vector with mean vector $\boldsymbol{0} = (0, \ldots, 0)^\intercal \in \mathbb{R}^d$ and covariance matrix $A \in \mathbb{R}^{d\times d}$ and let $\mathcal{N}$ be a multivariate normal random vector also with mean $\boldsymbol{0}$ and covariance matrix A. Let $\boldsymbol{V}^{(i)}, i \in \mathbb{N}$ be iid copies of $\boldsymbol{V}$ and put $\boldsymbol{W}(m) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \boldsymbol{V}^{(i)}$. If there is an $\epsilon > 0$ such for that every $j = 1, \ldots, d$ the random variable V_j has a finite moment generating function $M_{V_j}(s)$ on the interval $[-\epsilon, \epsilon]$, then we have the limit*

$$\lim_{m\to\infty} \mathbb{E}\left(\max_{j=1,\ldots,d} |y_j| \exp(W_j(m))\right) = \mathbb{E}\left(\max_{j=1,\ldots,d} |y_j| \exp(\mathcal{N}_j)\right)$$

for all $\boldsymbol{y} \in \mathbb{R}^d$.

Proof. We have the convergence $\mathbf{W}(m) \xrightarrow{\mathcal{D}} \mathcal{N}$ by the Cramér–Wold device together with a suitable univariate central limit theorem. For a fixed $\mathbf{y} \in \mathbb{R}^d$ we can set $c := \max_{j=1,\dots,d} |y_j|$. We have the convergence

$$\max_{j=1,\dots,d} |y_j| \exp\left(W_j(m)\right) \xrightarrow{\mathcal{D}} \max_{j=1,\dots,d} |y_j| \exp(\mathcal{N}_j)$$

by the continuous mapping theorem. According to Lemma 10 we only need to show the uniform integrability of the sequence $(\max_{j=1,\dots,d} |y_j| \exp(W_j(m)))_{m\in\mathbb{N}}$. We have the inequality

$$0 \leq \max_{j=1,\dots,d} |y_j| \exp(W_j(m) \leq c \cdot \sum_{j=1}^{d} \exp(W_j(m)).$$

By using this inequality and Lemma 10 we only need to show the uniform integrability of $\left(\sum_{j=1}^{d} \exp\left(W_j(m)\right)\right)_{m\in\mathbb{N}}$. We have

$$\begin{aligned}
&\lim_{m\to\infty} \mathbb{E}\left(\sum_{j=1}^{d} \exp(W_j(m))\right) \\
&\overset{a}{=} \sum_{j=1}^{d} \lim_{m\to\infty} \mathbb{E}\left(\exp(W_j(m))\right) \\
&\overset{b}{=} \sum_{j=1}^{d} \lim_{m\to\infty} M_{V_j}\left(\frac{1}{\sqrt{m}}\right)^m \\
&\overset{c}{=} \sum_{j=1}^{d} \lim_{m\to\infty} \left(1 + \frac{a_{jj}/2 + m \cdot o(1/m)}{m}\right)^m \\
&\overset{d}{=} \sum_{j=1}^{d} \exp(a_{jj}/2) = \mathbb{E}\left(\sum_{j=1}^{d} \exp(\mathcal{N}_j)\right),
\end{aligned}$$

where a comes from the linearity of the expected value and the limes. b uses the properties of the exponential function and the definition of $W_j(m)$. In c we used the Taylor expansion from Lemma 11. d is an application of Lemma 4. The second part of Lemma 10 now implies the uniform integrability of the sequence $\left(\sum_{j=1}^{d} \exp(W_j(m))\right)_{n\in\mathbb{N}}$. □

Now that all preparatory results have been proven, the central limit theorem for the Hüsler–Reiss distributions remains.

Proof of Theorem 10. According to Lemma 12 with the choice $\mathbf{y} = \mathbf{e}_j$, the j-th unit vector, we have the limit

$$\lim_{m\to\infty} \mathbb{E}\left(\exp\left(W_j(m)\right)\right) = \mathbb{E}\left(\exp\left(\mathcal{N}_j\right)\right) > 0$$

for every $j = 1, \ldots, d$. Then there exists an $M \in \mathbb{N}$ such that for every $m \geq M$ and every $j = 1, \ldots, d$ the expression $b_j(m) := \log(\mathbb{E}(\exp(W_j(m))))$ is well defined and we also have $\lim_{m\to\infty} b_j(m) = \log(\mathbb{E}(\exp(\mathcal{N}_j))) =: b_j$ for all $j = 1, \ldots, d$.

For $m \geq M$ we can define D-norm generators $\mathbf{Z}(m)$ by

$$Z_j(m) = \frac{\exp(W_j(m))}{\mathbb{E}(\exp(W_j(m)))}, \quad j = 1, \ldots, d.$$

By definition of $X_j(m)$ in Equation (10) we obtain

$$\exp(X_j(m) - b_j(m)) = \exp(X) \cdot Z_j(m)$$

for all $j = 1, \ldots, d$. With this representation and with Theorem 9 applied to $\mathbf{Z}^{(1)} = \mathbf{1}$ almost surely and $\mathbf{Z}^{(2)} := \mathbf{Z}(m)$ we get the limit

$$\max_{i=1,\ldots,n} \mathbf{X}^{(i)}(m) - \mathbf{b}(m) - \log(\mathbf{n}) \xrightarrow{\mathcal{D}} \log(\mathbf{Y}(m)),$$

where $\mathbf{X}^{(i)}(m)$ are iid copies of $\mathbf{X}(m)$, the maximum is meant component-wise and where $\mathbf{Y}(m)$ is simple max-stable with

$$P\left(\mathbf{Y}(m) \leq \frac{1}{\mathbf{y}}\right) = \exp\left(-\mathbb{E}\left(\max_{j=1,\ldots,d} y_j Z_j(m)\right)\right)$$

for all $\mathbf{y} > \mathbf{0}$. If we take the limit $n \to \infty$ we obtain

$$\begin{aligned}\lim_{n\to\infty} P\left(\mathbf{Y}(n) \leq \frac{1}{\mathbf{y}}\right) &= \lim_{n\to\infty} \exp\left(-\mathbb{E}\left(\max_{j=1,\ldots,d} y_j Z_j(n)\right)\right) \\ &= \exp\left(-\mathbb{E}\left(\max_{j=1,\ldots,d} y_j \frac{\exp(\mathcal{N}_j)}{\mathbb{E}(\exp(\mathcal{N}_j))}\right)\right)\end{aligned}$$

for all $\mathbf{y} > \mathbf{0}$, where we used the continuity of the exponential function and Lemma 12. □

We have seen that the parametric model of Hüsler–Reiss distributions is one way to deal with high dimensions in extreme value theory. The number of parameter only grows quadratically in dimension, but its use is justified in the context of the central limit theorem. The next section features a completely different approach to dealing with high dimensionality.

2.6 Dimension reduction with principal component analysis

Before we introduce some new techniques to reduce dimension in extreme value theory, let's first have a look at what we might need it for in practice:

Example 8. *If wind speed is measured simultaneously at d locations, this produces a single observation $\boldsymbol{x} \in \mathbb{R}^d$. If we collect those data over a long time, we get a dataset*

$$M = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\} \subset \mathbb{R}^d$$

of historic wind data. We can also define a set $E \subset \mathbb{R}^d$ as those possible realization of windspeeds, we would call a storm. $M \cap E$ then becomes our historic storm data.

For each historic storm $\boldsymbol{x} \in M \cap E$ we can research whether people were hurt or infrastructure was damaged. This gives us a function

$$\ell : E \cap M \to \{1, 2, 3, 4\},$$

where 1 stands for no harm done, 2 stands for hurt people, 3 stands for damaged infrastructure and 4 stands for both people hurt and infracstucture damaged.

ℓ only classifies past storms. If we could extend it to a classifier

$$\widehat{\ell} : E \to \{1, 2, 3, 4\}$$

that will classify future storms correctly, this would be of great use. During the next storm we could plug the observed speed of wind $\boldsymbol{x} \in E$ into the function $\widehat{\ell}$ and know what to expect.

Learning the classifier $\widehat{\ell}$ is almost a classical classification problem, except only the extreme observations get a label ℓ. But extreme events are by nature rare, so the set M of all training data gets reduced to a much smaller set $E \cap M$ of relevant training data.

This is a problem, because a classifier should generalize from the training data, not memorize it. And the potential of overfitting is large, if there are few training data placed sparsely in many dimensions. It is therefore reasonable to first reduce the dimension of the problem with a transformation $\Phi : E \to E'$, where E' has a lower dimension than E and then to train a classifier $\widehat{\ell}_\Phi$ on the transformed training data

$$\{(\Phi(\mathbf{x}), \ell(\mathbf{x})) : \mathbf{x} \in M \cap E\}$$

and to set the final classifier as

$$\widehat{\ell}(\mathbf{x}) := \widehat{\ell}_\Phi(\Phi(\mathbf{x})), \text{ for all } \mathbf{x} \in E.$$

One the one hand every dimensions we remove reduces the risk of overfitting. But on the other hand every dimension we remove carries the risk of also

removing the information we need for correct classification. To make reasonable decisions in this trade-off scenario we first need a reasonable metric for the 'information loss'. After we have done that we will set our minds to find proper dimension reduction techniques with respect to that metric. This is the main purpose of this section.

Before we go into extreme value theory, let us have a look at dimension reduction in classical statistics:

Lemma 13. *Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\intercal$ be a square-integrable random vector. Further let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz bounded function with a constant L such that*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L \cdot \|\boldsymbol{x} - \boldsymbol{y}\|_2 , \tag{2.10}$$

holds for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, where $\|\cdot\|_2$ denotes the Euclidean norm. Then for every function $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ such that $\Phi(\boldsymbol{X})$ is a square integrable random vector we have

$$|\mathbb{E}(f(\boldsymbol{X})) - \mathbb{E}(f(\Phi(\boldsymbol{X})))| \leq L \cdot \sqrt{\mathbb{E}\left(\|\boldsymbol{X} - \Phi(\boldsymbol{X})\|_2^2\right)}.$$

This lemma shows that $\mathbb{E}\left(\|\mathbf{X} - \Phi(\mathbf{X})\|_2^2\right)$ is a reasonable metric for the information lossfrom replacing $\mathbf{X}$ with $\Phi(\mathbf{X})$.

Proof. It is elementary to show:

$$\begin{aligned} |\mathbb{E}(f(\mathbf{X})) - \mathbb{E}(f(\Phi(\mathbf{X})))| &= |\mathbb{E}(f(\mathbf{X}) - f(\Phi(\mathbf{X})))| \\ &\leq \mathbb{E}(|f(\mathbf{X}) - f(\Phi(\mathbf{X}))|) \\ &\leq L \cdot \mathbb{E}\left(\|\mathbf{X} - A\mathbf{X}\|_2\right). \end{aligned}$$

Recall Jensen's inequality (see Section 5 in the book by Billingsley (1979)), which says that $g(\mathbb{E}(Y)) \leq \mathbb{E}(g(Y))$ for all convex functions g and all integrable random variables Y. Applying this to the convex function $y \mapsto y^2$ and the random variable $Y = \|\mathbf{X} - \Phi(\mathbf{X})\|_2$ we get

$$|\mathbb{E}(f(\mathbf{X})) - \mathbb{E}(f(\Phi(\mathbf{X})))| \leq L \cdot \mathbb{E}\left(\|\mathbf{X} - A\mathbf{X}\|_2\right) \leq L \cdot \sqrt{\mathbb{E}\left(\|\mathbf{X} - \Phi(\mathbf{X})\|_2^2\right)}.$$

□

Now let us only consider Φ that are linear affine, i.e. $\Phi(\mathbf{x}) = A \cdot \mathbf{x} + \mathbf{b}$ for a square matrix A and a vector $\mathbf{b}$. We now wish to optimize the function

$$(A, \mathbf{b}) \mapsto \mathbb{E}\left(\|\mathbf{X} - (A\mathbf{X} + \mathbf{b})\|_2^2\right)$$

under the condition $\operatorname{rank}(A) := \dim(\operatorname{image}(A)) = d'$, where $d' = 0, 1, \ldots, d$. Making no restriction of the dimension leads to the trivial minimizer $A = I$, the $d \times d$ identity matrix.

Lemma 14. *Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\intercal$ be a square integrable random vector with mean vector $\mu_{\boldsymbol{X}} = \mathbb{E}(\boldsymbol{X})$ and covariance matrix $\Sigma = \mathrm{Cov}(\boldsymbol{X})$, where Σ has the ordered eigenvalue*

$$\lambda_1 \geq \cdots \geq \lambda_d \geq 0$$

and corresponding orthogonal eigenvectors

$$\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(d)}.$$

For every $d' \leq d$ a minimizer of the optimization problem

$$\min_{\substack{A \in \mathbb{R}^{d \times d} \\ \mathrm{rank}(A) \leq d'}} \min_{\boldsymbol{b} \in \mathbb{R}^d} \mathbb{E}\left(\|\boldsymbol{X} - (A\boldsymbol{X} + \boldsymbol{b})\|_2^2 \right) \tag{2.11}$$

is the choice $A = P_{d'}$, where $P_{d'}$ is the orthogonal projection onto the subspace

$$V_{d'} := \mathrm{span}(\{\boldsymbol{v}^{(j)} : j = 1, \ldots, d'\})$$

and $\boldsymbol{b} = \mu_{\boldsymbol{X}} - A \cdot \mu_{\boldsymbol{X}}$.

The minimal value is $\sum_{j=d'+1}^{d} \lambda_j$.

Proof. This proof consists of four parts.

(i) We will find a closed form of

$$\min_{\mathbf{b} \in \mathbb{R}^d} \mathbb{E}\left(\|\mathbf{X} - (A\mathbf{X} + \mathbf{b})\|_2^2 \right)$$

for arbitrary matrices A by solving the inner minization problem in Equation (2.11).

(ii) We will show that we can always replace A by the orthogonal projection onto the space image(A) in the outer minization problem in Equation (2.11).

(iii) We will refer to the literature for why $P_{d'}$ is the best orthogonal project for outer optimization problem.

(iv) We will show that the minimal value is $\sum_{j=d'+1}^{d} \lambda_j$

Part (i) can be shown with the well known fact that for a square integrable random variable Y the expression $b \mapsto \mathbb{E}\left((Y-b)^2\right)$ is minimized by the choice $b = \mathbb{E}(Y)$. For a fixed square matrix A we can define a random vector $\mathbf{Y} = \mathbf{X} - A\mathbf{X}$ and get

$$\begin{aligned} \mathbb{E}\left(\|\mathbf{X} - A\mathbf{X} - \mathbf{b}\|_2^2 \right) &= \mathbb{E}\left(\|\mathbf{Y} - \mathbf{b}\|_2^2 \right) \\ &= \sum_{j=1}^{d} \mathbb{E}\left((Y_j - b_j)^2 \right). \end{aligned}$$

Each summand of this can be minimized separately to get the minizer $\mathbf{b} = \mathbb{E}(\mathbf{Y}) = \mu_{\mathbf{X}} - A \cdot \mu_{\mathbf{X}}$. We then have

$$\min_{\mathbf{b}\in\mathbb{R}^d} \mathbb{E}\left(\|\mathbf{X} - (A\mathbf{X} + \mathbf{b})\|_2^2\right) = \mathbb{E}\left(\|\mathbf{X} - \mu_{\mathbf{X}} - A \cdot (\mathbf{X} - \mu_{\mathbf{X}})\|_2^2\right)$$

for arbitrary A.

For Part (ii) let A be an arbitrary square matrix, let $V := \text{image}(A)$ be its image space and let P_V be the orthogonal projection onto V. For every $\mathbf{x} \in \mathbb{R}^d$ the vector $P_V\mathbf{x} - A\mathbf{x}$ is an element of V, because V is linear subspace of $\mathbb{R}^d$. Therefore $P_V(P_V\mathbf{x} - A\mathbf{x}) = P_V\mathbf{x} - A\mathbf{x}$. Consequently we have

$$(P_V\mathbf{x} - A\mathbf{x})^\intercal \cdot (\mathbf{x} - P_V\mathbf{x}) = (P_V\mathbf{x} - A\mathbf{x})^\intercal \cdot \underbrace{(P_V^\intercal\mathbf{x} - P_V^\intercal P_V\mathbf{x})}_{=\mathbf{0}} = 0.$$

This means $P_V\mathbf{x} - A\mathbf{x}$ is orthogonal to the residual $\mathbf{x} - P_V\mathbf{x}$ after the orthogonal projection. This implies

$$\begin{aligned}\|\mathbf{x} - A\mathbf{x}\|_2^2 &= \|\mathbf{x} - P_V\mathbf{X} + P_V\mathbf{x} - A\mathbf{x}\|_2^2 \\ &= \|\mathbf{x} - P_V\mathbf{x}\|_2^2 + \|P_V\mathbf{x} - A\mathbf{x}\|_2^2 \geq \|\mathbf{x} - P_V\mathbf{x}\|_2^2 .\end{aligned}$$

We can replace the constant $\mathbf{x}$ by the random $\mathbf{X} - \mu_{\mathbf{X}}$ and integrate over the both sides of this inequality to get

$$\mathbb{E}\left(\|\mathbf{X} - \mu_{\mathbf{X}} - A(\mathbf{X} - \mu_{\mathbf{X}})\|_2^2\right) \geq \mathbb{E}\left(\|\mathbf{X} - \mu_{\mathbf{X}} - P_V(\mathbf{X} - \mu_{\mathbf{X}})\|_2^2\right).$$

So for the outer minimization problem in Equation (2.11) it is never wrong to replace A by the orthogonal projection onto image(A).

For (iii) it is convenient to set $\mathbf{X}_c := \mathbf{X} - \mu_{\mathbf{X}}$. For every orthogonal projection P we have

$$\text{const} = \mathbb{E}\left(\|\mathbf{X}_c\|_2^2\right) = \mathbb{E}\left(\|\mathbf{X}_c - P\mathbf{X}_c\|_2^2\right) + \mathbb{E}\left(\|P\mathbf{X}_c\|_2^2\right),$$

so minimizing $\mathbb{E}\left(\|\mathbf{X}_c - P\mathbf{X}_c\|_2^2\right)$ is equivalent to the maximization of $\mathbb{E}\left(\|P\mathbf{X}_c\|_2^2\right)$. We will investigate this term further with the trace operator tr, that maps a square matrix to the sum of its diagonal elements. It is well known, that

$$\text{tr}(AB) = \text{tr}(BA),$$

if the row-dimension of A is the column-dimension of B and vice versa. Also we will also identify vectors as $1 \times d$ matrices and real numbers als 1×1 matrices. The trace of a 1×1 matrix is also the single entry it has. For every projection matrix P we have

$$\begin{aligned}\mathbb{E}\left(\|P\mathbf{X}_c\|_2^2\right) &= \mathbb{E}\left(\mathbf{X}_c^\intercal P^\intercal P\mathbf{X}_c\right) \\ &= \mathbb{E}\left(\mathbf{X}_c^\intercal P\mathbf{X}_c\right) \\ &= \mathbb{E}\left(\text{tr}\left(\mathbf{X}_c^\intercal P\mathbf{X}_c\right)\right) \\ &= \mathbb{E}\left(\text{tr}\left(P\mathbf{X}_c\mathbf{X}_c^\intercal\right)\right) \\ &= \text{tr}\left(P \cdot \mathbb{E}\left(\mathbf{X}_c\mathbf{X}_c^\intercal\right)\right) = \text{tr}(P \cdot \Sigma),\end{aligned}$$

where we used the linearity of the trace operator and that $\mathbb{E}(\mathbf{X}_c\mathbf{X}_c^\intercal) = \Sigma$. With similar reasoning we can show that $\mathbb{E}\left(\|\mathbf{Y} - P\mathbf{Y}\|_2^2\right) = \operatorname{tr}((I-P)\cdot\Sigma)$, as the matrix $(I-P)$ is a projection matrix as well.

Paragraph 16.4 in the book by Puntanen et al. (2013) can be used to show that of all rank d' projection matrices the projection $P_{d'}$ onto the subspace $V_{d'}$ maximizes the term $\operatorname{tr}(P\cdot\Sigma)$.

For Part (iv) it is possible to confirm that the projection matrix $P_{d'}$ onto the space $V_{d'}$ and the residual $R_{d'} = I - P_{d'}$ can be represented by

$$\begin{aligned} P_{d'} &= \sum_{j=1}^{d'} \mathbf{v}^{(j)}(\mathbf{v}^{(j)})^\intercal \\ R_{d'} &= \sum_{j=d'+1}^{d} \mathbf{v}^{(j)}(\mathbf{v}^{(j)})^\intercal. \end{aligned}$$

We then end up with

$$\begin{aligned} \mathbb{E}\left(\|\mathbf{Y} - P_{d'}\mathbf{Y}\|_2^2\right) &= \operatorname{tr}(R_{d'}\cdot\Sigma) \\ &= \sum_{j=d'+1}^{d} \operatorname{tr}(\mathbf{v}^{(j)}(\mathbf{v}^{(j)})^\intercal\cdot\Sigma) \\ &= \sum_{j=d'+1}^{d} \operatorname{tr}((\mathbf{v}^{(j)})^\intercal\cdot\Sigma\cdot\mathbf{v}^{(j)}) \\ &= \sum_{j=d'+1}^{d} \operatorname{tr}(\lambda_j) = \sum_{j=d'+1}^{d} \lambda_j. \end{aligned}$$

□

The underlying theory of Lemma 14 is principal component analysis of the covariance matrix. We will now develop a way to apply this to reduce the dimension in multivariate extreme value theory. The interesting part is that both the dimension of the vector $\mathbf{X}$ as well as the dimension of an abstract D-norm generator $\mathbf{Z}$ is reduced. For that we will have to look at the following lemma:

Lemma 15. *Let $\Phi : [0,\infty)^d \to [0,\infty)^d$ be a continuous function that is homogeneous of order 1 in the sense that $\Phi(\lambda\boldsymbol{x}) = \lambda\Phi(\boldsymbol{x})$ for all $\lambda \geq 0$ and all $\boldsymbol{x} \geq \boldsymbol{0}$.*

Let $\boldsymbol{X}$ be a random vector that fulfills one of the equivalent statements of the Rosetta Stone theorem with corresponding D-norm generator $\boldsymbol{Z}$.

For every function $h : [0,\infty)^d \to [0,\infty)$ that is continuous and homogeneous of order 1 we have the following limits:

$$\lim_{t\to\infty} t \cdot P(h(\boldsymbol{X}) > t) = \mathbb{E}(h(\boldsymbol{Z}))$$
$$\lim_{t\to\infty} t \cdot P(h(\Phi(\boldsymbol{X})) > t) = \mathbb{E}(h(\Phi(\boldsymbol{Z}))).$$

If h is also Lipschitz-bounded in the sense that there exists a constant L such that $|h(\boldsymbol{x}) - h(\boldsymbol{y})| \leq L \cdot \|\boldsymbol{x} - \boldsymbol{y}\|_2$, where $\|\cdot\|_2$ denotes the Euclidean norm, than we further have:

$$|\mathbb{E}(h(\boldsymbol{Z})) - \mathbb{E}(h(\Phi(\boldsymbol{Z})))| \leq \mathbb{E}(\|\boldsymbol{Z} - \Phi(\boldsymbol{Z})\|_2) \leq \sqrt{\mathbb{E}\left(\|\boldsymbol{Z} - \Phi(\boldsymbol{Z})\|_2^2\right)}. \quad (2.12)$$

Proof. The limits are a consequence of the Rosetta Stone theorem and the fact that $h \circ \Phi$ is a continuous function that is homogeneous of order 1 as well. The inequalities in Equation (2.12) can be shown exactly as in the proof of Lemma 13. □

One noteworthy detail in Equation (2.12) is that the term $\mathbb{E}(\|\mathbf{Z} - \Phi(\mathbf{Z})\|_2)$ does not depend on the specific choice of the generator, as $\mathbf{z} \mapsto \|\mathbf{Z} - \Phi(\mathbf{Z})\|_2$ is homogeneous of order 1 and we can apply Theorem 3. This quantity solely depends on the underlying D-norm. However this does not apply to the term $\mathbb{E}\left(\|\mathbf{Z} - \Phi(\mathbf{Z})\|_2^2\right)$ as we will see in the following example:

Example 9. *Let $\Phi : [0,\infty)^d \to [0,\infty)^d$ be an arbitrary continuous function that is homogeneous of order 1 and let $\boldsymbol{Z}$ be an arbitratry D-norm generator. Further let U be a random variable following the uniform distribution on the interval $(0, 2)$. Then for every continuous function h that is homogeneous of order 1 we have*

$$\mathbb{E}(h(U \cdot \boldsymbol{Z})) = \mathbb{E}(U \cdot h(\boldsymbol{Z})) = \mathbb{E}(U) \cdot \mathbb{E}(h(\boldsymbol{Z})) = \mathbb{E}(h(\boldsymbol{Z})),$$

where we used that U is also independent of the random variable $h(\boldsymbol{Z})$ and U has expectation 1, which is easy to confirm. According to Example 3 the random vectors $\boldsymbol{Z}$ and $U \cdot \boldsymbol{Z}$ generate the same D-norm. Also with the special choice $h(\boldsymbol{x}) := \|\boldsymbol{x} - \Phi(\boldsymbol{x})\|_2$ we get

$$\mathbb{E}(\|U \cdot \boldsymbol{Z} - \Phi(U \cdot \boldsymbol{Z})\|_2) = \mathbb{E}(\|\boldsymbol{Z} - \Phi(\boldsymbol{Z})\|_2).$$

However it turns out that

$$\begin{aligned}\mathbb{E}\left(\|U \cdot \boldsymbol{Z} - \Phi(U \cdot \boldsymbol{Z})\|_2^2\right) &= \mathbb{E}\left(U^2 \cdot \|\boldsymbol{Z} - \Phi(\boldsymbol{Z})\|_2^2\right) \\ &= \mathbb{E}(U^2) \cdot \mathbb{E}\left(\|\boldsymbol{Z} - \Phi(\boldsymbol{Z})\|_2^2\right) = \frac{4}{3} \cdot \mathbb{E}\left(\|\boldsymbol{Z} - \Phi(\boldsymbol{Z})\|_2^2\right),\end{aligned}$$

where we used that U^2 is independent of $\|\boldsymbol{Z} - \Phi(\boldsymbol{Z})\|_2^2$ and U^2 has the expected value $4/3$.

The switch from $\boldsymbol{Z}$ to $U \cdot \boldsymbol{Z}$ has changed neither the underlying D-norm nor the expression $\mathbb{E}(\|\boldsymbol{Z} - \Phi(\boldsymbol{Z})\|_2)$, but its upper bound $\sqrt{\mathbb{E}\left(\|\boldsymbol{Z} - \Phi(\boldsymbol{Z})\|_2^2\right)}$ was increased by factor $\sqrt{4/3} > 1$. Obviously an upper bound is less useful, the higher it is.

In the following we will restrict ourselves to a D-norm generator $\mathbf{Z}$ that fulfills $\sum_{j=1}^d Z_j = d$ almost surely. Such a generator exists according to Corollary 4 and all generators that fulfill $\sum_{j=1}^d Z_j = d$ almost surely follow the same unique distribution according to Lemma 2.

Let P be a $d \times d$ projection matrix with rank k with the property $\mathbf{1}^\intercal \cdot P = \mathbf{0}^\intercal$. If we replace $\mathbf{Z}$ by $\mathbf{Z}' := P \cdot (\mathbf{Z} - \mathbf{1}) + \mathbf{1}$, then we have

$$\mathbb{E}(\mathbf{Z}') = P \cdot \underbrace{\mathbb{E}(\mathbf{Z}' - \mathbf{1})}_{=\mathbf{0}} + \mathbf{1} = \mathbf{1} \text{ for all } j = 1, \ldots, d$$

and

$$\sum_{j=1}^d Z_j' = \mathbf{1}^\intercal \mathbf{Z}' = \underbrace{\mathbf{1}^\intercal P}_{=\mathbf{0}^\intercal} \cdot (\mathbf{Z} - \mathbf{1}) + \underbrace{\mathbf{1}^\intercal \cdot \mathbf{1}}_{=d} = d \text{ almost surely.}$$

At a first glance it would seem that $\mathbf{Z}'$, which realizes on the rank(P)-dimensional affine space $\{\mathbf{1} + P\mathbf{x} : \mathbf{x} \in \mathbb{R}^d\}$ is a normed D-norm generator as well, but there is no guarantee that $\mathbf{Z}' \geq 0$ holds almost surely.

We will fix this by introducing $\mathbf{Z}'' = \max(\mathbf{Z}', \mathbf{0})$, where the maximum is meant componentwise. Essentially, whenever one of the random variables Z_j' realizes as a negative number it is replaced by 0. It only takes little effort to prove that

$$\left\|\mathbf{Z}'' - \mathbf{Z}\right\|_2^2 \leq \left\|\mathbf{Z}' - \mathbf{Z}\right\|_2^2$$

holds almost surely.

Now will introduce a homogeneous function Φ_P that depends on the underlying projection matrix P. We set

$$\Phi_P(\mathbf{x}) := \max\left(P \cdot \left(\mathbf{x} - \frac{\sum_{j=1}^d x_j}{d} \cdot \mathbf{1}\right) + \frac{\sum_{j=1}^d x_j}{d} \cdot \mathbf{1}, \mathbf{0}\right),$$

where the maximum is once again meant component-wise. Observe that this is in fact continuous and homogeneous of order 1. Also we have $\Phi(\mathbf{Z}) = \mathbf{Z}''$ almost surely, as $\mathbf{Z}$ was chosen as a generator that fulfills $\frac{\sum_{j=1}^d Z_j}{d} = 1$ almost surely.

In terms of dimensions we should note that

$$\text{image}(\Phi_P) \subset \{\max(\mathbf{y}, \mathbf{0}) : \mathbf{y} \in \underbrace{\text{image}(P) + (\mathbb{R} \cdot \mathbf{1})}_{\substack{\text{vectorspace of} \\ \text{dimension rank}(P)+1}}\}.$$

We also have

$$\Phi_P(\mathbf{Z}) = \mathbf{Z}'' \in \{\max(\mathbf{y}, \mathbf{0}) : \mathbf{y} \in \underbrace{\text{image}(P)}_{\substack{\text{vectorspace of} \\ \text{dimension rank}(P)}} + \{\mathbf{1}\}\} \text{ almost surely.}$$

So by replacing $\mathbf{X}$ with $\Phi(\mathbf{X})$ we have reduced the degrees of freedom of $\mathbf{X}$ and by replacing $\mathbf{Z}$ with $\Phi(\mathbf{Z})$ we have have reduced the degrees of freedom of $\mathbf{Z}$. In general $\Phi(\mathbf{X})$ has a different tail behavior than $\mathbf{X}$ in the sense that

$$\lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t) \neq \lim_{t\to\infty} t \cdot P(h(\Phi(\mathbf{X})) > t)$$

in general, but if h is Lipschitz-bounded with constant L, we get

$$\begin{aligned}
&|\lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t) - \lim_{t\to\infty} t \cdot P(h(\Phi(\mathbf{X})) > t)| \\
=&|\mathbb{E}(h(\mathbf{Z})) - \mathbb{E}(h(\Phi(\mathbf{Z})))| \\
\leq& L \cdot \sqrt{\mathbb{E}\left(\|\mathbf{Z} - \Phi(\mathbf{Z})\|_2^2\right)} \\
\leq& L \cdot \sqrt{\mathbb{E}\left(\|(\mathbf{Z} - \mathbb{E}(\mathbf{Z}) - P(\mathbf{Z} - \mathbb{E}(\mathbf{Z})))\|_2^2\right)}.
\end{aligned}$$

So a low value for $\mathbb{E}\left(\|(\mathbf{Z} - \mathbb{E}(\mathbf{Z}) - P(\mathbf{Z} - \mathbb{E}(\mathbf{Z})))\|_2^2\right)$ guarantees that the tail-behavior of the lower-dimensional $\Phi(\mathbf{X})$ is a good approximation for the tail-behavior of the original $\mathbf{X}$. Principal component analysis helps us find such projection matrices:

Corollary 10. *Let $\boldsymbol{Z}$ be a D-norm generator that fulfills $\sum_{j=1}^d Z_j = d$ almost surely. Further let Σ be its Covariance matrix with eigenvalues*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$$

and corresponding orthonormal eigenvectors $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(d)}$. Further let d' be a dimension strictly less than d. Then of all orthogonal projection P with $\text{rank}(P) = d'$ the projection $P_{d'}$ onto the vectorspace $V_{d'} = \text{span}(\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(k)})$ minimizes the expression

$$\mathbb{E}\left(\|(\boldsymbol{Z} - \boldsymbol{1}) - P(\boldsymbol{Z} - \boldsymbol{1})\|_2^2\right).$$

The minimum value is $\sum_{j=j'+1}^d \lambda_j$.

This is an immediate consequence of Lemma 14.

Corollary 11. *If have $\mathbb{E}\left(\|(\boldsymbol{Z} - \boldsymbol{1}) - P_{d'}(\boldsymbol{Z} - \boldsymbol{1})\|_2^2\right) > 0$ in the previous corollary, then we also have $\boldsymbol{1}^\intercal P_k = \boldsymbol{0}^\intercal$.*

Proof. First we will show that $\Sigma\mathbf{1} = \mathbf{0}$, where Σ was the covariance matrix of the D-norm generator $\mathbf{Z}$ that fulfills $\sum_{j=1}^{d} Z_j = d$ almost surely. Observe that

$$\mathbf{1}^\intercal \Sigma \mathbf{1} = \mathbb{E}\left(\left(\sum_{j=1}^{d}(Z_j - 1)\right)^2\right) = \mathbb{E}(0) = 0.$$

At first this looks weaker than $\Sigma\mathbf{1} = \mathbf{0}$. The matrix Σ however is symmetric and positive semidefinite and therefore there exists a matrix A with the property $A^\intercal A = \Sigma$. And we get

$$\|A\mathbf{1}\|_2^2 = \mathbf{1}^\intercal \Sigma \mathbf{1} = 0$$

and consequently $A\mathbf{1} = \mathbf{0}$ and finally $\Sigma\mathbf{1} = A^\intercal A\mathbf{1} = A^\intercal\mathbf{0} = \mathbf{0}$.

$P_{d'}$ is the projection onto the space $\operatorname{span}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(d')})$, which are eigenvectors to eigenvalues $\geq \lambda_{d'}$. If $\lambda_{d'}$ was 0, then we would have $0 = \sum_{j=k+1}^{d} \lambda_j = \mathbb{E}\left(\|(\mathbf{Z} - \mathbf{1}) - P_k(\mathbf{Z} - \mathbf{1})\|_2^2\right)$, which would violate the condition of this corollary. Consequently all $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(d')}$ are eigenvectors of Σ to eigenvalues greater zero and consequently perpendicular to $\mathbf{1}$, which is an eigenvector to the eigenvalue zero. Therefore $\mathbf{1}^\intercal P_{d'} = \mathbf{0}^\intercal$. □

We will now illustrate dimension reduction with two examples. They are the extreme cases of tail-dependence. First the case of complete tail-dependence in Example 10 and then the case of tail-independence in Example 11.

Example 10. *The D-norm that corresponds to complete tail-dependence is the norm* $\|\boldsymbol{x}\|_D = \max_{j=1,\ldots,d} |x_j|$. *A generator* $\boldsymbol{Z}$ *with* $P(\boldsymbol{Z} = \boldsymbol{1}) = 1$ *generates this D-norm and fulfills* $\sum_{j=1}^{d} Z_j = d$ *almost surely. The covariance matrix of* $\boldsymbol{Z}$ *is the matrix that has the value* 0 *in every entry.*

No matter what projection matrix P *we chose, we get*

$$\boldsymbol{Z}' = \boldsymbol{1} + P \cdot \underbrace{(\boldsymbol{Z} - \boldsymbol{1})}_{=\boldsymbol{0} \text{ almost surely}} = \boldsymbol{1} = \boldsymbol{Z}$$

almost surely. In a way we can no longer reduce the dimension of $\boldsymbol{Z}$*, as* $\boldsymbol{Z}$ *already realizes on the '0-dimensional' set* $\{\boldsymbol{1}\}$*. But we still can reduce the dimension of* $\boldsymbol{X}$*, where* $\boldsymbol{X}$ *is a random vector on* $[0, \infty)^d$ *with*

$$\lim_{t\to\infty} t \cdot P(h(\boldsymbol{X}) > t) = \mathbb{E}(h(\boldsymbol{Z})) = h(\boldsymbol{1})$$

for every non-negative continuous function h *that is homogeneous of order* 1*. By choosing* $P = 0$*, the matrix that has the values* 0 *in every entry, we get:*

$$\Phi_P(\boldsymbol{X}) = \frac{\sum_{j=1}^{d} X_j}{d} \cdot \boldsymbol{1}.$$

The random vector $\Phi_P(\boldsymbol{X})$ *only realizes on the one-dimensional set*

$$\{\lambda \cdot \boldsymbol{1} : \lambda \geq 0\},$$

but it has the same tail-behavior as $\boldsymbol{X}$ *in the sense of*

$$\lim_{t\to\infty} t \cdot P(h(\boldsymbol{X}) > t) = \mathbb{E}(h(\boldsymbol{Z})) = \mathbb{E}(h(\boldsymbol{Z}')) = \lim_{t\to\infty} t \cdot P(h(\Phi_P(\boldsymbol{X}) > t)$$

for every non-negative, continuous h *that is homogeneous of order* 1.

Example 11. *The D-norm that corresponds to tail-independence is the norm* $\|\boldsymbol{x}\|_D = \sum_{j=1}^d |x_j|$. *A generator* $\boldsymbol{Z}$ *with* $P(\boldsymbol{Z} = d \cdot \boldsymbol{e}_j) = 1/d$, *for all* $j = 1, \ldots, d$ *and where* $\boldsymbol{e}_j$ *is the* j*-th unit vector, generates this D-norm and fulfills* $\sum_{j=1}^d Z_j = d$ *almost surely. It turns out that the covariance matrix* Σ *of* $\boldsymbol{Z}$ *has the entries:*

$$\sigma_{ij} = \begin{cases} d-1 & \text{if } i = j \\ -1 & \text{else.} \end{cases}$$

It takes little effort to show that

$$\Sigma \boldsymbol{1} = \boldsymbol{0} \text{ and}$$

$$\Sigma \boldsymbol{x} = d \cdot \boldsymbol{x} \text{ if } \sum_{j=1}^d x_j = 0.$$

This implies that the ordered eigenvalues of Σ *are*

$$\lambda_1 = \cdots = \lambda_{d-1} = d$$
$$\lambda_d = 0.$$

The projections P_k *in Corollary 10 are not unique (as we can find different orthonormal bases of the* $(d-1)$*-dimensional eigenspace to the eigenvector* $\lambda = d$*, but in any case those* P_k *fulfill*

$$\mathbb{E}\left(\|(\boldsymbol{Z} - \boldsymbol{1}) + P_k \cdot (\boldsymbol{Z} - \boldsymbol{1})\|_2^2\right) = \sum_{j=k+1}^d \lambda_j = (d - k + 1) \cdot d.$$

One can say that each successive reduction by one dimension 'costs the same'.

We will end this section with a variation of principal component analysis. Again we will have a continuous transformation $\Phi : [0,\infty)^d \to [0,\infty)^d$ that is homogeneous of order 1. Again for every projection matrix P we have one such transformation Φ_P. But this Φ_P is defined differently than before:

Definition 11. *Let* P *be a projection matrix. Then we will define the square space projection* Φ_P *by*

$$\Phi_P(\boldsymbol{x}) := (\max(y_1, 0)^2, \ldots, \max(y_d, 0)^2)^\intercal,$$

where $y_1, \ldots, y_d$ *are the entries of the vector* $\boldsymbol{y} := P\sqrt{\boldsymbol{x}}$.

The square space projections are obviously continuous and homogeneous of order 1. Also we have

$$\mathbb{E}\left(\left\|\sqrt{\mathbf{Z}}-\sqrt{\Phi_P(\mathbf{Z})}\right\|_2^2\right)\leq\mathbb{E}\left(\left\|\sqrt{\mathbf{Z}}-P\sqrt{\mathbf{Z}}\right\|_2^2\right)$$

for every D-norm generator $\mathbf{Z}$. If $\mathbf{X}$ is a vector $[0,\infty)^\intercal$, such that the equivalent statements of the Rosetta Stone theorem hold for $\mathbf{X}$ and $\mathbf{Z}$, then we also have:

$$\begin{aligned}
d &= \lim_{t\to\infty} t\cdot P\left(\sum_{j=1}^{d} X_j > t\right)\\
&= \lim_{t\to\infty} t\cdot P\left(\left\|\sqrt{\mathbf{X}}\right\|_2^2 > t\right)\\
&= \mathbb{E}\left(\left\|\sqrt{\mathbf{Z}}\right\|_2^2\right)\\
&= \mathbb{E}\left(\left\|P\sqrt{\mathbf{Z}}\right\|_2^2+\left\|\sqrt{\mathbf{Z}}-P\sqrt{\mathbf{Z}}\right\|_2^2\right)\\
&= \mathbb{E}\left(\left\|P\sqrt{\mathbf{Z}}\right\|_2^2\right)+\mathbb{E}\left(\left\|\sqrt{\mathbf{Z}}-P\sqrt{\mathbf{Z}}\right\|_2^2\right)\\
&= \lim_{t\to\infty} t\cdot P\left(\left\|P\sqrt{\mathbf{X}}\right\|_2^2 > t\right)+\lim_{t\to\infty} t\cdot P\left(\left\|\sqrt{\mathbf{X}}-P\sqrt{\mathbf{X}}\right\|_2^2 > t\right).
\end{aligned}$$

Essentially the mass in the tail is split up between the projection and the remainder.

Corollary 12. *Let $\boldsymbol{X}$ be a random vector on $[0,\infty)^d$ that fulfills one of the equivalent statements of the Rosetta Stone theorem with co-extremal matrix*

$$C=(c_{ij})_{1\leq i,j\leq d}=\left(\lim_{t\to\infty} t\cdot P(X_iX_j>t^2)\right)_{1\leq i,j\leq d}$$

Let

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d\geq 0$$

be the eigenvalues of C with corresponding orthonormal eigenvectors $\boldsymbol{v}^{(1)},\ldots,\boldsymbol{v}^{(d)}$.

Further let d' be a dimension less than d. Then of all orthogonal projection P with $\operatorname{rank}(P)=d'$ the projection $P_{d'}$ onto the vectorspace $V_k=\operatorname{span}(\boldsymbol{v}^{(1)},\ldots,\boldsymbol{v}^{(d')})$ minimizes the expression

$$\lim_{t\to\infty} t\cdot P\left(\left\|\sqrt{\boldsymbol{X}}-P\sqrt{\boldsymbol{X}}\right\|_2^2 > t\right).$$

The minimum value is $\sum_{j=k+1}^{d}\lambda_j$.

Proof. Let $\mathbf{Z}$ be a generator of the corresponding D-norm. Prior to this corollary we have shown that

$$\mathbb{E}\left(\left\|P\sqrt{\mathbf{Z}}\right\|_2^2\right) + \lim_{t\to\infty} t \cdot P\left(\left\|\sqrt{\mathbf{X}} - P\sqrt{\mathbf{X}}\right\|_2^2 > t\right) = d = \text{const},$$

for all projections P, so the minimization problem of the right summand is equivalent to the maximization problem of the left summand in this equation.

Exactly as in the proof of Lemma 14 we can show that

$$\mathbb{E}\left(\left\|P\sqrt{\mathbf{Z}}\right\|_2^2\right) = \operatorname{tr}\left(P \cdot \mathbb{E}\left(\sqrt{\mathbf{Z}} \cdot \sqrt{\mathbf{Z}}\right)\right) = \operatorname{tr}(P \cdot C),$$

where C is the co-extremal matrix. At this point we can proceed just like in the proof of Lemma 14. □

Working with the co-extremal matrix C instead of Σ has an advantage: Each entry c_{ij} of this matrix only depends on the joint extremal behavior of X_i and X_j. It can be estimated with a structure variable estimator (see Section 3.1) or with a local threshold procedure (see Section 3.4).

For the matrix Σ however there is no obvious way to estimate it. Most likely one could adapt the global threshold procedure for a consistent estimator. This is troublesome, as our ultimate goal is to handle datasets with many dimension, datasets with missing values, merged datasets from different sources and the global threshold procedure does not fare well in those scenarios.

As for the root-space principal component analysis there is an open problem: Is there a monotone function f such that we have

$$\mathbb{E}(\|\mathbf{Z} - \Phi_P(\mathbf{Z})\|) \leq f\left(\mathbb{E}\left(\left\|\sqrt{\mathbf{Z}} - \sqrt{\Phi_P(\mathbf{Z})}\right\|_2^2\right)\right)?$$

Other instances of when people used dimension reduction in multivariate extremes include the following:

Example 12. *Very similar to our dimension reduction techniques is the approach in a work by Drees and Sabourin (2019). They work with the assumption that there exists a linear subspace $V \subset \mathbb{R}^d$ such that*

$$\nu([0,\infty)^d \setminus V) = 0,$$

where ν is the measure from the Rosetta Stone theorem and aim to indentify V from the the data.

Under their assumption the measure ν is uniquely identified by its restriction

$$\nu|_{V\cap([0,\infty)^d)\setminus\{\mathbf{0}\})}$$

and they don't lose any information about the multivariate tail-behavior by switching from $[0,\infty)^d$ to the lower dimensional $[0,\infty)^d \cap V$.

In this thesis we did not make this assumption about the underlying tail-dependence here and consequently had to pay for every reduction of dimension.

A similar approach is used in a work by Clémençon et al. (2017), where they call an observation an anomaly if it falls into a set $t \cdot M$, where M is a cone with $\nu(M) \approx 0$ as this is at odds with the limit

$$\lim_{t \to \infty} t \cdot P(\boldsymbol{X} \in t \cdot M) = \nu(M) \approx 0.$$

Example 13. *Chautru (2015) reduces dimension by picking subsets (clusters) of indices $I \subset \{1, \ldots, d\}$ and modelling the dependence structure between $(X_i)_{i \in I}$ rather than the dependence structure of $(X_i)_{i \in \{1,\ldots,d\}}$.*

2.7 Exploratory extreme value analysis

In Section 1.1 we introduced tail-behavior not as a mathematical object, but as mathematical properties that help us answer questions about the appearance of extreme events. Following this philosophy tail-dependence between the components $X_1, \dots, X_d$ of a random vector $\mathbf{X}$ should be mathematical properties that help us answer question about the joint appearance of extreme events. The D-norm framework is tailored to approach problems like the following:

Example 14. *Let $\boldsymbol{X} = (X_1, \dots, X_d)^\intercal$ be a random vector with continuous marginal distributions $F_j(x) = P(X_j \leq x), j = 1, \dots, d$, such that*

$$\boldsymbol{X}' := \left(\frac{1}{1 - F_1(X_1)}, \dots, \frac{1}{1 - F_d(X_d)} \right)^\intercal$$

fulfills one of the equivalent statements of the Rosetta Stone theorem with D-norm $\|\cdot\|_D$ and D-norm generator $\boldsymbol{Z} = (Z_1, \dots, Z_d)$.

We now want to approximate $P(X_1 > x_1, X_2 > x_2)$, where the individual probabilites $p_1 = P(X_1 > x_1)$ and $p_2 = P(X_2 > x_2)$ are small. We use the Rosetta Stone theorem to get:

$$\begin{aligned}
\frac{1}{p_1} \cdot P(X_1 > x_1, X_2 > x_2) &= \frac{1}{p_1} \cdot P\left(X_1' > \frac{1}{p_1}, X_2' > \frac{1}{p_2} \right) \\
&= \frac{1}{p_1} \cdot P\left(\min\left(X_1', \frac{p_2}{p_1} X_2' \right) > \frac{1}{p_1} \right) \\
&\approx \mathbb{E}\left(\min\left(Z_1, \frac{p_2}{p_1} Z_2 \right) \right) \\
&= \mathbb{E}\left(Z_1 + \frac{p_2}{p_1} Z_2 - \max\left(Z_1, \frac{p_2}{p_1} Z_2 \right) \right) \\
&= 1 + \frac{p_2}{p_1} - \left\| \left(1, \frac{p_2}{p_1} \right)^\intercal \right\|_D
\end{aligned}$$

and therefore $P(X_1 > x_1, X_2 > x_2) \approx p_1 + p_2 - \|(p_1, p_2)^\intercal\|_D$.

In this example we were given a concrete problem, which was 'What is the value of $P(X_1 > x_1, X_2 > x_2)$?' and our solution was an extrapolation justified by the Rosetta Stone theorem. In contrast we might be given the task to 'explore' the data.

In this section we will introduce some tools to do exploratory multivariate extreme value analysis. A task that falls unter this umbrella is the following:

Definition 12. *We will call it the 'Cluster Problem' to find clusters of indices in $\{1, \dots, d\}$ such that extreme events tend to not hit more than one cluster at the same time. A solution to the Cluster Problem has to include a reasonable metric what constitutes a good clustering and a strategy or an algorithm to find one.*

We call it a binary clustering if the set $\{1, \ldots, d\}$ splits into a disjoint union of exactly two clusters.

The following example shows that binary clusterings can be turned into more general clusterings:

Example 15. *Assume we have two binary clusterings (I_+^1, I_-^1) and (I_+^2, I_-^2) of the same index set $I = \{1, \ldots, d\}$. Then we can produce a clustering on I with $2^2 = 4$ clusters by intersecting the binary clusterings, by which we mean the following:*

$$\begin{aligned} I_{++} &= I_+^1 \cap I_+^2 \\ I_{+-} &= I_+^1 \cap I_+^2 \\ I_{-+} &= I_+^1 \cap I_-^2 \\ I_{--} &= I_-^1 \cap I_-^2. \end{aligned}$$

If an extreme event now hits for example I_{++} at the same time as it hits I_{+-}, then it also hits I_+^2 at the same time as it hits I_-^2, which whould happen rarely if (I_+^2, I_-^2) is a good binary clustering. The resulting non-binary clustering inherits the quality of the underlying binary clusterings.

Note that Example 15 is not useful if either one binary clustering is trivial (e.g. $I_+ = \emptyset$) or if the binary clusters are too similar to one another with the extreme case $I_+^1 = I_+^2$. The optimal case would be something like

$$\begin{aligned} \frac{|I_+^i|}{d} = \frac{|I_-^i|}{d} = 1/2 \text{ for all } i = 1, 2 \text{ and} \\ \frac{|I_{++}|}{d} = \frac{|I_{+-}|}{d} = \frac{|I_{-+}|}{d} = \frac{|I_{--}|}{d} = 1/4. \end{aligned}$$

If we introduce vectors $\mathbf{y}^{(1)}, \mathbf{y}^{(2)} \in \{-1, 1\}^d$ with

$$j \in I_+^{(i)} \Leftrightarrow y_j^{(i)} > 0 \text{ for all } j = 1, \ldots, d \text{ and all } i = 1, 2, \tag{2.13}$$

then the optimal case above is equivalent to

$$\begin{aligned} \mathbf{y}^{(i)} \perp \mathbf{1} \text{ for } i = 1, 2 \\ \mathbf{y}^{(1)} \perp \mathbf{y}^{(2)}. \end{aligned}$$

So to produce good non-binary clusters it is sufficient to look for orthogonal vectors $\mathbf{y} \in \{-1, 1\}^d$ that are also perpendicular to $\mathbf{1} = (1, \ldots, 1)^\intercal$ and which also produce good individual binary clusters. Those are pretty strong restrictions. The size of our search-space $\{-1, 1\}^d$ grows exponentially in d. Also the condition $\sum_{j=1}^d y_j^{(i)} = 0$ can only be fulfilled if d is an even number.

If we lessen our restrictions we actually end up in a scenario, where principal component analysis is our solution.

First we lift the restriction $\mathbf{y} \in \{-1, 1\}^d$ to $\mathbf{y} \in \mathbb{R}^d$. Those vectors still create binary clusterings by Equation 2.13. We still want orthogonal vectors $\mathbf{y}^{(i)}$ and we still require that there exists a $\mathbf{x} > \mathbf{0}$ with $\mathbf{x}^\intercal \mathbf{y}^{(i)} = 0$ for all i.

As we will later see the eigenvalue/eigenvector decomposition of the co-extremal matrix C will provide such vectors. But first we need some theory.

Lemma 16. *Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\intercal$ be a random vector that fulfills one of the equivalent statements of the Rosetta Stone theorem and which has the co-extremal matrix $C = (c_{ij})_{1 \leq i,j \leq d}$. Further let $\boldsymbol{y} = (y_1, \ldots, y_d)^\intercal$ be an arbitrary vector in $\mathbb{R}^d$. Then we have*

$$\boldsymbol{y}^\intercal C \boldsymbol{y} = \lim_{t \to \infty} t \cdot P\left(\left(\sum_{j=1}^{d} y_j \cdot \sqrt{X_j} \right)^2 > t \right). \tag{2.14}$$

The Co-extremal matrix is positive semidefinite.

Proof. We have

$$\begin{aligned} \mathbf{y}^\intercal C \mathbf{y} &= \sum_{i=1}^{d} \sum_{j=1}^{d} y_i y_j \mathbb{E}(\sqrt{Z_i Z_j}) \\ &= \mathbb{E}\left(\sum_{i=1}^{d} \sum_{j=1}^{d} y_i y_j \sqrt{Z_i Z_j} \right) = \mathbb{E}\left(\left(\sum_{j=1}^{d} y_j \cdot \sqrt{Z_j} \right)^2 \right). \end{aligned}$$

Note that $\mathbf{z} \mapsto (\sum_{j=1}^{d} y_j \cdot \sqrt{z_j})^2$ is non-negative, continuous and homogeneous of order 1. Therefore the Rosetta Stone theorem is applicable, which leads us to Equation (2.14). Because this equation holds for all $\mathbf{y}$, the co-extremal matrix is positive semidefinite. □

For a given $\mathbf{y} \in \mathbb{R}^d$ Equation (2.13) gives a binary clustering (I_+, I_-). We can write the following:

$$\left(\sum_{j=1}^{d} y_j \cdot \sqrt{X_j} \right)^2 = \left(\sum_{i \in I_+} |y_i| \sqrt{X_j} - \sum_{i \in I_-} |y_i| \sqrt{X_j} \right)^2.$$

Extreme events that only hit indices I_+ lead to a high positive value inside the brackets, while extreme events that only hit indices in I_- lead to a low negative value. Extreme events that hit both clusters in general don't cancel the value to zero, but should go together with a reduced probability $P\left(\left(\sum_{j=1}^{d} y_j \cdot \sqrt{X_j} \right)^2 > t \right)$ and therefore a low value of $\mathbf{y}^\intercal C \mathbf{y}$ by Lemma 16.

At first it looks like $\mathbf{y}^\intercal C \mathbf{y}$ is a reasonable metric for the quality of the cluster that $\mathbf{y}$ generates by Equation (2.13). But only at first glance, because if what

happens if we replace $\mathbf{y}$ by $c \cdot \mathbf{y}$, where $c > 1$? We have the same binary clustering, but the value $\mathbf{y}^\intercal C\mathbf{y}$ is increased artificially!

Therefore $\mathbf{y}^\intercal C\mathbf{y}$ is a reasonable metric for the quality of a binary cluster if we further make the restriction $\|\mathbf{y}\|_2 = 1$, where $\|\cdot\|_2$ is the Euclidean norm.

Now let

$$\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_d \geq 0$$

be the eigenvalues of C with multiplicities and let $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(d)}$ be the corresponding orthogonal eigenvectors with $\left\|\mathbf{y}^{(j)}\right\|_2 = 1$ for all $j = 1, \ldots, d$.

Then we have orthogonal vectors $\mathbf{y}^{(j)}, j = 1, \ldots, d$ with

$$(\mathbf{y}^{(j)})^\intercal C \cdot \mathbf{y}^{(j)} = \lambda_j.$$

In this sequence the first elements produce the strongest binary clusterings. Under a weak assumption only the first eigenvector produces a trivial binary clustering, as we will see in the following Lemma:

Lemma 17. *Let C be a co-extremal matrix with $c_{ij} > 0$ for all $1 \leq i, j \leq d$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_d \geq 0$ and corresponding orthogonal eigenvectors with Euclidean norm 1. Then $\boldsymbol{y}^{(1)} > \boldsymbol{0}$ or $-\boldsymbol{y}^{(1)} > \boldsymbol{0}$.*

Proof. First we will define $\mathbf{v} := |\mathbf{y}^{(1)}|$, where the absolute value is defined component-wise. Then $\mathbf{v}$ has Euclidean norm 1 and it has a representation

$$\mathbf{v} = \sum_{j=1}^{d} a_j \mathbf{y}^{(i)},$$

where a_j are real numbers satisfying $\sum_{j=1}^{d} a_j^2 = 1$. This is possible because the change of Euclidean coordinates to coordinates in the orthonormal base $\{\mathbf{y}^{(j)} : j = 1, \ldots, d\}$ is an isometric linear transformation.

Then on the one hand we have

$$\mathbf{v}^\intercal C\mathbf{v} = \sum_{j=1}^{d}\sum_{i=1}^{d} a_i \cdot a_j \cdot \underbrace{(\mathbf{y}^{(j)})^\intercal C\mathbf{y}^{(i)}}_{=\delta(i,j)\cdot\lambda_j} = \sum_{j=1}^{d} a_j^2 \cdot \lambda_j \leq \lambda_1,$$

where $\delta(i, j) = 1$ if $i = j$ and $\delta(i, j) = 0$ else. On the other hand we have

$$\mathbf{v}^\intercal C\mathbf{v} = \sum_{j=1}^{d}\sum_{i=1}^{d} c_{ij}|y_i^{(1)}| \cdot |y_i^{(j)}| \geq \sum_{j=1}^{d}\sum_{i=1}^{d} c_{ij} \cdot y_i^{(1)} \cdot y_i^{(j)} = \lambda_1.$$

If there was a pair of indices (i, j) with $y_j \cdot y_i < 0$, then the second inequality would be strict, which would violate the first inequality. This means we either have $\mathbf{y}^{(i)} \geq \mathbf{0}$ or $-\mathbf{y}^{(i)} \geq \mathbf{0}$. Now we will show that those inequalities have to be strict. We do that by contradiction.

Assume there was an index j with $v_j = 0$. Then for every $\epsilon \in [0,1]$ we could define a vector $\mathbf{v}(\epsilon) = \sqrt{1-\epsilon^2}\mathbf{v} + \epsilon \cdot \mathbf{e}_j$, where $\mathbf{e}_j$ is the j-th unit vector. $\mathbf{v}(\epsilon)$ has Euclidean norm 1 and therefore

$$\mathbf{v}(\epsilon)^\intercal C \mathbf{v}(\epsilon) \leq \lambda_1 \tag{2.15}$$

with the same reasons as above. But we would have

$$\begin{aligned} &\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0} \mathbf{v}(\epsilon)^\intercal \cdot C \cdot \mathbf{v}(\epsilon) \\ &= \underbrace{\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0} (1-\epsilon^2)}_{=0} \cdot \mathbf{v}^\intercal C \mathbf{v} + \underbrace{\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0} \epsilon^2}_{=0} \cdot \mathbf{e}_j^\intercal C \mathbf{e}_j + \underbrace{\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0} \sqrt{1-\epsilon^2} \cdot \epsilon}_{=1} \cdot 2\mathbf{v}^\intercal C \mathbf{e}_j, \end{aligned}$$

but all components of the vector $C\mathbf{e}_j$ are positive (this is the first and only time we use the condition $c_{ij} > 0$ for all i, j) and thus $\mathbf{v}^\intercal C \mathbf{e}_j > 0$. So for small values of ϵ the inequality in Equation (2.15) is violated. By contradiction there cannot be such an index j. Therefore $\mathbf{v} > \mathbf{0}$ and either $\mathbf{y}^{(1)} = \mathbf{v} > \mathbf{0}$ or $-\mathbf{y}^{(1)} = \mathbf{v} > \mathbf{0}$. □

Under the condition of Lemma 17 all orthogonal eigenvectors $\mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(d)}$ that follow after the eigenvector $\mathbf{y}^{(1)}$ produce binary clusterings, that are non-trivial (because they are all perpendicular $|\mathbf{y}^{(1)}| > \mathbf{0}$) and dissimilar to each other (because of orthogonality) and which are ordered by strength, with the strongest being produces by $\mathbf{y}^{(2)}$ and the weakest produced by $\mathbf{y}^{(d)}$.

Example 16. *If each index $j \in \{1, \ldots, d\}$ corresponds to a position on a 2-dimensional map, the binary clustering produced by an (eigen)vector $\boldsymbol{y}$ can visualized the following way: Plot the map and at each of the d locations plot a symbol, where the color of the symbol represents if $y_j > 0$ or $y_j < 0$ and the size of the symbol is proportional to $|y_j|$.*

This was actually in done in Figure 2.8 with the colours black and white on a map of Germany. The 3 plots in the right column are the first three eigenvectors of the co-extremal matrix C of extreme temperature drops in winter.

From a cluster analysis standpoint the following happens: The first eigenvector produces the trivial binary clustering 'All Locations vs Empty Set', the second eigenvector produces the binary clustering 'North-East Germany vs South-West Germany', while the third eigenvector procuces the binary clustering 'North-West vs Central Germany'. Doing all intersections as in Example 15 we a clustering of north-west Germany vs central Germany vs south-west Germany.

This is not the first instance principal component analysis was used in extreme value analysis. Cooley and Thibaud (2018) introduced a matrix that contains measures of pairwise extremal dependence, too, so they could find a discrete model thats produces the same matrix. In their data analysis part they described eigenvectors with geographic directions (north, south, east and west), too.

Definition 13. *If we can match explanatory variables/features with eigenvectors, there still remains the problem, whether this is just a coincidence or the effect of a causal relationship. We will call this the 'Attribution Problem'.*

One can solve the Attribution Problem with absolute certainty only in a controlled environment, like laboratory experiments or computer simulations. Investigating extreme events with those tools is not unheard of, see the following two examples:

Example 17. *The sinking of the M.V.*[1] *Derbyshire faced the court with many technical questions like how much pressure could hitting waves exert on hold covers under different conditions. Heffernan and Tawn (2003) based their extreme value analysis for this question on data produced by a* $1:65$ *replica of the Derbyshire placed in a programmable test tank.*

Example 18. *Sippel et al. (2015) explore the idea of combining extreme value theory with computer simulations to investigate rare weather events.*

[1] motor vessel

2.8 Comparing extremal and non-extremal dependence

This main goal of this section is to introduce a novel way to compare the joint behavior of extremal events with the joint behavior of non-extremal events of a random vector $\mathbf{X} = (X_1, \ldots, X_d)^\intercal$.

The following example will introduce the concept.

Example 19. *Let there be* 4 *cities* A, B, C *and* D *such that between both* A *and* B *and also between* C *and* D *you can travel by plane and by train. Let* T_{AB} *be the time it takes to travel between* A *and* B *by train and let* P_{AB} *be the price it takes to travel between* A *and* B *by plane.* T_{CD} *and* P_{CD} *are corresponding values if you travel between* C *and* D.

We cannot compare T_{AB} *with* P_{AB}. *One is time and the other is currency. However we would expect that both of them are monotone functions of the distance between city* A *and city* B. *A longer distance means both more time to travel by train and more kerosine burnt by the plane that someone has to pay for with their ticket price. When this expectation is true, then Plane-Ticket-Price* P *is a monotone function of Train-Travel-Time* T.

But what if it turns out that

$$T_{AB} < T_{CD} \text{ and} \tag{2.16}$$

$$P_{AB} > P_{CD}? \tag{2.17}$$

This does not violate the laws of logic, physics or economics, it just defies our expectation. We also get an immediate push to investigate this further. Are there in differences in infrastructure or terrain, that negatively affect train-travel-time between C *and* D*? We could have a look at a good map and find out. Or is there an economy of scale effect for air-travel between city* C *and* D*? We could have a look at what types of planes are used in air travel between those pair of cities, how full they are, etc and interview someone from the industry.*

If we are given a larger data-set about train-travel-times and plane-ticket-price between many different cities, we should look for instances of four cities (A, B, C, D) *that fulfill Equation* (2.16) *and* (2.17). *We will call this an instance of reversion.*

If our larger data-set contains no reversion and additionally we have $T_{AB} \neq T_{CD}$, *when* (A, B) *is a different pair than* (C, D), *then there exists a monotone increasing function* f *with*

$$f(T_{AB}) = P_{AB} \text{ for every pair of cities } (A, B). \tag{2.18}$$

So our exploratory analysis has two possible outcomes: We find reversions, which are interesting or we find a monotone function f *that fulfills Equation* (2.18). *A possible application of such a function* f *is the following:*

Assume that in our data-set there is a city Z *such that the values*

$$P_{AZ}, P_{BZ}, P_{CZ}, \ldots$$

are known but the values

$$T_{AZ}, T_{BZ}, T_{CZ}, \dots$$

are not. Then we can predict these with $\widehat{P}_{AZ} = f(T_{AZ}), \widehat{P}_{BZ} = f(T_{BZ}), \widehat{P}_{CZ} = f(T_{CZ}), \dots$.

These predictors are valid in the following sense: Let A *be a city different than* Z*. If* L_1, L_2 *and* U_1, U_2 *are the cities with*

$$\begin{aligned} T_{L_1L_2} &= \max\{T_{XY} : X, Y \text{ are cities with }, X \neq Z, Y \neq Z, T_{XZ} \leq T_{AZ}\} \\ T_{U_1U_2} &= \min\{T_{XZ} : X, Y \text{ are cities with }, X \neq Z, Y \neq Z, T_{XZ} \geq T_{AZ}\}, \end{aligned} \tag{2.19}$$

and the are no reversions in the complete dataset, then we have $P_{AZ} \in [P_{L_1L_2}, P_{U_1U_2}]$ *and also*

$$\widehat{P}_{AZ} = f(T_{AZ}) \in [f(T_{L_1L_2}), f(T_{U_1U_2})] = [P_{L_1L_2}, P_{U_1U_2}],$$

so both the true and the predicted value fall into the same interval. Obviously this only works if the completed dataset has no reversions and the sets in (2.19) *are not empty.*

Every point raised and explored in this example will reoccur in this section.

We will investigate a random vector $\mathbf{X} = (X_1, \dots, X_d)^\intercal$ that has continuous marginal distributions $F_j(x) = P(X_j \leq x)$ and one of the equivalent statements of the Rosetta Stone theorem applies to the random vector

$$\left(\frac{1}{1 - F_1(X_1)}, \dots, \frac{1}{1 - F_d(X_d)}\right)^\intercal.$$

We will reduce its extremal dependence structure to the co-extremal matrix

$$C = (c_{ij})_{1 \leq i,j \leq d} = \left(\lim_{t \to \infty} t \cdot P\left(\frac{1}{(1 - F_i(X_i))(1 - F_j(X_j))} > t^2\right)\right)_{1 \leq i,j \leq d}$$

and the non-extremal dependence structure we will reduce to the matrix

$$R := (r_{ij})_{1 \leq i,j \leq d} := (\mathrm{Corr}(F_i(X_i), F_j(X_j)))_{1 \leq i,j \leq d}.$$

Even though both c_{ij} and r_{ij} are numerical values, comparing them in a naive way suffers from the following pitfalls:

1. One cannot compare the entries of those two matrices numerically. A correlation of 0.05 cannot be called 'weaker' than a co-extremality of 0.95 *without context*. This might seem paradox, because a big motivation for multivariate extreme value theory are scenarios where the joint appearance of extremes events is *more likely* than anticipated with *regular models* and 'more likely' sounds like a comparison. But this just means the regular models can be a bad context for comparison.

2. Extreme value theory and regular statistical theory work with different topologies. Take for example the bivariate Gaussian copula that has correlation coefficient ρ as parameter. Without getting technical regular statisticians can treat $\rho \to 1$ as a continuous transition of the underlying copulae, while for extreme value theory all Gaussian copulae with $\rho < 1$ are tail-independent, which at $\rho = 1$ discontinuously switches to the strongest tail-dependence there is, skipping everything in between (see e.g. Donelly and Embrechts (2010))

We cannot compare correlation with co-extremality just like we could not compare currency with time in our introductory Example 19. What we can do however is compare correlation with correlation and co-extremality with co-extremality and introduce the concept of reversion:

Definition 14. *We will call an instance of four indices (i, j, k, ℓ) a reversion, if they fulfill:*

$$r_{ij} > r_{k\ell} \text{ and} \tag{2.20}$$
$$c_{ij} < c_{k\ell}. \tag{2.21}$$

The set

$$\{(r_{ij}, c_{ij}), 1 \leq i < j \leq d\} \subset \mathbb{R}^2 \tag{2.22}$$

we will call the reversion diagram.

See Figure 2.1 for some very simple reversion diagrams. Those are for illustrating purposes and are not based on real data. Figure 2.5 contains a reversion diagram based on real world data.

Investigating reversions is a way to investigate the difference between the extremal dependence and the non-extremal dependence of components of $\mathbf{X}$ that falls in none of the pitfalls described above.

Finding reversions is straightforward in theory. Either we simply search the finite set $\{(i, j, k, \ell), 1 \leq i, j, k, \ell \leq d\}$ for reversions or we plot the reversion diagram. If the points of the reversion diagram can be connected with a monotonously increasing curve, as in the top diagram of Figure 2.1, there are no reversions. If not, every instance where one point is to the bottom right of another point stands for a reversion. In the bottom diagram of Figure 2.1 we can see this between the two points with the respective x-coordinates 0.4 and 0.6.

In practice however we can only plot

$$\{(\widehat{r}_{ij}, \widehat{c}_{k\ell}) : 1 \leq i, j, k, \ell \leq d\} \subset \mathbb{R}^2, \tag{2.23}$$

where $\widehat{r}$ and $\widehat{c}$ are estimators for the true values. An example of this can be seen in Figure 2.5.

While an individual instance of where (2.20) and (2.21) are fulfilled can be interesting, from a data analysis point of view there is no deeper structure

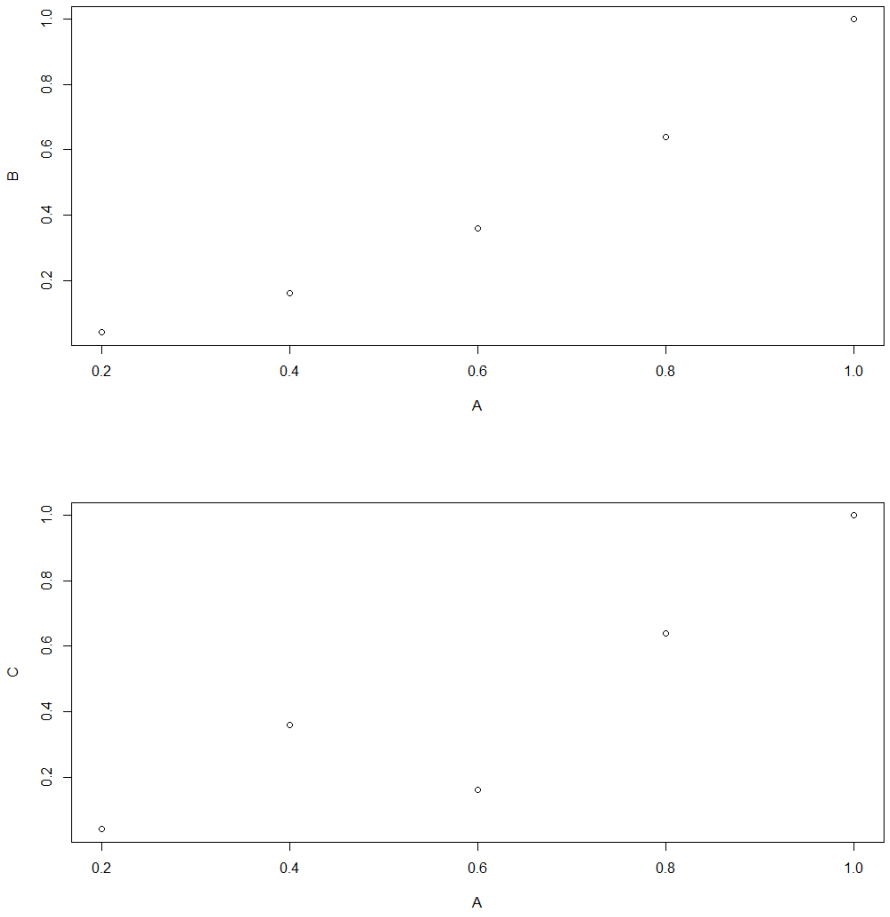


Figure 2.1: Very simple exemplary reversion diagrams.

than 'we have found this instance in our data'. But the collection of all those instances should be investigated with methods of data analysis for the following reasons:

R.1 The instances are set in four dimensions (2 pairs of indices), which is hard to visualize.

R.2 There might be too many instances to investigate separately

R.3 Individual reversions might be the result of estimation errors.

R.4 Our goal might be to find the cause of those instances (e.g. geographic features with a strong effect on regular dependence but no effect on extremal dependence) instead of cataloging the individual instances.

The very last point is something that occured to us in our introductory example. Planes can fly straight above uneven terrain, while train tracks are seldom constructed 'as the crow flies'. One way to look for such geographic features is the following: We will do a principal component analysis of the matrix R and a principal component analysis of the matrix C and visualize the result side by side (for the details of the visualization see Example 16 and for a concrete case see Figure 2.8). Differences in the visualization of this parallel principal component analysis will give hints to the underlying effect that causes reversions. The mathematics behind that we will explain below:

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ be the eigenvalues of the matrix $R = (r_{ij})_{1 \leq i,j \leq d}$ with corresponding eigenvectors $\mathbf{q}^{(h)}, h = 1, \ldots, d$, while $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_d \geq 0$ are the eigenvalues of $C = (c_{ij})_{1 \leq i,j \leq d}$ with corresponding eigenvalue $\mathbf{p}^{(h)}, h = 1, \ldots, d$.

Then we have $R = Q \cdot \operatorname{diag}(\lambda_1, \ldots, \lambda_d) \cdot Q^{\intercal}$, where Q is the matrix that has the vectors $\mathbf{q}^{(h)}$ as columns. This matrix equality is equivalent to

$$r_{ij} = \sum_{h=1}^{d} \lambda_h q_i^{(h)} q_j^{(h)} \text{ for all } 1 \leq i, j \leq d,$$

which is the same equality, just expressed as an equality of every entry of the matrix R. So an eigenvector $\mathbf{q}^{(h)}$ with $q_i^{(h)} \cdot q_j^{(h)} < 0$ is a negative summand in the representantion of r_{ij}. In a visualization of an eigenvector (every single plot of Figure 2.8) this can be seen that i has a white symbol and j has a black symbol or vice versa. Also the negative effect on the dependence grows with the size of the symbols in the visualization.

If we also do this for a second pair of indices (k, ℓ) and also for the matrix C this results in

$$r_{ij} - r_{k\ell} = \sum_{h=1}^{d} \lambda_h (q_i^{(h)} q_j^{(h)} - q_k^{(h)} q_\ell^{(h)}),$$
$$c_{ij} - c_{k\ell} = \sum_{h=1}^{d} \mu_h (p_i^{(h)} p_j^{(h)} - p_k^{(h)} p_\ell^{(h)}) \text{ for all } 1 \leq i, j, k, \ell \leq d.$$

So if (i, j, k, ℓ) is an instance of a reversion, then there has to be a difference between the two sequences of eigenvectors $\mathbf{q}^{(h)}, 1 \leq h \leq d$ and $\mathbf{p}^{(h)}, 1 \leq h \leq d$ or a difference in the two sequences of eigenvalues $\lambda_h, 1 \leq h \leq d$ and $\mu_h, 1 \leq h \leq d$ that assign weight to the eigenvectors.

Example 20. *In Figure 2.8 we have done a parallel principal component analysis of the correlation matrix and the co-etremal matrix of temperature drops in winter in Germany between* 16 *different weather stations and then visualized the first three eigenvectors of each matrix. There is a noticeable difference in the third eigenvector: For extremal dependence Sylt is seperated more strongly from Central Germany than it is in regular dependence.*

Later in Section 2.9 we will see that Sylt is also responsible for many 'isometry violations', which is a concept similar to reversions.

Isometry violations are a derivation of reversions:

Definition 15. *Let $j = 1, \ldots, d$ be geographic locations with a matrix $D = (\text{dist}_{ij})_{1 \leq i,j \leq d}$, where dist_{ij} is the geographic distance between i and j. Further let A be an arbitrary $\mathbb{R}^{d \times d}$ matrix.*

We will call an instance of four indices (i, j, k, ℓ) a violation of (monotone) isometry if they fulfill:

$$\text{dist}_{ij} < \text{dist}_{k\ell} \text{ and}$$
$$a_{ij} < a_{k\ell}.$$

The set

$$\{(\text{dist}_{(i,j)}, a_{ij}), 1 \leq i < j \leq d\} \subset \mathbb{R}^2 \tag{2.24}$$

we will call the (monotone) isometry diagram.

In theory finding violations of isometry is as straightforward as was finding reversions. Either we search the finite set $\{(i, j, k, \ell), 1 \leq i, j, k, \ell \leq d\}$ for violations or we plot the set from Equation (2.24). If the points are on a monotonously decreasing curve, there are no violations of monotone isometry. If not, every instance where one point is to the bottom left of another point stands for a violation. Again we have to keep in mind that if the entries a_{ij} of A are quantities that have to be estimated from data, then we can only plot

$$\{(\text{dist}(i, j), \widehat{a}_{ij}), 1 \leq i < j \leq d\} \subset \mathbb{R}^2,$$

where $\widehat{a}_{ij}$ are estimators for a_{ij}. Examples of isometry diagrams can be seen in Figure 2.6.

The concept of reversion diagrams and isometry diagrams are exceedingly simple and can be applied to any combination of pairwise quantities. However we will not digress further from multivariate extreme value theory and continue with a thought experiment:

Example 21. *Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\intercal$ be a random vector of dimension d. Set $\boldsymbol{X}' = (X_1, \ldots, X_{d'})^\intercal$, where $d' < d$.*

Assume that we know the correlation matrix R of $\boldsymbol{X}$, but only the co-extremal matrix C' of $\boldsymbol{X}'$.

If we find $\boldsymbol{X}'$ to be free of reversions, then there exists a montone function f that fulfills

$$c_{ij} = f(r_{ij}) \text{ for every pair of indices } (i,j) \text{ with } 1 \leq i,j \leq d'$$

and we can predict the missing co-extremalities as

$$\widehat{c}_{ij} = f(r_{ij}) \text{ for every pair of indices } (i,j) \text{ with } i > d' \text{ or } j > d'.$$

For the validity of this prediction we refer to the thoughts about Price-Prediction in Example 19. The main point is that the complete vector $\boldsymbol{X}$ also has to be free of reversions, too.

We can do similar things if instead of the correlation matrix R we have a matrix of distances D and we find the vector $\boldsymbol{X}'$ to have a monontone isometric extremal dependence function. By that we mean there exists a monotone function g that fulfills

$$c_{ij} = g(\mathrm{dist}_{ij}) \text{ for every pair of indices } (i,j) \text{ with } 1 \leq i,j \leq d'$$

and we can now predict the missing co-extremalities as

$$\widehat{c}_{ij} = g(\mathrm{dist}_{ij}) \text{ for every pair of indices}(i,j) \text{ with } i > d' \text{ or } j > d'.$$

Again the validity of this prediction depends on whether the whole vector $\boldsymbol{X}$ has a monotone isometric dependence structure.

The scenario of Example 21 is not too farfetched. In general it takes less data to reliably estimate dependence in regular statistics than in extreme value statistics.

Figure 2.2: Weather Station Locations

2.9 Weather data analysis

The previous sections featured some new concepts of exploratory extreme value analysis (reversion diagrams, isometry diagrams, parallel principal component analysis) and it is therefore reasonable to apply those to an example from the real world.

Our data originates from the DWD, the German Weather Service and is free to the public. Germany consists of 16 states and we included one weather station per state as can be seen in illustration 2.2.

Our mathematical modelling is that for every day there is a random vector $\mathbf{X} = (X_1, \ldots, X_{16})$, such that X_i is the difference between the average temperature of the current day and the average temperature of the previous day at the i-th location.

For a normal year there are 365 random vectors of this kind, but it would not be reasonable to assume them iid. In spring there would be a bias towards increasing temperatures and in autumn the reverse, so we can't assume identical distributions. If there was independence and by using telescopic sums we would have the following equation:

$$\mathrm{Var}(T_{in} - T_{i0}) = \mathrm{Var}\left(\sum_{j=1}^{n} X_i^{(j)}\right) = \sum_{j=1}^{n} \mathrm{Var}\left(X_i^{(j)}\right),$$

where T_{in}, T_{i0} stand for the average temperatures at location i and at one day we declare day zero and the n-th day after day zero. However it would not reasonable for this quantity to approach infinity as $n \to \infty$.

Our assumption for the data analysis is that there are two random vectors $\mathbf{X}_{\text{winter}}$ and $\mathbf{X}_{\text{summer}}$, such that if we evaluate the temperature differences during December/January/February, we get iid observations of $\mathbf{X}_{\text{winter}}$ by skipping at least 5 days between. By this we mean that for example $(T_{\text{Berlin, Jan 2nd}} - T_{\text{Berlin, Jan 1st}})$ is independent of $(T_{\text{Berlin, Jan 8th}} - T_{\text{Berlin, Jan 7th}})$.

We assume the same for $\mathbf{X}_{\text{summer}}$ in the months of June/July/August.

The reason for choosing a lag of 6 was a look at the autocovariance functions of the differenced data during those months, see Figure 2.3.

If $F_j, j = 1, \ldots, 16$ are the marginal distribution functions of $\mathbf{X}_{\text{winter}}$ and $G_j, j = 1, \ldots, 16$ of $\mathbf{X}_{\text{summer}}$ respectively we also assume that the four derived random vectors

$$\left(\frac{1}{1 - F_j(X_{j,\text{winter}})}\right)_{j=1}^{16}, \quad \left(\frac{1}{F_j(X_{j,\text{winter}})}\right)_{j=1}^{16},$$
$$\left(\frac{1}{1 - G_j(X_{j,\text{summer}})}\right)_{j=1}^{16}, \quad \left(\frac{1}{G_j(X_{j,\text{summer}})}\right)_{j=1}^{16}$$

are in the max-domain of attraction of 4 simple max-stable random vectors with co-extremal matrices $C_{\text{up, winter}}, C_{\text{down,winter}}, C_{\text{up,summer}}, C_{\text{down, summer}}$.

Also we have matrices $R_{\text{winter}}, R_{\text{summer}}$ with entries

$$R_{ij,\text{winter}} = \text{Corr}(F_i(X_{i,\text{winter}}), F_j(X_{j,\text{winter}}))$$
$$R_{ij,\text{summer}} = \text{Corr}(G_j(X_{j,\text{summer}}), G_j(X_{j,\text{winter}}), 1 \leq i, j \leq 16.$$

And with no randomness we also have the matrix D, where the entry D_{ij} stands for the distance between weather station i and weather station j.

See Figure 2.4 for which of these quantities measures the strength of which phenomenon.

The purpose of this data analysis is to illustrate the exploratory approaches to multivariate extreme value theory described in Section 2.7 and Section 2.8. Consequently any claim about German Weather announced in this section is not valid unless the assumptions are checked more rigorously.

The correlation matrices were measured with Spearman's rank correlation, the co-extremalities with the local threshold approach as described in Section 3.4.

So one instance of an reversion diagram would be in Figure 2.5. With the 6 different matrices we can potentially plot $\binom{6}{2} = 15$ different reversion diagrams. To not get overwelmed we simply plot the isometry diagrams (see Figure 2.6). The extreme temperature shifts (both ups and downs) in winter are of interest, because in those two diagrams there are many outliers above and below the central diagonal clouds, which is a strong violation of isometry. We will investigate this violation further. We use the same diagram type, but instead of plotting one point per pair of location we plot an ASCII-symbol that indicates the relative positions on the map (a little horizontal bar when they are east/west from

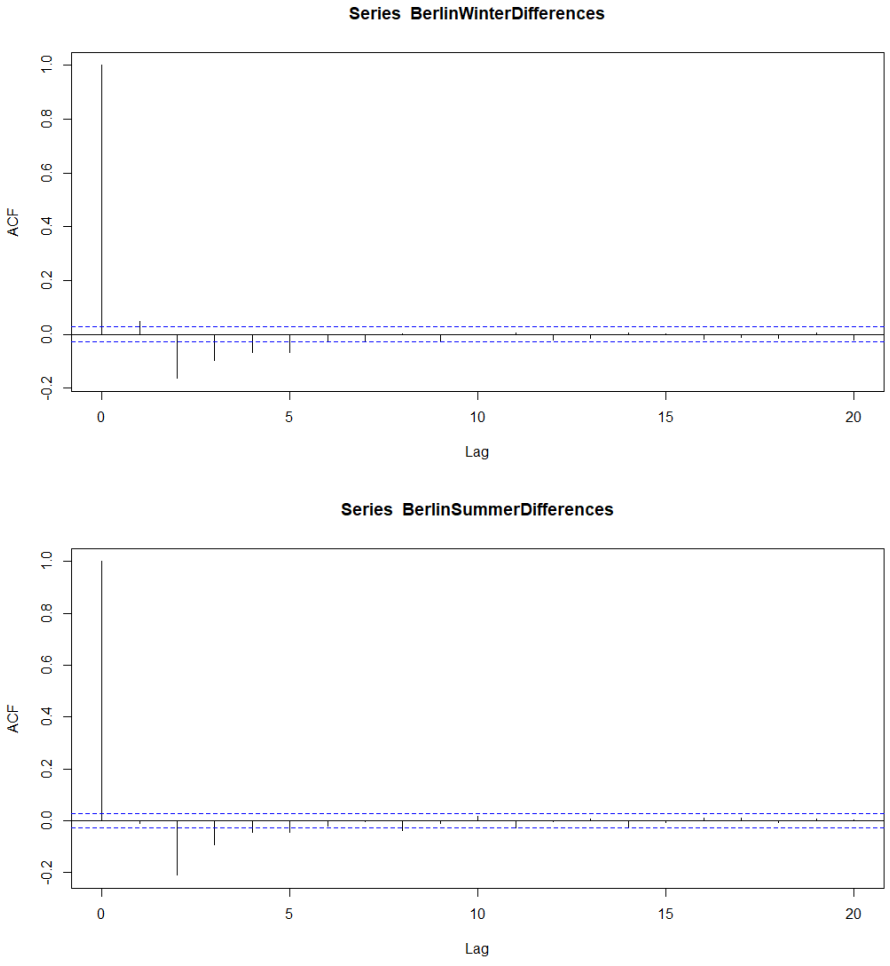


Figure 2.3: Autocorrelation of Temperature differences

Season / Phenomenon	Winter	Summer
increasing/decreasing temperatures at location i go together with increasing/decreasing temperatures at location j	$R_{ij,\text{winter}}$	$R_{ij,\text{summer}}$
extreme increases of temperature at location i go together with extreme increases of temperature at location j	$C_{ij,\text{up,winter}}$	$C_{ij,\text{up,summer}}$
extreme decreases of temperature at location i go together with extreme decreases of temperature at location j	$C_{ij,\text{down,winter}}$	$C_{ij,\text{down,summer}}$

Figure 2.4: Dependency measures and corresponding phenomena

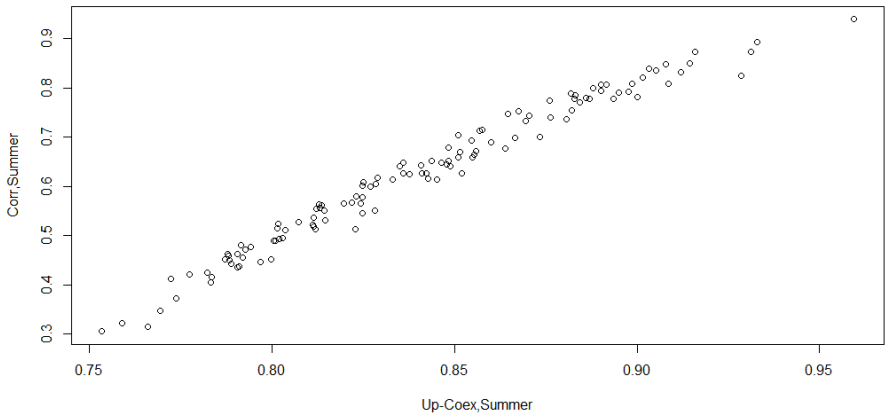


Figure 2.5: An example of a reversion diagram

one another, a little vertical bar when they are north/south from one another, and two diagonal symbols for northwest/southeast and northeast/southwest). The result is shown in Figure 2.7. A careful inspection reveals the outliers to the top right of the data cloud are predominantly east/west pairs, while none of the outlieres to the bottom left are. This indicates that a distance along the north/south axis reduces the extremal dependence more than the same distance but along the east/west axis. For the sake of illustration we will also apply parallel principal component analysis on the matrices $C_{ij,\text{down,winter}}$ and R_{winter}.

The third eigenvector of the winter-correlation makes a clean split between weather stations of the east and weather stations of the west. The third eigenvector of the winter-co-extremality however makes a split between the weather stations of the east and the weather stations of the north-west with a strong

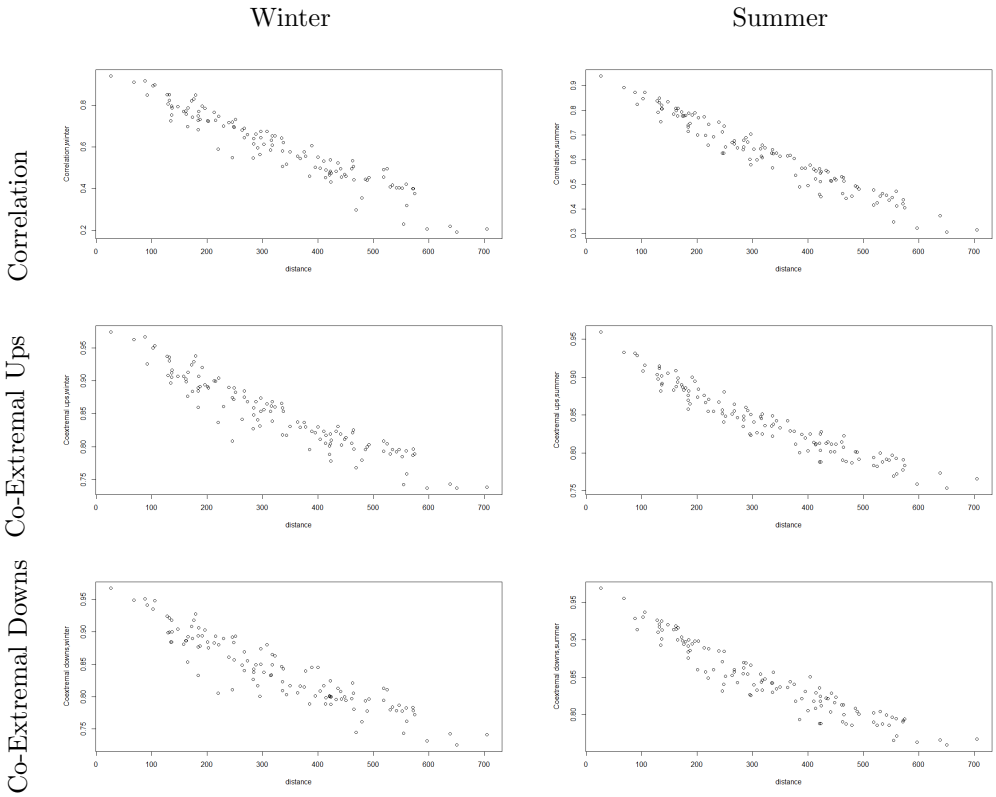


Figure 2.6: Isometry diagrams

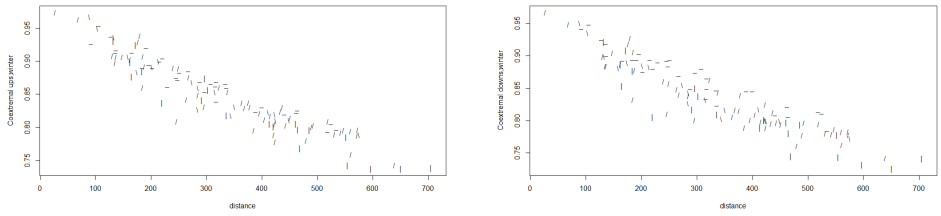


Figure 2.7: Isometry diagram with directions

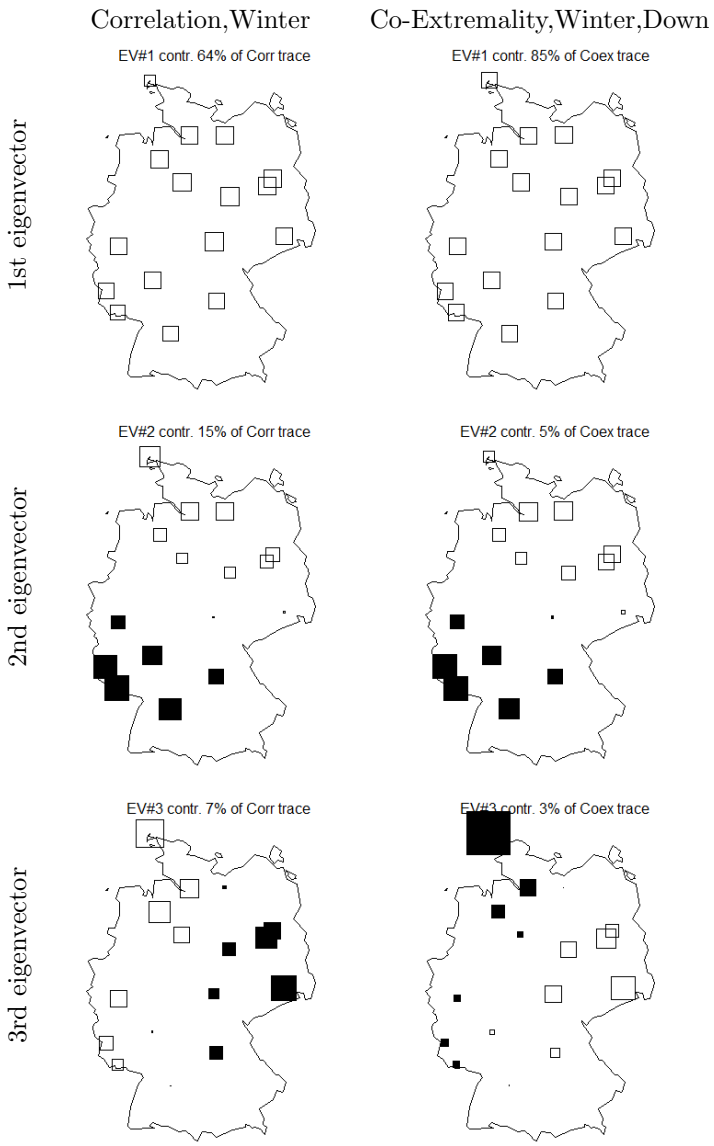


Figure 2.8: Parallel Principal Component Analysis

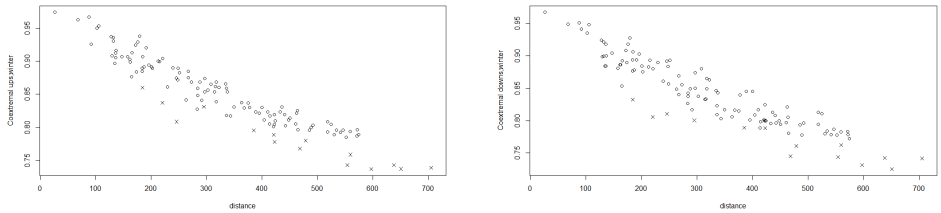


Figure 2.9: Isometry diagrams where pairs (Sylt, ...) are marked with an x

emphasis on the island Sylt.

To investigate this further we will the isometry diagrams of winter-co-extremality again, but this time we highlight the points, where Sylt is involved. This results in the diagrams in Figure 2.9.

We can immediately see that Sylt is responsible for almost all outliers to the bottom left of the data clouds. But keep in mind that removing Sylt from the data analysis would not change the fact that most pairs of locations that are east/west are to the top right of the data clouds in Figure 2.7.

We have used the exploratory tools of the previous sections to get the following results: The difference of temperature from one day to the other has a positive spatial correlation in Germany. The correlation becomes weaker the further two places are apart. The same holds for co-extremality, but in winter you have to go farther east or west to experience the same co-extremality as you would experience from going north or south. The island Sylt is responsible for the strongest cases of isometry violation, but not all of them.

Chapter 3

Multivariate peaks-over-threshold statistics

We have seen in Chapter 1 that if $\mathbf{X}$ is in the max-domain of attraction of some other random vector, then there are quantitites $H = \mathbb{E}(h(\mathbf{Z})))$ that govern the joint appearance of extremes in different components of $\mathbf{X}$. Example 14 and the whole of Chapter 2 give applications for when we know those quantities. But in practice we have to estimate those and that is what this chapter is all about.

Section 3.1 introduces some non-parametric estimators for those quantitites. What these estimators have in common that we have to pick a parameter k, such that n/k is our threshold for making the distinction between extreme and non-extreme observations. Picking a suitable k is a non-trivial task: Literature references are given at the end of Section 3.1 for how other authors approach this problem in the multivariate case. The author's personal approach is stated in the parts denoted by 'threshold strategy'. It is not about finding optimal solutions, but solutions that fall into a tolerable level of inaccuracy.

Section 1.5 has told us that the Rosetta Stone theorem is rarely applicable without marginal transformations. The 'direct estimators' from Section 3.1 do not cover this, but we can extend them to consistent 'indirect estimators' in Section 3.2. We also cover threshold strategies for those cases.

The consistency of the indirect estimators is proven in Section 3.3. This section also contains a valuable observation for how the estimation of the margins affect the behavior of D-norm estimators.

Section 3.4 introduces the concept of local and global thresholds. A pair of examples shows that local thresholds are neither superior to global thresholds nor inferior, but an argument is made why it is more reasonable to use local thresholds, especially in high dimensions.

The split-and-merge-procedure presented in Section 3.5 splits up the sample into different blocks and then evaluates the estimators on each bin separately.

This allows us to investigate the inherent variability of the estimators to produce confidence intervals and to check if our data is really iid.

The chapter concludes with Section 3.6, which features a simulation study to approach some open problems in estimation and statistical inference of extremal dependence structure.

3.1 Direct estimators and threshold strategies

When a random vector $\mathbf{X}$ fulfills one of the equivalent statements of the Rosetta Stone theorem, then its multivariate tail-behavior can be characterized by the values $E(h(\mathbf{Z}))$, where $\mathbf{Z}$ is a corresponding D-norm generator.

But how can $\mathbb{E}(h(\mathbf{Z}))$ be estimated from iid observations $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$?

The Rosetta Stone theorem motivates two possbilities, which we introduce in Definition 16 and Definition 17. Both of them fall under the broad term peak-over-threshold approach, as they rely on observations that exceed some kind of threshold.

For the sake of completeness we will not forget that there is another prominent estimation approach in extreme value theory, the so-called block-maxima approach, which we will briefly cover in Example 26 at the end of this section.

Definition 16. *The following procedure results in what we will call the structure-variable estimator for $H := \mathbb{E}(h(\boldsymbol{Z}))$.*

1. *Choose a number $k = k(n)$, such that $t = n/k$ is a large threshold.*
2. *Our estimator is $\widehat{H} = \frac{1}{k} \sum_{i=1}^{n} 1_{h(\boldsymbol{X}^{(i)}) > n/k}$.*

The motivation for this estimator is $H = \lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t)$ and the following approximations:

$$H \approx \frac{n}{k} \cdot P(h(\mathbf{X}) > n/k) \approx \frac{n}{k} \cdot \frac{1}{n} \sum_{i=1}^{n} 1_{h(\mathbf{X}^{(i)})>n/k} = \frac{1}{k} \sum_{i=1}^{n} 1_{h(\mathbf{X}^{(i)})>n/k}.$$

The estimator follows a rescaled binomial distribution and has the following bias:

$$\mathbb{E}(\widehat{H}) - H = \frac{n}{k} \cdot P(h(\mathbf{X}) > n/k) - H, \tag{3.1}$$

which is a function of the threshold n/k and which converges to 0 as $n/k \to \infty$. The variance of the estimator turns out to be:

$$\begin{aligned} \mathrm{Var}(\widehat{H}) &= \frac{1}{k} \cdot \Big(\frac{n}{k} \cdot P(h(\mathbf{X}) > n/k)\Big) \cdot (1 - P(h(\mathbf{X}) > n/k)) \\ &= \frac{1}{k} \cdot (H + \mathrm{Bias}) \cdot (1 - P(h(\mathbf{X}) > n/k)), \end{aligned} \tag{3.2}$$

which on the long run behaves like $\frac{H}{k}$.

Unfortunately without further assumptions on how the bias depends on n/k we can't minimize the mean squared error MSE with respect to k to find the 'best' value of k, which would look like the following:

$$k_{\mathrm{best}}(n) = \arg\min_{k} \mathrm{MSE}(n, k).$$

Knowing the bias in advance is not a realistic assumption. We need another strategy. For that it is important to recall the notation $h_{\max}$ in Definition 4 and Corollary 2.

Threshold Strategy 1 (Direct structure-variable estimator)**.** *When choosing the value for k in the structure-variable estimator for H, there are two conflicting interests: On the one hand n/k has to be high so the bias is low. On the other hand k has to be high, because even if the estimator hits the true value in mean the variance is still proportional to $1/k$.*

One solution to this dilemma is the following: We should figure out what variance for the estimator is tolerable in our practical situation and call it Var_{tol} and then choose $k = \frac{h_{\max}}{\mathrm{Var}_{tol}}$. By choosing a k as low as we can tolerate it in terms of variance we are doing our best to diminish the bias.

The Rosetta Stone theorem motivates another estimator for H. Because $H = \mathbb{E}(h(\mathbf{Z}))$ is the integral of a probability distribution, we can estimate it as an arithmetic mean. The D-norm generator $\mathbf{Z}$ is an abstract object, but (vi) of the Rosetta Stone theorem indicates that we can treat the angular components of extreme observations as if they were observations of a generator $\mathbf{Z}$ of the underlying D-norm.

Definition 17. *The following procedure results in what we will call the peaks-over-threshold estimator for $H := \mathbb{E}(h(\boldsymbol{Z}))$.*

1. *Determine the radii $r_i := (X_1^{(i)} + \cdots + X_d^{(i)})/d$*
2. *Choose a number $k = k(n)$, such that $t = n/k$ is a large threshold.*
3. *Determine the set $M := \{i : 1 \leq i \leq n, r_i > t\}$.*
4. *Our estimator is $\widehat{H} := \frac{1}{|M|} \sum_{i \in M} h\left(\frac{\boldsymbol{X}^{(i)}}{r_i}\right)$*

Let's investigate the behavior of this estimator:

Theorem 11. *Let $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$ be iid random vectors on $[0, \infty)^d$, let k be an arbitrary positive number and let h be an arbitrary non-negative, continuous function that is homogeneous of order 1. If we set $\widehat{H}$ as in Definition 17, then we have*

$$\mathbb{E}\left(\widehat{H}\middle| |M| = m\right) = \mathbb{E}\left(h\left(\frac{\boldsymbol{X}^{(1)}}{r_1}\right)\middle| r_1 > n/k\right) \tag{3.3}$$

$$\mathrm{Var}\left(\widehat{H}\middle| |M| = m\right) = \frac{1}{m} \cdot \mathrm{Var}\left(h\left(\frac{\boldsymbol{X}^{(1)}}{r_1}\right)\middle| r_1 > n/k\right) \leq \frac{|h_{\max} - h_{\min}|}{4m} \tag{3.4}$$

for every $m \geq 1$, where $h_{\max}$ and $h_{\min}$ are notations from Definition 4.

Proof. Let $P_{n/k}$ be the conditional distribution of $h\left(\frac{\mathbf{X}^{(1)}}{r_1}\right)$ under the condition $r_1 > n/k$ and let $\tilde{h}_1, \ldots, \tilde{h}_n$ be iid random variables following $P_{n/k}$, which also are independent of the sample $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$. By putting

$$h_i = \begin{cases} h\left(\frac{\mathbf{X}^{(i)}}{r_i}\right) & \text{if } r_i > t \\ \tilde{h}_i & \text{else} \end{cases}$$

for $i = 1, \ldots, n$ we end up with iid random variables $(h_i)_{i=1,\ldots,n}$ following the distribution $P_{n/k}$, which are independent of the indicator functions $(1_{r_i>t})_{i=1,\ldots,n}$. Also we have the equation

$$\widehat{H} = \frac{1}{|M|}\sum_{i=1}^{n} h\left(\frac{\mathbf{X}^{(i)}}{r_i}\right) \cdot 1_{r_i>t} = \frac{1}{|M|}\sum_{i=1}^{n} h_i \cdot 1_{r_i>t}$$

because whenever there is a difference between $h\left(\frac{\mathbf{X}^{(i)}}{r_i}\right)$ and h_i, the indicator function $1_{r_i>t}$ becomes 0.

Now let m be an arbitrary positive integer. Then we get

$$\begin{aligned}
\mathbb{E}\left(\widehat{H}\middle||M| = m\right) \cdot P(|M| = m) &= \mathbb{E}\left(\widehat{H} \cdot 1_{|M|=m}\right) \\
&= \frac{1}{m}\sum_{i=1}^{n} \mathbb{E}(h_i \cdot 1_{r_i>t} \cdot 1_{|M|=m}) \\
&= \frac{1}{m}\sum_{i=1}^{n} \mathbb{E}(h_i) \cdot \mathbb{E}(1_{r_i>t} \cdot 1_{|M|=m}) \\
&= \mathbb{E}(h_1)\mathbb{E}\left(\sum_{i=1}^{n} 1_{r_i>t}/m \cdot 1_{|M|=m}\right) \\
&= \mathbb{E}(h_1)\mathbb{E}\left(|M|/m \cdot 1_{|M|=m}\right) \\
&= \mathbb{E}(h_1)\mathbb{E}(1_{|M|=m}) = \mathbb{E}(h_1) \cdot P(|M| = m),
\end{aligned}$$

which shows Equation (3.3). We will also evaluate the variance of $\widehat{H}_m$. For that we have to keep in mind that $\mathbb{E}(X^2) = \mathrm{Var}(X) + E(X)^2$ for every square integrable random variable X.

We evaluate the following:

$$
\begin{aligned}
&\mathbb{E}\left(\widehat{H}^2 \middle| |M| = m\right) \cdot P(|M| = m) = \mathbb{E}\left(\widehat{H}^2 \cdot 1_{|M|=m}\right) \\
&\overset{(a)}{=} \frac{1}{m^2} \sum_{1 \le i,j \le n} \mathbb{E}\left(h_i \cdot h_j \cdot 1_{r_i > t} \cdot 1_{r_j > t} \cdot 1_{|M|=m}\right) \\
&\overset{(b)}{=} \frac{1}{m^2} \sum_{1 \le i,j \le n} \mathbb{E}(h_i \cdot h_j) \cdot \mathbb{E}\left(1_{r_i > t} \cdot 1_{r_j > t} \cdot 1_{|M|=m}\right) \\
&\overset{(c)}{=} \frac{1}{m^2} \left(\sum_{1 \le i,j \le n} \mathbb{E}(h_i) \cdot \mathbb{E}(h_j) \cdot \mathbb{E}(1_{r_i > t} \cdot 1_{r_j > t} \cdot 1_{|M|=m}) + \sum_{i=1}^{n} \mathrm{Var}(h_i) \cdot \mathbb{E}(1_{r_i > t} \cdot 1_{|M|=m}) \right) \\
&\overset{(d)}{=} \mathbb{E}(h_1)^2 \cdot \mathbb{E}\left(\sum_{1 \le i,j \le n} 1_{r_i > t} \cdot 1_{r_j > t} / m^2 \cdot 1_{|M|=m} \right) + \frac{\mathrm{Var}(h_1)}{m} \cdot \mathbb{E}\left(\sum_{i=1}^{n} 1_{r_i > t} / m \cdot 1_{|M|=m} \right) \\
&= \mathbb{E}(h_1)^2 \cdot \mathbb{E}(|M|^2 / m^2 \cdot 1_{|M|=m}) + \frac{\mathrm{Var}(h_1)}{m} \cdot \mathbb{E}(|M| / m \cdot 1_{|M|=m}) \\
&= \left(\mathbb{E}\left(\widehat{H} \middle| |M| = m \right)^2 + \frac{\mathrm{Var}(h_1)}{m} \right) \cdot P(|M| = m).
\end{aligned}
$$

In step (a) we simply evaluated the square of a single sum as a double sum. For (b) we used that every random variable that depends on the $h_i, i = 1, \ldots, n$ (including $h_i \cdot h_j$) is independent from every random variable that depends on the indicator function $1_{r_i > t}, i = 1, \ldots, d$ (including $(1_{r_i > t} \cdot 1_{r_j > t} \cdot 1_{|M|=m})$. In step (c) we used that

$$
\mathbb{E}(h_i h_j) = \begin{cases} \mathbb{E}(h_1)^2 & \text{if } i \neq j \\ \mathbb{E}(h_1)^2 + \mathrm{Var}(h_1) & \text{if } i = j. \end{cases}
$$

Step (d) simply uses the linearity of the expected value several times.

All those steps together imply that the variance of the conditional distribution of the estimator (the condition is $|M| = m$) is equal to $\frac{\mathrm{Var}(h_1)}{m}$. □

Note that we did not derive a closed expression for the variance of $\widehat{H}$. However the term $\mathrm{Var}(h_1)$ is bounded by $(h_{\max} - h_{\min})^2/4$, so when we evaluate the point estimator $\widehat{H}$ we get $|M|$ as a side product and

$$
\left[\widehat{H} - 3 \cdot \frac{h_{\max} - h_{\min}}{2\sqrt{|M|}}, \widehat{H} + 3 \cdot \frac{h_{\max} - h_{\min}}{2\sqrt{|M|}} \right]
$$

is a confidence interval for the expected value of the estimator with the 3σ rule derived from Chebyshev's inequality (see also Section 3.5). In the proof of the previous theorem we can see that the bias is equal to $\mathbb{E}(h_1) - H$ which according to the Rosetta Stone theorem together with the Portmanteau lemma converges to 0 as the threshold n/k converges to ∞.

Also according to the Rosetta Stone theorem the probability $P(r_i > n/k) \approx k/n \cdot \mathbb{E}((Z_1 + \cdots + Z_d)/d) = k/n$, so the value of $|M|$ should be somewhere around k as $n/k \to \infty$.

Threshold Strategy 2 (Direct peaks-over-threshold estimator)**.** *When choosing the value for k in the peaks-over-threshold estimator for H, there are two conflicting interests. On the one hand n/k has to be high, so the bias is low. On the other hand k should be high, so $|M|$ hits large values with increased probability, which reduces the conditional variance.*

One solution to this dilemma is the following: We should figure out what variance of the estimator is tolerable in our practical situation and call it Var_{tol} and the choose $k = \frac{(h_{\max}-h_{\min})^2}{4\cdot \mathrm{Var}_{tol}}$. If we then evaluate the estimator and it turns out $|M| \geq k$ then we are fine. If not, we can ask ourselves, if $\frac{(h_{\max}-h_{\min})^2}{4\cdot |M|}$ is still a tolerable variance.

Threshold strategies other authors used in multivariate extremes include the following:

Example 22. *Jeon and Smith (2012) included a short discussion about thresholds. Their contribution was that for a given model you can do a simulation study to find the threshold that minimizes the Mean-Squared-Error, but concluded that in general finding the optimal threshold (with respect to Mean-Squared-Error) is an open problem.*

Example 23. *Einmahl et al. (2009) approached to threshold choice problem for estimating $\|\boldsymbol{x}\|_D = \lim_{t\to\infty} t \cdot P(\max_{j=1,\ldots,d} |x_j| X_j > t)$ with a structure-variable estimator $\widehat{\|\boldsymbol{x}\|}_D$ in the following way: Instead of minmizing the true Mean Squared Error*

$$\mathbb{E}\left(\left(\widehat{\|\boldsymbol{x}\|}_D - \|\boldsymbol{x}\|_D\right)^2\right),$$

with respect to the underlying threshold n/k, they suggested minizing

$$\mathbb{E}\left(\left(2 \cdot \widehat{\|\boldsymbol{x}\|}_D - \widehat{\|2 \cdot \boldsymbol{x}\|}_D\right)^2\right)$$

with a bootstrap procedure.

Example 24. *Davis and Wan (2019) present an algorithm for finding a threshold. The basis for this is a result similar to Equation (1.5) in the Rosetta Stone theorem, which says that for increasing thresholds $t = n/k$ the angular component $\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}$ 'becomes' independent from the radial component $\|\boldsymbol{X}\|$. The algorithm checks if this holds for different threshold t and chooses k accordingly.*

Example 25. *Fan et al. (2015) go into a similar direction. For a given combination of*

1. *a threshold t and*

2. a model of how the data should behave above t

they introduce a test-statistic T to test the Null-hypothesis: 'Above t the data follows this model.' They suggest p-values of this test should then be incoroporated into the choice of the threshold.

Ultimately the purpose of thresholds is to divide the data into an extreme and a non-extreme part. This sharp division is avoided, when we only investigate maxima:

Example 26 (Direct block-maxima)**.** *Let $\boldsymbol{X}$ be a random vector that fulfills one of the equivalent statements of the Rosetta Stone theorem. Then there is a simple max-stable random vector $\boldsymbol{Y}$ with*

$$\boldsymbol{M}_n := \left[\frac{1}{n} \cdot \max_{i=1,\dots,n} X_j^{(i)}\right]_{j=1}^{d} \xrightarrow{\mathcal{D}} [Y_j]_{j=1}^{d}.$$

If we have $N = n \cdot m$ independent samples $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(N)}$, we can split this sample into m blocks of size n, e.g. the first block consists of $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(n)}$, the second one consists of $\boldsymbol{X}^{(n+1)}, \dots, \boldsymbol{X}^{(2n)}$, etc. For each block we get one observation of the rescaled block-maxima vector $\boldsymbol{M}_n$.

If we treated those iid observations of $\boldsymbol{M}_n$ as if they were iid observations of $\boldsymbol{Y}$ to infer the distribution of $\boldsymbol{Y}$, we end up with what we will call the multivariate block-maxima method (see also Example 29).

There are no thresholds in the block-maxima methodology. So it looks like we could have saved our effort of thinking about threshold strategies by only using the block-maxima method. But what about choosing block size n and number of blocks m? What's a reasonable choice for those? The higher n, the closer the distribution of $\mathbf{M}_n$ is to the true distribution $\mathbf{Y}$. The higher m, the more iid observations we have for inference. But $m \cdot n$ is bounded by our sample size N. Suddenly we have two conflicting interests when choosing n and m and have to start with the strategic thinking again. This thesis is not about block-maxima methods, so we won't do this. But we will to block-maxima again in Example 29.

3.2 Indirect estimators

In the previous section we only investigated random vectors $\mathbf{X}$ that fulfill one of the equivalent statements of the Rosetta Stone theorem. In Section 1.5 we learned that this is a rather restricting assumption. A much weaker assumption is that $\mathbf{X}$ has continuous marginal distributions $F_j(x) = P(X_j \leq x), j = 1, \ldots, d$ and that the transformed random vector

$$\mathbf{X}' = \left(\frac{1}{1 - F_j(X_j)} \right)_{j=1}^{d}$$

fulfills one of the equivalent statements of the Rosetta Stone theorem. As the transformation $x \mapsto \frac{1}{1-F_j(x)}$ is monotonous the underlying D-norm governs the joint appearance of extremes in different components of $\mathbf{X}'$, but also the joint appearance of extremes in different components of $\mathbf{X}$.

Our new task is to estimate the value $H = \lim_{t\to\infty} t \cdot P(\mathbf{X}' > t)$ for non-negative, continuous functions that are homogeneous of order 1 from iid observations $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ of $\mathbf{X}$ without further assumptions.

If we knew the functions F_j in advance we could transform the iid observations $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ of $\mathbf{X}'$ to

$$(\mathbf{X}')^{(i)} := \left(\frac{1}{1 - F_j(X_j^{(i)})} \right)_{j=1}^{d},$$

which results in iid observations $(\mathbf{X}')^{(1)}, \ldots, (\mathbf{X}')^{(1)}$ of $\mathbf{X}'$. We can then use the direct estimators from the previous sections. But in general we don't know F_j and have to use surrogates $\widehat{F}_j$. This could for example be the empirical distribution function of X_j or it could be a parametric model fitted to the observations of X_j.

We can still apply the transformation

$$(\mathbf{X}'')^{(i)} := \left(\frac{1}{1 - \widehat{F}_j(X_j^{(i)})} \right)_{j=1}^{d}, \tag{3.5}$$

and evaluate the direct estimators on those observations.

Definition 18. *The following procedure results in what we will call the indirect structure-variable estimator for* $H := \mathbb{E}(h(\boldsymbol{Z}))$.

0. *Transform all observations with Equation* (3.5). *For the sake of notation we will keep using the term* $\boldsymbol{X}^{(i)}$ *for the transformed data.*
1. *Choose a number* $k = k(n)$, *such that* $t = n/k$ *is a large threshold.*
2. *Our estimator is* $\widehat{H} = \frac{1}{k} \sum_{i=1}^{n} 1_{h(\boldsymbol{X}^{(i)}) > n/k}$.

Definition 19. *The following procedure results in what we will call the indirect peaks-over-threshold estimator for* $H := \mathbb{E}(h(\boldsymbol{Z}))$.

0. *Transform all observations with Equation* (3.5). *For the sake of notation we will keep using the term* $\boldsymbol{X}^{(i)}$ *for the transformed data.*
1. *Determine the radii* $r_i := (X_1^{(i)} + \cdots + X_d^{(i)})/d$
2. *Choose a number* $k = k(n)$, *such that* $t = n/k$ *is a large threshold.*
3. *Determine the set* $M := \{i : 1 \leq i \leq n, r_i > t\}$.
4. *Our estimator is* $\widehat{H} := \frac{1}{|M|} \sum_{i \in M} h\left(\frac{\boldsymbol{X}^{(i)}}{r_i}\right)$

Note that if $\widehat{F}_j = F_j$ holds for all $j = 1, \ldots, d$, the indirect estimator is the direct estimator applied to the true realizations of $\mathbf{X}'$ and inherits all the properties from the previous section. But $(\mathbf{X}'')^{(1)}, \ldots, (\mathbf{X}'')^{(n)}$ are neither guaranteed to be iid, nor to follow the same distribution as $\mathbf{X}'$. This makes investigating the properties of the indirect estimators especially hard.

But the following result shows us that consistency is achieved under extremely mild assumptions.

Theorem 12. *(Consistency of various estimators) Let* $\boldsymbol{X}$ *be a random vector with continuous marginal distribution functions* $F_j(x) = P(X_j \leq x), x \in \mathbb{R}, j = 1, \ldots, d$ *such that* $\boldsymbol{X}' := (\frac{1}{1-F_1(X_1)}, \ldots, \frac{1}{1-F_d(X_d)})^\intercal$ *fulfills one of the statements of the Rosetta Stone theorem and let* $h : [0, \infty)^d \to [0, \infty)$ *be continuous and homogeneous of order* 1. *Then both the indirect structure-variable estimator in Definition 18 and the indirect peaks-over-treshold estimator in Definition 19 produce a consistent sequence of estimators for the quantitiy*

$$H = \lim_{t \to \infty} t \cdot P(h(\boldsymbol{X}') > t)$$

under the condition that the sample consists of iid repetitions of $\boldsymbol{X}$, *the sequence* $k = k(n)$ *fulfills* $n/k \to \infty$ *and* $k \to \infty$ *and for every* $j = 1, \ldots, d$ *the random function* $\widehat{F}_j$ *always realizes as an montonously increasing, right continuous function that fulfills*

$$\frac{n}{k} \cdot (1 - F_j) \circ (1 - \widehat{F}_j)^{(-1)} \left(\frac{k}{n} \cdot x\right) \to x \text{ in probability as } n \to \infty \tag{3.6}$$

for all $x > 0$.

Those assumptions on $\widehat{F}_j$ are really mild as we will see in the following lemma:

Lemma 18. *Let* F *be a continuous distribution function. Then for every iid sample of size* n *the empirical distribution function* $\widehat{F}$ *is monotonously increasing, right continuous and fulfills the limit in Equation* (3.6) *for all* $x > 0$ *as long as* $k = k(n)$ *is a sequence with* $k(n) \to \infty$ *and* $k(n)/n \to 0$ *as* $n \to \infty$.

Proof. Let x be an arbitrary positive real number. Further let $(X_i)_{i\in\mathbb{N}_0}$ be sequence of iid random variable with distribution function F. Because F is continuous we can assume without loss of generality that there are not ties in the sequence. Now let $\widehat{F}$ be the empirical distribution function of the sample $X_1, \ldots, X_n$ (note that the $i = 0$ is excluded).

For all n that are large enought the term $\frac{k}{n} \cdot x$ falls into the interval $(0, 1)$ and we get

$$(1 - \widehat{F})^{(-1)} \left(\frac{k}{n} \cdot x \right) = X_{n-[k\cdot x],n},$$

the $n - [k \cdot x]$-th order statistic. We further get

$$(1 - F)(X_{n-[k\cdot x],n}) = P(X_0 > X_{n-[k\cdot x],n})$$

for every realization of $X_1, \ldots, X_n$. This probability is

$$\frac{([kx] + 1) \cdot n!}{(n + 1)!} = \frac{[kx] + 1}{n + 1}$$

for combinatorical reasons: There are $(n + 1)!$ permutations of the indices $\{0, \ldots, n\}$ and the order of the iid sequence $X_0, \ldots, X_n$ corresponds to one of those permutations without preference. In $([kx]+1) \cdot n!$ cases the indix 0 falls in one of the last $[kx] + 1$ slots.

If we now multiply this probability with $\frac{n}{k}$ we get

$$\frac{n}{k} \cdot \frac{[kx] + 1}{n + 1} = \underbrace{\frac{n}{n + 1}}_{\to 1} \cdot \underbrace{\frac{[kx] + 1}{k}}_{\in [x, x+2/k]} \to x$$

as $n \to \infty$ and $k(n) \to \infty$. This can be done for all $x > 0$. □

We will prove Theorem 12 in the next section. We will now have a look at the role of the parameter k for indirect estimators. For Theorem 12 both k and n/k have to converge to infinity as n increases. In a finite-sample world (a world where the sample size n is constant) this obviously poses two conflicting interests. Again we need threshold strategies for picking k. We suggest the following strategy:

Threshold Strategy 3. *(Indirect estimators) Just like in Strategies 1 and 2 we should first realize what variance is tolerable in the practical situation and vall it* Var_{tol}*. For the indirect structure-variable estimator we should only consider values*

$$k \geq \frac{h_{\max}}{\mathrm{Var}_{tol}}$$

and for the indirect peaks-over-threshold estimator we should only consider values

$$k \geq \frac{(h_{\max} - h_{\min})^2}{4 \cdot \mathrm{Var}_{tol}}.$$

The rationale behind this is the following: The indirect estimators are essentially direct estimators after applying the transformation

$$(x_j)_{j=1}^d \mapsto \left(\frac{1}{(1-\widehat{F}) \circ (1-F)^{(-1)}\left(\frac{1}{x_j}\right)} \right)_{j=1}^d \tag{3.7}$$

to every observation. If there was a value k, such that the direct estimator does not fulfill the needs of our practical application, but the indirect estimator does, then this would imply the distortion in Equation (3.7) *somehow works in our favor. This is not impossible (see also Example 27 below), but it definitely not an effect we should count on without further explaination.*

The following example givese a case where the indirect estimator is better than the direct estimator, so a case where distortion works in our favor.

Example 27. *Let $\boldsymbol{X} = (X, X)^\intercal$ be a bivariate random vector such that X has a continuous distribution function F. $(\frac{1}{1-F(X_1)}, \frac{1}{1-F(X_2)})$ fulfills the equivalent statements of the Rosetta Stone theorem with the D-norm generator $\boldsymbol{Z} = \boldsymbol{1}$ almost surely. We then want to investigate the structure-variable estimator for $h(z_1, z_2) = \max(z_1, z_2)$. The true value is $H = \mathbb{E}(h(\boldsymbol{Z})) = 1$.*

So let $X^{(i)}, i = 1, \ldots, n$ be iid copies of X. Consequently $\boldsymbol{X}^{(i)} = (X^{(i)}, X^{(i)})^\intercal$ are iid copies of $\boldsymbol{X}$. Further let k be a positive integer. The direct structure-variable estimator turns out to be

$$\frac{1}{k}\sum_{i=1}^n 1_{h\left(\frac{1}{1-F(X^{(i)})}, \frac{1}{1-F(X^{(i)})}\right) > n/k} = \frac{1}{k}\sum_{i=1}^n 1_{F(X^{(i)}) > 1-\frac{k}{n}}.$$

This is a rescaled sum of iid Bernoulli random variables with parameter $\frac{k}{n}$. If we instead use the empirical distribution functions $\widehat{F}_j, j = 1, 2$ it first turns out that $\widehat{F}_1 = \widehat{F}_2$ and also with probability 1 there are no ties and therefore

$$\{\widehat{F}_j(X^{(i)}) : i = 1, \ldots, n\} = \{i/n : i = 1, \ldots, n\}$$

almost surely by the nature of the empirical distribution function. The indirect estimator (that uses the empirical marginal distributions functions) estimator turns out to be

$$\begin{aligned}\frac{1}{k}\sum_{i=1}^n 1_{h\left(\frac{1}{1-F(X^{(i)})}, \frac{1}{1-F(X^{(i)})}\right) > n/k} &= \frac{1}{k}\sum_{i=1}^n 1_{\widehat{F}_1(X^{(i)}) > 1-\frac{k}{n}} \\ &= \frac{1}{k}|\{i/n : i = 1, \ldots, n, i/n > 1 - k/n\}| = 1\end{aligned}$$

almost surely. Both estimators have no bias, but by switching from the direct estimator to the indirect estimator the variance was reduced to 0.

Example 27 fits well into another result from literature.

Example 28. *Bücher (2014) compared the behavior of two possible estimators for*

$$\left\| \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \right\|_D = \lim_{t \to \infty} t \cdot P(\max(X_1 \cdot |z_1|, X_2 \cdot |z_2| > t).$$

In our terminology those estimators were the direct and the indirect structure-variable estimators, where the indirect estimator uses the empirical marginal distribution functions.

Under certain conditions on $\boldsymbol{X}$ he found out that both estimators estimators are asymptotically normal, but the asymptotic variance of the indirect estimator is always less or equal than the the asymptotic variance of the direct estimator.

This effect holds for all $\boldsymbol{z} \in \mathbb{R}^2$.

We will return to this example in Threshold Strategy 4.

In this section we switched from direct to indirect estimators because the assumption that $\mathbf{X}$ fulfills the equivalent statements of the Rosetta Stone theorem is too restricting. So the block-maxima methods from Example 26 also needs to adapt to our weaker assumptions:

Example 29 (Indirect block-maxima method)**.** *Let $\boldsymbol{X}$ be a random vector with continuous marginal distribution functions $F_j, j = 1, \ldots, d$. If the random vector $\left(\frac{1}{1-F_1(X_1)}, \ldots, \frac{1}{1-F_d(X_d)}\right)^\intercal$ fulfills one of the equivalent statements of the Rosetta Stone theorem, then there is a simple max-stable random vector $\boldsymbol{Y}$ with*

$$\boldsymbol{M}_n := \left[\frac{1}{n} \cdot \max_{i=1,\ldots,n} \frac{1}{1 - F_j(X_j^{(i)})} \right]_{j=1}^{d} \xrightarrow{\mathcal{D}} [Y_j]_{j=1}^{d}$$

as $n \to \infty$, where the convergence is meant in distribution. If we have $N = n \cdot m$ independent samples $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}$, we can split this sample into m blocks of size n, e.g. the first block consists of $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$, the second one consists of $\boldsymbol{X}^{(n+1)}, \ldots, \boldsymbol{X}^{(2n)}$, etc. For each block we get one observation of the rescaled block-maxima vector $\boldsymbol{M}_n$.

If we knew the marginal distribution functions F_j in advance, we could apply the determistic transformation

$$(x_1, \ldots, x_j)^\intercal \mapsto \left(\frac{1}{1 - F_1(X_1)}, \ldots, \frac{1}{1 - F_d(X_d)} \right)^\intercal$$

and have independent blocks to use the direct block-maxima method from Example 26 on. But in general those marginal distributions have to be estimated themselves with $\widehat{F}_j$ (e.g. the empirical marginal distribution function). After applying the random transformation

$$(x_1, \ldots, x_j)^\intercal \mapsto \left(\frac{1}{1 - \widehat{F}_1(X_1)}, \ldots, \frac{1}{1 - \widehat{F}_d(X_d)} \right)^\intercal$$

the blocks that are no longer guaranteed to be independent.

If we still proceed as in Example 26 this results in what we will call the indirect block-maxima method. For limit results see the following literature: Cooley et al. (2009) for the bivariate case and an introduction to the λ-Madogram, Bücher and Segers (2014) for a limit results in more than 2 dimensions and also Bücher et al. (2018) as a most recent contribution. All those papers work with empirical copula processes, which means that observations $(X_1, \ldots, X_d)$ are replaced by $(\widehat{F}_1(X_1), \ldots, \widehat{F}_d(X_d))$, where $\widehat{F}_j$ are the empirical distribution functions.

3.3 Proof of consistency for various estimators

Let $\mathbf{X}$ be a random vector with continuous marginal distributions $F_j, j = 1, \ldots, d$ such that $\mathbf{X}' := (\frac{1}{1-F_1(X_1)}, \ldots, \frac{1}{1-F_d(X_d)})^\intercal$ fulfills one of the statements of the Rosetta Stone theorem. Further let ν be the measure with

$$\lim_{t\to\infty} t \cdot P(\mathbf{X}' \in t \cdot A) = \nu(A)$$

for every measurable continuity set M.

Estimating the values $\nu(A)$ for different sets A is an important task as not only we can express the underlying D-norm by

$$\left\| \frac{1}{\mathbf{z}} \right\|_D = \nu([0, \mathbf{z})^\complement)$$

for every $\mathbf{z} > \mathbf{0}$ (see Lemma 1), but we also have

$$H := \lim_{t\to\infty} t \cdot P(h(\mathbf{X}') > t) = \nu(\{\mathbf{x} : \mathbf{x} \geq \mathbf{0}, h(\mathbf{x}) > 1\})$$

for every non-negative, continuous function h that is homogeneous of order 1. So the complete tail-dependence structure given in the Rosetta Stone theorem is hidden in the values $\nu(A)$, the measure ν assigns to certain sets A.

For any pair (n, k) of integers and any measurable continuity set A we will define a non-parametric estimator for $\nu(A)$ by

$$\nu_{n,k}(A) := \frac{1}{k} \cdot \left| \left\{ i : 1 \leq i \leq n, \left((\mathbf{X}'')^{(i)} \right)^\intercal \in \frac{n}{k} \cdot A \right\} \right|, \tag{3.8}$$

where $\mathbf{X}^{(i)}, i = 1, \ldots, n$ are iid copies of $\mathbf{X}$, which are transformed to $(\mathbf{X}'')^{(i)}, i = 1, \ldots, n$ with Equation (3.5).

For convencience let us write $k = k(n)$ as a function of n and put

$$\nu_n := \nu_{n,k(n)}$$

The following result is well known and it is quite important as the value $\nu(A)$ is equal to $\left\| \frac{1}{\mathbf{z}} \right\|_D$, the underlying D-norm evaluated at the position $\mathbf{z} > \mathbf{0}$.

Lemma 19. *Under the conditions of Theorem 12 and under the condition that $\widehat{F}_j, j = 1, \ldots, d$ are chosen as the empirical distribution functions $\nu_n(A)$ is a consistent estimator of $\nu(A)$, whenever A is a set of the form*

$$A = \{\boldsymbol{x} : \boldsymbol{x} \in [0, \infty)^d, \text{ there is an index } j \text{ with } x_j \geq z_j\}, \tag{3.9}$$

where $\boldsymbol{z}$ is a vector in $(0, \infty)^d$.

The proof for the bivariate case can be looked up in the book by de Haan and Ferreira (2006)[Theorem 7.2.1], but a careful inspection of it reveals that it can be repeated almost line by line for dimension $d > 2$ and therefore Lemma 19 holds.

We will now show a slightly more general result than Lemma 19, which allows us to use different marginal estimators than the empirical marginal distribution function without changing the fact that the resulting estimator is consistent.

Lemma 20. *Under the conditions of Theorem 12 $\nu_n(A)$ is a consistent estimator of $\nu(A)$, whenever A is the set in Equation (3.9) for a vector $\boldsymbol{z} \in (0,\infty)^d$.*

Proof. Let F_j be the true distribution function of the j-th margin, while $\widehat{F}_j$ is the surrogate the indirect estimator uses. Then we have

$$\begin{aligned}
& \frac{1}{1-\widehat{F}_j(X_j)} \geq \frac{n}{k} \cdot z \\
\Leftrightarrow \quad & 1-\widehat{F}_j(X_j) \leq \frac{k}{n} \cdot 1/z \\
\Leftrightarrow \quad & 1-F_j(X_j) \leq (1-F_j) \circ (1-\widehat{F}_j)^{(-1)} \left(\frac{k}{n} \cdot 1/z \right) \\
\Leftrightarrow \quad & \frac{1}{1-F_j(X_j)} \geq \frac{n}{k} \cdot \frac{1}{\frac{n}{k}(1-F_j) \circ (1-\widehat{F}_j)^{(-1)} \left(\frac{k}{n} \cdot 1/z \right)}
\end{aligned}$$

for all $z > 0$. For any $z_j > 0$ we define the random variable $\widehat{z}_j$ by

$$\widehat{z}_j := \frac{1}{\frac{n}{k}(1-F_j) \circ (1-\widehat{F}_j)^{(-1)} \left(\frac{k}{n} \cdot 1/z_j \right)}, \tag{3.10}$$

which implicitely depends upon n and $k = k(n)$ and which converges in probability to z_j as $n \to \infty$ because we required Equation (3.6) to hold.

So let $\mathbf{z} = (z_1, \ldots, z_d)^\intercal$ be a vector in $(0,\infty)^d$ and let the set A be defined by Equation (3.9). Then we have

$$\begin{aligned}
\nu_n(A) &= \frac{1}{k} \sum_{i=1}^{n} 1 \left(\text{there is an index } j \text{ with } \frac{1}{1-\widehat{F}_j(X_j^{(i)})} \geq \frac{n}{k} \cdot z_j \right) \\
&= \frac{1}{k} \sum_{i=1}^{n} 1 \left(\text{there is an index } j \text{ with } \frac{1}{1-F_j(X_j^{(i)})} \geq \frac{n}{k} \cdot \widehat{z}_j \right)
\end{aligned}$$

by virtue of the previous equivalences. The estimator

$$\widehat{\left\| \frac{1}{\mathbf{z}'} \right\|}_D := \frac{1}{k} \sum_{i=1}^{n} \underbrace{1 \left(\text{there is an index } j \text{ with } \frac{1}{1-F_j(X_j^{(i)})} \geq \frac{n}{k} \cdot z_j' \right)}_{\text{iid indicator functions}} \tag{3.11}$$

is a consistent estimator of $\left\| \frac{1}{\mathbf{z}'} \right\|_D$ for every $\mathbf{z}' > \mathbf{0}$. The argument by de Haan and Ferreira (2006)[Theorem 7.2.1], where they used the characteristic functions can be adapted to more than 2 dimensions to prove this. Alternatively $\widehat{\left\| \frac{1}{\mathbf{z}'} \right\|}_D$ is a direct structure-variable estimator on the data transformed with the true marginal distribution function. Bias and variance are covered in Equation (3.1) and (3.1): They converge to 0 and one could use Chebyshev's inequality (see Section 3.5) to show consistency.

The indirect structure-variable estimator turns out to be

$$\nu_n(A) = \widehat{\left\| \frac{1}{\widehat{\mathbf{z}}} \right\|}_D. \tag{3.12}$$

It randomly chooses a $\widehat{\mathbf{z}}$ close to the true $\mathbf{z}$ and then evaluates the true direct estimator not at $\frac{1}{\mathbf{z}}$ but at $\frac{1}{\widehat{\mathbf{z}}}$.

To show that this is a consistent estimators, we will resort to an ϵ-argument, which essentially does the same as the local uniform convergence argument by de Haan and Ferreira (2006)[Theorem 7.2.1].

So let ϵ be a positive constant. We will show that the probability

$$P(|\nu_n(A) - \nu(A)| > 2\epsilon) = P\left(\left| \widehat{\left\| \frac{1}{\widehat{\mathbf{z}}} \right\|}_D - \left\| \frac{1}{\mathbf{z}} \right\|_D \right| > 2\epsilon \right)$$

converges to 0 as $n \to \infty$. As $\mathbf{z}' \mapsto \left\| \frac{1}{\mathbf{z}'} \right\|_D$ is a function that is monotonous in every component and continuous at $\mathbf{z}$, we can find a $\mathbf{z}_+$ and $\mathbf{z}_-$ such that $\mathbf{0} < \mathbf{z}_+ < \mathbf{z} < \mathbf{z}_-$ and additionally

$$\left[\left\| \frac{1}{\mathbf{z}_-} \right\|_D, \left\| \frac{1}{\mathbf{z}_+} \right\|_D \right] \subset \left[\left\| \frac{1}{\mathbf{z}} \right\|_D - \epsilon, \left\| \frac{1}{\mathbf{z}} \right\|_D + \epsilon \right].$$

At first we we get

$$\begin{aligned}
&P\left(\left| \widehat{\left\| \frac{1}{\widehat{\mathbf{z}}} \right\|}_D - \left\| \frac{1}{\mathbf{z}} \right\|_D \right| > 2\epsilon \right) = P\left(\widehat{\left\| \frac{1}{\widehat{\mathbf{z}}} \right\|}_D \notin \left[\left\| \frac{1}{\mathbf{z}} \right\|_D - 2\epsilon, \left\| \frac{1}{\mathbf{z}} \right\|_D + 2\epsilon \right] \right) \\
\leq & P\left(\widehat{\left\| \frac{1}{\widehat{\mathbf{z}}} \right\|}_D \notin \left[\widehat{\left\| \frac{1}{\mathbf{z}_-} \right\|}_D, \widehat{\left\| \frac{1}{\mathbf{z}_+} \right\|}_D \right] \right) \\
& + P\left(\left[\widehat{\left\| \frac{1}{\mathbf{z}_-} \right\|}_D, \widehat{\left\| \frac{1}{\mathbf{z}_+} \right\|}_D \right] \not\subset \left[\left\| \frac{1}{\mathbf{z}_-} \right\|_D - \epsilon, \left\| \frac{1}{\mathbf{z}_+} \right\|_D + \epsilon \right] \right) \\
& + \underbrace{P\left(\left[\left\| \frac{1}{\mathbf{z}_-} \right\|_D - \epsilon, \left\| \frac{1}{\mathbf{z}_+} \right\|_D + \epsilon \right] \not\subset \left[\left\| \frac{1}{\mathbf{z}} \right\|_D - 2\epsilon, \left\| \frac{1}{\mathbf{z}} \right\|_D + 2\epsilon \right] \right)}_{=0},
\end{aligned}$$

which implies for monotonicity reasons

$$\begin{aligned}
P\left(\left| \widehat{\left\| \frac{1}{\widehat{\mathbf{z}}} \right\|}_D - \left\| \frac{1}{\mathbf{z}} \right\|_D \right| > 2\epsilon \right) \leq & P(\widehat{\mathbf{z}} \notin [\mathbf{z}_+, \mathbf{z}_-]) \\
& + P\left(\widehat{\left\| \frac{1}{\mathbf{z}_-} \right\|}_D < \left\| \frac{1}{\mathbf{z}_-} \right\|_D - \epsilon \right) \\
& + P\left(\widehat{\left\| \frac{1}{\mathbf{z}_+} \right\|} > \left\| \frac{1}{\mathbf{z}_+} \right\|_D + \epsilon \right).
\end{aligned} \tag{3.13}$$

Because $\widehat{z}_j$ converges to the true z_j in probability for all $j = 1, \ldots, d$ and the direct estimators are consistent estimators, all three probabilities on the right hand side of Equation (3.13) converge to 0 as $n \to \infty$. Consequently $\nu_n(A)$ is a consistent estimator for $\nu(A)$. □

The notation in this proof already suggests that $\widehat{\nu}(A)$ estimates $\left\|\frac{1}{\mathbf{z}}\right\|_D$, the underlying D-norm evaluated at the position $\frac{1}{\mathbf{z}}$. And it does so consistently.

Threshold Strategy 4. *(Choice of marginal estimators for D-norm estimation)*

If we estimate the value of the D-norm

$$\left\|\frac{1}{\boldsymbol{z}}\right\| = \nu(A)$$

using the estimator

$$\widehat{\left\|\frac{1}{\widehat{\boldsymbol{z}}}\right\|} = \nu_n(A)$$

with the notation from the proof of Lemma 20, what estimators $\widehat{F}_j$ should we use for the true marginal distribution functions F_j? Should we use the empirical marginal distribution functions, which are rescaled ranks, or should we use more sophisticated (parametric) estimators?

This is a non-trivial problem. Example 28 and the reference in there imply that in dimension $d = 2$ even if we somehow picked $\widehat{F}_j = F_j, j = 1, 2$, the true marginal distribution functions, the resulting estimator would have an asymptotic variance greater or equal than the asymptotic variance of the rank-based estimator. Intuitively any 'move' away from ranks into the 'direction' of true distribution function will eventually be harmful for the variance of the estimator.

On the other hand Equation (3.13) *finds an upper bound for the probability* $P\left(\left|\left\|\frac{1}{\boldsymbol{z}}\right\|_D - \widehat{\left\|\frac{1}{\widehat{\boldsymbol{z}}}\right\|}\right| > 2\epsilon\right)$.

The upper bound consists of three summands, two of which do not depend upon the marginal estimators at all, while the remaining summand is the probability that $\widehat{\boldsymbol{z}}$ falls into a small rectangle around the true value $\boldsymbol{z}$. This probability mainly depends upon how well the marginal estimators model the univariate tail behavior (see Equation (3.10)*, which defines the components of $\widehat{\boldsymbol{z}}$).*

To reduce the upper bound in Equation (3.13) *it is thefore reasonable to incorporate more sophisticated estimators $\widehat{F}_j$ from univariate extreme value theory.*

Obviously minimizing an upper bound is not the same as minizing the quantity itself. Let us return to Example 27 with $\mathbf{z} = (1, 1)$ for a case where this is most obvious:

If $\widehat{F}_j$ is the true distribution function F_j for every $j = 1, 2$, then the $\widehat{\mathbf{z}}$ coincides with the true $\mathbf{z}$, and the probability $P(\widehat{\mathbf{z}} \in [\mathbf{z}_+, \mathbf{z}_-])$ is zero, the absolute minimum. So for every $\epsilon > 0$, the upper bound in Equation (3.13) is mimized for the choice of $\widehat{F}_j := F_j$.

But if $\widehat{F}_j$ is chosen as the empirical marginal distribution function, then the estimator in Example 27 is almost surely equal to the true value and therefore the left-hand side of Equation (3.13) is equal to 0, the absolute minimum for every $\epsilon > 0$.

Estimating the values of the D-norm

$$\begin{aligned}\left\|\frac{1}{\mathbf{z}}\right\|_D = \mathbb{E}\left(\max_{j=1,\dots,d} Z_j/z_j\right) &= \lim_{t\to\infty} t \cdot P\left(\max_{j=1,\dots,d} X_j/z_j > t\right) \\ &= \nu\left(\left\{\mathbf{x} : \mathbf{x} \geq \mathbf{0}, \max_{j=1,\dots,d} x_j/z_j > 1\right\}\right)\end{aligned}$$

for different vectors $\mathbf{z} > \mathbf{0}$ is a very central part of multivariate extreme value statistics, but not everything. The quantities

$$\mathbb{E}(h(\mathbf{Z})) = \lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t) = \nu(\{\mathbf{x} : \mathbf{x} \geq \mathbf{0}, h(\mathbf{x}) \geq 1\})$$

for other non-negative continuous functions h that are homogeneous of order 1 are of interest, too, as we have seen in Chapter 2.

To show the consistency of estimators for those quantitites we need some auxiliary results. The first of those is that $(\nu_n(A))_{n\in\mathbb{N}}$ is a consistent sequence of estimators for $\nu(A)$, whenever A is a rectangle aligned with the Cartesian coordinate grid.

Lemma 21. *$\nu_n([\boldsymbol{y}, \boldsymbol{z}))$ is a consistent estimator of $\nu([\boldsymbol{y}, \boldsymbol{z}))$ for every $\boldsymbol{y} \in [0,\infty)^d, \boldsymbol{y} \neq \boldsymbol{0}$ and $\boldsymbol{z} > \boldsymbol{y}$.*

Proof. We will show that the indicator function $1_{[\mathbf{y},\mathbf{z})}$ can be written as a linear combination of certain other indicator functions. For a subset $I \subseteq \{1,\dots,d\}$ define the set A_I by

$$\begin{aligned}A_I = \{\mathbf{x} : \mathbf{x} \in [0,\infty)^d, &\text{ there is an index } j \in I \text{ with } x_j \geq y_j \\ &\text{or there is an index } j \notin I \text{ with } x_j \geq z_j\}\end{aligned}$$

Also define $I_+ := \{j : y_j > 0\}$ as the set of indices, where $\mathbf{y}$ is positive. Note that I_+ is not empty. We will prove that the following equation holds for every $\mathbf{x} \in [0,\infty)^d$:

$$1_{[\mathbf{y},\mathbf{z})}(\mathbf{x}) = \sum_{I\subseteq I_+} (-1)^{|I|+1} 1_{A_I}(\mathbf{x}). \tag{3.14}$$

For that we will investigate three cases. For the first case let us assume that there is an index j such that $x_j \geq z_j$. Then we have $1_{A_I}(\mathbf{x}) = 1$ for all subsets I and thus

$$\sum_{I\subseteq I_+} (-1)^{|I|+1} 1_{A_I}(\mathbf{x}) = \sum_{I\subseteq I_+} (-1)^{|I|+1} = 0,$$

where the second step comes from Combinatorics.

For the second case let us assume that there is an index j such that $x_j < y_j$. This implies $j \in I_+$. Observe that the expression $1_{A_I}(\mathbf{x})$ no longer depends on whether $j \in I$ or not. Therefore we get

$$\sum_{I \subseteq I_+} (-1)^{|I|+1} 1_{A_I}(\mathbf{x}) = \sum_{\substack{I \subseteq I_+ \\ j \notin I}} (-1)^{|I|+1} \underbrace{(1_{A_I}(\mathbf{x}) - 1_{A_{I \cup \{j\}}}(\mathbf{x}))}_{=0} = 0.$$

For the third case let us assume $\mathbf{x} \in [\mathbf{y}, \mathbf{z})$. Then because $x_i \geq y_i$ for all i, we have $\mathbf{x} \in A_I$ for all non-empty subsets $I \subsetneq I_+$. However because $x_i < z_i$ for all i, we have $\mathbf{x} \notin A_\emptyset$. Then we get

$$\sum_{I \subseteq I_+} (-1)^{|I|+1} 1_{A_{\mathbf{x}_I}}(\mathbf{x}) = \sum_{\emptyset \subsetneq I \subseteq I_+} (-1)^{|I|+1} = 1$$

again from Combinatorics.

With these three cases we have shown Equation (3.14) to hold for all $\mathbf{x} \geq \mathbf{0}$. Observe that Lemma 19 applies to every set A_I that appears in the sum, which leads us to the conclusion that

$$\nu_n([\mathbf{y}, \mathbf{z})) = \sum_{I \subseteq I_+} (-1)^{|I|+1} \nu_n(A_I)$$

is a consistent estimator for

$$\nu([\mathbf{y}, \mathbf{z})) = \sum_{I \subseteq I_+} (-1)^{|I|+1} \nu(A_I).$$

□

These rectangles will be our building blocks to approximate continuous functions.

Lemma 22. *If f is a bounded, continuous function on $[0, \infty)^d$ such that f vanishes on a neighborhood around $\mathbf{0}$, then $\int f \, \mathrm{d}\nu_n$ is a consistent estimator for $\int f \, \mathrm{d}\nu$.*

Proof. Without loss of generality we can assume $\|f\|_\infty := \sup |f| = 1$. Let $\epsilon, \delta > 0$ be two arbitrary positive real number. We will show that

$$P\left(\left|\int f \, \mathrm{d}\nu_n - \int f \, \mathrm{d}\nu\right| > 9\epsilon\right) < 3\delta \tag{3.15}$$

holds for all n that are large enough.

For that purpose we will partition the set $[0, \infty)^d$ into three subsets A, B and C and integrate the function f over those sets seperately. To define those sets we need a vector $\mathbf{y} > \mathbf{0}$ with $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in [\mathbf{0}, \mathbf{y}]$. Such a $\mathbf{y}$ exists because we assumed f to vanish in a neighborhood around $\mathbf{0}$. Further let $\mathbf{z} > \mathbf{y}$ be a vector

with $\nu([0,\infty)^d \setminus [\mathbf{0},\mathbf{z})) < \epsilon$. Such a $\mathbf{z}$ exists because $\nu([0,\mathbf{z})^{\complement}) = \left\|\frac{1}{\mathbf{z}}\right\|_D \to 0$ as $\mathbf{z} \to \infty$.

Then let us define A, B and C by

$$\begin{aligned} A &= [\mathbf{0},\mathbf{y}), \\ B &= [\mathbf{0},\mathbf{z}) \setminus A, \\ C &= [0,\infty)^d \setminus (A \cup B). \end{aligned}$$

Firstly we have

$$\int_A f \,\mathrm{d}\nu_n - \int_A f \,\mathrm{d}\nu = \int_A 0 \,\mathrm{d}\nu_n - \int_A 0 \,\mathrm{d}\nu = 0.$$

Secondly

$$\left| \int_C f \,\mathrm{d}\nu_n - \int_C f \,\mathrm{d}\nu \right| \leq \int_C \|f\|_\infty \,\mathrm{d}\nu_n + \int_C \|f\|_\infty \,\mathrm{d}\nu = \nu_n(C) + \nu(C).$$

According to Lemma 19 the term $\nu_n(C)$ is a consistent estimator for $\nu(C)$ and because $\nu(C) < \epsilon$ we can assume $P(|\int_C f \,\mathrm{d}\nu_n - \int_C f \,\mathrm{d}\nu| > 3\epsilon) < \delta$ for all n that are large enough.

For the set B we will assume that there exists a function g that is a linear combination of indicator functions of rectangles of the form $[\mathbf{x}_\ell, \mathbf{x}_u)$ that are completely contained in B and the approximation

$$\sup_{\mathbf{x}\in B} |f(\mathbf{x}) - g(\mathbf{x})| \cdot \nu(B) < \epsilon \tag{3.16}$$

holds. Essentially such a g can be constructed by putting a Cartesian grid on the set B and assigning each cell the value of f in the center of the cell. Because f is continuous on the compact set $[\mathbf{0},\mathbf{z}]$, it is uniformly continuous on B and it is only a matter of decreasing the maximal cell size until inequality (3.16) is achieved.

Because of Lemma 21 we know that $\int g \,\mathrm{d}\nu_n$ is a consistent estimator for $\int g \,\mathrm{d}\nu$ and thus for n high enough we get $P(|\int g \,\mathrm{d}\nu_n - \int g \,\mathrm{d}\nu| > \epsilon) < \delta$. Also for n large enough we get $P(\nu_n(B) > 2\nu(B)) < \delta$ and, thus,

$$P\left(\Bigg| \underbrace{\int_B (f-g) \,\mathrm{d}\nu_n}_{\leq \|f-g\|_\infty \cdot \nu_n(B)} \Bigg| > 2\epsilon\right) < \delta \tag{3.17}$$

for those n. Also Equation (3.16) implies

$$P\left(\Bigg| \underbrace{\int_B (f-g) \,\mathrm{d}\nu}_{\leq \|f-g\|_\infty \cdot \nu(B)} \Bigg| > \epsilon\right) = 0.$$

Consequently we have

$$
\begin{aligned}
&P\left(\left|\int_B f\,\mathrm{d}\nu_n - \int_B f\,\mathrm{d}\nu\right| > 6\epsilon\right)\\
=&P\left(\left|\int_B f\,\mathrm{d}\nu_n - \int_B g\,\mathrm{d}\nu_n + \int_B g\,\mathrm{d}\nu_n - \int_B g\,\mathrm{d}\nu + \int_B g\,\mathrm{d}\nu - \int f\,\mathrm{d}\nu\right| > 6\epsilon\right)\\
\leq&\underbrace{P\left(\left|\int_B f\,\mathrm{d}\nu_n - \int_B g\,\mathrm{d}\nu_n\right| > 2\epsilon\right)}_{\leq\delta} + \underbrace{P\left(\left|\int_B g\,\mathrm{d}\nu_n - \int_B g\,\mathrm{d}\nu\right| > 3\epsilon\right)}_{\leq\delta}\\
&+\underbrace{P\left(\left|\int_B f\,\mathrm{d}\nu - \int_B g\,\mathrm{d}\nu\right| > \epsilon\right)}_{=0}
\end{aligned}
$$

for all n that are large enough. Ultimately we get

$$
\begin{aligned}
&P\left(\left|\int f\,\mathrm{d}\nu_n - \int f\,\mathrm{d}\nu\right| > 9\epsilon\right)\\
\leq&P\left(\left|\int_B f\,\mathrm{d}\nu_n - \int_B f\,\mathrm{d}\nu\right| > 6\epsilon\right) + P\left(\left|\int_C f\,\mathrm{d}\nu_n - \int_C f\,\mathrm{d}\nu\right| > 3\epsilon\right) \leq 2\delta + \delta
\end{aligned}
$$

for all n that are large enough. Because these constructions can be done for all $\epsilon, \delta > 0$, the estimator is consistent.

□

The following lemma is essential for proving the consistency of peak-over-threshold estimators.

Lemma 23. *Let f be a continuous, bounded function and let the set D be defined by*

$$D := \{\boldsymbol{x} : \boldsymbol{x} \in [0,\infty)^d, x_1 + \cdots + x_d > 1\}.$$

Then $\int_D f\mathrm{d}\nu_n$ is a consistent estimator for $\int_D f\mathrm{d}\nu$.

Proof. Without loss of generality we can assume that $|f|$ is bounded by 1. Let ϵ, δ be two positive real numbers where without loss of generality we can assume $\epsilon \in (0,1)$. We will show that for all n that are large enough we have

$$P\left(\left|\int_D f\,\mathrm{d}\nu_n - \int_D f\,\mathrm{d}\nu\right| > (4d+2)\epsilon\right) < \delta \tag{3.18}$$

Define two rescaled versions of D, $D_{+\epsilon}$ and $D_{-\epsilon}$ by

$$
\begin{aligned}
D_{+\epsilon} &= \left\{\mathbf{x} : \mathbf{x} \in [0,\infty)^d, x_1 + \cdots + x_d > \frac{1}{1+\epsilon}\right\}\\
D_{-\epsilon} &= \left\{\mathbf{x} : \mathbf{x} \in [0,\infty)^d, x_1 + \cdots + x_d > \frac{1}{1-\epsilon}\right\}.
\end{aligned}
$$

Note that $D_{-\epsilon} \subset D \subset D_{+\epsilon}$. Then we can construct functions f_+ and f_- on $[0,\infty)^d$ that fulfill the following properties:

- f_+ and f_- are continuous,
- $|f_+|$ and $|f_-|$ are bounded by 1,
- $f_+(\mathbf{x}) = 0$ for all $\mathbf{x} \notin D_{+\epsilon}$,
- $f_+(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x} \in D$,
- $f_-(\mathbf{x}) = 0$ for all $\mathbf{x} \notin D$,
- and $f_-(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x} \in D_{-\epsilon}$.

One example for such a construction is

$$f_+(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{for } \sum_{j=1}^d x_j > 1 \\ 0 & \text{for } \sum_{j=1}^d x_j < \frac{1}{1+\epsilon} \\ f\left(\frac{\mathbf{x}}{\sum_{j=1}^d x_j}\right) \cdot ((1+\epsilon) \cdot \sum_{j=1}^d x_j - 1) & \text{else.} \end{cases}$$

Note that f_+ and f_- coincide outside the set $D_{+\epsilon} \setminus D_{-\epsilon}$ and therefore

$$\begin{aligned} \left| \int f_+ \, \mathrm{d}\nu - \int f_- \, \mathrm{d}\nu \right| &\leq \int_{D_{+\epsilon} \setminus D_{-\epsilon}} \|f_+ - f_-\|_\infty \, \mathrm{d}\nu \\ &\leq 2 \cdot (\nu(D_{+\epsilon}) - \nu(D_{-\epsilon})) \\ &= 2 \cdot \left(\mathbb{E}\left((1+\epsilon) \sum_{j=1}^d Z_j \right) - E\left((1-\epsilon) \sum_{j=1}^d Z_j \right) \right) = 4 \cdot d \cdot \epsilon, \end{aligned}$$

where $\mathbf{Z}$ is an arbitrary generator of the underlying D-norm and where we used Lemma 1.

Because $f_- \leq f \cdot 1_D \leq f_+$ we have

$$\begin{aligned} \int_D f \, \mathrm{d}\nu &\in \left[\int f_- \, \mathrm{d}\nu, \int f_+ \, \mathrm{d}\nu \right] \\ \text{and } \int_D f \, \mathrm{d}\nu_n &\in \left[\int f_- \, \mathrm{d}\nu_n, \int f_+ \, \mathrm{d}\nu_n \right]. \end{aligned}$$

Because of Lemma 22 we have

$$P\left(\left[\int f_- \, \mathrm{d}\nu_n, \int f_+ \, \mathrm{d}\nu_n \right] \subseteq \left[\int f_- \, \mathrm{d}\nu - \epsilon, \int f_+ \, \mathrm{d}\nu + \epsilon \right] \right) > 1 - \delta$$

for all n that are large enough.

In this event both $\int_D f \mathrm{d}\nu$ and $\int_D f \mathrm{d}\nu_n$ are located in the interval

$$\left[\int f_- \,\mathrm{d}\nu - \epsilon, \int f_+ \,\mathrm{d}\nu + \epsilon\right],$$

which has a length of at most $(4d+2)\epsilon$. This implies Equation (3.18) for all n that are large enough.

□

Lemma 24. *Let h be a continuous, non-negative function on $[0,\infty)^d$ that is homogeneous of order 1 and let $A = h^{-1}((1,\infty))$. Then $\nu_n(A)$ is a consistent estimator of $\nu(A)$.*

Proof. This proof will be similar to the proof of Lemma 23. We set $f = 1_A$. We then have $\nu(A) = \int f \,\mathrm{d}\nu$ and $\nu_n(A) = \int f \,\mathrm{d}\nu_n$. The function f is not continuous so we can't apply Lemma 22 directly. But for every $\epsilon \in (0,1)$ we can introduce the sets

$$A_{+\epsilon} := h^{-1}\left(\left(\frac{1}{1+\epsilon}, \infty\right)\right)$$
$$A_{-\epsilon} := h^{-1}\left(\left(\frac{1}{1-\epsilon}, \infty\right)\right).$$

We now can introduce two functions f_+, f_- by

$$f_+(\mathbf{x}) := \begin{cases} 1 & \text{for } h(\mathbf{x}) > 1 \\ 0 & \text{for } h(\mathbf{x}) < \frac{1}{1+\epsilon} \\ \frac{1+\epsilon}{\epsilon} \cdot h(\mathbf{x}) - \frac{1}{\epsilon} & \text{else} \end{cases}$$

and

$$f_-(\mathbf{x}) := \begin{cases} 1 & \text{for } h(\mathbf{x}) > \frac{1}{1-\epsilon} \\ 0 & \text{for } h(\mathbf{x}) < 1 \\ \frac{1-\epsilon}{\epsilon} \cdot h(\mathbf{x}) - \frac{1-\epsilon}{\epsilon} & \text{else.} \end{cases}$$

Both f_+ and f_- are continuous and the coincide outside the set $A_{+\epsilon} \setminus A_{-\epsilon}$. This leads us to

$$\begin{aligned} \left|\int f_+ \,\mathrm{d}\nu - \int f_- \,\mathrm{d}\nu\right| &= \left|\int_{A_{+\epsilon}\setminus A_{-\epsilon}} f_+ - f_- \,\mathrm{d}\,\nu\right| \\ &\leq \|f_+ - f_-\|_\infty \, \nu(A_{+\epsilon} \setminus A_{-\epsilon}) \\ &\leq 2 \cdot (\nu(A_{+\epsilon}) - \nu(A_{-\epsilon})) \\ &= 2 \cdot (\mathbb{E}((1+\epsilon)\cdot h(\mathbf{Z})) - \mathbb{E}((1-\epsilon)\cdot h(\mathbf{Z}))) \quad = 4 \cdot \epsilon \cdot H, \end{aligned}$$

where $\mathbf{Z}$ is an arbitrary generator of the underlying D-norm and $H = \mathbb{E}(h(\mathbf{Z}))$.

Just like in the proof of Lemma 23 the constant value $\int f \,\mathrm{d}\nu$ falls into the constant interval $\left[\int f_- \,\mathrm{d}\nu, \int f_+ \,\mathrm{d}\nu\right]$ and the random value $\int f \,\mathrm{d}\nu_n$ falls into the random interval $\left[\int f_- \,\mathrm{d}\nu_n, \int f_+ \,\mathrm{d}\nu_n\right]$. Because the endpoints of the random interval converge to the endpoints of the constant interval in probability we have for all n that are high enough

$$\begin{aligned}&P(|\nu_n(A) - \nu(A)| \leq (4H+2)\epsilon)\\ \geq &P\left(\int f \,\mathrm{d}\nu_n \in \left[\int f_- \,\mathrm{d}\nu - \epsilon, \int f_+ \,\mathrm{d}\nu + \epsilon\right]\right) \geq (1-\delta),\end{aligned}$$

where we used that the inveral in the second line contains the true value $\nu(A)$ and has a lenght of at most $(4H+2)\epsilon$.

These constructions can be done for every $\epsilon \in (0,1)$ and every $\delta > 0$, so $\nu_n(A)$ is a consistent estimator for $\nu(A)$. □

So far the integral $\int f \,\mathrm{d}\nu_n$ has been treated like an abstract object and the only thing we have investigated is the random difference between it and the non-random integral $\int f \,\mathrm{d}\nu$. But this will change in the following proof: We will find the connection to the peaks-over-threshold estimators.

Proof of Theorem 12. Note that for every realization of $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ the measure $\nu_n := \nu_{n,k(n)}$ in Equation (3.8) is a sum of n point masses. So if we define the set $A = h^{-1}((1,\infty))$ we have can evaluate $\nu_n(A)$ as

$$\nu_n(A) = \frac{1}{k}\sum_{i=1}^{d} 1_{h((\mathbf{X}'')^{(i)})>n/k},$$

which is exactly the indirect structure-variable estimator. Because of Lemma 24 we know that this produces a consistent sequence of estimators for $\nu(h^{-1}((1,\infty)))$. In the proof of the Rosetta Stone theorem (the step denoted by '(v) ⇒ (i)') we have seen that this quantity is equal to $H = \lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t)$.

To prove the consistency of the peaks-over-threshold estimator we need a generator $\mathbf{Z}$ of the underlying D-norm that fulfills

$$\mathbf{Z} \in B := \{\mathbf{x} : x_1 + \cdots + x_d = d\}$$

almost surely. Such ta $\mathbf{Z}$ exists by virtue of Corollary 4. By the proof of Lemma 2 we have $P(\mathbf{Z} \in A) = \nu((1,\infty)\cdot A)$ for all Borel subsets $A \subseteq B$. We can rewrite this to

$$\mathbb{E}(1_A(\mathbf{Z})) = \int_D 1_A\left(\frac{\mathbf{x}}{(x_1+\cdots+x_d)/d}\right)\mathrm{d}\nu,$$

where the set D is defined just like in Lemma 23. Linearity of both the expected value and the integral this leads us to

$$\mathbb{E}(f(\mathbf{Z})) = \int_D f\left(\frac{\mathbf{x}}{(x_1+\cdots+x_d)/d}\right)\mathrm{d}\nu, \tag{3.19}$$

whenever f is a linear combination of indicator functions of measurable subsets $A \subset B$. Now we can use monotone convergence on both sides of Equation (3.19) to show that Equation (3.19) holds for arbitrary continuous, non-negative functions $f : B \to [0, \infty)$.

By choosing $f(\mathbf{x}) = h(\mathbf{x})$ for all $\mathbf{x} \in B$ we get

$$\begin{aligned}\mathbb{E}(h(\mathbf{Z})) = \mathbb{E}(f(\mathbf{Z})) &= \int_D f\left(\frac{\mathbf{x}}{(x_1 + \cdots + x_d)/d}\right) \mathrm{d}\nu \\ &= \int_D h\left(\frac{\mathbf{x}}{(x_1 + \cdots + x_d)/d}\right) \mathrm{d}\nu.\end{aligned}$$

Again for every realization of $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ the measure ν_n consists of finitely many point masses, which we can use to evaluate the integral

$$\begin{aligned}\int_D h\left(\frac{d \cdot \mathbf{x}}{x_1 + \cdots + x_d}\right) \mathrm{d}\nu_n &= \sum_{i:\mathbf{X}^{(i)} \in D} h\left(\frac{d \cdot \mathbf{X}^{(i)}}{r_i}\right) \cdot \nu_n(\{\mathbf{X}^{(i)}\}) \\ &= \frac{1}{k} \cdot \sum_{i \in M} h\left(\frac{d \cdot \mathbf{X}^{(i)}}{r_i}\right).\end{aligned}$$

According to Lemma 23 this is a consistent estimator for $\mathbb{E}(h(\mathbf{Z}))$. But in the peaks-over-threshold estimator we have the random term $\frac{1}{|M|}$ instead of the constant term $\frac{1}{k}$. What remains to be shown is $|M|/k \to_p 1$.

With the same arguments as before

$$\frac{|M|}{k} = \sum_{i \in M} 1/k = \int_D 1 \, \mathrm{d}\nu_n$$

is a consistent estimator for $\mathbb{E}(1) = 1$. □

3.4 Local and global thresholds

This section is concerned about estimating $\mathbb{E}(h(\mathbf{Z})) = \lim_{t\to\infty} t \cdot P(h(\mathbf{X}) > t)$, where h is as usual a continuous, non-negative function that is homogeneous of order 1, but which only depends on some components of its input. One example for this is $h(x_1, \ldots, x_d) = \max(x_1, x_2)$. This function only depends on the input x_1 and x_2 and is not affected by $x_3, \ldots, x_d$. More formally we are referring to functions of the following type:

$$h(\mathbf{x}) = g(\mathbf{x}'), \tag{3.20}$$

where there is a lower dimension $d' < d$, indices $j_1 < j_2 < \cdots < j_{d'}$ and where we use the notation $\mathbf{x}' := (x_{j_1}, \ldots, x_{j_{d'}})$ and where g is a non-negative continuous function on $[0, \infty)^{d'}$, that is homogeneous of order 1.

The direct structure-variable estimator $\widehat{H}_{\text{struct}}$ for $H = \mathbb{E}(h(\mathbf{Z}))$ ignores the additional components as can be seen in the following:

$$\begin{aligned}\widehat{H}_{\text{struct}} &= \frac{1}{k}\sum_{i=1}^{n} 1_{h(\mathbf{X}^{(i)})>n/k} \\ &= \frac{1}{k}\sum_{i=1}^{n} 1_{g(\mathbf{X}'^{(i)})>n/k}.\end{aligned}$$

However, for the direct peaks-over-threshold estimator it does make a difference, whether we involve all d dimensions or only the necessary subset of d' dimensions. The global threshold procedure is the direct estimator as we already know it:

$$\begin{aligned}r_i &:= \frac{1}{d}\sum_{j=1}^{d} X_j^{(i)}, i = 1, \ldots, n, \\ M &:= \{i : 1 \le i \le d, r_i > n/k\}, \\ \widehat{H}_{\text{PoT}} &:= \frac{1}{|M|}\sum_{i\in M} h\left(\frac{\mathbf{X}^{(i)}}{r_i}\right).\end{aligned}$$

The local threshold procedure is defined by:

$$\begin{aligned}r'_i &:= \frac{1}{d'}\sum_{k=1}^{d'} X_{j_k}^{(i)}, i = 1, \ldots, n, \\ M' &:= \{i : 1 \le i \le d, r'_i > n/k\}, \\ \widehat{H'}_{\text{PoT}} &:= \frac{1}{|M'|}\sum_{i\in M} g\left(\frac{\mathbf{X}'^{(i)}}{r_i}\right).\end{aligned}$$

The following two examples will show, that comparing the performance of the local threshold procedure with the performance of the global threshold procedure is not easy.

Example 30. *Let $\boldsymbol{Z}$ be 4-dimensional D-norm generator defined by the following probabilities:*

$$P(\boldsymbol{Z} = (2, 2, 0, 0)^\intercal) = P(\boldsymbol{Z} = (0, 0, 2, 2)^\intercal) = 1/2, \tag{3.21}$$

while U is a random variable independent of $\boldsymbol{Z}$ and uniformly distributed on $(0, 1)$. According to Corollary 3 the random vector $\boldsymbol{X} = \frac{1}{U} \cdot \boldsymbol{Z}$ fulfills the equivalent statements of the Rosetta Stone theorem with the D-norm generator $\boldsymbol{Z}$.

We now want to estimate $\mathbb{E}(\max(Z_1, Z_2)) = 1$ with a direct peaks-over-threshold estimator. Using a local threshold on $X_1 + X_2$, all normed exceedances are of the form $(1, 1)$ and the estimator becomes

$$\frac{1}{|M'|} \sum_{i \in M'} \max(X_1^{(i)}/r_i', X_2^{(i)}/r_i') = \frac{1}{|M'|} |M'| \cdot \max(1, 1) = 1,$$

for any choice of threshold as long as $M' \neq \emptyset$.

If we use a global threshold on $X_1 + X_2 + X_3 + X_4$, then for every $i \in M$ there is a 50% chance that the normed exceedance is $(2, 2, 0, 0)$ and a 50% chance that it is $(0, 0, 2, 2)$. The estimator becomes:

$$\frac{1}{|M|} \sum_{i \in M} \max(X_1^{(i)}/r_i, X_2^{(i)}/r_i) = \frac{2}{|M|} \cdot \sum_{i \in M} 1_{X_1^{(i)} > 0}.$$

If we condition the estimator on $|M| = m \in \mathbb{N}$ we can repeat the steps in the proof of Theorem 11 to show that under this condition the estimator is distributed like $2/m \cdot \mathcal{B}(m, 1/2)$, where $\mathcal{B}(m, 1/2)$ stands for a binomially distributed random variable with parameters m and $p = 1/2$.

So while the bias of zero remained unchanged, the best possible variance of 0 became $1/m$ (under the condition $|M| = m$) by switching from local to global thresholds.

Example 31. *Let $\boldsymbol{Z}$ be 3 dimensional D-norm generator defined by the following probabilities:*

$$P(\boldsymbol{Z} = (1.5, 0, 1.5)^\intercal) = P(\boldsymbol{Z} = (0, 1.5, 1.5)^\intercal) = P(\boldsymbol{Z} = (1.5, 1.5, 0)^\intercal) = 1/3,$$

while U is a random variable independent of $\boldsymbol{Z}$ and uniformly distributed on $(0, 1)$. Once again $\boldsymbol{X} = \frac{1}{U} \cdot \boldsymbol{Z}$ fulfills the equivalent statements of the Rosetta Stone theorem with D-norm generator $\boldsymbol{Z}$.

This time we want to estimate $\mathbb{E}(\max(Z_1, Z_2)) = 1.5$ with a peaks-over-threshold procedure. Using global thresholds, every normed excedance is one of the following types:

$$(1.5, 0, 1.5)^\intercal, (0, 1.5, 1.5)^\intercal, (1.5, 1.5, 0)^\intercal$$

In any case the maximum of the first two components is 1.5*, which means the global threshold procedure produces the constant* 1.5*, which is the true value of what we want to estimate. A local threshold estimator procedure can only be worse than that.*

Threshold Strategy 5. *(Local or global threshold) We have seen a case, where switching from a global to a local threshold is a mistake, but we have also seen a case where switching from a global to a local threshold is a mistake as well. It is therefore reasonable to chose that kind of threshold procedure, where the upper bound of the variance conditioned on the number of exceedances* $|M| = m$*, which is* $\frac{h_{\max}-h_{\min}}{4m}$ *is minimal. This is always the local threshold procedure, as we will see in Lemma 25.*

Lemma 25. *Let* $g : [0,\infty)^{d'} \to [0,\infty)$*,* $h : [0,\infty)^{d} \to [0,\infty)$ *be a continuous functions that are homogeneous of order* 1 *and that are connected by Equation* (3.20)*. Then we have*

$$h_{\max} - h_{\min} \geq \frac{d}{d'} \cdot (g_{\max} - g_{\min}),$$

where we used notation from Definition 4.

Proof. Let $I = \{1,\dots,d\}$ be the set of indices and $I' = \{i_j : j = 1,\dots,d'\}$ be the set of indices that are not lost in the projection $\mathbf{x} \mapsto \mathbf{x}'$. The inequality $h_{\max} \geq g_{\max}$ can be shown by

$$\begin{aligned} h_{\max} &= \max\left\{ h(\mathbf{x}) : \mathbf{x} \geq \mathbf{0}, \sum_{j=1}^{d} x_j = d \right\} \\ &\geq \max\left\{ h(\mathbf{x}) : \mathbf{x} \geq \mathbf{0}, x_j = 0 \text{ for all } j \notin I', \sum_{j=1}^{d} x_j = d \right\} \\ &= \max\left\{ g(\mathbf{x}') : \mathbf{x}' \geq \mathbf{0}, \sum_{j=1}^{d'} x'_j = d \right\} = \frac{d}{d'} \cdot g_{\max}, \end{aligned}$$

where in the last step we used that g is non-negative and homogeneous of order 1.

Also $h_{\min} = 0$, because we can construct a vector $\mathbf{x} \geq \mathbf{0}$ that places its sum of d on components with indices $\notin I'$. □

In fact a scenario like in Example 30 happens quite naturally. If you investigate the extremal dependence of components $X_i, i \in I$ of a random vector such that $I = A \dot\cup B$, the disjoint union of A and B and $(X_i)_{i\in A}$ is independent of $(X_i)_{i\in B}$. In that case extreme events in group A will contribute to the total of global threshold exceedances without carrying information about the extremal dependence within B and vice versa. This will unnecessarily increase the variance of estimators as shown in the example.

3.5 The split-and-merge-procedure

One problem in estimating quantitites in multivariate extreme value theory are the conflicting interests for choosing the parameter k, which determines the threshold n/k.

Threshold Strategies 1 and 2 were founded on clear relationships between the parameter k and upper bounds for variance of the direct estimators. For indirect estimators we have Threshold Strategy 3, but it only tells us what not to do. It is not constructive.

This is where the split-and-merge-procedure steps in. In this procedure we can freely speculate about the relationship between the parameter (in our case k) of a point estimator and its variance. No matter how wrong we are in the speculation phase, in the end we get an interval estimator for the expected value of the point estimator with a guaranteed coverage probability, but of random length. If our assumptions about the variability of our point estimator are wrong, we are will notice it by the length of the confidence intervals.

The math behind the split-and-merge-procedure is simple as we will see.

Definition 20. *Let L be a natural number. We call it the split-and-merge-procedure to split up an iid sample $X^{(1)}, \ldots, X^{(n)}$ into L non-overlapping blocks of size $\lfloor n/L \rfloor$ and evaluate an estimator on every block separately, which results in different point estimators $\widehat{H}_1, \ldots, \widehat{H}_L$. The arithmetic mean*

$$\widehat{H} := \frac{1}{L} \sum_{\ell=1}^{L} \widehat{H}_\ell$$

we call the merged point estimator. The split-and-merge-interval is defined by

$$I_L = \left[\widehat{H} - 3 \cdot \frac{\sqrt{S_L^2 + \Delta}}{\sqrt{L}}, \widehat{H} + 3 \cdot \frac{\sqrt{S_L^2 + \Delta}}{\sqrt{L}} \right], \tag{3.22}$$

where

$$S_L^2 := \frac{1}{L-1} \sum_{\ell=1}^{L} \left(\widehat{H}_\ell - \widehat{H} \right)^2 \tag{3.23}$$

is the sample variance of the point estimators and $\Delta := 3 \cdot \sqrt{\frac{|h_{\max} - h_{\min}|^4}{12L}}$ and where $h_{\max}$ and $h_{\min}$ are values such that

$$P(\widehat{H}_1 \in [h_{\min}, h_{\max}]) = 1.$$

Theorem 13 (Split-and-merge-interval as a confidence interval)**.** *The interval in Equation* (3.22) *has a probability of at least* $71/90 \approx 78.8\%$ *to cover the expected value $E(\widehat{H}_1)$.*

To prove Theorem 13 we need some inequalities for the second and forth centered moment of a bounded random variable.

Lemma 26. *Let X be a random variable that only takes values in the closed interval $[a, b]$. Then we have*

$$\mathbb{E}((X-\mu)^2) \leq \frac{1}{4}(b-a)^2 \text{ and}$$
$$\mathbb{E}((X-\mu)^4) \leq \frac{1}{12}(b-a)^4,$$

where $\mu = \mathbb{E}(X)$.

Proof. Without loss of generality we can assume $a = 0$ and $b = 1$. The proof for the variance is straightforward once we realize that μ is the minimizer of the quadratic function $y \mapsto \mathbb{E}((X-y)^2)$, which leads us to the following inequality:

$$\mathbb{E}((X-\mu)^2) = \min_{y\in\mathbb{R}} \mathbb{E}((X-y)^2) \leq \mathbb{E}(\underbrace{|X-1/2|}_{\leq 1/2}^2) \leq \frac{1}{4}$$

The proof for the fourth centered moment is more complicated. First we will show that for a given expected value μ a Bernoulli random variable maximizes the fourth centered moments. Let X be a random variable that only takes values in the interval $[0, 1]$ and which has an expected value of μ. Let us define a second random variable X' with the following conditional probabilities:

$$P(X' = 1|X = x) = x,$$
$$P(X' = 0|X = x) = 1 - x \text{ for all } x \in [0, 1].$$

This results in a Bernoulli random variable X' as we have

$$\begin{aligned} P(X' \notin \{0, 1\}) &= \int_0^1 P(X' \notin \{0, 1\}|X = x)\,\mathrm{d}P(X = x) \\ &= \int_0^1 0\,\mathrm{d}P(X = x) = 0. \end{aligned}$$

Also we get

$$\mathbb{E}(X') = \int_0^1 \mathbb{E}(X'|X = x)\,\mathrm{d}P(X = x) = \int_0^1 x\,\mathrm{d}P(X = x) = \mathbb{E}(X) = \mu.$$

Because $x \mapsto (x-\mu)^4$ is a convex function we also have

$$(x-\mu)^4 \leq x \cdot (1-\mu)^4 + (1-x) \cdot (0-\mu)^4$$

for all $x \in [0, 1]$. This leads us to

$$\begin{aligned}\mathbb{E}((X-\mu)^4) &= \int_0^1 (x-\mu)^4 \,\mathrm{d}P(X=x) \\ &\leq \int_0^1 x\cdot(1-\mu)^4 + (1-x)\cdot(0-\mu)^4 \,\mathrm{d}P(X=x) \\ &= \int_0^1 \mathbb{E}((X'-\mu)^4|X=x)\,\mathrm{d}P(X=x) = \mathbb{E}((X'-\mu)^4).\end{aligned}$$

Therefore the Bernoulli random variable with parameter $p=\mu$ maximizes the fourth centered moment under all random variables X on $[0,1]$ with expected value μ. The fourth centered moment of a Bernoulli random variable is

$$\mu\cdot(1-\mu)^4 + (1-\mu)\cdot\mu^4 = -3\mu^4+6\mu^3-4\mu^2+\mu.$$

On the interval $[0,1]$ this polynomial has 5 extremal points. The trivial minima at $\mu=0$ and $\mu=1$, the local minimum at $\mu=1/2$ and the two local maxima at $\mu=\frac{1}{2}\pm\frac{\sqrt{3}}{6}$. At both of those points the function has the value $\frac{1}{12}$. □

The main point of the split-and-merge-procedure is that we can estimate $\mathrm{Var}(\widehat{H}_1)$ with the sample variance S_L^2. For that we also need to know some properties about the sample variance:

Lemma 27 (Properties of the sample variance)**.** *S_L^2 is an unbiased estimator for $\mathrm{Var}(\widehat{H}_1)$ and if $L\geq 3$ it has a variance of less or equal than $\frac{|h_{\max}-h_{\min}|^4}{12L}$.*

Proof. The bias of 0 is elementary to show and will be omitted here. It is more tedious to show that the variance of the sample variance is given by

$$\mathrm{Var}(S_L^2) = \frac{1}{L}\cdot\left(\mu_4 - \frac{L-3}{L-1}\sigma^4\right),$$

where μ_4 is the fourth centered moment of $\widehat{H}_1$. The calculations for that be looked up in a work by Cho and Cho (2009), where it was formulated for sampling with replacement from a finite set. Their calculations hold nonetheless for arbitrary random variables. From Lemma 26 we know that the forth centered moment is less or equal $\frac{|h_{\max}-h_{\min}|^4}{12}$. This then implies

$$\mathrm{Var}(S_L^2) \leq \frac{1}{L}\cdot\mu_4 \leq \frac{|h_{\max}-h_{\min}|^4}{12L}.$$

□

For the proof of Theorem 13 we also need two very basic inequalities: Chebyshev's inequality and Cantelli's inequality. They can be looked up in Chapter 1, Section 5 in the book by Billingsley (1979) and essentially say, that if X is a

square integrable random variable with mean μ and standard deviation σ, then for every positive value α we have

$$P(|X-\mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2} \text{ (Chebyshev)}$$
$$P(X-\mu \geq \alpha) \leq \frac{\sigma^2}{\sigma^2+\alpha^2} \text{ (Cantelli)}.$$

With the special choice of $\alpha = 3\sigma$ those inequalities become

$$P(|X-\mu| \geq 3\sigma) \leq \frac{1}{9}$$
$$P(X-\mu \geq 3\sigma) \leq \frac{1}{10},$$

which we will use in the following proof.

Proof of Theorem 13. Let μ be the expected value of $\widehat{H}_1$. Obviously $\mu = \mathbb{E}(\widehat{H})$ for the merged estimator as well. We will introduce the following random interval:

$$\tilde{I}_L = \left[\widehat{H} - 3 \cdot \sqrt{\mathrm{Var}\left(\widehat{H}\right)}, \widehat{H} + 3 \cdot \sqrt{\mathrm{Var}\left(\widehat{H}\right)}\right]. \tag{3.24}$$

We can apply Chebyshev's inequality to get $\mu \in \tilde{I}_L$ with a probability of at least $8/9$. Consequently we have

$$P(\mu \notin I_L) = \underbrace{P(\mu \notin I_L, \mu \notin \tilde{I}_L)}_{\leq 1/9} + P(\mu \notin I_L, \mu \in \tilde{I}_L).$$

The event $\mu \notin I_L, \mu \in \tilde{I}_L$ implies that $\tilde{I}_L \not\subset I_L$. Comparing the definitions of those two intervals in the Equations (3.22) and (3.24) we can see that $\tilde{I}_L \not\subset I_L$ implies the inequality

$$\frac{\sqrt{S_L^2 + \Delta}}{\sqrt{L}} < \sqrt{\mathrm{Var}(\widehat{H})},$$

which together with the results of Lemma 27 and the definition of Δ leads to

$$S_L^2 - \mathbb{E}(S_L^2) = S_L^2 - \mathrm{Var}(\widehat{H}) \cdot L < -\Delta \leq -3 \cdot \sqrt{\mathrm{Var}(S_L^2)}.$$

According to Cantelli's inequality this only happens with a probability of at most $1/10$. Consequently $P(\mu \notin I_L) \leq 1/9 + 1/10 = 19/90$, which means the split-and-merge-interval has a coverage probability of at least $71/90$. □

The split-and-merge-procedure can be applied to every estimator that falls into a finite interval $[h_{\min}, h_{\max}]$. But we can also use it for a threshold strategy for indirect estimators that is more constructive than Threshold Strategy 3, that only tells us what not to do.

Threshold Strategy 6. *(Split-and-merge-procedure) First we should figure out what width of a confidence interval is tolerable in our practical situation and call it w_{tol}.*

The number of blocks L we set to

$$L := \left\lceil \frac{36}{\sqrt[3]{12}} \cdot \left(\frac{|h_{\max} - h_{\min}|}{w_{tol}} \right)^{4/3} \right\rceil$$

and just like in Definition 20 we define Δ by

$$\Delta := 3 \cdot \sqrt{\frac{|h_{\max} - h_{\min}|^4}{12L}}$$

for this value L.

Then we need a function $f : k \mapsto f(k)$, of which we suspect that $\mathrm{Var}(\widehat{H}) \leq f(k)$. In the case of the structure-variable estimator that might be $f(k) = \frac{h_{\max}}{k}$ inspired by Threshold Strategy 1.

The parameter k is then picked as

$$k := f^{-1}(\Delta).$$

Then we split the dataset into L overlapping blocks of the same size, evaluate the point estimators $\widehat{H}_\ell, \ell = 1, \ldots, L$ individually on each block using the parameter k, determine the merged estimator $\widehat{H}$ and the sample variance of the estimator S_L^2.

The random length of the split-and-merge-interval is $6 \cdot \frac{\sqrt{S_L^2 + \Delta}}{\sqrt{L}}$. If our suspicion about the relationship between the parameter k and the variance of the point estimator is correct, then the empirical variance S_L^2 will tend to below Δ. Whenever we have $S_L^2 \leq \Delta$ we also have

$$6 \cdot \frac{\sqrt{S_L^2 + \Delta}}{\sqrt{L}} \leq 6 \cdot \sqrt{\frac{2\Delta}{L}} = 6 \cdot \sqrt{6 \cdot \sqrt{\frac{|h_{\max} - h_{\min}|^4}{12L^3}}} \leq 6 \cdot \sqrt{6 \cdot \sqrt{\frac{w^4}{36^3}}} = w.$$

The following example will illustrate the split-and-merge-procedure.

Example 32. *In this scenario we have an iid sample $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$ of a bivariate random vector $\boldsymbol{X} = (X_1, X_2)^\intercal$ in the max-domain of attraction of a max-stable distribution, which has a dependence structure given by a D-norm $\|\cdot\|_D$ with generator $\boldsymbol{Z} = (Z_1, Z_2)^\intercal$. We want to estimate the value $H = \mathbb{E}(\max(Z_1, Z_2)) = \|(1,1)^\intercal\|_D$. We are also in a scenario, where a confidence interval of length 0.1 would be acceptable.*

First and foremost $h_{\min} = 1$ and $h_{\max} = 2$. The number of blocks will therefore be

$$L = \left\lceil \frac{36}{\sqrt[3]{12}} \cdot \left(\frac{|2-1|}{0.1} \right)^{4/3} \right\rceil = \left\lceil \frac{36}{\sqrt[3]{12}} \cdot 10^{4/3} \right\rceil = 339.$$

We would then set Δ to

$$\Delta = 3 \cdot \sqrt{\frac{|2-1|^4}{12 \cdot 339}} \approx 0.0470.$$

If we use the structure-variable estimator on each block we would suspect that $f(k) = \frac{h_{\max}}{k}$ is an upper bound for the variance and we would set $k = \frac{2}{\Delta} \approx 42.5$. If we use the peaks-over-threshold estimator on each block we would suspect that $f(k) = \frac{h_{\max} - h_{\min}}{4k}$ is an upper bound for the variance and we would set $k = \frac{2-1}{4\Delta} = 5.32$.

In each of those cases our threshold in each block would be $t = \frac{\lfloor N/L \rfloor}{k}$.

This example shows us that even for moderate requirements on the interval-estimator we end up with a number L in the hundreds. This shows that the split-and-merge-procedure is not a 'one size fits all' solution, but it was never meant to be to begin with. Its purpose is to cover a gap in between the theory of multivariate extremes, the theory of univariate extremes and statistical practice.

Asymptotic normality of estimators for the quantitites $H = \lim_{t \to \infty} t \cdot P(h(\mathbf{X}') > t)$, where $\mathbf{X}'$ is defined with Equation (3.2), has been proven for the case that h has the form $h(\mathbf{x}) = \max_j |z_j| x_j$, the estimator is the structure-variable estimator, the marginal distributions are estimated with ranks and that there is second order convergence (see Einmahl et al. (2012) and Bücher et al. (2014)). To the knowledge of the author asymptotic normality has never been proven for the peaks-over-threshold estimator, for when a different kind of function h is picked, for when there is no second-order convergence or for when the marginal distributions are not estimated with ranks.

Especially the last part is problematic: If multivariate extreme value theory restricts itself to only working with ranks, then it can never incorporate results from univariate extreme value theory. Threshold Strategy 4 made a case for why better estimation of the marginal tail behavior leads to better upper bounds for error probabilities for D-norm estimators.

The split-and-merge-procedure allows one to combine marginal estimators without knowing their joint asymptotics and the effect of their joint asymptotics on the asymptotic behavior of the the estimator $\widehat{H}$ and still produce a confidence interval.

As for the the 'multivariate extreme value theory will not incorporate univariate extreme value theory', there is a noteworthy observation to be made: In the opposite direction researchers in univariate extreme value theory have realized that they can use multivariate extreme value theory to their advantage:

Example 33. *Clémençon and Dematteo (2016) investigated the scenario, where they have iid observations $\boldsymbol{X}^{(i)}, i = 1, \ldots, n$ of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\intercal$, where all the components have they same tail index $\alpha > 0$. To estimate α one can apply the Hill-estimator on the iid sequence $(X_1^{(1)}, \ldots, X_1^{(n)})$, but also on the iid sequence $(X_2^{(1)}, \ldots, X_2^{(n)})$, etc. This results in one Hill-estimator per component, that is $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_d$. They figure out the best way to merge those in-*

dividual estimators into a single estimator depends on the underlying extremal dependence structure of the components of $\boldsymbol{X}$.

Example 34. *Kim and Lee (2017) treat the very same problem, while in their solution they apply the Hill-estimator on different linear combinations of the form*

$$\sum_{j=1}^{d} \lambda_j \cdot X_j, \text{ where } \lambda_j \geq 0 \text{ for all } j \text{ and } \sum_{j=1}^{d} \lambda_j = 1.$$

Let us also discuss an interesting side product of the split-and-merge-procedure: For each dimension we end up with L iid estimators for the marginal distribution, one estimator per block. This lets us evaluate the inherent variability of the marginal estimators. If it is too high and we have made a parametric assumption about the univariate tail behavior, then this indicates our parametric assumption might be wrong. Or it might mean that the 'identically distributed' part of 'iid' does not hold. Let's investigate the second scenario in a non-extreme setup.

Example 35. *Let* $X_1, \ldots, X_n$ *be an iid sequence of random variables and let* $Y_1, \ldots, Y_n$ *be another iid sequence of random variables independent of the first sequence. Further let there be two sequences of real numbers* $T_{x1}, \ldots, T_{xn}$ *and* $T_{y1}, \ldots, T_{yn}$*, both of which are monotonously increasing.*

Then $X_i + T_{xi}$ *is independent of* $Y_i + T_{yi}$ *for all* $i = 1, \ldots, n$*, so the dependence structure (the underlying copula) does not change by adding the trends. But if we observe the sequence* $(X_i + T_{xi}, Y_i + T_{yi})$ *and treat it as an iid sample, we would think the components are positively correlated, because large observations in the first component tend to appear for higher indices* i *and for higher indices* i*, the second component tends to be higher as well.*

The standard way to detect the trends T_{xi} and T_{y_i} in Example 35 would be to apply a moving average filter, which is not too different from splitting the sample into blocks and doing a statistical procedure (taking the average) on each block.

In extreme value statistic one could imagine something similar to trends in the margins: For example the probability to exceed a certain threshold increases over time. Or the value at risk for a certain probability level increases over time. If the direction of the trends in all margins is the same, statistical procedures would tend to overestimate the extremal dependence. For example the first component will exceed its threshold more often during times, where the second component exceeds its threshold more often as well.

As the split-and-merge-procedure produces both estimates for the marginal distributions as well a measure of extremal dependence, it is more robust for changes in the marginal distributions. One could even go so far as to say, that a change in the univariate extremal behavior of the marginal distributions is a 'quantitative change', while a change in the extremal dependence structure is a

'qualitative change' in the sense that switching from tail-independence of two components to tail-dependence is of deeper nature.

The split-and-merge-procedure could be a standard tool in investigating the changes of the extremal behavior of random vectors, both in margins and in copula. If we use it for that purpose, the parameter L can be chosen more freely than in Example 32.

3.6 Simulation study

This simulation study will investigate several open questions.

Q.1 Equations (3.1) and (3.2) and Theorem 11 show that the bias of the direct estimators is a function of the height of the threshold n/k, while the variance essentially depends on the parameter k. Does this also hold for the indirect estimators?

Q.2 The structure-variable estimators from the Definitions 16 and 18 are fundamentally different from the peaks-over-threshold estimators from the Definitions 17 and 19, but they both aim to estimate the same values. Which is the better option?

Q.3 Let $\mathbf{Z} = (Z_1, Z_2)$ be a bivariate lognormal D-norm generator. It generates the 2-dimensional Hüsler–Reiss D-norm with parameter $v = \mathrm{Var}(\log(Z_1) - \log(Z_2))$. Now on the one hand we have

$$v = -8 \cdot \log(\mathbb{E}(\sqrt{Z_1 Z_2})), \tag{3.25}$$

and on the other hand we have

$$v = \left(2 \cdot \Phi^{-1}\left(\frac{\mathbb{E}(\max(Z_1, Z_2))}{2}\right)\right)^2, \tag{3.26}$$

where Φ is the cumulative distribution function of the standard normal distribution. Equation (3.25) is a consequence of Equation (2.5) with $\lambda_1 = \lambda_2 = 1/2$, while Equation (3.26) is a consequence from Equation (1) in the work by Huser and Davison (2013).

Now we can estimate both the co-extremality $c = \mathbb{E}(\sqrt{Z_1 Z_2})$ and the extremal coefficient $e = \mathbb{E}(\max(Z_1, Z_2))$ and end up with two different ways to infer the parameter v:

$$\begin{aligned} \widehat{v}_c &:= -8 \cdot \log(\widehat{c}) \\ \widehat{v}_e &:= \left(2 \cdot \Phi^{-1}\left(\frac{\widehat{e}}{2}\right)\right)^2. \end{aligned} \tag{3.27}$$

The question is: Which of these options is better?

To investigate these questions we will draw samples from the two classes of bivariate distributions:

1. This class of distributions is parametrized by $v \geq 0$. We generate two independent random variables, U and Z_2, where U is uniformly distributed on $(0, 1)$ and $\log(Z_2)$ follows a normal distribution with variance v and mean $-v/2$. According to Corollary 3 the random vector $(X_1, X_2) = (1/U, Z_2/U)$ fulfills the equivalent statements of the Rosetta Stone theorem with a bivariate Hüsler–Reiss D-norm.

2. This class of distributions is parametrized by $\rho \in [0,1]$. We generate three independent random variables, U, V and B, where U und V are uniformly distributed on $(0,1)$ and B is a Bernoulli random variable with parameter ρ. If $B = 1$ we set $(X_1, X_2) = (1/U, 1/U)$ and if $B = 0$ we set $(X_1, X_2) = (1/U, 1/V)$. One can confirm that the tail-behavior of (X_1, X_2) is governed by the D-norm that is generated by (Z_1, Z_2) that has the following probabilites:

$$\begin{aligned} P(Z_1 = 1, Z_2 = 1) &= \rho, \\ P(Z_1 = 2, Z_2 = 0) &= (1-\rho)/2, \\ P(Z_1 = 0, Z_2 = 2) &= (1-\rho)/2. \end{aligned}$$

This model is known as the Marshall–Olkin model (one can check that the min-stable multivariate distributions by Marshall and Olkin (1967) can be 'flipped' to multivariate max-stable distribution with this underlying D-norm).

To investigate Q.1 we will pick different thresholds $n/k = 10, 100, 1000$ and different values of $k = 10, 100, 1000$. For each combination we will determine bias and variance of the estimator. As for the random vectors from which we will draw n samples: 3 of them will be in the max-domains of attraction of different Hüsler–Reiss models und another 3 will be in the max-domains of attraction of different Marshall–Olkin models. The parameters are chosen in a way such that $e := \mathbb{E}(\max(Z_1, Z_2)) \in \{1.25, 1.5, 1.75\}$, hence the 3 possbilities. As for the estimators themselves, one time we will use the multivariate peaks-over-threshold method, another time we will use the structure-variable estimator. The margins are estimated with ranks all the time.

Our results are visualized in the Figures 3.1 and 3.2. Every individual plot contains 9 data points. The parameter k can be read on the horizontal axis, while datapoints that have the same threshold n/k are connected with a line.

Note that for $e = 1.75$ the bias varies in n/k, but barely in k, but for the other cases the bias changes both in n/k and k in approximately the same order of magnitude. This is a resounding 'no' to Q.1 when it comes to bias.

Now for analyzing variance: Here we want to check if the variance of the indirect estimator is proportional to $1/k$' holds. Therefore we multiply the variances by k before plotting them on the vertical axis in the Figures 3.3 and 3.4.

The lines in the plots are not perfectly horizontal, but $k \cdot \mathrm{Var}(\widehat{H})$ remains remarkably stable within the setting of this simulation study. So we have a conditional 'yes' to Q.1, when it comes to variance.

Question Q.2 is about whether the structure-variable estimator or the peaks-over-threshold estimator is better. A simple comparison of the mean squared error $\mathrm{MSE}_{\mathrm{SV}}(n,k)$ with the mean squared error $\mathrm{MSE}_{\mathrm{PoT}}(n,k)$ for different values of n and k would be naive: The role of k in the structure-variable estimator is a different role than it has in the peaks-over-threshold estimator.

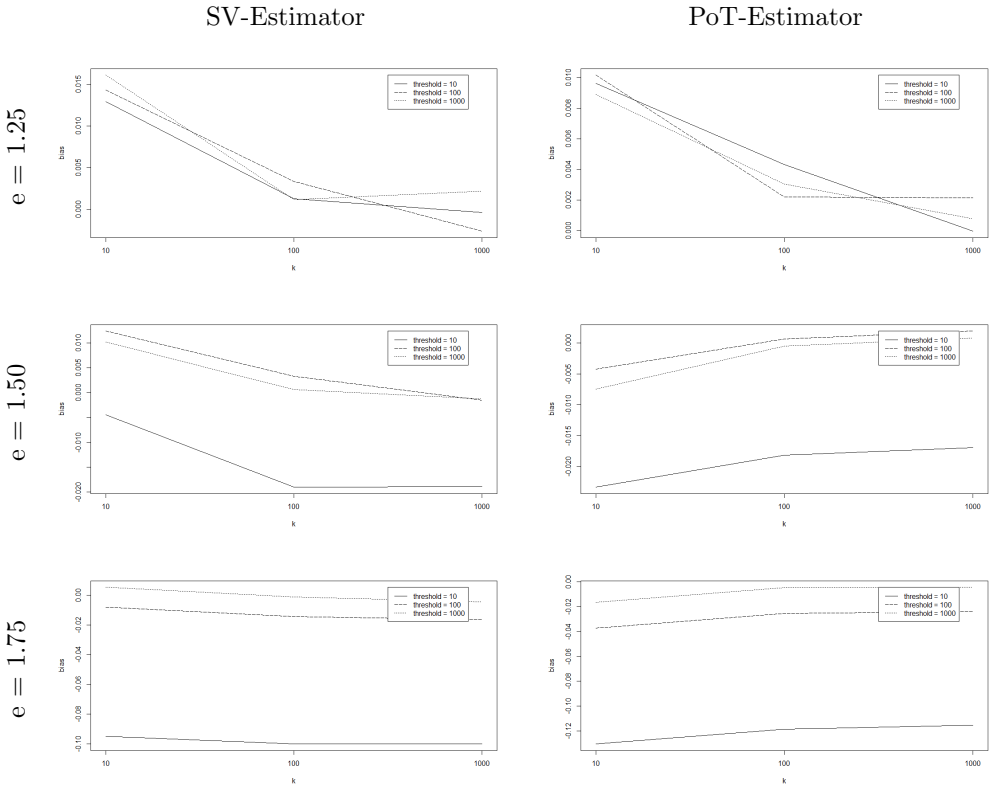


Figure 3.1: Bias (Hüsler–Reiss model, indirect estimator)

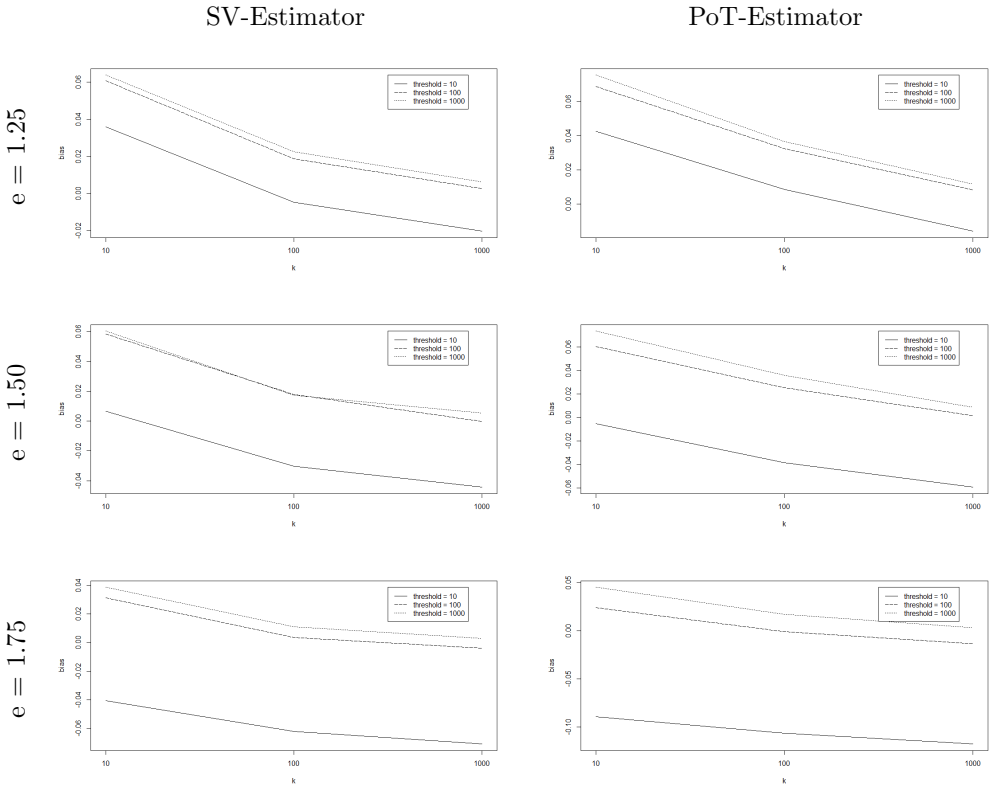


Figure 3.2: Biases, (Marshall–Olkin model, indirect estimator)

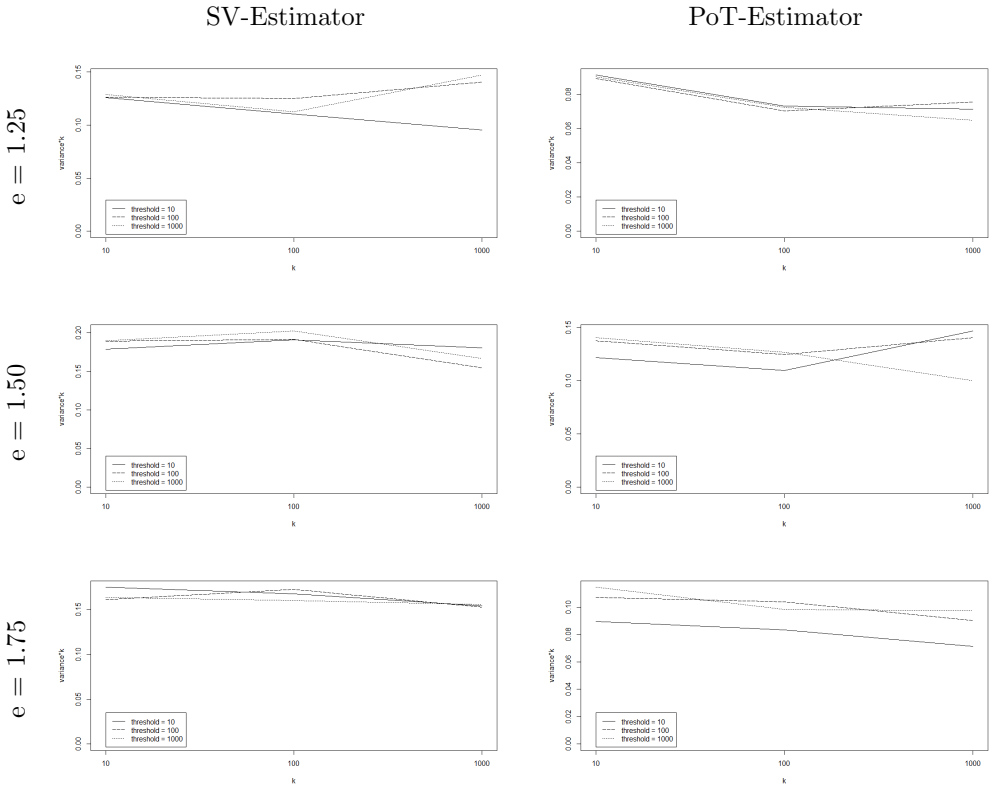


Figure 3.3: Variances, (Hüsler–Reiss model, indirect estimator)

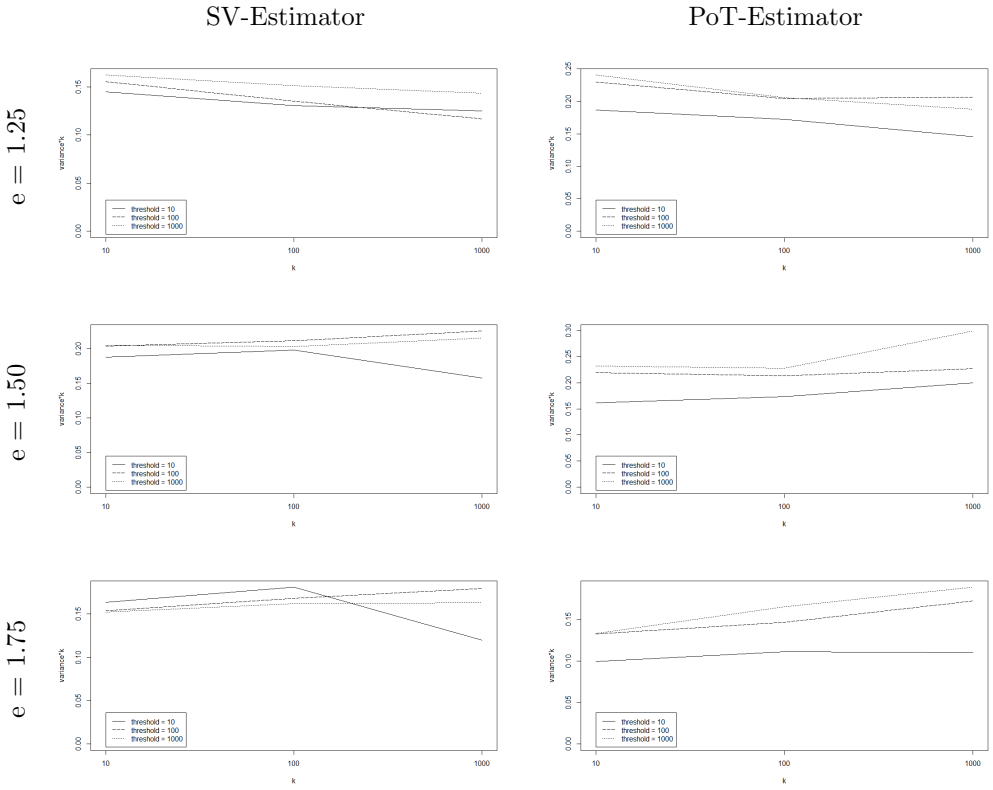


Figure 3.4: Variances, (Marshall–Olkin model, indirect estimator)

An alternative would be to compare $\min_k \mathrm{MSE}_{\mathrm{SV}}(n,k)$ to $\min_k \mathrm{MSE}_{\mathrm{PoT}}(n,k)$, but in practice our choice of k will always be suboptimal, so we also need comparisons between suboptimal choices of k in the structure-variable estimator versus suboptimal choices of k in the peaks-over-threshold estimators.

To do that we will plot the squared bias versus the variance of the estimators into the same plot for different values of k. We have done this in the Figures 3.5 and 3.6 with sample size is now fixed to $n = 10000$ and our models are the usual 6 models we have used in this simulation study, which are the Hüsler–Reiss model and the Marshall–Olkin model with $e \in \{1.25, 1.50, 1.75\}$.

Note that for the indirect estimators the structure-variable estimators are toe-to-toe with the peaks-over-threshold estimators. With a few exceptions both types of estimator follow the same bias-variance-tradeoff-curve.

But for the direct estimators we realize that if a the peaks-over-threshold estimator has the same variance as a structure-variable estimator, it beats the structure-variable estimator in bias. And if it has the same bias a structure-variable estimators, it beats the structure-variable estiamtor in variance. Keep in mind that the diagrams use the natural logarithm, the inverse of the exponential function, to scale the points. So a shift of ≈ 2.3 to the right or to the top in the diagram means multiplication with factor 10. We can check the diagrams again to see that switching from the structure-variable estimator to the peaks-over-threshold estimator improves bias or variance by orders of magnitute.

To approach question Q.3 we will generate samples from the bivariate random vector in the max-domain of the Hüsler–Reiss distribution and evaluate the estimators $\widehat{v}_e$ and $\widehat{v}_c$ from Equation (3.27) by plotting their squared biases versus their variances to visualize our results in Figure 3.7.

We can't say if the black symbols reliably beat their white counterparts or vice versa in these diagrams.

The following answers are by nature only valid within in the limits of the simulation study:

A.1 Regarding Q.1: The variance of the indirect estimator seems to be proportional to $1/k$. The idea that the bias only depends on n/k, but not on k does not carry over to indirect estimators.

A.2 Regarding Q.2: For indirect estimators with rank-based estimation of the margins switching from structure-variable estimation to peaks-over-thresholds is no improvement. For direct estimators it is - we can improve bias or variance by orders of magnitude without hurting the other.

It is possible that if we don't estimate the margins with ranks, but with more sophisticated methods that the properties of the indirect estimators become closer to the properties of the direct estimators and peaks-over-threshold is preferable then, too.

But if we only use ranks and only estimate the D-norm evaluated at one point, the most conservative choice is the structure-variable estimator. This estimator is better understood in the literature and there are results

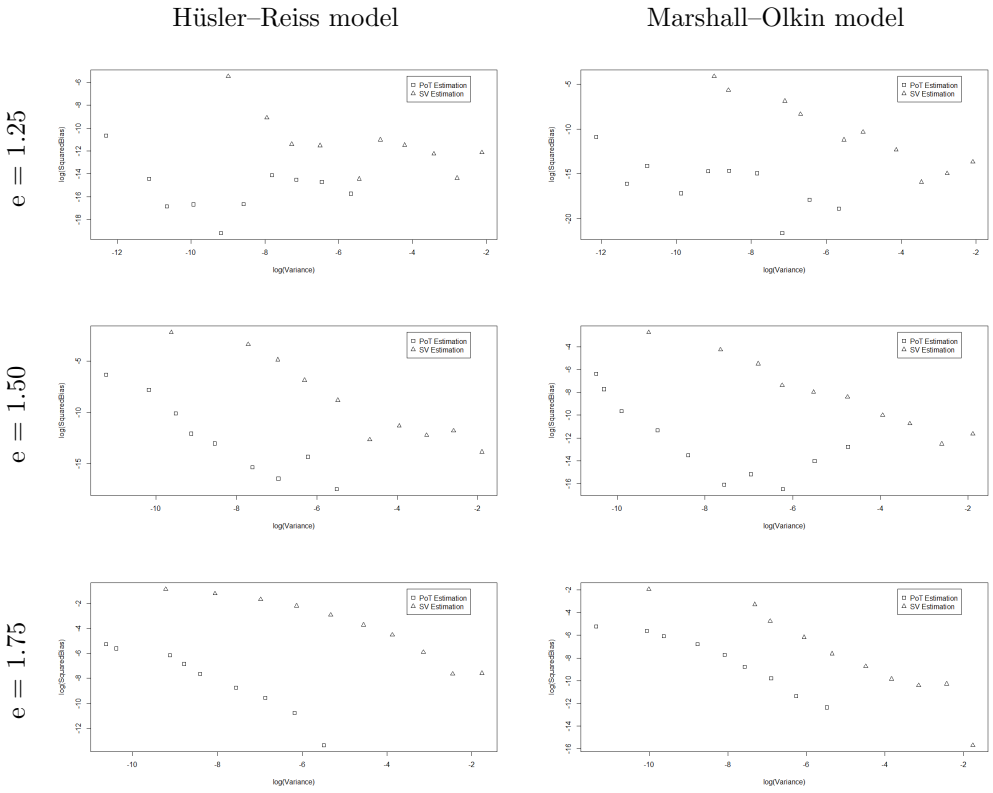


Figure 3.5: Bias versus Variance, Direct Estimator

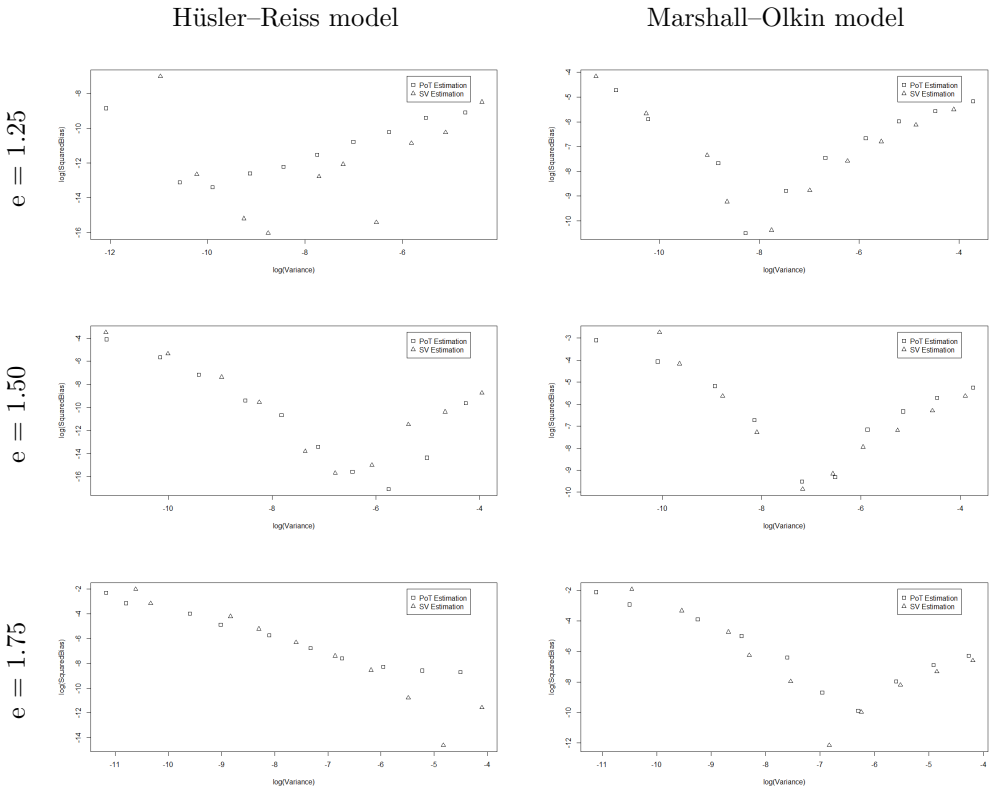


Figure 3.6: Bias versus Variance, Indirect Estimator

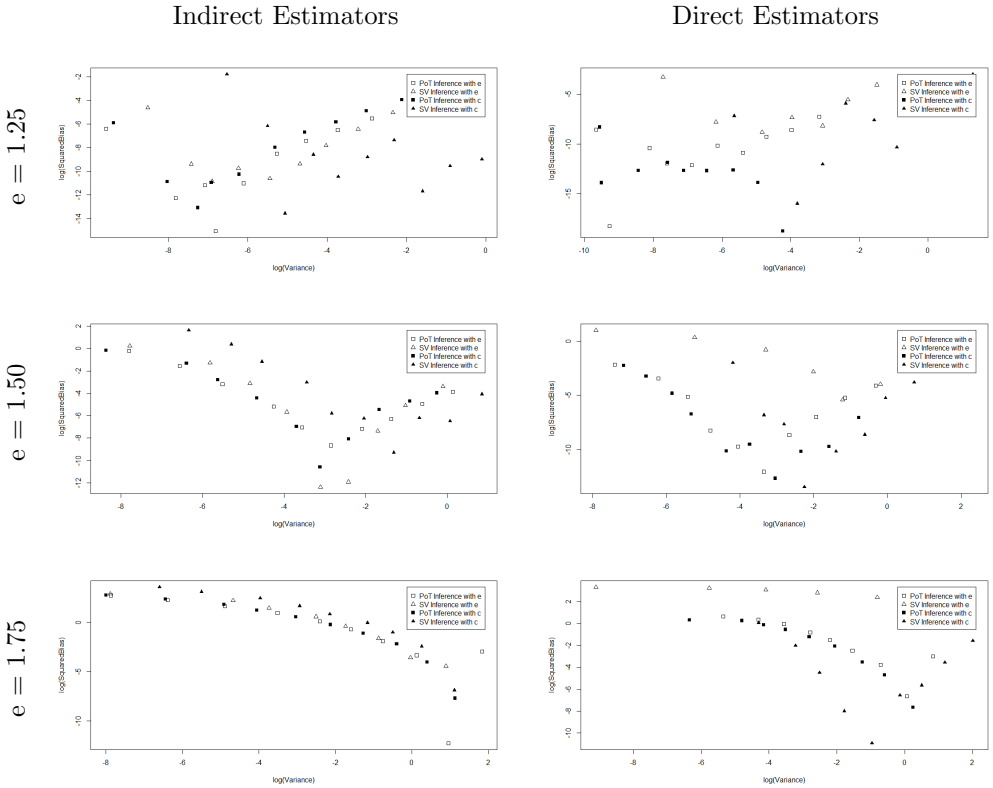


Figure 3.7: Inference of the Hüsler–Reiss parameter.

like asymptotic normality under certain second order conditions, see Einmahl et al. (2012) and Bücher et al. (2014). It should therefore be the default.

A.3 Regarding Q.3. In our study we could not find a consistent improvement by switching from $\widehat{v}_e$ to $\widehat{v}_c$. We should go with what is better understood in the literature (see the references in A.2) and that is estimating e and then transforming it into an estimator $\widehat{v}_e$ for the unknown variogram v.

Bibliography

Billingsley, P. (1968). *Convergence of Probability Measures* (1 ed.). Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Billingsley, P. (1979). *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Bücher, A. (2014). A note on nonparametric estimation of bivariate tail dependence. *Stat. & Risk Mod. 31*(2), 151–162.

Bücher, A. and J. Segers (2014). Extreme value copula estimation based on block maxima of a multivariate stationary time series. *Extremes 17*(3), 495–528.

Bücher, A., J. Segers, and S. Volgushev (2014). When uniform weak convergence fails: Empirical processes for dependence functions and residuals vias epi- and hypographs. *Ann. o. Stat. 42*(4), 1598–1634.

Bücher, A., S. Volgushev, and N. Zou (2018). On second order conditions in the multivariate block maxima and peak over threshold method. https://arxiv.org/abs/1808.10828.

Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics 9*, 383–418.

Cho, E. and M. J. Cho (2009). Variance of sample variance with replacement. *International Journal of Pure and Applied Mathematics 52*(1), 43–47.

Clémençon, S. and A. Dematteo (2016). On tail index estimation based on multivariate data. *J. of Nonparam. Statistics. 28*(1), 152–176.

Clémençon, S., N. Goix, and A. Sabourin (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis 161*, 12–31.

Cooley, D., J. Diebolt, A. Guillou, and P. Naveau (2009). Modelling pairwise dependence of maxima in space. *Biometrika 96*(1), 1–17.

Cooley, D. and E. Thibaud (2018). Decompositions of dependence for high-dimensional extremes. arXiv:1612.07190 [stat.ME].

Davis, R. A. and P. Wan (2019). Threshold selection for multivariate heavy-tailed data. *Extremes 22*(1), 131–166.

de Haan, L. and A. Ferreira (2006). *Extreme Value Theory: An Introduction.* Springer Series in Operations Research and Financial Engineering. New York: Springer. see `http://people.few.eur.nl/ldehaan/EVTbook.correction.pdf` and `http://home.isa.utl.pt/~anafh/corrections.pdf` for corrections and extensions.

Donelly, C. and P. Embrechts (2010). The devil is in the tails: Actuarial mathematics and the subprime mortgage crisis. *ASTIN Bulletin. 40*(1), 1–33.

Drees, H. and A. Sabourin (2019). Principal component analysis for multivariate extremes. arXiv:1906.11043 [math.ST].

Einmahl, J. H. J., A. Krajina, and J. Segers (2012). An M-estimator for tail dependence in aribtrary dimensions. *Ann. Statist. 40*(3), 1764–1793.

Einmahl, J. H. J., J. Li, and R. Y. Liu (2009). Thresholding events of extreme in simultaneous monitoring of multiple risks. *Journal of the American Statistical Association 104*(487), 982–992.

Falk, M. (2013). On idempotent D-norms. *Journal of Multivariate Analysis 139*, 283–294.

Falk, M. (2019). *Multivariate Extreme Value Theory and D-norms.* New York: Springer.

Fan, Y., J. Lee, and S. Sisson (2015). Bayesian threshold selection for extremal models using measures of surprise. *Computational Statistics and Data Analysis. 85*, 84–99.

Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics* (1 ed.). Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Heffernan, J. E. and J. A. Tawn (2003). An extreme value analysis for the investigation into the sinking of the m. v. derbyshire. *Appl. Statist. 52*(3), 337–354.

Huser, R. and A. C. Davison (2013). Composite likelihood estimation for the Brown-Resnick process. *Biometrika 100*(2), 511–518.

Jeon, S. and R. Smith (2012). Dependence structure of spatial extremes using threshold approach. arXiv:1209.6344v1 [stat.ME].

Jessen, A. H. and T. Mikosch (2006). Regularly varying functions. *Publications de l'Institut Mathematique, Nouvelle Serie 80*(94), 171–192.

Kabluchko, Z., M. Schlather, and L. de Haan (2009). Stationary max-stable fields associated to negative definite functions. *Ann. Probab. 37*(5), 2042–2065.

Kim, M. and S. Lee (2017). Estimation of the tail exponent of multivariate regular variation. *AISM 69*(5), 945–968.

Krupskii, P., H. Joe, D. Lee, and M. G. Genton (2018). Extreme-value limit of the convolution of exponential and multivariate normal distributions: Links to the Hüsler-Reiß distribution. *J. Multivariate Anal. 163*, 80–95.

Marshall, A. W. and I. Olkin (1967). A multivariate exponential distribution. *J.o.t.Am.Stat.Asso. 62*(317), 30–44.

Puntanen, S., G. P. H. Styan, and J. Isotalo (2013). *Formulas Useful for Linear Regression Analysis and Related Matrix Theory*. Springer Briefs in Statistics. Springer.

Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*, Volume 4 of *Applied Probability*. New York: Springer. First Printing.

Resnick, S. I. (2008). *Extreme Values, Regular Variation, and Point Processes*. Springer Series in Operations Research and Financial Engineering. New York: Springer.

Schlather, M. and J. A. Tawn (2002). Inequalities for the extremal coefficients of multivariate extreme value distributions. *Extremes 5*(1), 87–102.

Segers, J. (2012). Max-stable models for multivariate extremes. *REVSTAT 10*(1), 61–82.

Sippel, S., M. T. Black, A. J. Dittus, L. Harrington, N. Schaller, and F. E. Otto (2015). Combining large model ensembles with extreme value statistics to improve attribution statements of rare events. *Weather and Climate Extremes 9*, 25–35.

Acknowledgements

Ich möchte meinem Betreuer Prof. Dr. Michael Falk danken - für die Chance, das hier umzusetzen, und für eine Balance aus Freiräumen und Förderung.

Ich möchte meiner Familie in Deutschland und in Österreich danken - die Unterstützung war klasse und die Urlaube erholsam.

Außerdem möchte meinen Kollegen am Lehrstuhl danken, insbesondere Karin Krumpholz und Silke Korbl - dafür, dass ich bei bürokratischen Sachen nie allein gelassen wurde.

Ich danke ebenfalls Gerhard Osius und dem Deutschen Wetterdienst.

I also thank Daniel Cooley and an anonymous reviewer that rejected a paper of mine with a very elaborate reply. It showed me that someone had to develop the theory in Chapter 3 and that someone turned out to be me.