Multivariate threshold exceedances and extremal graphical models in the presence of multiple extreme directions

Anas Mourahib
Louvain Institute of Data Analysis and Modeling in economics and statistics (LIDAM)
Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), UCLouvain

Jury:
Prof. Segers Johan (Supervisor, UCLouvain, KU Leuven)
Prof. Kiriliouk Anna (Supervisor, University of Namur)
Prof. Denuit Michel (President, UCLouvain)
Prof. Pircalabelu Eugen (Secretary, UCLouvain)
Prof. Engelke Sebastian (External member, University of Geneva)
Prof. Röttger Frank (External member, Eindhoven University of Technology)

August 5, 2025

Better, Stronger, Faster

Personne 1 : On peut s'attaquer à un homme, on peut l'arrêter, le jeter au fond d'un cachot... Mais comment poignarder une idée — surtout une idée nouvelle ?

Personne 2 : Il est le dernier de sa race. Il est courageux, il n'a peur de rien. Il ne connaît pas la pitié... et il est cruel.

Personne 1 : Comment il s'appelle ?

Personne 2 : Acronas. Oui, c'est bien ça... Acronas. Acornas... Acornas.

Silence.

Personne 1 : Qu'est-ce que t'on dit ?

Personne 2 : Échecs et mat.

Freely adapted from a scene in the film Moby Dick (1956), directed by John Huston.

Acknowledgments

I want to thank some people. I start first with my supervisors, Anna and Johan. Thank you both for the support. Anna, I'm always impressed by how efficient you are at work and how your ideas are different from the classic ones. I guess I'm your first PhD student, and I'm sure you will have many. Johan, I'm impressed and jealous of how much you are involved in all the details of your research. I guess there's one reason, which is the fact that you love what you do. I have learned many things from both of you. Big thanks!

I would like to sincerely thank the jury members for their time and valuable feedback on my thesis. My gratitude goes to Professors Michel Denuit, Eugen Pircalabelu, Sebastian Engelke, and Frank Röttger.

A special thanks to Sebastian for welcoming me for a research visit, and to the colleagues in Geneva — Manuel, Gloria, Olivier, Zhongwei, and Flore — for the insightful discussions and warm atmosphere during my stay.

Next, I want to thank some of my friends. Shout out to the special one, Achraf, my man. You have always thought about the most notorious plans of life. Big respect, bro, and stay like you are. Shout out to the one I have known the longest, Manal. We have together spent the best years at Ibn Tofail University. Good luck to you for your PhD. Shout out to the one and only one, Ayoub. Thank you, bro! I also had the chance to meet many people in the institute. Ensiyeh, you are a wonderful person with a big heart. Stay like you are and thanks a lot. Aigerim, you don't smile a lot, but when you do, you do it from your heart. Thanks. Mon fils, Keivain, je t'aime, mais quand tu dis que tu veux faire du rap avec ton talent, j'imagine quelqu'un qui veut faire "facial Van Damm" avec un jean en pantalon. Shout out to the GOAT, Haotian. Congrats on your new position, man. Shuang, Vanessa, Jeongjin — thanks a lot for the wonderful times. Stéphane, it was a pleasure to start and finish the PhD in the same period as you. I'm sure you will have a wonderful career. Good luck, man! Meng, thanks a lot for the support. I had the pleasure to share the office with you, and you helped me a lot. Thanks, my friend! Shout out to Kamal. Merci fréro pour le support. À un certain moment, je ne savais plus comment attraper le guidon et tu m'as montré comment. Tu m'as vraiment beaucoup inspiré, l'ami! Much love! Finalement, merci Sibylle. On s'est rencontrés récemment, mais je me sens si bien avec toi.

Much love to my two best friends, Mohammed and Valentina. Mohammed, I met you three years ago, more or less. Valentina was not here yet in Belgium. Before she came, I was afraid that I would be kicked out. But then Valentina came, and I had the chance to be the third wheel. Thank you both, guys, for the support. You guys took care of me here. I love you!

I want to take some time to present Acronas. This is my rap, a.k.a. I've written some lyrics that you can find online, and I hope one day I can work more on this

passion.

Just like Walter White said in Breaking Bad: La familia es todo. Thanks to the ones I love the most. Thanks to the best brother, Ayoub. You have taught me a lot, personally and also in Maths and Science. Good luck to you on your new life, bro. My father, Abdelkbir — I have learned hard work from you. Big respect. I love you. To the two who love me better than they love themselves, my mother Assia and my grandmother Zineb. Zineb, I love you — thanks for the support. Assia, I love you the most. I hope I can make you proud!

To my future me: Things you own end up owning you. Stay indifferent.

Abstract

Consider a random vector representing risk factors, and suppose that we are interested in extreme scenarios. The threshold exceedances method is widely used in extreme value theory. It uses the fact that, asymptotically, exceedances over a high threshold can be modeled using a multivariate generalized Pareto distribution. In the literature, statistical practice of this method has been discussed only when all risks are large simultaneously. This condition is not realistic, for example, when some of the risks are nearly independent. To address this limitation, we first develop a parametric model that accommodates cases where only some risk factors are extreme, while others remain moderate. We introduce the concept of extreme directions to describe these cases, and prove that our construction encompasses the full range of possible max-stable dependence structures. Second, we propose an estimation procedure for this model, leading to an algorithm capable of identifying extreme directions. Third, we extend the framework of extremal graphical models by incorporating both multiple extreme directions and extremal independence, offering a more flexible and realistic approach to modeling extreme events. Finally, we apply our developed methods to two real-world datasets: first, to discharge measurements at stations along the Danube River, and second, to financial portfolio losses from stocks listed on the NYSE, AMEX, and NASDAQ.

Contents

Chapter 1

Introduction

To smoothly explain the idea of the thesis, we consider in this chapter a daily river discharge dataset for different locations from the Danube river; see Figure 1.1. We want to study the extremal dependence of river discharge across these stations. To do so, multivariate extreme value theory provides two main methods: the block maxima approach and the threshold exceedances approach (also known as the peaks-over-threshold approach). In the following, let $\boldsymbol{X} = (X_1, \ldots, X_d) \in (0, \infty)^d$ be a d-variate random vector where each entry can be seen as the daily river discharge at one station. Suppose that we observe n daily observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, copies of $\boldsymbol{X}$, and suppose that, using some declustering techniques, these observations are independent and identically distributed.

Univariate extreme value theory is already a well-developed field; see for instance de Haan and Ferreira [24], Beirlant et al. [4]. Therefore, typically, in multivariate extreme value theory, provided that each marginal distribution F_j of X_j (for $j = 1, \ldots, d$) is continuous, each marginal distribution is estimated first, then standardized to a unit-Pareto distribution via $1/(1 - F_j(X_j))$.

For simplicity, we assume throughout that the margins of $\boldsymbol{X}$ have been transformed accordingly. In practice, during data analysis, this transformation must be based on the estimated marginal distributions.

1.1 Block maxima approach

In this approach, asymptotically, standardized componentwise maxima of $\boldsymbol{X}$ are modeled using a distribution G; see Figure 1.2. More precisely, we assume the existence of a distribution G with non-degenerate margins such that

$$\Pr\Big(\max_{i=1,\ldots,n} X_{i1} \leqslant nx_1, \ldots, \max_{i=1,\ldots,n} X_{id} \leqslant nx_d\Big) \;\to\; G(x_1, \ldots, x_d), \quad n \to \infty, \tag{1.1.1}$$

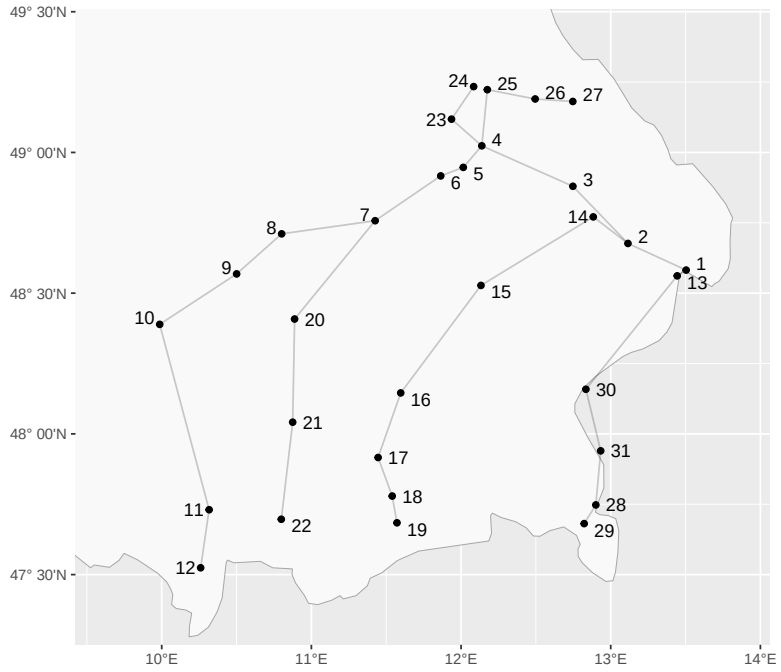


Figure 1.1
River flow data from the Danube river dataset available in the `R` package `graphicalExtremes`; see Engelke et al. [40]. Dataset contains $d = 31$ stations.

for every continuity point $(x_1, \ldots, x_d) \in (0, \infty)^d$ of G. The distribution G is called a *multivariate extreme value* distribution, or *max-stable* distribution [4, 18, 24].

In practice. As its name indicates, observations are first divided into blocks (e.g., yearly, monthly, etc.), depending on the application and then a max-stable distribution is fitted to the componentwise maxima over these blocks.

As statisticians, we do not wish to impose model assumptions that are verified only by a "niche" class of distributions. Fortunately, Hypothesis (1.1.1) holds for a wide variety of distributions; see, for instance, Beirlant et al. [4, Chapter 2] for $d = 1$ and Segers [87] for $d \geqslant 2$.

For $d = 1$, the family of max-stable distributions is parametrized by a finite-dimensional family, unlike the case $d \geqslant 2$, where the dependence between components makes the model infinite-dimensional. Hence the central question: "How can we describe this dependence structure?" One answer, inspired by probability theory, is via copulas; see Jaworski et al. [61, Chapter 6]. In addition, extreme value theory introduces other tools—most notably the *exponent measure* and the *angular measure*—which we will present in this introduction.

Notation. In the following, bold symbols will refer to multivariate quantities. Let $\mathbf{0}_d$ be the d-variate vector with 0 components. Write $\mathbb{E} = [\mathbf{0}_d, \boldsymbol{\infty})$. Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and $\mathbb{S}$ denote the positive unit sphere associated to this norm, that is, $\mathbb{S} = \{\boldsymbol{x} \in \mathbb{E} : \|\boldsymbol{x}\| = 1\}$.

The *exponent measure* of G is the unique Borel measure μ concentrated on $\mathbb{E}$ such that

$$G(\boldsymbol{x}) = \exp\{-\mu(\mathbb{E} \setminus [\mathbf{0}, \boldsymbol{x}])\}, \qquad \boldsymbol{x} \in \mathbb{E}. \tag{1.1.2}$$

The exponent measure μ describes the extremal behavior of the risk factors $\boldsymbol{X}$. In fact, for a Borel set $B \subset \mathbb{E}$ verifying certain topological conditions and regularity conditions with respect to μ (see, e.g., Section 2.2.1), we have

$$\lim_{n\to\infty} n\,\mathsf{P}(\boldsymbol{X} \in nB) = \mu(B). \tag{1.1.3}$$

The *angular measure* ϕ associated to G with respect to the norm $\|\cdot\|$ is

$$\phi(B) = \mu\Big\{\boldsymbol{x} \in \mathbb{E} : \|\boldsymbol{x}\| > 1,\ \boldsymbol{x}/\|\boldsymbol{x}\| \in B\Big\}, \qquad B \in \mathcal{B}(\mathbb{S}). \tag{1.1.4}$$

The angular measure describes the extremal behavior of the radial part associated to the risk factors $\boldsymbol{X}$. In fact, for a Borel set $B \subset \mathbb{S}$ verifying certain regularity conditions with respect to ϕ, Equations (1.1.3) and (1.1.4) yield

$$\lim_{n\to\infty} n\,\mathsf{P}\Big(\|\boldsymbol{X}\| > n,\ \boldsymbol{X}/\|\boldsymbol{X}\| \in B\Big) = \phi(B).$$

1.2 Threshold exceedances approach

In this approach, observations of $\boldsymbol{X}$ for which a high river discharge occurs at least at one station are modeled using a non-degenerate distribution H. More precisely, we assume the existence of a distribution H such that, for every $\boldsymbol{y} > \mathbf{0}_d$,

$$\mathsf{P}\left(\boldsymbol{X}/u \leqslant \boldsymbol{y} \mid \max(\boldsymbol{X}) > u\right) \to H(\boldsymbol{y}), \quad u \to \infty. \tag{1.2.1}$$

The distribution H is called a multivariate Pareto (mp) distribution. Note that in the literature [83], a small transformation of (1.2.1) results in another distribution, $\tilde{H}$, called a multivariate **generalized** Pareto (mgp) distribution. The choice between mp and mgp distributions depends on the context and on personal preference. In Chapter 2, we use mgp distributions, while in Chapter 4, following the literature on extremal graphical models, we use mp distributions. We can always express one distribution in terms of the other via a simple transformation.

Similarly to max-stable distributions, for $d \geqslant 2$, mp distributions cannot be parameterized by a finite-dimensional space, and the dependence structure between

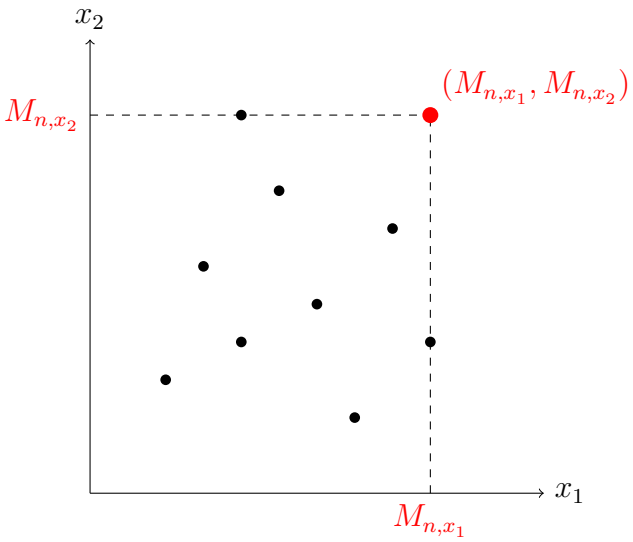


Figure 1.2
Bivariate block maxima approach applied to a single block. The red point represents the componentwise maxima, i.e., $M_{n,x_j} = \max(X_{1j}, \ldots, X_{nj})$, for $j = 1, 2$.

the components of H can be described by the exponent measure in (1.1.2) and the angular measure in (1.1.4). The two families—max-stable distributions and mgp distributions—are related but not in a one-to-one correspondence. In fact, two different max-stable distributions can be associated with the same mgp distribution [84, Section 3]. The mp distribution describes only the subset of points—those that are extreme in at least one margin. This is observed explicitly in the univariate case, where an mp distribution is parameterized by two parameters, while a max-stable distribution has three parameters.

1.3 Extreme directions

Consider now the two stations 1 and 12 from the Danube river dataset. These two stations are relatively far apart spatially; see Figure 1.1. Therefore, it is possible that a high river discharge occurs at one station without the other. This phenomenon is illustrated in Figure 1.4, where for many observations we see that high river discharge occurs only at station 1.

More generally, in dimension d, we will call an *extreme direction* a subset $J \subset \{1, \ldots, d\}$ of stations where high river discharge occurs at each of these stations without occurring at the remaining $\{1, \ldots, d\} \setminus J$ stations.

The concept of extreme directions can also be interesting in other fields. In finance, given a set of stocks, we might be interested in clusters of stocks that

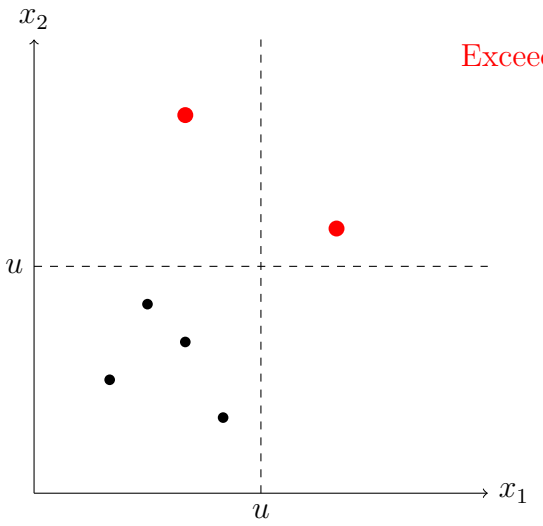


Figure 1.3
Bivariate threshold exceedances approach. The two red points exceed the threshold u.

exhibit high price changes, excluding the rest of the stocks. In climate science, we may look for groups of countries where heatwaves co-occur, excluding the rest of the countries.

Throughout, suppose that the random vector $\boldsymbol{X}$ satisfies (1.1.1) with exponent measure μ as in (1.1.4). Simpson et al. [90] proposes the following definition for the concept of extreme directions.

Definition 1.3.1. The set J is an extreme direction of $\boldsymbol{X}$ or μ if

$$\mu(\left\{\boldsymbol{x} \in \mathbb{E} : \, x_j > 0 \text{ iff } j \in J\right\}) > 0.$$

Other terms are used in the literature to refer to a group of variables (stations in our example) that can be large simultaneously while the remaining ones are not. Fomichov and Ivanovs [45] opt for "concomitant extremes", Simpson et al. [90] choose to refer to this as "cones". Our choice of the term "extreme direction" is motivated by the following geometrical intuition: in a bivariate setting where each variable is associated with an axis, as in Figure 1.2, if $\{1\}$ is an extreme direction, then following the axis x_1 (or the direction 1) leads to something extreme. Similarly for the other two potential extreme directions, $\{1, 2\}$ and $\{2\}$.

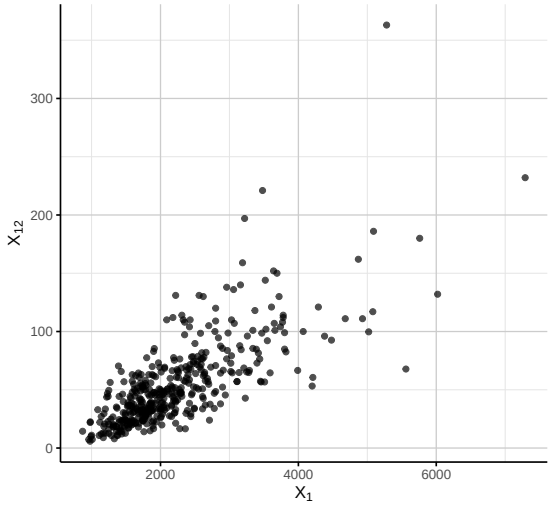


Figure 1.4
Pairwise scatter plots between the stations 1 and 12.

1.4 Extremal conditional independence and threshold exceedance approach

As in classic statistics, one challenge in modeling multivariate extremal dependence is the dimension d. As the number d of stations becomes larger, inference becomes more complicated given the limited number of observations in the tail. One solution is to look for sparsity by studying the concept of extremal conditional independence.

Studying conditional independence in the context of extreme value theory is not straightforward. Given stations X_1, X_2, X_3, we first need to define "X_1 and X_3 exhibit extremal independence given X_2". A straightforward first idea is to use max-stable distributions. However, it turns out that this is not so useful. In fact, under certain conditions, Papastathopoulos and Strokorb [79] show that if (Z_1, Z_2, Z_3) is a max-stable random vector, then $Z_1 \perp\!\!\!\perp Z_3 \mid Z_2$ reduces simply to $Z_1 \perp\!\!\!\perp Z_3$.

Using the multivariate Pareto random vector $(Y_1, Y_2, Y_3) \sim H$ in (1.2.1) to define extremal conditional independence is also challenging. In fact, the support of $\boldsymbol{Y}$, given by $\mathcal{L} = \left\{\boldsymbol{x} \in [0,\infty)^3 : \max(\boldsymbol{x}) > 1\right\}$ [83], is not a product space, and therefore defining conditional independence of $\boldsymbol{Y}$ is impossible.

Engelke and Hitz [34] propose to define extremal conditional independence using the conditioned multivariate Pareto random vector $\boldsymbol{Y}^{(k)} = \boldsymbol{Y} \mid Y_k > 1$, with support $\mathcal{L}^{(k)} = \{\boldsymbol{x} \in \mathcal{L} : x_k > 1\}$, for $k = 1, 2, 3$. We say that Y_1 and Y_3 exhibit

conditional independence given Y_2, and write $Y_1 \perp\!\!\!\perp_e Y_3 \mid Y_2$ if

$$Y_1^{(k)} \perp\!\!\!\perp Y_3^{(k)} \mid Y_2^{(k)}, \qquad \forall k \in \{1,2,3\}\,. \tag{1.4.1}$$

Note that a similar definition is used to define $\boldsymbol{Y}_A \perp\!\!\!\perp_e \boldsymbol{Y}_C \mid \boldsymbol{Y}_B$ for disjoint sets $A, B, C \subset V$; see Engelke and Hitz [34, Definition 1]. Thanks to (1.2.1), conditional independence (1.4.1) of $\boldsymbol{Y}$ can be interpreted as extremal conditional independence of $\boldsymbol{X}$. Once this is defined, conditional extremal independencies are encoded using a graph $\mathcal{G} = (\{1,\dots,d\}, E)$ [68]. A d-variate Pareto random vector $\boldsymbol{Y}$ is then said to be an *extremal graphical model with respect to the graph* $\mathcal{G}$ if

$$Y_i \perp\!\!\!\perp_e Y_j \mid \boldsymbol{Y}_{\{1,\dots,d\}\setminus\{i,j\}}, \qquad (i,j) \notin E.$$

1.5 Limitations of statistical practice of threshold exceedances approach

The threshold exceedances approach is well studied in the literature for both the univariate and multivariate cases. Many parametric models have been proposed in the literature, for instance the Hüsler–Reiss model, the logistic model, the extremal Hüsler–Reiss graphical model; see Smith et al. [93], Huser and Davison [57], Engelke and Hitz [34]. One drawback of these models is that they suppose only the extreme direction $\{1,\dots,d\}$, meaning they only cover the case where all components of the random vector $\boldsymbol{X}$ are extreme simultaneously. Figure 1.5 shows simulations from a max-stable Hüsler–Reiss model (left) and a logistic model (right). Although we choose to plot a log-transformation of the simulation (for better visualization), we can still see that observations are close to the diagonal line $x = y$, which suggests a unique extreme direction $\{1,2\}$. Such a restriction on the structure of extreme directions is not realistic, for example, when some components exhibit some sort of extremal independence—stations 1 and 12 in Figure 1.4.

The max-linear model [46, 32] and the asymmetric logistic model [96] cover the case of other extreme directions. However, the first model is discrete in the sense that it does not have a density, and thus makes statistical inference complicated. The second one is too specific in the sense that it cannot represent a large variety of max-stable distributions.

1.6 Outline of the thesis

The main contribution of this thesis concerns statistical practice of the threshold-exceedances approach in the presence of extreme directions different from $\{1,\dots,d\}$. Such extreme directions will be called non-standard extreme directions.

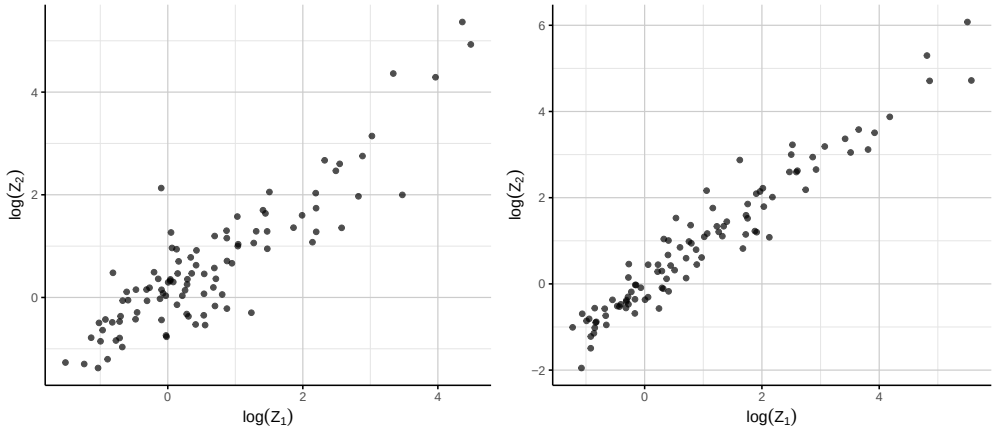


Figure 1.5
Sample of size $n = 100$ from a max-stable Hüsler–Reiss model (left) and logistic model (right).

Chapter 2. The first paper, co-authored with Anna Kiriliouk and Johan Segers (published in *Extremes*, 2024), proposes a method for constructing multivariate generalized Pareto distributions, or equivalently, max-stable distributions, with non-standard extreme directions. This construction yields a family of parametric models called mixture models; for instance, the mixture Hüsler–Reiss model and the mixture logistic model, which can be seen as the classic Hüsler–Reiss and logistic models, respectively, generalized to capture non-standard extreme directions. Figure 1.6 shows a simulation from a max-stable distribution associated with this model. This comes after we characterize extreme directions in Definition 1.3.1 in terms of the angular measure, the exponent measure, and the associated multivariate generalized Pareto distribution. We then show the generality of our model in the sense that any multivariate generalized Pareto distribution, or equivalently any max-stable distribution, can be produced by our construction (see Theorem 2.4.8). Finally, we propose an algorithm to simulate from our model and implement it in `R` (see `https://github.com/AnasMourahib/MGPD-simulation`).

Chapter 3. The second paper, co-authored with Anna Kiriliouk and Johan Segers, is available as a preprint on arXiv at `https://arxiv.org/pdf/2506.15272` and has been submitted to *Technometrics*. This paper focuses on the question of estimation of the mixture model and, implicitly, identification of the extreme directions. It turns out that this is not straightforward: by construction, setting some parameters of the mixture model to values on the border of the parameter space is the key to obtaining non-standard extreme directions. We are facing the

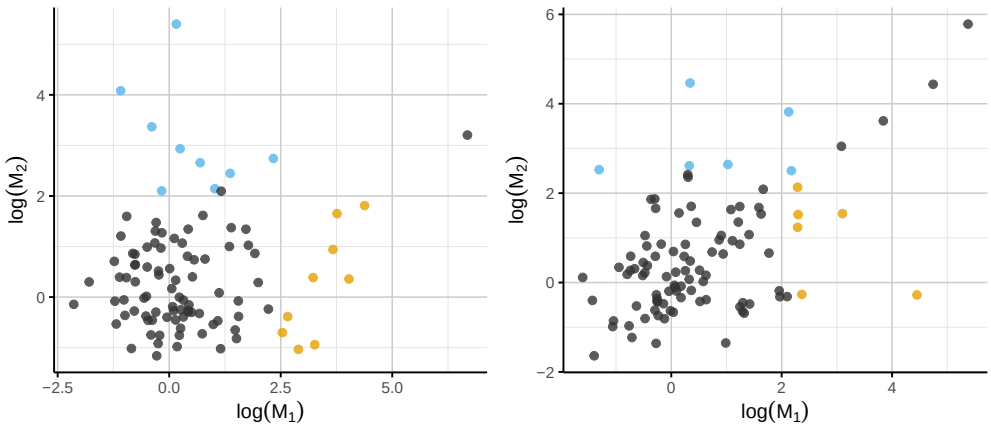


Figure 1.6
Log-transformed sample of size $n = 100$ from a max-stable mixture Hüsler–Reiss model (left) and mixture logistic model (right). Yellow points correspond to the extreme direction $\{1\}$, blue points to $\{2\}$, and gray points to $\{1, 2\}$.

classic difficulty of parametric estimation when the true parameters are on the border of the parametric space [97].

We introduce a penalized estimator based on the least-squares estimator proposed in Einmahl et al. [33]. Subsequently, we develop a data-driven algorithm to identify the extreme directions present in a given sample and compare it with the DAMEX algorithm [49]. As our estimator involves tuning parameters, we demonstrate that their selection effectively reduces to the choice of a single penalization parameter, which we determine via cross-validation. Finally, we apply our method to two real-world datasets: first, to discharge measurements at stations along the Danube River, and second, to financial portfolio losses from stocks listed on the NYSE, AMEX, and NASDAQ. In both applications, we identify the sets of variables that can become large simultaneously.

Chapter 4. The third paper, co-authored with Sebastian Engelke and Johan Segers (in progress), lies in the framework of extremal graphical models. As we pointed out in Section 1.4, many parametric extremal graphical models exist in the literature, for instance, the chain-logistic model Smith et al. [92], Smith [91], Coles and Tawn [17], and the Hüsler–Reiss graphical model Engelke and Hitz [34]. However, most of these models suppose the existence of a unique extreme direction $\{1, \ldots, d\}$ and extremal dependence between all the margins. To our knowledge, the only extremal graphical model that covers multiple extreme directions is the one constructed in Engelke et al. [38, Example 8.8] where the underlying graph

structure is a tree.

The goal of this paper is to give a construction method for parametric extremal graphical models that, on the one hand, covers non-standard extreme directions and, on the other hand, allows for extremal independence[1] between some groups of variables. We first state our main result, which provides a construction of extremal graphical models with non-standard extreme directions and possible extremal independence among groups of variables. Next, we apply this construction to define the *mixture Hüsler–Reiss graphical model* with respect to a graph and detail its parameterization; in the special case where the graph is a forest, inference reduces to bivariate procedures. We propose a data-driven algorithm to learn both the extreme directions and the forest dependence structure. We present a simulation study evaluating the algorithm, demonstrating good finite-sample performance under appropriate choices of tuning parameters. Finally, in this chapter, we apply our algorithm to two real-world datasets. The first concerns discharge measurements at stations along the Danube River, while the second involves financial portfolio losses from stocks listed on the NYSE, AMEX, and NASDAQ. In both applications, we focus on modeling the extremal dependence among the associated variables.

1.7 Code and data application

In this thesis, most of the implementations are done in *C* or *R*. See the GitHub repositories:

- `https://github.com/AnasMourahib/Penalized_least-squares_estimator`
- `https://github.com/AnasMourahib/MGPD-simulation`

In this thesis, we use two datasets to apply the statistical methods constructed in Chapters 3 and 4:

- **River discharge data.** We use daily river discharge data from the summer months between 1960 and 2010 at 31 stations along the Danube River; see Figure 1.1. A declustering technique from Asadi et al. [1] is used to remove temporal dependence, yielding $n = 428$ observations. The data is available in the *R* package *graphicalExtremes*.
- **Industry portfolios data.** We use the value-weighted returns for $d = 10$ industry portfolios, downloaded from the Kenneth French Data Library[2]:

[1] Two components j_1 and j_2 are considered *asymptotically independent* if their upper tail dependence coefficient $\chi_{j_1 j_2}$ equals zero. For further details on the coefficient χ, see Remark 2.3.8.

[2] See `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_10_ind_port.html` for more details about the data and the definition of each portfolio.

"1 = Nondurables", "2 = Durables", "3 = Manufacturing", "4 = Energy", "5 = HiTech", "6 = Telecom", "7 = Shops", "8 = Health", "9 = Utilities", "10 = Other". The individual stocks that make up the industry portfolios are taken from all listed firms on the *NYSE*, *AMEX*, and *NASDAQ*. The data is available between July 1926 and March 2025, yielding a sample size of $n = 51\,922$.

Chapter 2

Multivariate generalized Pareto distributions along extreme directions

Anas Mourahib[1] Anna Kiriliouk[2] Johan Segers[3]

Contents

Abstract

When modeling a vector of risk variables, extreme scenarios are often of special interest. The peaks-over-thresholds method hinges on the notion that, asymptotically, the excesses over a vector of high thresholds follow a multivariate generalized

[1]UCLouvain, LIDAM/ISBA, anas.mourahib@uclouvain.be.
[2]UNamur, NaXys, anna.kiriliouk@unamur.be.
[3]KU Leuven, Department of Mathematics, jjjsegers@kuleuven.be.

Pareto distribution. However, existing literature has primarily concentrated on the setting when all risk variables are always large simultaneously. In reality, this assumption is often not met, especially in high dimensions.

In response to this limitation, we study scenarios where distinct groups of risk variables may exhibit joint extremes while others do not. These discernible groups are derived from the angular measure inherent in the corresponding max-stable distribution, whence the term *extreme direction*. We explore such extreme directions within the framework of multivariate generalized Pareto distributions, with a focus on their probability density functions in relation to an appropriate dominating measure.

Furthermore, we provide a stochastic construction that allows any prespecified set of risk groups to constitute the distribution's extreme directions. This construction takes the form of a smoothed max-linear model and accommodates the full spectrum of conceivable max-stable dependence structures. Additionally, we introduce a generic simulation algorithm tailored for multivariate generalized Pareto distributions, offering specific implementations for extensions of the logistic and Hüsler–Reiss families capable of carrying arbitrary extreme directions. *Note.*

The content of this chapter is based on the paper:

Anas Mourahib, Anna Kiriliouk, and Johan Segers (2024). *Multivariate generalized Pareto distributions along extreme directions. Extremes*, 1–34.

2.1 Introduction

The statistical description of extremes is a long-standing subject of research, with the ultimate goal of extrapolating beyond the range of the data and estimating probabilities of events more extreme than those observed in the past. The foundations of univariate extreme value theory were established many decades ago through the derivation of the asymptotic distributions of suitably normalized block maxima and excesses over high thresholds. Modeling multivariate extreme events is a bigger challenge because, contrary to the univariate setting, the class of multivariate extreme-value distributions is infinite-dimensional. Over the last thirty years, numerous parametric extreme-value models have been proposed with the goal of capturing the tail dependence structure of a random d-dimensional vector $\boldsymbol{X} = (X_1, \ldots, X_d)$; see, e.g., Beirlant et al. [4, Chapter 9] for an early overview or Cooley et al. [20], Ballani and Schlather [3] and Opitz [77] for more recent examples. However, many such models are limited to the setting where all components of $\boldsymbol{X}$ can become large simultaneously. While this may be sufficient when the dimension d of $\boldsymbol{X}$ is low, more often than not, one encounters the setting where some variables may be large while at the same time, the others are of smaller order. In finance, for

example, characterizing groups of stocks within a portfolio that have the potential of suffering major losses simultaneously is of great importance. Managing flood risk requires identifying and modeling groups of gauging stations where extreme water discharges may occur simultaneously.

Groups of variables that may become large without the others being so—in an asymptotic sense to be made precise below—will be called *extreme directions*. More specifically, if $J \subseteq \{1, \ldots, d\}$ is such that the components $(X_j, j \in J)$ can be large simultaneously while the other components $(X_j, j \in \{1, \ldots, d\} \setminus J)$ are small, then J defines an extreme direction. The terminology is in line with a characterization of the support of the angular measure of the multivariate extreme-value distribution to which $\boldsymbol{X}$ is attracted. In the literature, such groups are also referred to as extreme clusters [14], extreme features [59], or concomitant extremes [45]. Many parametric extreme-value families are limited to the single extreme direction $J = \{1, \ldots, d\}$. The overarching goal of this paper is to explore the case when other extreme directions than $\{1, \ldots, d\}$ exist, and this in terms of several well-known objects in extreme-value theory, and multivariate generalized Pareto (mgp) distributions in particular.

In the literature, tail dependence is described mathematically through various concepts, such as the exponent measure, the angular measure, and the stable tail dependence function, among others. When the focus is on modeling multivariate threshold exceedances, the mgp distribution is a natural candidate [83]. In the univariate case, peaks-over-threshold modeling with the generalized Pareto distribution is well-established [23], while in the multivariate set-up, statistical modeling is quite recent [85, 66, 54, 34, 53]. Asymptotically, the class of mgp distributions represents proper multivariate probability distributions on a region where at least one variable is large. Moreover, parametric mgp models can be constructed quite easily via the *generators* proposed in Kiriliouk et al. [66].

Current theory for mgp distributions does not cover extreme directions other than $J = \{1, \ldots, d\}$, a gap we aim to fill. From their generators, we derive expressions for mgp densities with respect to an appropriate dominating measure. Furthermore, we propose a stochastic construction based on smoothed max-linear models, wherein the strong dependence within factors inherent in max-linear models is substituted with a more lenient dependency framework. Our construction is designed to accommodate all mgp models, regardless of their extreme directions.

Learning extreme directions from data has become an active area of research. Goix et al. [48, 50] focus on the identification of extreme directions, providing a sparse representation of the support of the angular measure. Chiapino and Sabourin [12] and Chiapino et al. [13] study an alternative algorithm with better performance when the number of extreme directions is high. Jalalzai and Leluc [59] introduce an approach based on empirical risk minimization. Chiapino et al. [14] propose

a Dirichlet mixture model for the angular measure of a heavy-tailed distribution whose support has been identified and apply it to anomaly detection. Simpson et al. [90] exploit the concept of hidden regular variation to identify extreme directions. Finally, Meyer and Wintenberger [73, 74] frame extreme directions within what they call sparse regular variation.

The study of extreme directions is related to that of clustering and dimension reduction methods for multivariate extremes. For example, Chautru [11] and Janßen and Wan [60] identify extreme directions via spherical k-means clustering of the angles $\boldsymbol{X}/\|\boldsymbol{X}\|$ provided $\|\boldsymbol{X}\|$ is large. Fomichov and Ivanovs [45] provide theoretical results in favor of this approach and extend it to a spherical k-principal components algorithm. Clustering angles is also considered in Medina et al. [72], based on a random k-nearest neighbors graph. Finally, Cooley and Thibaud [19] and Drees and Sabourin [31] propose to perform principal component analysis of the angles $\boldsymbol{X}/\|\boldsymbol{X}\|$. These and other methods are surveyed in Engelke and Ivanovs [35].

Our contribution differs from those above from several perspectives. First, our focus is on extreme directions in the context of multivariate threshold exceedances. Second, our aim is not to learn extreme directions, but rather to introduce parametric mgp models capable of generating extreme directions other than $\{1, \dots, d\}$. For multivariate extreme-value distributions, the max-linear factor model [32] and the asymmetric logistic model [96] already allow for this possibility. In the recent literature, only Simpson et al. [90] and Chiapino et al. [14] propose mixture models for multivariate extreme-value distributions. The power of our model construction resides in the fact that it is capable of representing mgp distributions with general extreme directions. Moreover, for specific parametric families, our model is straightforward to simulate from.

The paper is organized as follows. Section 2.2 recalls essential background on multivariate extreme-value theory. Section 2.3 defines and characterizes extreme directions in terms of the angular measure, the exponent measure, and the mgp distribution. We also compute the density of an mgp distribution with arbitrary extreme directions in terms of its generators, extending results in Kiriliouk et al. [66]. A model construction tool based on a mixture of mgp generators is presented in Section 2.4. Apart from enjoying a convenient interpretation as a smoothed max-linear model, the construction is able to reproduce any mgp distribution. For the sake of illustration, we work out two examples based on the logistic and the Hüsler–Reiss families. Moreover, the construction underlies a simulation algorithm for mgp distributions with non-standard extreme directions (Section 2.5). As a by-product, we revisit known simulation algorithms in Dombry et al. [28] and Engelke and Hitz [34] that can be used to simulate mgp random vectors. Section 2.6 brings the paper to a close by outlining several avenues for future research. All

proofs are deferred to the appendices.

2.2 Background

Notation. Let d be a positive integer and write $D = \{1, \dots, d\}$. Throughout, bold symbols will refer to multivariate quantities. Let $\mathbf{0}$ and $\mathbf{1}$ denote vectors of zeroes and ones, respectively, of a dimension clear from the context. A d-dimensional vector $\boldsymbol{x}$ will be written as $\boldsymbol{x} = (x_1, \dots, x_d)$ and for a non-empty set $J \subseteq \{1, \dots, d\}$, write $\boldsymbol{x}_J = (x_j)_{j \in J}$. Mathematical operations on vectors such as addition, multiplication and comparison are considered component-wise apart from $\boldsymbol{a} \not\leqslant \boldsymbol{b}$ which denotes $a_j > b_j$ for at least one $j \in D$.

We also fix notation for some sets that will be used throughout the paper. For two vectors $\boldsymbol{a}$, $\boldsymbol{b}$, the set $[\boldsymbol{a}, \boldsymbol{b}]$ will denote the Cartesian product $\prod_{j=1}^{d} [a_j, b_j]$ and so forth for $(\boldsymbol{a}, \boldsymbol{b})$, $[\boldsymbol{a}, \boldsymbol{b})$, $(\boldsymbol{a}, \boldsymbol{b}]$. Write $\mathbb{E} = [0, \infty)^d \setminus \{\mathbf{0}\}$. Let $\mathcal{B}(F)$ denote the Borel sets over a topological space F. For a general norm $\| \cdot \|$ on $\mathbb{R}^d$, let $\mathbb{S} = \{\boldsymbol{x} \in \mathbb{E} : \ \|\boldsymbol{x}\| = 1\}$ denote the intersection of the unit sphere with the positive orthant. For a non-empty set $J \subseteq D$, let

$$\mathbb{A}_J := \left\{ \boldsymbol{x} \in [-\infty, \infty)^d : \ x_j > -\infty \text{ iff } j \in J \right\}, \tag{2.2.1}$$

and let $\pi_J : \mathbb{A}_J \to \mathbb{R}^J$ be the projection $\boldsymbol{x} \mapsto \boldsymbol{x}_J$ and with $\mathbb{R}^J$ the set of real vectors $(x_j)_{j \in J}$ indexed by J. Clearly, $\mathbb{R}^J$ is isomorphic to $\mathbb{R}^{|J|}$, with $|J|$ the number of elements of J, but it will be convenient to keep track of the index set explicitly, not only its cardinality.

The arrow $\rightsquigarrow$ denotes convergence in distribution (weak convergence). If μ is a measure, then a Borel set B is μ-continuous if $\mu(\partial B) = 0$, with ∂B the topological boundary of B. The distribution or law of a random vector $\boldsymbol{X}$ is denoted by $\mathcal{L}(\boldsymbol{X})$. Finally, the Lebesgue measure on $\mathbb{R}^m$ is denoted by λ_m.

2.2.1 Multivariate extreme value distributions

Consider a random vector $\boldsymbol{\xi}$ on $\mathbb{R}^d$ with joint cumulative distribution function (cdf) F and margins $F_1, \dots, F_d$. We say that F is in the domain of attraction of a multivariate extreme-value (mev) distribution G and we write $F \in \mathrm{DA}(G)$ if there exist sequences $\boldsymbol{a}_n \in (0, \infty)^d$ and $\boldsymbol{b}_n \in \mathbb{R}^d$ such that

$$\lim_{n \to \infty} F^n(\boldsymbol{a}_n \boldsymbol{x} + \boldsymbol{b}_n) = G(\boldsymbol{x}),$$

for every point $\boldsymbol{x} \in \mathbb{R}^d$ where G is continuous. Equivalently, if $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n$ are i.i.d. random vectors with common cdf F, then

$$\frac{\max_{i=1,\dots,n} \boldsymbol{\xi}_i - \boldsymbol{b}_n}{\boldsymbol{a}_n} \rightsquigarrow G, \qquad n \to \infty. \tag{2.2.2}$$

The margins of an mev distribution are univariate extreme-value distributions themselves. Thus, for $j \in D$, the marginal distribution G_j belongs to a three-parameter family and is

$$G_j(x_j) = \begin{cases} \exp\left[-\left\{1+\gamma_j\left(x_j-\mu_j\right)/\alpha_j\right\}^{-1/\gamma_j}\right], & \text{if } \gamma_j \neq 0, \\ \exp\left[-\exp\left\{-\left(x_j-\mu_j\right)/\alpha_j\right\}\right], & \text{if } \gamma_j = 0, \end{cases}$$

for some shape parameter $\gamma_j \in \mathbb{R}$, location parameter $\mu_j \in \mathbb{R}$, and scale parameter $\alpha_j > 0$; further, x_j needs to be such that $1+\gamma_j(x_j-\mu_j)/\alpha_j > 0$.

For $d \geqslant 2$, the family of d-dimensional mev distributions no longer constitutes a finite-dimensional parametric family, in contrast to the univariate case. This is due to the nature of the dependence structure between the components of G, or equivalently, the extremal dependence between the components of $\boldsymbol{\xi}$. Many functions and measures that describe this dependence are discussed in the literature [see, e.g., 4, Chapter 8]. In this paper, three such objects will appear on the forefront: the exponent measure, the angular measure [25] and the stable tail dependence function [29].

To concentrate on the dependence structure of G, it is convenient to standardize its margins. The choice of common scale is a matter of convention and convenience. In this section, we will opt for the unit-Fréchet distribution, i.e., $\alpha_j = \gamma_j = \mu_j = 1$ for all $j \in D$, in line with the theory as originally exposed in [25]. In Section 2.2.2, however, we will choose the Gumbel distribution instead ($\alpha_j = 1$ and $\gamma_j = \mu_j = 0$ for all $j \in D$), in order to link up with the additive model formulation for the mgp in Rootzén et al. [84]. Switching between different marginal scales is an easy and sometimes illuminating exercise.

An mev distribution G with general margins admits the representation

$$G(\boldsymbol{x}) = G^*\left(-1/\ln G_1(x_1), \ldots, -1/\ln G_d(x_d)\right),$$

where G^* is an mev distribution with unit-Fréchet margins. For $\boldsymbol{\xi}_i$ as in (2.2.2), if $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})$ is defined by

$$X_{ij} = -1/\ln F_j\left(\xi_{ij}\right), \qquad i \in \mathbb{N},\ j \in D, \tag{2.2.3}$$

then we have $\max_{i=1,\ldots,n} \boldsymbol{X}_i/n \rightsquigarrow G^*$ as $n \to \infty$.

The *exponent measure* of G^* is the unique Borel measure μ concentrated on $\mathbb{E}$ such that

$$G^*(\boldsymbol{x}) = \exp\left\{-\mu\left(\mathbb{E} \setminus [\boldsymbol{0}, \boldsymbol{x}]\right)\right\}, \qquad \boldsymbol{x} \in [\boldsymbol{0}, \boldsymbol{\infty}]. \tag{2.2.4}$$

The measure μ is finite on Borel sets whose topological closure does not contain the origin, i.e., is bounded away from the origin. Moreover, $\mu\left(\mathbb{E} \setminus [\boldsymbol{0}, \boldsymbol{x}]\right) = \infty$ as

soon as $x_j = 0$ for some $j \in D$. The exponent measure μ is (-1)-homogeneous in the sense that $\mu(c\,\cdot) = c^{-1}\mu(\,\cdot\,)$ for all $c > 0$. Moreover, for any μ-continuous Borel set $B \subset \mathbb{E}$ bounded away from the origin, we have

$$\lim_{n\to\infty} n\,\mathsf{P}(\boldsymbol{X} \in nB) = \mu(B). \tag{2.2.5}$$

Another way to describe the dependence structure of G is via the *stable tail dependence function (stdf)*,

$$\ell(\boldsymbol{x}) = \mu\left(\mathbb{E} \setminus [\boldsymbol{0}, \boldsymbol{1}/\boldsymbol{x}]\right) = -\ln G^*(\boldsymbol{1}/\boldsymbol{x}), \qquad \boldsymbol{x} \in [\boldsymbol{0}, \boldsymbol{\infty}). \tag{2.2.6}$$

The margins are constrained by $\ell(0, \ldots, 0, x_j, 0, \ldots, 0) = x_j$ for $x_j \geqslant 0$ and $j \in D$. Moreover, ℓ is homogeneous in the sense that for $c \geqslant 0$ and $\boldsymbol{x} \in [\boldsymbol{0}, \boldsymbol{\infty})$, we have $\ell(c\boldsymbol{x}) = c\ell(\boldsymbol{x})$.

The *angular measure* ϕ of G^* with respect to a general norm $\|\cdot\|$ on $\mathbb{R}^d$ is

$$\phi(B) = \mu\left(\{\boldsymbol{x} \in \mathbb{E} :\ \|\boldsymbol{x}\| > 1,\ \boldsymbol{x}/\|\boldsymbol{x}\| \in B\}\right), \qquad B \in \mathcal{B}(\mathbb{S}). \tag{2.2.7}$$

Equations (2.2.5) and (2.2.7) yield

$$\lim_{n\to\infty} n\,\mathsf{P}(\|\boldsymbol{X}\| > n,\ \boldsymbol{X}/\|\boldsymbol{X}\| \in B) = \phi(B),$$

for any ϕ-continuous Borel set $B \subseteq S$.

2.2.2 Multivariate generalized Pareto distributions

The peaks-over-thresholds method is grounded in the fact that asymptotically, vectors of component-wise maxima follow an mev distribution if and only if exceedances over a high threshold vector follow an mgp distribution. Recall $\boldsymbol{\xi}$ and G in (2.2.2) and suppose $G(\boldsymbol{0}) > 0$. Convergence (2.2.2) is satisfied if and only if there exists a distribution H such that

$$\mathcal{L}\left(\frac{\boldsymbol{\xi} - \boldsymbol{b}_n}{\boldsymbol{a}_n} \vee \boldsymbol{\zeta} \,\middle|\, \boldsymbol{\xi} \not\leqslant \boldsymbol{b}_n\right) \rightsquigarrow H, \qquad n \to \infty,$$

where $\boldsymbol{\zeta}$ is the vector of lower endpoints of $G_1, \ldots, G_d$ and $\vee$ denotes the component-wise maximum. The conditioning event $\boldsymbol{\xi} \not\leqslant \boldsymbol{b}_n$ is to be read as $\exists j \in D : \xi_j > b_{nj}$, that is, at least one variable within $\boldsymbol{\xi}$ exceeds a high threshold. We write $H = \mathrm{GP}(G)$. As in Rootzén et al. [84], we consider the standardized case where G has Gumbel margins, i.e., $\gamma_j = \mu_j = 0$ and $\alpha_j = 1$ for all $j \in D$. The distribution of an mgp random vector $\boldsymbol{Y} \sim H$ is then determined by the stdf ℓ.

Recall the exponent measure μ of G^*. By Rootzén and Tajvidi [83, Proposition 3.1], the law of $\mathrm{e}^{\boldsymbol{Y}}$ can be expressed in terms of μ on the collection of sets $\mathbb{A} = \{[\boldsymbol{0}, \boldsymbol{y}],\ \boldsymbol{y} \in [\boldsymbol{0}, \boldsymbol{\infty})\}$, via

$$\mathsf{P}[\mathrm{e}^{\boldsymbol{Y}} \leqslant \boldsymbol{y}] = \frac{\mu\left([\boldsymbol{0}, \boldsymbol{y}] \cap \{\boldsymbol{x} \geqslant \boldsymbol{0} :\ \boldsymbol{x} \nleqslant \boldsymbol{1}\}\right)}{\mu\left(\{\boldsymbol{x} \geqslant \boldsymbol{0} :\ \boldsymbol{x} \nleqslant \boldsymbol{1}\}\right)}.$$

Since a probability measure on $[\boldsymbol{0}, \boldsymbol{\infty})$ is determined by its values on $\mathbb{A}$, we get

$$\mathsf{P}\left[\mathrm{e}^{\boldsymbol{Y}} \in B\right] = \frac{\mu\left(B \cap \{\boldsymbol{x} \geqslant \boldsymbol{0} :\ \boldsymbol{x} \nleqslant \boldsymbol{1}\}\right)}{\mu\left(\{\boldsymbol{x} \geqslant \boldsymbol{0} :\ \boldsymbol{x} \nleqslant \boldsymbol{1}\}\right)}, \qquad B \in \mathcal{B}\left([\boldsymbol{0}, \boldsymbol{\infty})\right). \tag{2.2.8}$$

In words, the law of $\mathrm{e}^{\boldsymbol{Y}}$ is equal to the restriction of μ to $\{\boldsymbol{x} \geqslant \boldsymbol{0} :\ \boldsymbol{x} \nleqslant \boldsymbol{1}\}$, normalized to a probability measure. Conversely, since $\mu\left(\left\{\boldsymbol{x} \geqslant \boldsymbol{0} :\ x_j > 1\right\}\right) = -\ln G_j^*(1) = 1$ by (2.2.6), we obtain the normalization constant via $\mathsf{P}\left[Y_j > 0\right] = 1/\mu\left(\{\boldsymbol{x} \geqslant \boldsymbol{0} :\ \boldsymbol{x} \nleqslant \boldsymbol{1}\}\right)$ for all $j \in D$, so that μ can be reconstructed from $\boldsymbol{Y}$ by

$$\mu\left(B \cap \{\boldsymbol{x} \geqslant \boldsymbol{0} :\ \boldsymbol{x} \nleqslant \boldsymbol{1}\}\right) = \mathsf{P}\left[\mathrm{e}^{\boldsymbol{Y}} \in B\right] / \mathsf{P}\left[Y_j > 0\right], \qquad B \in \mathcal{B}\left([\boldsymbol{0}, \boldsymbol{\infty})\right). \tag{2.2.9}$$

Rootzén et al. [84, Section 4] propose three stochastic representations for H. In view of their importance to the construction device and the simulation algorithm to follow, we review these representations in some detail.

Spectral random vector $\boldsymbol{S}$. Let $\boldsymbol{S}$ be a random vector in $[-\infty, 0]^d$ such that $\mathsf{P}[\max(\boldsymbol{S}) = 0] = 1$, $\mathsf{P}[S_j > -\infty] > 0$ for all $j \in D$ and $\mathsf{E}[\mathrm{e}^{S_1}] = \ldots = \mathsf{E}[\mathrm{e}^{S_d}]$. Then $\boldsymbol{S}$ is said to be the *spectral random vector* [84, Theorem 6] associated to the mgp H if

$$\ell(\boldsymbol{y}) = \mathsf{E}[\max\left(\boldsymbol{y}\mathrm{e}^{\boldsymbol{S}}\right)] / \mathsf{E}[\mathrm{e}^{S_1}], \qquad \boldsymbol{y} \in [\boldsymbol{0}, \boldsymbol{\infty}). \tag{2.2.10}$$

We write $H = \mathrm{GP}_{\boldsymbol{S}}(\mathcal{L}(\boldsymbol{S}))$, while $Y \sim H$ admits the representation

$$\boldsymbol{Y} = E + \boldsymbol{S}, \tag{2.2.11}$$

with $E = \max(\boldsymbol{Y})$ a unit exponential random variable independent of the spectral random vector $\boldsymbol{S} = \boldsymbol{Y} - \max(\boldsymbol{Y})$; see Rootzén et al. [84, Theorem 7].

The distribution of the random vector $\boldsymbol{S}$ can thus be obtained directly from the mgp H. Conversely, any random vector $\boldsymbol{S}$ with the above properties generates an mgp H. Still, the constraint $\max(\boldsymbol{S}) = 0$ almost surely is not satisfied by common families of distributions, which raises the question how to construct appropriate $\boldsymbol{S}$. The next representation resolves this issue, thereby facilitating the construction of an mgp.

Generator $\boldsymbol{T}$. Let $\boldsymbol{T}$ be a random vector in $[-\infty,\infty)^d$ such that $\mathsf{P}[\max(\boldsymbol{T}) > -\infty] = 1$, $\mathsf{P}[T_j > -\infty] > 0$ for all $j \in D$ and $\mathsf{E}[\mathrm{e}^{T_1-\max(\boldsymbol{T})}] = \ldots = \mathsf{E}[\mathrm{e}^{T_d-\max(\boldsymbol{T})}]$. The random vector $\boldsymbol{T}$ is called a $\boldsymbol{T}$-*generator* of H [84, Proposition 8], if we have the equality in distribution

$$\boldsymbol{S} \stackrel{\mathsf{d}}{=} \boldsymbol{T} - \max(\boldsymbol{T}), \tag{2.2.12}$$

where $\boldsymbol{S}$ is the spectral random vector associated to H. We will write $H = \mathrm{GP}_{\boldsymbol{T}}(\mathcal{L}(\boldsymbol{T}))$. Equations (2.2.11) and (2.2.12) yield that if $\boldsymbol{T}$ is a $\boldsymbol{T}$-generator of H and if $\boldsymbol{Y} \sim H$, then

$$\boldsymbol{Y} \stackrel{\mathsf{d}}{=} \boldsymbol{T} - \max(\boldsymbol{T}) + E. \tag{2.2.13}$$

Any random vector $\boldsymbol{T}$ with the above properties yields an mgp random vector $\boldsymbol{Y}$ via (2.2.13). In comparison to $\boldsymbol{S}$, it has one constraint less to satisfy; the construction in (2.2.12) guarantees that $\max(\boldsymbol{S})$ vanishes almost surely. On the other hand, the $\boldsymbol{T}$-generator of an mgp distribution H is not unique: adding the same random variable to all components of $\boldsymbol{T}$ does not change the distribution of $\boldsymbol{T} - \max(\boldsymbol{T})$ and produces the same mgp H.

A drawback of both representations $\boldsymbol{S}$ and $\boldsymbol{T}$ is that they are not stable with respect to marginalization. If $\boldsymbol{Y} \sim H$ is a d-dimensional mgp random vector and if $J \subset D$ is non-empty, the conditional distribution of $\boldsymbol{Y}_J$ given $\max \boldsymbol{Y}_J > 0$ is a $|J|$-dimensional mgp, say $H_{|J}$ (which is different from H_J, the unconditional distribution of $\boldsymbol{Y}_J$, which is in general not a mgp). However, if $\boldsymbol{S}$ is the spectral random vector of H, then $\boldsymbol{S}_J$ is *not* the spectral random vector of $H_{|J}$. The following representation resolves this issue, at the price of a change of measure.

Generator $\boldsymbol{U}$. Let $\boldsymbol{U}$ be a random vector on $[-\infty,\infty)^d$ such that $\mathsf{E}[\mathrm{e}^{U_1}] = \ldots = \mathsf{E}[\mathrm{e}^{U_d}]$ and $0 < \mathsf{E}[\mathrm{e}^{U_1}] < \infty$. The random vector $\boldsymbol{U}$ is said to be a $\boldsymbol{U}$-*generator* of H [84, Proposition 9] if

$$\ell(\boldsymbol{y}) = \mathsf{E}\left[\max\left(\boldsymbol{y}\mathrm{e}^{\boldsymbol{U}}\right)\right] / \mathsf{E}\left[\mathrm{e}^{U_1}\right], \qquad \boldsymbol{y} \in [\boldsymbol{0},\boldsymbol{\infty}). \tag{2.2.14}$$

We write $H = \mathrm{GP}_{\boldsymbol{U}}(\mathcal{L}(\boldsymbol{U}))$. If $\boldsymbol{U}$ is a generator of H and if $c \in \mathbb{R}$, then $\boldsymbol{U} + c$ is also a generator of H. This means that without loss of generality, we can and will assume that

$$\mathsf{E}\left[\mathrm{e}^{U_j}\right] = 1, \qquad j \in D. \tag{2.2.15}$$

This constraint still does not make $\boldsymbol{U}$ identifiable from H: if the random variable V satisfies $\mathsf{E}\left[\mathrm{e}^{V}\right] = 1$ and is independent of $\boldsymbol{U}$, then the random vector $(U_1 + V, \ldots, U_d + V)$ yields another $\boldsymbol{U}$-generator for the same mgp H.

If $\boldsymbol{U}$ is a generator of $\boldsymbol{Y} \sim H$, then, for non-empty $J \subset D$, the random vector $\boldsymbol{U}_J$ is a generator of $\boldsymbol{Y}_J \mid \max(\boldsymbol{Y}_J) > 0$, which thus follows a $|J|$-variate mgp [84, Proposition 18]. In Proposition 2.5.2 below, we establish a link between $\boldsymbol{U}$ and

the extremal functions in Dombry et al. [28]. Equation (2.2.14) represents the stdf ℓ as a D-norm with generator $\boldsymbol{U}$ in the sense of Falk [43]. In Kiriliouk et al. [66, Section 7], common parametric families of multivariate extreme value models are described in terms of their $\boldsymbol{U}$-generator.

The spectral random vector $\boldsymbol{S}$ can be obtained from the generator $\boldsymbol{U}$ via

$$\mathsf{P}[\boldsymbol{S} \in B] = \frac{\mathsf{E}\left[\mathbb{1}\left\{\boldsymbol{U} - \max(\boldsymbol{U}) \in B\right\} \mathrm{e}^{\max(\boldsymbol{U})}\right]}{\mathsf{E}\left[\mathrm{e}^{\max(\boldsymbol{U})}\right]}, \qquad B \in \mathcal{B}\left([-\boldsymbol{\infty}, \mathbf{0})\right). \quad (2.2.16)$$

Equations (2.2.12) and (2.2.16) link representation $\boldsymbol{S}$ to $\boldsymbol{T}$ and to $\boldsymbol{U}$, respectively. Likewise, a $\boldsymbol{T}$-generator can be obtained from the $\boldsymbol{U}$-generator via

$$\mathsf{P}[\boldsymbol{T} \in B] = \frac{\mathsf{E}\left[\mathbb{1}\left\{\boldsymbol{U} \in B\right\} \mathrm{e}^{\max(\boldsymbol{U})}\right]}{\mathsf{E}\left[\mathrm{e}^{\max(\boldsymbol{U})}\right]}, \qquad B \in \mathcal{B}\left([-\boldsymbol{\infty}, \boldsymbol{\infty})\right). \quad (2.2.17)$$

Note that a spectral random vector $\boldsymbol{S}$ is also a $\boldsymbol{T}$-generator, since $\max(\boldsymbol{S}) = 0$ and thus $\mathrm{GP}_{\boldsymbol{T}}(\mathcal{L}(\boldsymbol{S})) = \mathrm{GP}_{\boldsymbol{S}}(\mathcal{L}(\boldsymbol{S}))$. On the other hand, the representation via $\boldsymbol{U}$ is of another nature: if $\boldsymbol{V}$ is a random vector whose distribution satisfies both the constraints to be a $\boldsymbol{T}$-generator and a $\boldsymbol{U}$-generator, then the mgp distributions $\mathrm{GP}_{\boldsymbol{T}}(\mathcal{L}(\boldsymbol{V}))$ and $\mathrm{GP}_{\boldsymbol{U}}(\mathcal{L}(\boldsymbol{V}))$ are different in general; see Kiriliouk et al. [66, Section 7] for examples.

Let ϕ be the angular measure of G^* with respect to a norm $\|\cdot\|$ on $\mathbb{R}^d$ as in Equation (2.2.7). Let $\boldsymbol{V}$ be a random vector on the $\|\cdot\|$-unit sphere $\mathbb{S}$ with distribution $\phi/\phi(\mathbb{S})$. The random vector $\boldsymbol{U}^\phi = \ln(\boldsymbol{V})$ is then a $\boldsymbol{U}$-generator of H. For non-empty $J \subseteq D$, we have

$$\mathsf{P}\left[U_j^\phi > -\infty \text{ iff } j \in J\right] = \phi\left(\left\{\boldsymbol{w} \in \mathbb{S} : \; w_j > 0 \text{ iff } j \in J\right\}\right)/\phi(\mathbb{S}),$$

an equality that makes the link to extreme directions of the underlying random vector $\boldsymbol{\xi}$, as treated next.

2.3 Extreme directions

Let the random vector $\boldsymbol{X} = (X_j)_{j \in D}$ be defined on the basis of the probability integral transform applied to the components of a random vector $\boldsymbol{\xi}$ as in (2.2.3), and assume the multivariate regular variation condition $n\,\mathsf{P}\left[\boldsymbol{X} \in n\,\cdot\right] \to \mu(\,\cdot\,)$ in the sense of (2.2.5) holds. We are interested in scenarios where all variables X_j for j in some subset J of D are large while the other variables are not; a precise meaning will be given below. Such a group of variables J will be called an extreme direction, defined formally in Section 2.3.1. We will provide various characterizations of such

extreme directions: by the angular measure, by the asymptotic probability of $\boldsymbol{X}$ to belong to certain sets called thickened rectangles and cones, and finally by the mgp distribution. Next, in Section 2.3.2, we express the conditional probability density function of an mgp random vector given a certain extreme direction in terms of other densities that are easier to compute.

Upper tail dependence coefficients [15, 86]

$$\chi_J = \lim_{q\to\infty} q\,\mathsf{P}\left[\forall j \in J : X_j > q\right], \qquad \varnothing \neq J \subseteq D, \tag{2.3.1}$$

are well-known summary measures of extremal dependence. However, we would like to emphasize that they are not sufficient to characterize extreme directions. In fact, even though the equality $\chi_J = 0$ implies that the variables X_j for $j \in J$ cannot be large simultaneously, the inequality $\chi_J > 0$ does not exclude the possibility of extreme directions properly contained in J. For instance, for a pair $J = \{j_1, j_2\}$, the inequality $\chi_{j_1 j_2} > 0$ implies that X_{j_1} and X_{j_2} can be large simultaneously but does not rule out the possibility that X_{j_1} is large while X_{j_2} is not.

2.3.1 Definition and characterizations

Recall the exponent measure μ of G^* as defined in (2.2.4) and $\mathbb{E} = [\mathbf{0}, \boldsymbol{\infty}) \setminus \{\mathbf{0}\}$. In this section, let $J \subseteq D$ be a fixed non-empty set and $\mathbb{E}_J := \left\{\boldsymbol{x} \in \mathbb{E} :\ x_j > 0 \text{ iff } j \in J\right\}$. The collection $\{\mathbb{E}_J :\ \varnothing \neq J \subseteq D\}$ forms a partition of $\mathbb{E}$. The following definition was introduced in Simpson et al. [90] in order to identify extreme directions of the random vector $\boldsymbol{X}$ defined as in (2.2.3).

Definition 2.3.1. The set J is said to be an extreme direction of $\boldsymbol{X}$ or μ if $\mu(\mathbb{E}_J) > 0$.

A trivial but useful fact is that μ has no mass outside $\bigcup_{J\in\mathcal{R}(\mu)} \mathbb{E}_J$, where $\mathcal{R}(\mu)$ is the collection of extreme directions of μ. Further, even if a certain set $J \subseteq D$ is not an extreme direction of $\boldsymbol{X}$, it is still possible that J is a superset or a subset of an extreme direction. In contrast, if $\chi_J = 0$, then also $\chi_{\bar{J}} = 0$ whenever $\bar{J} \subseteq J$. Remark 2.3.8 below points out a connection between tail dependence coefficients and extreme directions.

Throughout this section, let $\|\cdot\|$ be a fixed arbitrary norm on $\mathbb{R}^d$ and let ϕ denote the angular measure of G^* with respect to this norm. Recall the set $\mathbb{S} = \{\boldsymbol{x} \in \mathbb{E} :\ \|\boldsymbol{x}\| = 1\}$ and let $\mathbb{S}_J := \left\{\boldsymbol{w} \in \mathbb{S} :\ w_j > 0 \text{ iff } j \in J\right\}$. The following proposition gives a characterization of extreme directions in terms of the angular measure ϕ.

Proposition 2.3.2. *The set J is an extreme direction of $\boldsymbol{X}$ if and only if $\phi(\mathbb{S}_J) > 0$.*

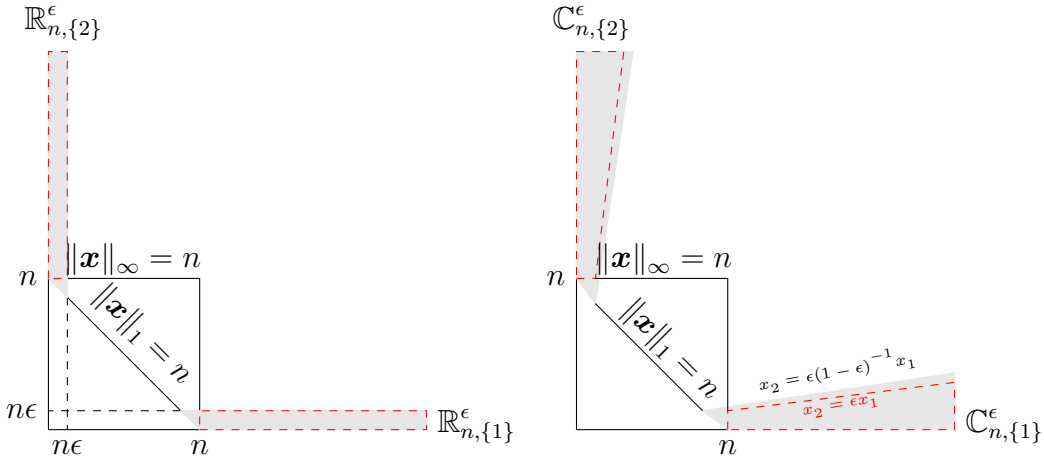


Figure 2.1
Thickened rectangles as in (2.3.2) (left) and thickened cones as in (2.3.3) (right) associated with the L_1-norm (simple gray region) and L_∞-norm (gray region with dashed red line boundary).

The special case of Proposition 2.3.2 for the L_∞-norm will be used below for the characterization of extreme directions in terms of the mgp distribution.

Let $\epsilon > 0$ and $n \in \mathbb{N}$. We will interpret extreme directions in terms of the asymptotic probability of $\boldsymbol{X}$ to be contained in certain thickened rectangles and cones associated with $\|\cdot\|$, defined as

$$\mathbb{R}^{\epsilon}_{n,J} = \left\{\boldsymbol{x} \in \mathbb{E} : \ \|\boldsymbol{x}\| > n, \ \boldsymbol{x}_J/n \in (\epsilon,\infty)^J, \ \boldsymbol{x}_{D\setminus J}/n \in [0,\epsilon]^{|D\setminus J|}\right\}, \tag{2.3.2}$$

$$\mathbb{C}^{\epsilon}_{n,J} = \left\{\boldsymbol{x} \in \mathbb{E} : \ \|\boldsymbol{x}\| > n, \ \boldsymbol{x}_J/\|\boldsymbol{x}\| \in (\epsilon,\infty)^J, \ \boldsymbol{x}_{D\setminus J}/\|\boldsymbol{x}\| \in [0,\epsilon]^{|D\setminus J|}\right\}, \tag{2.3.3}$$

respectively, and illustrated in Figure 2.1.

The latter sets were introduced in Goix et al. [50, 48] for the L_∞-norm. For $\epsilon > 0$, the Borel set $\mathbb{R}^{\epsilon}_{1,J}$ is μ-continuous as defined in Section 2.2. Hence we can make the following link with Proposition 2.3.2.

Lemma 2.3.3. *We have*

$$\phi(\mathbb{S}_J) = \lim_{\epsilon\downarrow 0} \lim_{n\to\infty} n\,\mathsf{P}\left[\boldsymbol{X} \in \mathbb{R}^{\epsilon}_{n,J}\right]. \tag{2.3.4}$$

In contrast, a standardization by $\|\boldsymbol{x}\|$ rather than n is used to define thickened cones. Thus, for $\epsilon > 0$, the Borel set $\mathbb{C}^{\epsilon}_{1,J}$ is not necessarily μ-continuous. To obtain a similar limit to (2.3.4) for the thickened cones, we should exclude at most countably many values of $\epsilon > 0$ for which $\mathbb{C}^{\epsilon}_{1,J}$ is not μ-continuous.

Lemma 2.3.4. *There exists an at most countable set $N \subset [0,\infty)$ such that*

$$\phi(\mathbb{S}_J) = \lim_{\substack{\epsilon \downarrow 0 \\ \epsilon \notin N}} \lim_{n\to\infty} n\,\mathsf{P}\left[\boldsymbol{X} \in \mathbb{C}^{\epsilon}_{n,J}\right]. \tag{2.3.5}$$

Since the sets $\mathbb{C}^{\epsilon}_{n,J}$ depend in a monotone way on ϵ, the formula (2.3.5) also holds without excluding any values of ϵ, provided the limit over n is replaced by a limsup.

Proposition 2.3.2 and Lemmas 2.3.3 and 2.3.4 imply that the set J is an extreme direction of $\boldsymbol{X}$ if and only if the equivalent conditions 3. and 4. in Theorem 2.3.6 are satisfied.

Finally, we characterize extreme directions in terms of the mgp distribution and its representations. Let $\tilde{G}(\boldsymbol{x}) = G^*\left(\exp\left(x_1\right), \ldots, \exp\left(x_d\right)\right)$ for $\boldsymbol{x} = (x_1, \ldots, x_d)$, be the Gumbel standardization of G^* and consider the associated mgp distribution $H = \mathrm{GP}(\tilde{G})$ as in Section 2.2.2. Recall representations $\boldsymbol{S}$, $\boldsymbol{T}$ and $\boldsymbol{U}$ of H in Eqs. (2.2.10)–(2.2.14). Let $\mathbb{S}_\infty = \{\boldsymbol{x} \geqslant \boldsymbol{0} : \|\boldsymbol{x}\|_\infty = 1\}$ denote the positive L_∞-sphere and $\mathbb{S}_{J,\infty} := \left\{\boldsymbol{w} \in \mathbb{S}_\infty : \ w_j > 0 \text{ iff } j \in J\right\}$ for non-empty $J \subseteq D$. Let $\boldsymbol{Y} \sim H$. Applying Eq. (2.2.8) to the set $\mathbb{E}_J$ gives

$$\mathsf{P}\left[Y_j > -\infty \text{ iff } j \in J\right] = \frac{\phi_\infty\left(\mathbb{S}_{J,\infty}\right)}{\phi_\infty(\mathbb{S}_\infty)},$$

where ϕ_∞ is the angular measure of G^* with respect to the L_∞-norm. Proposition 2.3.2 applied to the L_∞-norm yields that J is an extreme direction of $\boldsymbol{X}$ if and only if $\mathsf{P}\left[Y_j > -\infty \text{ iff } j \in J\right] > 0$. The following lemma expresses this probability in terms of representations $\boldsymbol{S}$, $\boldsymbol{T}$ and $\boldsymbol{U}$.

Lemma 2.3.5. *For non-empty $J \subseteq D$, we have*

$$\begin{aligned}
\mathsf{P}\left[Y_j > -\infty \textit{ iff } j \in J\right] &= \mathsf{P}\left[S_j > -\infty \textit{ iff } j \in J\right] && (2.3.6)\\
&= \mathsf{P}\left[T_j > -\infty \textit{ iff } j \in J\right] && (2.3.7)\\
&= \mathsf{E}\left[\mathrm{e}^{\max(\boldsymbol{U}_J)}\,\mathbb{1}\left\{U_j > -\infty \textit{ iff } j \in J\right\}\right] / \mathsf{E}[\mathrm{e}^{\max \boldsymbol{U}}]. && (2.3.8)
\end{aligned}$$

We finally get the following theorem which summarizes this section.

Theorem 2.3.6. *For non-empty $J \subseteq D$, the following statements are equivalent:*

1. *The set J is an extreme direction.*
2. $\phi(\mathbb{S}_J) > 0$.
3. $\lim\limits_{\epsilon\downarrow 0} \lim\limits_{n\to\infty} n\,\mathsf{P}\left[\boldsymbol{X} \in \mathbb{R}^{\epsilon}_{n,J}\right] > 0$.
4. $\lim\limits_{\epsilon\downarrow 0} \limsup\limits_{n\to\infty} n\,\mathsf{P}\left[\boldsymbol{X} \in \mathbb{C}^{\epsilon}_{n,J}\right] > 0$.
5. $\mathsf{P}[Y_j > -\infty \textit{ iff } j \in J] > 0$.
6. $\mathsf{P}[U_j > -\infty \textit{ iff } j \in J] > 0$.

7. $\mathsf{P}[T_j > -\infty \text{ iff } j \in J] > 0$. 8. $\mathsf{P}[S_j > -\infty \text{ iff } j \in J] > 0$.

Example 2.3.7 (Max-linear model)**.** Let r be a positive integer and write $R = \{1, \ldots, r\}$. Let $Z_1, \ldots, Z_r$ be independent unit-Fréchet random variables. Let $A = (a_{jk})_{j \in D; k \in R} \in [0,1]^{d \times r}$ be a coefficient matrix with row sums $\sum_{k=1}^{r} a_{jk} = 1$ for all $j \in D$ and column sums $\sum_{j=1}^{d} a_{jk} > 0$ for all $k \in R$. Consider the random vector

$$\boldsymbol{M} = \left(\max_{k \in R} \{a_{1k} Z_k\}, \ldots, \max_{k \in R} \{a_{dk} Z_k\} \right). \tag{2.3.9}$$

Its distribution is a max-linear model; see, e.g., Wang and Stoev [100] and Einmahl et al. [32]. As in linear factor models, the random variables $Z_1, \ldots, Z_r$ can be thought of as latent factor variables generating the observable random vector. The random vector $\boldsymbol{M}$ has an mev distribution G^* with unit-Fréchet margins. The angular measure of G^* with respect to $\| \cdot \|$ is a discrete measure with mass $h_k = \|\boldsymbol{a}_{\cdot k}\|$ at atom $\boldsymbol{w}_k = \boldsymbol{a}_{\cdot k}/h_k$, where $\boldsymbol{a}_{\cdot k}$ denotes the k-th column of A for $k \in R$. By Proposition 2.3.2, the extreme directions of $\boldsymbol{M}$ are the sets $J_k = \left\{ j \in D : \ a_{jk} > 0 \right\}$ for $k \in R$. The construction at the end of Section 2.2.2 yields a generator $\boldsymbol{U}$ of the mgp distribution associated to (2.3.9) with distribution

$$\mathcal{L}(\boldsymbol{U}) = \frac{1}{d} \sum_{k \in R} \|\boldsymbol{a}_{\cdot k}\|_1 \, \delta_{\ln\left(\frac{\boldsymbol{a}_{\cdot k}}{\|\boldsymbol{a}_{\cdot k}\|_1} \right)} (\cdot),$$

with $\delta_{\boldsymbol{x}}(\cdot)$ denoting a unit point mass at $\boldsymbol{x}$. Hence, inequality (6) in Theorem 2.3.6 is satisfied if and only if J is one of the signatures J_k, $k \in R$.

Remark 2.3.8. The tail dependence coefficient χ_J in (2.3.1) satisfies

$$\chi_J = \mu\left(\left\{ \boldsymbol{x} \in \mathbb{E} : \ \boldsymbol{x}_J \in (1, \infty)^J \right\} \right) = \sum_{\substack{\bar{J} \subseteq D \\ \bar{J} \supseteq J}} \mu\left(\left\{ \boldsymbol{x} \in \mathbb{E} : \ \boldsymbol{x}_{\bar{J}} \in (1, \infty)^{|\bar{J}|}, \ \boldsymbol{x}_{D \setminus \bar{J}} \in [0,1]^{|D \setminus \bar{J}|} \right\} \right).$$

We have $\chi_J > 0$ if and only if there exists $\bar{J} \supseteq J$ such that the corresponding term on the right-hand side is positive, and by homogeneity of μ, this is equivalent to $\bar{J}$ being a subset of an extreme direction. We find that $\chi_J > 0$ if and only if J is a subset of an extreme direction.

2.3.2 Density of an mgp distribution with multiple extreme directions

We provide density expressions for potential use in statistical inference on mgp models. As in Section 2.2.2, let $\boldsymbol{Y}$ be an mgp random vector with distribution $H = \mathrm{GP}(G)$ where G is an mev distribution with Gumbel margins. Further, let $\boldsymbol{T}$

and $\boldsymbol{U}$ be its generators as defined in Section 2.2. With respect to a dominating measure υ, defined as a sum of Lebesgue measures λ_m of various dimensions $m \in D$, we express the density of $\boldsymbol{Y}$ in terms of the densities of $\boldsymbol{U}$ and $\boldsymbol{T}$. This will be helpful since the latter densities are known for several parametric families, for instance, the logistic and the Hüsler–Reiss model [66, Sections 7.1 and 7.2]. Furthermore, in view of (2.2.9), the density of the exponent measure μ on $\{\boldsymbol{x} \geqslant \boldsymbol{0} : \boldsymbol{x} \not\leqslant \boldsymbol{1}\}$ is proportional to the mgp density, after appropriate marginal scaling; by homogeneity, this determines the exponent measure density everywhere.

Let λ_k be the k-dimensional Lebesgue measure and let $\upsilon = \upsilon_d$ be the measure on $[-\infty, \infty)^d$ defined by

$$\upsilon(\,\cdot\,) := \sum_{J:\varnothing\neq J\subseteq D} \lambda_{|J|}(\pi_J(\,\cdot\, \cap \mathbb{A}_J)), \tag{2.3.10}$$

where we recall the set $\mathbb{A}_J$ and the projection π_J defined in and right after Eq. (2.2.1). For Borel sets B in $[-\infty, \infty)^d$, we have thus

$$\upsilon(B) = \sum_{J:\varnothing\neq J\subseteq D} \lambda_{|J|}\left(\left\{\boldsymbol{x} \in \mathbb{R}^{|J|} : \exists\, \boldsymbol{y} \in B \text{ such that } y_j = x_j \text{ if } j \in J \text{ and } y_j = -\infty \text{ if } j \notin J\right\}\right).$$

Theorem 2.3.9. *If the generator $\boldsymbol{U}$ of $\boldsymbol{Y}$ has density $f_{\boldsymbol{U}}$ with respect to υ, then the mgp random vector $\boldsymbol{Y}$ has density $h_{\boldsymbol{Y}}$ with respect to υ equal to*

$$h_{\boldsymbol{Y}}(\boldsymbol{y}) = \mathbb{1}\{\max(\boldsymbol{y}) > 0\} \frac{1}{\mathsf{E}\left[\mathrm{e}^{\max(\boldsymbol{U})}\right]} \int_{r\in\mathbb{R}} f_{\boldsymbol{U}}\left(r + \boldsymbol{y}\right) e^r \,\mathrm{d}r.$$

Theorem 2.3.10. *If the generator $\boldsymbol{T}$ has density $f_{\boldsymbol{T}}$ with respect to υ, then the mgp random vector $\boldsymbol{Y}$ has density $h_{\boldsymbol{Y}}$ with respect to υ equal to*

$$h_{\boldsymbol{Y}}(\boldsymbol{y}) = \mathbb{1}\{\max(\boldsymbol{y}) > 0\}\, \mathrm{e}^{-\max(\boldsymbol{y})} \int_{r\in\mathbb{R}} f_{\boldsymbol{T}}(r + \boldsymbol{y})\,\mathrm{d}r.$$

The proofs of Theorems 2.3.9 and 2.3.10 are identical to the ones of Theorems 12 and 13 in Rootzén et al. [84]. In the next section, we will propose a model for the $\boldsymbol{U}$-generator that leads to mgp distributions with general extreme directions and whose density can be computed by means of Theorem 2.3.9.

2.4 A model with extreme directions

We will provide a stochastic construction that allows any collection $\mathcal{R}$ of non-empty subsets of $D = \{1, \ldots, d\}$ with the property $\bigcup\{J : J \in \mathcal{R}\} = D$ to be the extreme directions (Definition 2.3.1) of a random vector. We do so by considering a smoothed version of the max-linear model that we will call mixture model (Section 2.4.1). We

show that this model determines an mev distribution, of which we compute the stdf, the extreme directions, and the $\boldsymbol{U}$-generator of the associated mgp distribution. Importantly, the mixture model covers all mev distributions and thus also all mgp distributions. Density expressions will be provided too. Next, in Section 2.4.2, we work out two parametric examples for the mixture model based on the logistic and the Hüsler–Reiss dependence structures.

2.4.1 Definition and properties

Definition 2.4.1. Let r be a positive integer and write $R = \{1, \dots, r\}$. Let $\boldsymbol{Z}^{(1)}, \dots, \boldsymbol{Z}^{(r)}$ be independent random vectors in $\mathbb{R}^d$. For all $k \in R$, suppose that $\boldsymbol{Z}^{(k)} = (Z_1^{(k)}, \dots, Z_d^{(k)})$ has an mev distribution with unit-Fréchet margins, stdf $\ell^{(k)}$ and a single extreme direction D. Let $A = (a_{jk})_{j \in D; k \in R} \in [0,1]^{d \times r}$ be a coefficient matrix with unit row sums $\sum_{k=1}^{r} a_{jk} = 1$ for all $j \in D$ and positive column sums $\sum_{j=1}^{d} a_{jk} > 0$ for all $k \in R$. The d-variate random vector

$$\boldsymbol{M} = (M_1, \dots, M_d) = \left(\max_{k \in R} \{a_{1k} Z_1^{(k)}\}, \dots, \max_{k \in R} \{a_{dk} Z_d^{(k)}\} \right), \tag{2.4.1}$$

will be called the $(d \times r)$ *mixture model* with coefficient matrix A and factors $\boldsymbol{Z}^{(1)}, \dots, \boldsymbol{Z}^{(r)}$.

Remark 2.4.2. In (2.4.1), for a given $j \in D$, only those $k \in R$ that satisfy $a_{jk} > 0$ matter for the value of the maximum. Therefore, only the subvectors $\boldsymbol{Z}_{J_k}^{(k)}$ need to be given, where

$$J_k := \left\{ j \in D : \ a_{jk} > 0 \right\}, \tag{2.4.2}$$

is the k-th *signature* of the coefficient matrix A. The condition on $\boldsymbol{Z}_{J_k}^{(k)}$ is then that its only extreme direction is J_k. Note that J_k is non-empty by the assumption that A has positive column sums.

Remark 2.4.3. In case $Z_1^{(k)} = \ldots = Z_d^{(k)}$ almost surely for all $k \in R$, the distribution of $\boldsymbol{M}$ is max-linear with unit-Fréchet margins and we are back to the factor model in Example 2.3.7. The mixture model (2.4.1) can thus be seen as a smoothed version of the max-linear factor model in the sense that complete dependence within each of the factors $\boldsymbol{Z}^{(1)}, \dots, \boldsymbol{Z}^{(r)}$ is relaxed to a possibly weaker dependence structure. Under the conditions of Proposition 2.4.11 below, the resulting mixture model has a density with respect to the measure υ in Eq. (2.3.10). We hope this property will facilitate statistical inference through likelihood-based methods.

For $k \in R$, let $\ell_{J_k}^{(k)}$ be the stdf associated with $\boldsymbol{Z}_{J_k}^{(k)}$, that is,

$$\ell_{J_k}^{(k)}(\boldsymbol{x}_{J_k}) = \ell^{(k)}\left(\tilde{\boldsymbol{x}}\right), \qquad \boldsymbol{x}_{J_k} \in [0, \infty)^{J_k}, \tag{2.4.3}$$

with $\tilde{\boldsymbol{x}} = \sum_{j \in J_k} x_j \boldsymbol{e}_j$ and $\boldsymbol{e}_j$ the j-th canonical unit vector in $\mathbb{R}^d$. If J_k is a singleton, then $\ell^{(k)}_{J_k}$ is the identity function on $[0, \infty)$. The following theorem was proved in Stephenson [94, Theorem 1] for the particular case where $J_1, \ldots, J_r$ are all different.

Theorem 2.4.4. *The random vector $\boldsymbol{M}$ in* (2.4.1) *follows an mev distribution denoted by $G_{\boldsymbol{M}}$ with unit-Fréchet margins and stdf*

$$\ell(\boldsymbol{x}) = \sum_{k \in R} \ell^{(k)}_{J_k} \left(\left(a_{jk} x_j \right)_{j \in J_k} \right), \qquad \boldsymbol{x} \in [0, \infty)^d. \tag{2.4.4}$$

In the context of anomaly detection, Chiapino et al. [14] define a mixture model for the angular measure of $G_{\boldsymbol{M}}$ in terms of probability measures on the faces of the unit simplex. In contrast, the mixture model in Definition 2.4.1 is obtained by an aggregation of lower-dimensional mev distributions.

Zero entries of the coefficient matrix A allow some groups of components of the random vector $\boldsymbol{M}$ to constitute an extreme direction. Recall J_k in (2.4.2).

Proposition 2.4.5. *The set of extreme directions of the random vector $\boldsymbol{M}$ in* (2.4.1) *is*

$$\mathcal{R} = \{ J \subseteq D : \ J = J_k \text{ for some } k \in R \}.$$

Remark 2.4.6. Given any collection $\mathcal{R} = \{J_1, \ldots, J_r\}$ of distinct non-empty subsets of D with the property $\bigcup_{k=1}^r J_r = D$, we can easily construct a $d \times r$ coefficient matrix A as in Definition 2.4.1 such that equality (2.4.2) holds for all k (put zeros and ones on the appropriate places and then normalize the rows). The set of extreme directions of the mixture model (2.4.1) is then exactly $\mathcal{R}$.

Let $\tilde{G}_{\boldsymbol{M}}$ be the Gumbel standardization of $G_{\boldsymbol{M}}$, that is, $\tilde{G}_{\boldsymbol{M}}$ is the cdf of $(\ln M_j)_{j \in D}$. We will investigate the mgp distribution

$$H = \mathrm{GP}(\tilde{G}_{\boldsymbol{M}}) \tag{2.4.5}$$

as defined in Section 2.2.2. We will refer to H as the *mgp distribution associated with the mixture model $\boldsymbol{M}$* and to $\boldsymbol{Y} \sim H$ as the *mgp random vector associated with the mixture model $\boldsymbol{M}$*. For a non-empty set $J \subseteq D$, recall the set $\mathbb{A}_J$, the projection π_J defined in Eq. (2.2.1), and the representation $\boldsymbol{U}$ of the mgp distribution H as in Section 2.2.2. For $k \in R$, since the only extreme direction of the $\mathbb{R}^{J_k}$-valued random vector $\boldsymbol{Z}^{(k)}_{J_k}$ is J_k, Theorem 2.3.6 yields the existence of a generator $\boldsymbol{U}^{(k)}$ taking values in $\mathbb{R}^{J_k}$ which satisfies (2.2.15) and

$$\mathsf{E}\left[\max_{j \in J_k} \left\{ x_j \mathrm{e}^{U^{(k)}_j} \right\} \right] = \ell^{(k)}_{J_k} \left(\boldsymbol{x}_{J_k} \right), \qquad \boldsymbol{x}_{J_k} \in [0, \infty)^{J_k}. \tag{2.4.6}$$

The following proposition provides a $\boldsymbol{U}$-generator of H in terms of $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(r)}$.

Proposition 2.4.7. *For any* $\boldsymbol{m} \in (0,1)^r$ *such that* $\sum_{k=1}^r m_k = 1$*, the random vector* $\boldsymbol{U}$ *in* $[-\infty,\infty)^d$ *with distribution*

$$\mathsf{P}\left[\boldsymbol{U} \in \cdot\,\right] = \sum_{k\in R} m_k\, \mathsf{P}\left[\boldsymbol{U}^{(k)} + \left(\ln\left(a_{jk}/m_k\right)\right)_{j\in J_k} \in \pi_{J_k}\left(\mathbb{A}_{J_k} \cap \cdot\right)\right] \tag{2.4.7}$$

is a $\boldsymbol{U}$*-generator of the mgp distribution* H *in* (2.4.5)*, i.e.,* $H = \mathrm{GP}_{\boldsymbol{U}}(\mathcal{L}(\boldsymbol{U}))$*.*

The random vector $\boldsymbol{U}$ in (2.4.7) will be called a mixture generator with coefficient matrix A, generators $\boldsymbol{U}^{(1)},\dots,\boldsymbol{U}^{(r)}$ and mass vector $\boldsymbol{m}$. Different choices of the vector $\boldsymbol{m}$ yield different $\boldsymbol{U}$ generators, all of which however induce the *same* mgp distribution H, namely, the one associated to the mixture model. Among all possible choices for $\boldsymbol{m}$, the obvious one is $m_k = 1/r$ for all $k \in \{1,\dots,r\}$, but the formula allows for other possibilities. Further, note that m_k is *not* equal to the probability that the mgp vector $\boldsymbol{Y} \sim H$ lies in extreme direction J_k: instead, we have $\mathsf{P}[Y_j > -\infty \text{ iff } j \in J_k] = w_k$, a quantity which is proportional to $\ell^{(k)}_{J_k}((a_{jk})_{j\in J_k})$, as stated in Proposition 2.5.1 below.

Up to now, we introduced the mixture model and we investigated the associated mgp distribution. Conversely, the following theorem shows the generality of the construction in the sense that it accommodates all mgp distributions. We say that the coefficient matrix A has *mutually different signatures* if the sets $J_1,\dots,J_r$ in (2.4.2) are all distinct.

Theorem 2.4.8. *Any mev distribution with unit-Fréchet margins is the cdf of a mixture model whose coefficient matrix has mutually different signatures. As a consequence, any mgp distribution is associated with a mixture model whose coefficient matrix has mutually different signatures.*

The coefficient matrix and the factors of the mixture model in Theorem 2.4.8 are specified in the proof. Applying Theorem 2.4.8 to the mixture model itself yields the following corollary.

Corollary 2.4.9. *Let* $\boldsymbol{M}$ *be the mixture model in Definition 2.4.1. There exist an integer* $1 \leqslant s \leqslant r$ *and a* $(d \times s)$ *mixture model* $\boldsymbol{N}$ *whose coefficient matrix has mutually different signatures such that* $\boldsymbol{M}$ *and* $\boldsymbol{N}$ *have the same cdf and consequently the same associated mgp distribution.*

Remark 2.4.10. On the one hand, Corollary 2.4.9 suggests to restrict mixture models to the case where all columns of the coefficient matrix have different signatures. On the other hand, for statistical practice, it might be difficult to combine the dependence structures of two columns with the same signature and different stdf's.

Let $\boldsymbol{Y} \sim H$ be the mgp random vector associated with the mixture model $\boldsymbol{M}$. Using Theorem 2.3.9, we express the density of $\boldsymbol{Y}$ with respect to the measure υ in (2.3.10).

Proposition 2.4.11. *Consider the mixture model $\boldsymbol{M}$ in Definition 2.4.1 and recall its stdf ℓ in (2.4.4). For all $k \in R$, suppose that the generator $\boldsymbol{U}^{(k)}$ of $\boldsymbol{Z}^{(k)}$ is absolutely continuous with respect to $\lambda_{|J_k|}$ with Lebesgue density $f_{\boldsymbol{U}^{(k)}}$. The mgp random vector $\boldsymbol{Y}$ with distribution H in (2.4.5) is then absolutely continuous with respect to υ with density $h_{\boldsymbol{Y}}$ given for $\boldsymbol{y} \in [-\infty, \infty)^d$ with $\max(\boldsymbol{y}) > 0$ by*

$$h_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{k \in R} \frac{\mathbb{1}_{\mathbb{A}_{J_k}}(\boldsymbol{y})}{\ell(\mathbf{1})} \int_{z \in \mathbb{R}} f_{\boldsymbol{U}^{(k)}} \left(\left(y_j - \ln a_{jk} + z \right)_{j \in J_k} \right) e^z \, \mathrm{d}z. \tag{2.4.8}$$

Proposition 2.4.11 will be helpful in the next subsection when we let each of the factors $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(r)}$ follow a parametric model.

2.4.2 Parametric examples

We construct two parametric families of the mixture model $\boldsymbol{M}$ in Definition 2.4.1 with coefficient matrix A by choosing particular forms for the factors $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(r)}$: the logistic model and the Hüsler–Reiss model. In each case, we compute the stdf ℓ, the generator $\boldsymbol{U}$ and the density $h_{\boldsymbol{Y}}$ of the associated mgp distribution.

Mixture logistic model

For $k \in R$, recall J_k in (2.4.2) and suppose that the stdf associated with $\boldsymbol{Z}^{(k)}_{J_k}$ is

$$\ell^{(k)}_{J_k}(\boldsymbol{y}) = \left(\sum_{j \in J_k} y_j^{1/\alpha_k} \right)^{\alpha_k}, \tag{2.4.9}$$

for $\boldsymbol{y} \in [0, \infty)^{J_k}$, where $0 < \alpha_k < 1$. The only extreme direction of $\boldsymbol{Z}^{(k)}_{J_k}$ is J_k, so that, following Definition 2.4.1 and Remark 2.4.2, the *mixture logistic model* $\boldsymbol{M}$ is well-defined. Let $G_{\boldsymbol{M}}$ denote the cdf of $\boldsymbol{M}$; note that $G_{\boldsymbol{M}}$ is an mev distribution with unit-Fréchet margins. We compute its stdf using Theorem 2.4.4.

Corollary 2.4.12. *The stdf of the mixture logistic model $\boldsymbol{M}$ is*

$$\ell(\boldsymbol{y}) = \sum_{k \in R} \left\{ \sum_{j \in J_k} \left(a_{jk} y_j \right)^{1/\alpha_k} \right\}^{\alpha_k}, \qquad \boldsymbol{y} \in [0, \infty)^d. \tag{2.4.10}$$

Remark 2.4.13. In case the sets $J_1, \ldots, J_r$ are all different, the stdf ℓ in (2.4.10) coincides with the one of the asymmetric logistic model introduced in Tawn [96].

Next we construct a $\boldsymbol{U}$-generator of the mgp distribution $H = \mathrm{GP}(\tilde{G}_{\boldsymbol{M}})$, where $\tilde{G}_{\boldsymbol{M}}$ is the cdf of $(\ln M_1, \ldots, \ln M_d)$, an mev distribution with Gumbel margins.

Let $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(r)}$ be random vectors in $\mathbb{R}^{J_1}, \ldots, \mathbb{R}^{J_r}$, respectively, such that

$$\begin{aligned} &\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(r)} \text{ are independent,} \\ &U_j^{(k)} \overset{\mathrm{d}}{=} \ln\left(\frac{X_j^{\alpha_k}}{\Gamma(1-\alpha_k)}\right), \qquad k \in R,\ j \in J_k, \end{aligned} \tag{2.4.11}$$

where $X_1, \ldots, X_d$ are independent unit-Fréchet random variables and where Γ denotes the gamma function.

Proposition 2.4.14. *Let $\boldsymbol{m} \in (0,1)^r$ be such that $\sum_{k=1}^r m_k = 1$. The distribution* (2.4.7) *with coefficient matrix A, generators $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(r)}$ in* (2.4.11) *and mass vector $\boldsymbol{m}$ is the one of a $\boldsymbol{U}$-generator of the mgp distribution $H = \mathrm{GP}(\tilde{G}_{\boldsymbol{M}})$ for the mixture logistic model $\boldsymbol{M}$.*

Let $\boldsymbol{Y} \sim H$ be an mgp random vector associated with the mixture logistic model $\boldsymbol{M}$. The density of $\boldsymbol{Y}$ with respect to υ in (2.3.10) follows from Proposition 2.4.11.

Proposition 2.4.15. *The mgp random vector $\boldsymbol{Y} \sim H$ associated with the mixture logistic model $\boldsymbol{M}$ in* (2.4.9) *is absolutely continuous with respect to the measure υ with density*

$$h_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{k \in R} \mathbb{1}_{\mathbb{A}_{J_k}}(\boldsymbol{y}) \cdot \frac{(1/\alpha_k)^{|J_k|-1}\, \Gamma\left(|J_k| - \alpha_k\right) \prod_{j \in J_K} \left\{ \left(a_{jk} \mathrm{e}^{-y_j}\right)^{1/\alpha_k} \right\}}{\ell(\mathbf{1})\, \Gamma\left(1-\alpha_k\right) \left\{ \sum_{j \in J_k} \left(a_{jk} \mathrm{e}^{-y_j}\right)^{1/\alpha_k} \right\}^{|J_k| - \alpha_k}}, \tag{2.4.12}$$

for $\boldsymbol{y} \in [-\infty, \infty)^d \setminus [-\infty, 0]^d$, with ℓ as in (2.4.10).

Mixture Hüsler–Reiss model

For $k \in R$, recall J_k in (2.4.2) and suppose that $\boldsymbol{Z}_{J_k}^{(k)}$ has Hüsler–Reiss mev distribution [58] with variogram matrix[1] $\Gamma^{(k)} \in \mathbb{R}^{J_k \times J_k}$; see, e.g., Huser and Davison [57]. If J_k is a singleton, the stdf $\ell_{J_k}^{(k)}$ of $\boldsymbol{Z}_{J_k}^{(k)}$ reduces to the identity function on $[0, \infty)$. Otherwise,

$$\ell_{J_k}^{(k)}(\boldsymbol{y}) = \sum_{j \in J_k} y_j\, \Phi_{|J_k|-1}\left(\eta^{j,(k)}\left(\boldsymbol{y}\right); \Sigma^{j,(k)}\right), \qquad \boldsymbol{y} \in [0,\infty)^{J_k},$$

where for $j \in J_k$, $\eta^{j,(k)}\left(\boldsymbol{y}\right)$ is equal to $\left\{\ln(y_j/y_s) + \Gamma_{js}^{(k)}/2\right\}_{s \in J_k; s \neq j}$ with $y_j/0 = \infty$ if $y_j > 0$ and ∞ if $y_j = 0$, where the matrix $\Sigma^{j,(k)}$ is defined by

$$\Sigma^{j,(k)} := \frac{1}{2}\left\{\Gamma_{js}^{(k)} + \Gamma_{jt}^{(k)} - \Gamma_{st}^{(k)}\right\}_{s,t \in J_k \setminus \{j\}} \in \mathbb{R}^{(J_k \setminus \{j\}) \times (J_k \setminus \{j\})}$$

[1] A matrix $\Gamma \in \mathbb{R}^{d \times d}$ is a variogram matrix if it is symmetric, has zero diagonal, and satisfies $\boldsymbol{v}^\top \Gamma \boldsymbol{v} < 0$ for all $\mathbf{0} \neq \boldsymbol{v}^\top \perp \mathbf{1}$. To each positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is associated a variogram matrix $\Gamma \in \mathbb{R}^{d \times d}$ via $\Gamma_{st} = \Sigma_{ss} + \Sigma_{tt} - 2\Sigma_{st}$ for $s, t \in D$.

and finally, where $\Phi_m(\,\cdot\,;\Sigma)$ is the cdf of the centered m-variate normal distribution function with covariance matrix Σ. The only extreme direction of $\boldsymbol{Z}^{(k)}_{J_k}$ is J_k as required in Definition 2.4.1 and Remark 2.4.2. The resulting $\boldsymbol{M}$ in (2.4.1) is called the *mixture Hüsler–Reiss model*. Using Theorem 2.4.4, we compute the stdf of its cdf $G_{\boldsymbol{M}}$.

Proposition 2.4.16. *The stdf ℓ of $G_{\boldsymbol{M}}$ is*

$$\ell(\boldsymbol{y}) = \sum_{k\in R}\left\{\sum_{j\in J_k} a_{jk}y_j\,\Phi_{|J_k|-1}\left(\eta^{j,(k)}\left(\left(a_{jk}y_j\right)_{j\in J_k}\right);\Sigma^{j,(k)}\right)\right\}, \qquad \boldsymbol{y}\in[0,\infty)^d. \tag{2.4.13}$$

To construct a generator of the mgp distribution $H = \mathrm{GP}(\tilde{G}_{\boldsymbol{M}})$ of the mixture Hüsler–Reiss model, let $k\in R$ and let $\tilde{\Sigma}^{(k)}\in\mathbb{R}^{J_k\times J_k}$ be a positive definite covariance matrix with variogram matrix $\Gamma^{(k)}$; see, e.g., Segers [88] for a possible construction. Let $\boldsymbol{U}^{(1)},\dots,\boldsymbol{U}^{(r)}$ be random vectors in $\mathbb{R}^{J_1},\dots,\mathbb{R}^{J_r}$, respectively, such that

$$\begin{aligned}&\boldsymbol{U}^{(1)},\dots,\boldsymbol{U}^{(r)} \text{ are independent,}\\ &\boldsymbol{U}^{(k)}\sim\mathcal{N}_{|J_k|}\left(-\tfrac{1}{2}\operatorname{diag}\tilde{\Sigma}^{(k)},\,\tilde{\Sigma}^{(k)}\right), \qquad k\in R.\end{aligned} \tag{2.4.14}$$

Proposition 2.4.17. *Let $\boldsymbol{m}\in(0,1)^r$ be such that $\sum_{k=1}^r m_k = 1$. The distribution in (2.4.7) with coefficient matrix A, generators $\boldsymbol{U}^{(1)},\dots,\boldsymbol{U}^{(r)}$ in (2.4.14) and mass vector $\boldsymbol{m}$ is the one of a $\boldsymbol{U}$-generator of the mgp distribution H of the mixture Hüsler–Reiss model.*

Let $\boldsymbol{Y}\sim H$ be an mgp random vector for $H=\mathrm{GP}(\tilde{G}_{\boldsymbol{M}})$ and $\boldsymbol{M}$ following the mixture Hüsler–Reiss model. The following proposition gives the density of $\boldsymbol{Y}$ with respect to the measure υ in (2.3.10). Recall the sets $\mathbb{A}_J$ in (2.2.1).

Proposition 2.4.18. *The mgp random vector $\boldsymbol{Y}\sim H$ associated with the mixture Hüsler–Reiss model with variogram matrices $\Gamma^{(k)}$ for $k\in R$ and coefficient matrix A is absolutely continuous with respect to the measure υ with density*

$$h_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{k\in R}\mathbb{1}_{\mathbb{A}_{J_k}}(\boldsymbol{y})\;\cdot\;\frac{\lambda\left(\boldsymbol{y}_{J_k}-\left(\ln a_{jk}\right)_{j\in J_k};\Gamma^{(k)}\right)}{\ell(\mathbf{1})}, \tag{2.4.15}$$

for $\boldsymbol{y}\in[-\infty,\infty)^d\setminus[-\infty,0]^d$, with ℓ as in (2.4.13) and where for $k\in R$, the function $\lambda(\,\cdot\,;\Gamma^{(k)}):\mathbb{R}^{J_k}\to[0,\infty)$ is the density of the exponent measure of a Hüsler–Reiss mev distribution with Gumbel margins.

Remark 2.4.19. An expression for $\lambda(\,\cdot\,;\Gamma^{(k)})$ can be obtained for instance from Hentschel et al. [53, Definition 3.1]: writing $d(k) := \max(J_k)$, we have

$$\lambda\left(\boldsymbol{x};\Gamma^{(k)}\right) = \frac{\exp\left(-x_{d(k)}\right)}{\sqrt{(2\pi)^{|J_k|-1}\det\left(\Sigma^{d(k),(k)}\right)}} \cdot \exp\left(-\frac{1}{2}\left\|\left(\boldsymbol{x}_{J_k\setminus\{d(k)\}}\right)^\top - x_{d(k)}\mathbf{1}_{|J_k|-1} - \left(\boldsymbol{\mu}^{d(k)}\right)^\top\right\|^2_{\Theta^{d(k)}}\right)$$

with $\mathbf{1}_m = (1,\ldots,1) \in \mathbb{R}^m$ and

$$\Theta^{d(k)} = \left(\Sigma^{d(k),(k)}\right)^{-1}, \qquad \boldsymbol{\mu}^{d(k)} = \left\{-\frac{1}{2}\Gamma_{sd(k)}\right\}_{s\in J_k\setminus\{d(k)\}},$$

and, finally, $\|\boldsymbol{g}\|^2_M = \boldsymbol{g}^\top M\boldsymbol{g}$ for $M \in \mathbb{R}^{p\times p}$ and $\boldsymbol{g} \in \mathbb{R}^{p\times 1}$. In fact, replacing $d(k)$ by any other choice of index from J_k would give the same function.

2.5 Simulation from mixture model mgp distribution

In Section 2.5.1, we propose a generic algorithm to simulate random samples from the mgp distribution associated to the mixture model in Definition 2.4.1. The details of the algorithm for the mixture logistic model and the mixture Hüsler–Reiss model are worked out in Section 2.5.2.

2.5.1 Generic simulation algorithm

Recall the mixture model $\boldsymbol{M}$ in Definition 2.4.1. The goal of this section is to generate a random variable $\boldsymbol{Y}$ from the associated mgp distribution H in (2.4.5) and with density in Proposition 2.4.11.

Let $\boldsymbol{U}$ with mixture distribution in (2.4.7) be a $\boldsymbol{U}$-generator of H and let $\boldsymbol{T}$ be the $\boldsymbol{T}$-generator obtained from $\boldsymbol{U}$ via the measure transformation in (2.2.17). From (2.2.13), one can easily obtain $\boldsymbol{Y}$ from $\boldsymbol{T}$ and an independent unit-exponential random variable. The remaining question is how to simulate random samples from the law of $\boldsymbol{T}$. The following lemma expresses the random vector $\boldsymbol{T}$ as a mixture over $k \in R$ of random vectors $\boldsymbol{T}^{(k)}$ related to generators $\boldsymbol{U}^{(k)}$ in (2.4.6). Recall the notation introduced in the beginning of Section 2.2, in particular the sets $\mathbb{A}_J$ and the projection π_J in (2.2.1).

Proposition 2.5.1. *The $\boldsymbol{T}$-generator in* (2.2.17) *resulting from the mixture generator $\boldsymbol{U}$ in* (2.4.7) *satisfies*

$$\mathsf{P}[\boldsymbol{T} \in \cdot\,] = \sum_{k \in R} w_k \, \mathsf{P}\left[\boldsymbol{T}^{(k)} \in \pi_{J_k}\left(\mathbb{A}_{J_k} \cap \cdot\right)\right],$$

with J_k as in (2.4.2) *and where the distribution of the $|J_k|$-dimensional random vector $\boldsymbol{T}^{(k)}$ is*

$$\mathsf{P}\left[\boldsymbol{T}^{(k)} \in \cdot\,\right] = \frac{\mathsf{E}\left[\exp\left\{\max \boldsymbol{P}^{(k)}\right\} \mathbb{1}\left\{\boldsymbol{P}^{(k)} \in \cdot\,\right\}\right]}{\mathsf{E}\left[\exp\left\{\max \boldsymbol{P}^{(k)}\right\}\right]}, \qquad w_k = \frac{\ell_{J_k}^{(k)}\left(\left(a_{jk}\right)_{j \in J_k}\right)}{\ell\left(\mathbf{1}\right)}, \tag{2.5.1}$$

where $\boldsymbol{P}^{(k)} = \boldsymbol{U}^{(k)} + \left(\ln\left(a_{jk}/m_k\right)\right)_{j \in J_k}$ with $\boldsymbol{U}^{(k)}$ in (2.4.6), *$\ell_{J_k}^{(k)}$ in* (2.4.3) *and ℓ in* (2.4.4). *Moreover, if $\boldsymbol{U}^{(k)}$ has density $f_{\boldsymbol{U}^{(k)}}$ with respect to $\lambda_{|J_k|}$, then $\boldsymbol{T}^{(k)}$ has density*

$$f_{\boldsymbol{T}^{(k)}}(\boldsymbol{t}) = c^{(k)} \, g_k(\boldsymbol{t}) \sum_{j \in J_k} n_{j,k} \, q_{j,k}(\boldsymbol{t}), \qquad \boldsymbol{t} \in \mathbb{R}^{J_k}, \tag{2.5.2}$$

with $g_k(t) = \mathrm{e}^{\max(\boldsymbol{t})} / \sum_{j \in J_k} \mathrm{e}^{t_j}$ and

$$c^{(k)} := \frac{\sum_{j \in J_k} a_{jk}}{\ell_{J_k}^{(k)}\left(\left(a_{jk}\right)_{j \in J_k}\right)},$$

and, for $j \in J_k$,

$$n_{j,k} := \frac{a_{jk}}{\sum_{i \in J_k} a_{ik}}, \tag{2.5.3}$$

$$q_{j,k}\left(\boldsymbol{t}\right) := \frac{\mathrm{e}^{t_j} f_{\boldsymbol{U}^{(k)}}\left(\boldsymbol{t} - \left(\ln\left(a_{ik}/m_k\right)\right)_{i \in J_k}\right)}{a_{jk}/m_k}. \tag{2.5.4}$$

In view of Proposition 2.5.1, we can simulate random samples from the distribution of $\boldsymbol{T}$ once we can do so for the distributions of $\boldsymbol{T}^{(1)}, \ldots, \boldsymbol{T}^{(r)}$. The representation (2.5.2) features a mixture over exponentially tilted versions of the density of $\boldsymbol{U}^{(k)}$ times a uniformly bounded function. This opens the door to simulation via a rejection sampling algorithm.

As a consequence, we can simulate random samples from the distribution of $\boldsymbol{T}^{(k)}$ in (2.5.1) via rejection sampling and random vectors with densities $q_{j,k}$ for $j \in J_k$. The algorithm hence requires that $\boldsymbol{U}^{(k)}$ in (2.4.6) is absolutely continuous and that one can simulate from $q_{j,k}$ for all $k \in R$ and $j \in J_k$. The procedure is summarized in Algorithm 1.

The complexity of Algorithm 1 is driven by the expected number of times one has to simulate from densities $\left(q_{j,k}\right)_{k \in R, j \in J_k}$ in (2.5.4) and is given by $d/\ell(\mathbf{1})$ with

Algorithm 1 Simulation of an mgp random vector associated with a mixture model

Input: integers $d \geqslant 1$ and $r \geqslant 1$, a matrix $A \in [0,1]^{d \times r}$ as in Definition 2.4.1 and distributions $\mathcal{L}\left(\boldsymbol{U}^{(1)}\right), \ldots, \mathcal{L}\left(\boldsymbol{U}^{(r)}\right)$ satisfying (2.4.6).

Output: an mgp vector $\boldsymbol{Y}$ with distribution H generated by $\boldsymbol{U}$ in (2.4.7) with matrix A and generators $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(r)}$.

Define sign$[[k]] := J_k$ in (2.4.2) for $k \in \{1, \ldots, r\}$. ▷ sign is the list of signatures of A.
Define the d-dimensional vector $\boldsymbol{T} := (-\infty, \ldots, -\infty)$.
Simulate b from $\{1, \ldots, r\}$ such that $\mathsf{P}(b = k) = w_k$ in (2.5.1) for $k \in \{1, \ldots, r\}$.
accept:=FALSE.
while accept=FALSE **do**
    Simulate a from sign$[[b]]$ such that $\mathsf{P}(a = j) = n_{j,b}$ in (2.5.3) for $j \in$ sign$[[b]]$.
    Simulate $\boldsymbol{Q} \sim q_{a,b}$ in (2.5.4).
    Simulate $U_0 \sim \mathrm{Unif}(0,1)$.
    if $U_0 \leqslant \exp(\max \boldsymbol{Q}) / \sum_j \exp\left(Q_j\right)$ **then** accept:=TRUE.
    end if
end while
Update $\boldsymbol{T}_{\mathrm{sign}[[b]]} := \boldsymbol{Q}$.
Simulate a unit exponential random variable E.
Define $\boldsymbol{Y} := \boldsymbol{T} - \max(\boldsymbol{T}) + E$.
return $(\boldsymbol{Y})$.

ℓ in (2.4.4); see Appendix 2.C for the proof. This is equal to the complexity of the algorithm in Engelke and Hitz [34, Subsection 5.4] based on extremal functions $\boldsymbol{\mathcal{U}}^{(1)}, \ldots, \boldsymbol{\mathcal{U}}^{(d)}$ as defined in Dombry et al. [28] and characterized in Engelke and Hitz [34, Lemma 2] in terms of the mgp random vector $\boldsymbol{Y}$ by

$$\boldsymbol{\mathcal{U}}^{(j)} \overset{\mathrm{d}}{=} \exp\left(\boldsymbol{Y} - Y_j\right) \mid Y_j > 0, \qquad j \in D.$$

The following proposition connects these extremal functions to the densities $q_{j,k}$ in (2.5.4).

Proposition 2.5.2. *If $\boldsymbol{U}$ in (2.4.7) has density $f_{\boldsymbol{U}}$ with respect to υ, then for $j \in D$, we have*

$$\boldsymbol{\mathcal{U}}^{(j)} \overset{\mathrm{d}}{=} \exp\left(\boldsymbol{Q}^{(j)} - Q_j^{(j)}\right),$$

where $\boldsymbol{Q}^{(j)}$ is a d-variate random vector with density $\mathrm{e}^{t_j} f_{\boldsymbol{U}}(\boldsymbol{t})$ with respect to υ.

2.5.2 Parametric examples

We use the results in Section 2.5.1 to draw random samples from the mgp distributions associated with the mixture logistic and Hüsler–Reiss models from Section 2.4.2.

Mixture logistic model

We want to simulate the mgp random vector $\boldsymbol{Y}$ associated with the mixture logistic model defined in Section 2.4.2. Let $\boldsymbol{U}$ be the mixture generator in (2.4.7) with matrix A and generators $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(r)}$ in (2.4.11) with parameters $\alpha_1, \ldots, \alpha_r$. The $|J_k|$-variate random vector $\boldsymbol{U}^{(k)}$ has independent Gumbel components and is therefore absolutely continuous with respect to $\lambda_{|J_k|}$. Using Algorithm 1, it remains to generate random samples from the densities $q_{j,k}$ ($j \in J_k$) defined for $\boldsymbol{t} = (t_i)_{i \in J_k} \in \mathbb{R}^{J_k}$ by

$$q_{j,k}(\boldsymbol{t}) = \frac{\mathrm{e}^{t_j} f_{U_j^{(k)}}\left(t_j - \ln\left(a_{jk}/m_k\right)\right)}{a_{jk}/m_k} \cdot \prod_{i \in J_k \setminus \{j\}} f_{U_i^{(k)}}\left(t_i - \ln\left(a_{ik}/m_k\right)\right). \tag{2.5.5}$$

Proposition 2.5.3. *Let $k \in R$ and $j \in J_k$. Let $\boldsymbol{\psi} = (\psi_i)_{i \in J_k}$ have density $q_{j,k}$ in (2.5.5). The distribution of $\boldsymbol{\psi}$ is that of a random vector with independent components, such that*

$$\psi_i \overset{\mathrm{d}}{=} \begin{cases} U_i^{(k)} + \ln\left(a_{ik}/m_k\right), & \text{if } i \in J_k \setminus \{j\}, \\ -\alpha_k \ln N + \ln\left(a_{jk}/\left(m_k \Gamma(1-\alpha_k)\right)\right), & \text{if } i = j, \end{cases}$$

where N is a Gamma random variable, independent from $\boldsymbol{U}^{(k)}$, with shape and rate parameters $1 - \alpha_k$ and 1 respectively.

Mixture Hüsler–Reiss model

In the same way, we want to draw random samples from the mgp random vector $\boldsymbol{Y}$ associated with the mixture Hüsler–Reiss model in Section 2.4.2. Let $\boldsymbol{U}$ be the mixture generator in (2.4.7) with matrix A and generators $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(r)}$ in (2.4.14). Let $k \in R$. The $|J_k|$-variate random vector $\boldsymbol{U}^{(k)}$ has normal density $f_{\boldsymbol{U}^{(k)}}$. Using Algorithm 1, it remains to simulate from the densities $q_{j,k}$ ($j \in J_k$) in Equation (2.5.5), with the difference that the generators $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(r)}$ are defined via (2.4.14) instead of (2.4.11).

Proposition 2.5.4. *Let $j \in J_k$. Let $\boldsymbol{\zeta} = \left(\zeta_j\right)_{i \in J_k}$ have density $q_{j,k}$ in (2.5.5) with generators $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(r)}$ as in (2.4.14). With $\tilde{\Sigma}^{(k)}$ as in (2.4.14), we have*

$$\boldsymbol{\zeta} \sim \mathcal{N}\left(\left(\ln\left(a_{ik}/m_k\right) - \tilde{\Sigma}_{ii}^{(k)}/2 + \tilde{\Sigma}_{ij}^{(k)}\right)_{i \in J_k}, \tilde{\Sigma}^{(k)}\right).$$

Simulation example

The results in this section can be reproduced using the code available on `https://github.com/AnasMourahib/MGPD-simulation`. We use Algorithm 1 with inputs $d = r = 3$, triangular coefficient matrix

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \end{pmatrix},$$

and two different specifications of the generators $\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}$: first, those in (2.4.11) with dependence parameters $\alpha_1 = \alpha_2 = \alpha_3 = 0.5$ to simulate the mgp random vector $\boldsymbol{Y}$ associated with the mixture logistic model; second, those in (2.4.14) based on a variogram matrix with all off-diagonal elements equal to 1.38, to simulate the mgp random vector $\boldsymbol{Y}$ associated with the mixture Hüsler–Reiss model. Figure 2.2 displays the pairwise scatter plots of a sample of size 100 from the transformed mgp random vector

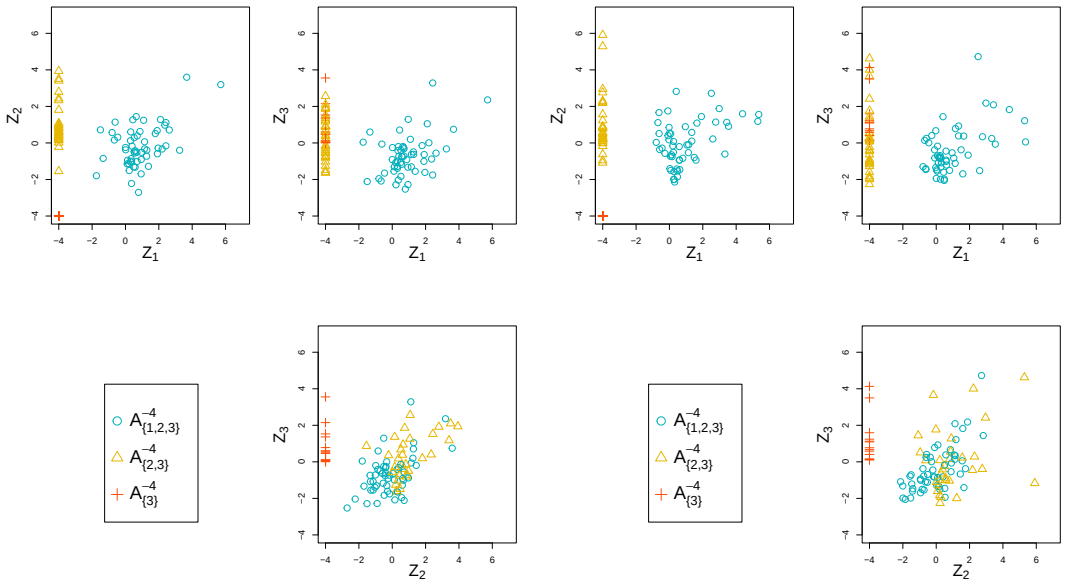
$$\boldsymbol{Z} = 4\left\{\exp\left(\boldsymbol{Y}/4\right) - 1\right\}, \tag{2.5.6}$$

associated with the mixture logistic and Hüsler–Reiss models in panels (a) and (b), respectively. The Box–Cox type transformation in (2.5.6) ensures that the vector of lower endpoints of $\boldsymbol{Z}$ is $\boldsymbol{\eta} = (-4, -4, -4)$ and facilitates visualization. By Proposition 2.4.5, the matrix A produces a mixture model with three distinct extreme directions $J \subset D$, namely $\{1, 2, 3\}$, $\{2, 3\}$, and $\{3\}$. This is visible in Figure 2.2 by a concentration of points on the three corresponding subfaces $\mathbb{A}_J^{-\boldsymbol{4}} = \left\{\boldsymbol{x} \in [-4, \infty)^3 : \ x_j > -4 \text{ iff } j \in J\right\}$.

To assess Algorithm 1, Table 2.1 compares the empirical proportions of points in the three sub-faces $\mathbb{A}_J^{-4}$ in a sample of size 1000 with the corresponding true probabilities in Eq. (2.5.1) for the logistic and Hüsler–Reiss mixture models.

2.6 Conclusion

The extreme directions of a random vector refer to those groups of variables that can be large simultaneously while those outside the group are not. As such scenarios are likely to occur in high dimensions or in spatial contexts, we have studied mgp models featuring arbitrary collections of extreme directions. Furthermore, we have provided a general stochastic representation of mev distributions with arbitrary extreme directions and we have computed their corresponding mgp distributions. The construction involved a mixture specification extending the max-linear model by including multivariate factors. The density of such an mgp distribution—and thus of the exponent measure of the mev distribution—with respect to an appropriate



(a) (b)

Figure 2.2
Sample of size 100 from the transformed mgp random vector $\boldsymbol{Z}$ in (2.5.6) associated with the mixture logistic model in (a) and the mixture Hüsler–Reiss model in (b).

dominating measure has been expressed in terms of those of its generators. In addition, we have provided a simulation algorithm for drawing samples from such mgp distributions. The formulas have been made explicit for two parametric models, extending the classical logistic and Hüsler–Reiss mev and mgp distributions to families with arbitrary extreme directions.

Our work raises intriguing questions for further research. The foremost among these is the challenge of statistical inference for mgp mixture models based on multivariate threshold excesses. This multifaceted problem encompasses parameter estimation, model selection, and the incorporation of sparsity constraints.

Additionally, our research opens doors to the application of mgp mixture models in the realm of graphical models for extremes. Whereas the foundations laid by Engelke and Hitz [34] involve a single extreme direction, the extension offered in Engelke et al. [38] incorporates general exponent measures, enabling a more comprehensive understanding of extreme events in diverse contexts.

Table 2.1
Empirical proportions and true probabilities of lying in each sub-face $\mathbb{A}_J^{-4}$ for the extreme directions $J = \{1,2,3\}$, $\{2,3\}$ and $\{3\}$ based on samples of size 1000 drawn from the logistic and Hüsler–Reiss mixture models by means of Algorithm 1.

	mixture logistic		**mixture Hüsler–Reiss**	
extreme direction J	empirical	true	empirical	true
$\{1,2,3\}$	0.559	0.555	0.560	0.553
$\{2,3\}$	0.297	0.286	0.291	0.288
$\{3\}$	0.144	0.158	0.148	0.157

2.A Proofs of Section 2.3

Proof of Proposition 2.3.2. The monotone convergence theorem and homogeneity of μ yield

$$\begin{aligned}\mu(\mathbb{E}_J) > 0 &\iff \forall r > 0 :\ \mu\left(\{\boldsymbol{x} \in \mathbb{E}_J :\ \|\boldsymbol{x}\| > r\}\right) > 0 \\ &\iff \mu\left(\{\boldsymbol{x} \in \mathbb{E}_J :\ \|\boldsymbol{x}\| > 1\}\right) > 0.\end{aligned}$$

Since $\mu\left(\{\boldsymbol{x} \in \mathbb{E}_J :\ \|\boldsymbol{x}\| > 1\}\right) = \phi(\mathbb{S}_J)$, the conclusion follows. □

Proof of Lemma 2.3.3. Let $\epsilon > 0$. The μ-continuity of $\mathbb{R}^{\epsilon}_{1,J}$ gives

$$n\,\mathsf{P}\left[\boldsymbol{X} \in \mathbb{R}^{\epsilon}_{n,J}\right] = n\,\mathsf{P}\left[n^{-1}\boldsymbol{X} \in \mathbb{R}^{\epsilon}_{1,J}\right] \to \mu\left(\mathbb{R}^{\epsilon}_{1,J}\right), \qquad n \to \infty. \tag{2.A.1}$$

By the dominated convergence theorem, letting ϵ tend to zero yields

$$\lim_{\epsilon \to 0} \mu\left(\mathbb{R}^{\epsilon}_{1,J}\right) = \mu\left(\{\boldsymbol{x} \in \mathbb{E}_J :\ \|\boldsymbol{x}\| > 1\}\right) = \phi\left(\mathbb{S}_J\right), \tag{2.A.2}$$

as $\{\boldsymbol{x} \in \mathbb{E}_J :\ \|\boldsymbol{x}\| > 1\} = \{\boldsymbol{x} \in \mathbb{E} :\ \|\boldsymbol{x}\| > 1, \boldsymbol{x}/\|\boldsymbol{x}\| \in \mathbb{S}_J\}$. Combining limits (2.A.1) and (2.A.2) gives (2.3.4). □

Proof of Lemma 2.3.4. Let $N \subset [0,\infty]$ be the set that contains those $\epsilon > 0$ for which the boundary of $\{\boldsymbol{w} \in \mathbb{S} :\ w_j > \epsilon \text{ iff } j \in J\}$ as a subset of $\mathbb{S}$ receives positive mass by ϕ. Since ϕ is a finite measure, N is at most countable. Restricting $\epsilon > 0$ to $(0,\infty) \setminus N$ and applying the same reasoning as in the proof of Lemma 2.3.3 yields (2.3.5). □

Proof of Lemma 2.3.5. From the representation $\boldsymbol{Y} = \boldsymbol{S} + E$ in (2.2.11), we deduce

$$\begin{aligned}\mathsf{P}\left[Y_j > -\infty \text{ iff } j \in J\right] &= \mathsf{P}\left[S_j + E > -\infty \text{ iff } j \in J\right] \\ &= \mathsf{P}\left[S_j > -\infty \text{ iff } j \in J\right].\end{aligned}$$

The identity (2.3.6) follows. Next, combining $\boldsymbol{S} = \boldsymbol{T} - \max(\boldsymbol{T})$ with $-\infty < \max(\boldsymbol{T}) < \infty$ almost surely yields (2.3.7). Finally, the random vector $\boldsymbol{U}$ is linked to $\boldsymbol{T}$ via (2.2.17), so that

$$\mathsf{P}\left[T_j > -\infty \text{ iff } j \in J\right] = \frac{\mathsf{E}\left[\mathrm{e}^{\max(\boldsymbol{U})}\, \mathbb{1}\left\{U_j > -\infty \text{ iff } j \in J\right\}\right]}{\mathsf{E}\left[\mathrm{e}^{\max(\boldsymbol{U})}\right]}.$$

Since $\mathsf{E}\left[\mathrm{e}^{\max(\boldsymbol{U})}\, \mathbb{1}\left\{U_j > -\infty \text{ iff } j \in J\right\}\right] > 0$ if and only if $\mathsf{P}\left[U_j > -\infty \text{ iff } j \in J\right] > 0$, equivalence (2.3.8) follows. □

2.B Proofs of Section 2.4

Proof of Theorem 2.4.4. Let $\boldsymbol{x} \in [0,\infty)^d$. If $x_j = 0$ for some $j \in D$, then $G^n_{\boldsymbol{M}}(n\boldsymbol{x}) = G_{\boldsymbol{M}}(\boldsymbol{x}) = 0$ for all $n \geqslant 1$. Suppose that $\boldsymbol{x} \in (0,\infty)^d$. By (2.4.1), $G_{\boldsymbol{M}}(\boldsymbol{x})$ is given by

$$\begin{aligned}
&\mathsf{P}\left[\max_{k\in R}\left\{a_{1k}Z_1^{(k)}\right\} \leqslant x_1, \ldots, \max_{k\in R}\left\{a_{dk}Z_d^{(k)}\right\} \leqslant x_d\right] = \mathsf{P}\left[\forall k \in R : \forall j \in J_k : a_{jk}Z_j^{(k)} \leqslant x_j\right] \\
&= \prod_{k\in R} \mathsf{P}\left[\forall j \in J_k : a_{jk}Z_j^{(k)} \leqslant x_j\right] = \prod_{k\in R} \exp\left\{-\ell^{(k)}_{J_k}\left(\left(a_{jk}/x_j\right)_{j\in J_k}\right)\right\} \\
&= \exp\left\{-\sum_{k\in R} \ell^{(k)}_{J_k}\left(\left(a_{jk}/x_j\right)_{j\in J_k}\right)\right\}. \qquad (2.B.1)
\end{aligned}$$

Homogeneity of $\ell^{(k)}_{J_k}$ in (2.4.3) for all $k \in R$ yields that $G^n_{\boldsymbol{M}}(n\boldsymbol{x}) = G(\boldsymbol{x})$ for all $n \geqslant 1$ so that $G_{\boldsymbol{M}}$ is an mev distribution. Moreover, the constraints $\sum_{k=1}^r a_{jk} = 1$ for all $j \in D$ in Definition 2.4.1 ensure that the margins of $G_{\boldsymbol{M}}$ are unit-Fréchet. Finally, from (2.2.6) and (2.B.1), we deduce that the stdf of $G_{\boldsymbol{M}}$ is given by (2.4.4). □

Proof of Proposition 2.4.5. Recall μ, the exponent measure of G. For $k \in R$, let $\mu^{(k)}_{J_k}$ be the exponent measure of the factor $\boldsymbol{Z}^{(k)}_{J_k}$, let $\pi_{J_k} : \mathbb{R}^d \to \mathbb{R}^{J_k}$ be the projection $\boldsymbol{x} \mapsto \boldsymbol{x}_{J_k}$ and write $\mathbb{F}_{J_k} := \left\{\boldsymbol{z} \in \mathbb{E} :\ \forall j \notin J_k, z_j = 0\right\}$. For $\boldsymbol{v} \in (0,1)^m$ and $A \subseteq [0,\infty)^m$, we write $\boldsymbol{v} \cdot A := \left\{(v_1a_1, \ldots, v_ma_m) :\ \boldsymbol{a} \in A\right\}$. Consider the measure

$$\eta(B) := \sum_{k\in R} \mu^{(k)}_{J_k}\left(B^{(k)}\right),$$

for Borel $B \subseteq \mathbb{E}$, where $B^{(k)} := \left(1/a_{jk}\right)_{j\in J_k} \cdot \pi_{J_k}\left(B \cap \mathbb{F}_{J_k}\right)$. For $\boldsymbol{x} \in [0,\infty]^d \setminus \{\mathbf{0}\}$,

$$\begin{aligned}
\eta\left(\mathbb{E}\setminus[\mathbf{0},\boldsymbol{x}]\right) &= \sum_{k\in R} \mu_{J_k}^{(k)}\left(\left(1/a_{jk}\right)_{j\in J_k} \cdot \pi_{J_k}\left(\left\{\boldsymbol{z}\in\mathbb{E} : \exists j \in D,\ z_j > x_j\right\} \cap \mathbb{F}_{J_k}\right)\right) \\
&= \sum_{k\in R} \mu_{J_k}^{(k)}\left(\left(1/a_{jk}\right)_{j\in J_k} \cdot \pi_{J_k}\left(\left\{\boldsymbol{z}\in\mathbb{E} : \exists j \in J_k,\ z_j > x_j\right\} \cap \mathbb{F}_{J_k}\right)\right) \\
&= \sum_{k\in R} \mu_{J_k}^{(k)}\left(\left[0,\left(x_j/a_{jk}\right)_{j\in J_k}\right]^c\right) = \mu\left(\mathbb{E}\setminus[\mathbf{0},\boldsymbol{x}]\right),
\end{aligned}$$

where the last equality is a result of (2.4.4). By uniqueness of the exponent measure, we deduce that $\eta = \mu$. Let $J \subseteq D$ be non-empty and recall $\mathbb{E}_J = \left\{\boldsymbol{x}\in\mathbb{E} : x_j > 0 \text{ iff } j \in J\right\}$. We find

$$\mu\left(\mathbb{E}_J\right) = \sum_{k\in R} \mu_{J_k}^{(k)}\left(\left(1/a_{jk}\right)_{j\in J_k} \cdot \pi_{J_k}\left(\mathbb{E}_J \cap \mathbb{F}_{J_k}\right)\right).$$

Let $k \in R$. We will show that the k-th term on the right-hand side is positive if and only if $J = J_k$. This will imply that $\mu(\mathbb{E}_J) > 0$ if and only if $J \in \{J_1,\ldots,J_r\} = \mathcal{R}$, proving the proposition in view of Definition 2.3.1.

- If $J\setminus J_k \neq \varnothing$, then $\mathbb{E}_J \cap \mathbb{F}_{J_k} = \varnothing$ and thus $\mu_{J_k}^{(k)}\left(\left(1/a_{jk}\right)_{j\in J_k} \cdot \pi_{J_k}\left(\mathbb{E}_J \cap \mathbb{F}_{J_k}\right)\right) = 0$.
- Otherwise, if $J \subseteq J_k$, we get $\mathbb{E}_J \cap \mathbb{F}_{J_k} = \mathbb{E}_J$ and thus

  $$\left(1/a_{jk}\right)_{j\in J_k} \cdot \pi_{J_k}\left(\mathbb{E}_J \cap \mathbb{F}_{J_k}\right) = \left(1/a_{jk}\right)_{j\in J_k} \cdot \pi_{J_k}\left(\mathbb{E}_J\right) = \pi_{J_k}\left(\mathbb{E}_J\right).$$

  Since J_k is the only extreme direction of $\boldsymbol{Z}_{J_k}^{(k)}$, we get $\mu_{J_k}^{(k)}\left(\pi_{J_k}\left(\mathbb{E}_J \cap \mathbb{F}_{J_k}\right)\right) > 0$ if and only if $J = J_k$. □

Proof of Proposition 2.4.7. For a non-empty set $J \subseteq D$, recall the set $\mathbb{A}_J$ defined in (2.2.1) and the projection $\pi_J : \mathbb{A}_J \to \mathbb{R}^J$ defined via $\boldsymbol{x} \mapsto \boldsymbol{x}_J$. Thus, for $J \subseteq D$ and $\boldsymbol{x} \in \mathbb{R}^J$, the point $\pi_J^{-1}(\boldsymbol{x})$ in $\mathbb{A}_J$ is defined by setting the components with indices $j \in J$ equal to x_j and the components with indices $j \in D \setminus J$ equal to $-\infty$.

Let $g : [-\infty,\infty)^d \to [0,\infty)$ be a Borel measurable function. By linearity of the expectation and using standard reasoning from measure theory, we get, for $\boldsymbol{U}$ with distribution (2.4.7),

$$\mathsf{E}\left[g\left(\boldsymbol{U}\right)\right] = \sum_{k\in R} m_k\, \mathsf{E}\left[g\left(\pi_{J_k}^{-1}\left(\boldsymbol{U}^{(k)} + \left(\ln\left(a_{jk}/m_k\right)\right)_{j\in J_k}\right)\right)\right]. \tag{2.B.2}$$

First, let $j_0 \in D$. Equation (2.B.2) applied to the function $g : [-\infty, \infty)^d \to [0, \infty)$ defined via $g(\boldsymbol{x}) = \mathrm{e}^{x_{j_0}}$ gives

$$\mathsf{E}\left[\mathrm{e}^{U_{j_0}}\right] = \sum_{k \in R: j_0 \in J_k} m_k \, \mathsf{E}\left[\exp\left\{U_{j_0}^{(k)} + \ln\left(a_{j_0 k}/m_k\right)\right\}\right] = \sum_{k \in R: j_0 \in J_k} m_k \cdot \left(a_{j_0 k}/m_k\right) = 1,$$

since $\mathsf{E}\left[\exp\left(U_j^{(k)}\right)\right] = 1$ for all $j \in J_k$ and since A has unit row sums.

Next, let $\boldsymbol{y} \in [0, \infty)^d$. Applying (2.B.2) to the function $g(\boldsymbol{x}) = \max\left(y_1 \mathrm{e}^{x_1}, \ldots, y_d \mathrm{e}^{x_d}\right)$ gives

$$\begin{aligned}
\mathsf{E}\left[\max\left(\boldsymbol{y}\mathrm{e}^{U}\right)\right] &= \sum_{k \in R} m_k \, \mathsf{E}\left[\max\left\{\boldsymbol{y}\mathrm{e}^{\pi_{J_k}^{-1}\left(\boldsymbol{U}^{(k)} + \left(\ln\left(a_{jk}/m_k\right)\right)_{j \in J_k}\right)}\right\}\right] \\
&= \sum_{k \in R} m_k \, \mathsf{E}\left[\max_{j \in J_k}\left\{y_j \frac{a_{jk}}{m_k}\mathrm{e}^{U_j^{(k)}}\right\}\right] \\
&= \sum_{k \in R} \ell_{J_k}^{(k)}\left(\left(a_{jk} y_j\right)_{j \in J_k}\right) = \ell(\boldsymbol{y}),
\end{aligned}$$

where we applied (2.4.4) in the last step. Hence, $\boldsymbol{U}$ satisfies (2.2.14) and (2.2.15) for the stdf ℓ of $G_{\boldsymbol{M}}$. Thus, $\boldsymbol{U}$ generates H. □

Proof of Theorem 2.4.8. Let G be an mev distribution with unit-Fréchet margins. Let μ be the exponent measure of G and put $\mathcal{R} := \{J \subseteq D : \ J \neq \varnothing, \ \mu\left(\mathbb{E}_J\right) > 0\}$ where we recall $\mathbb{E} = [0, \infty)^d \setminus \{0\}$ and $\mathbb{E}_J = \left\{\boldsymbol{x} \in \mathbb{E} : \ x_j > 0 \text{ iff } j \in J\right\}$. We write $\mathcal{R} = \{J_1, \ldots, J_r\}$ with $2^d \geqslant r = |\mathcal{R}| \geqslant 1$ and $R = \{1, \ldots, r\}$. Let $A = (a_{jk})_{j \in D; k \in R} \in [0, 1]^{d \times r}$ be the coefficient matrix whose (j, k)-th coordinate is equal to

$$a_{jk} = \begin{cases} \mu\left(\left\{\boldsymbol{y} \in \mathbb{E}_{J_k} : \ y_j > 1\right\}\right), & \text{if } j \in J_k, \\ 0, & \text{if } j \in D \setminus J_k. \end{cases}$$

For $\boldsymbol{v} \in (0, 1)^m$ and $A \subseteq [0, \infty)^m$, we write $\boldsymbol{v} \cdot A := \{(v_1 a_1, \ldots, v_m a_m) : \ \boldsymbol{a} \in A\}$. Fix $k \in R$ and consider the Borel measure $\mu_{J_k}^{(k)}$ on $(0, \infty)^{|J_k|}$ by

$$\mu_{J_k}^{(k)}(B) = \mu\left(\pi_{J_k}^{-1}\left(\left(a_{jk}\right)_{j \in J_k} \cdot B\right)\right),$$

with $\pi_{J_k} : \mathbb{E}_{J_k} \to (0, \infty)^{J_k}$ the natural coordinate projection. Let $\boldsymbol{Z}_{J_k}^{(k)}$ be a random vector on $\mathbb{R}^{J_k}$ with distribution

$$\mathsf{P}\left[\boldsymbol{Z}_{J_k}^{(k)} \leqslant \boldsymbol{x}\right] = \exp\left\{-\mu_{J_k}^{(k)}\left([\boldsymbol{0}, \boldsymbol{x}]^c\right)\right\}, \qquad \boldsymbol{x} \in [0, \infty]^{J_k} \setminus \{\boldsymbol{0}\}.$$

The construction of $\mu_{J_k}^{(k)}$ and $\left(a_{jk}\right)_{j \in J_k}$ as above ensures that $\boldsymbol{Z}_{J_k}^{(k)}$ has an mev distribution with unit-Fréchet margins and single extreme direction J_k. Following

Definition 2.4.1 and Remark 2.4.2, consider the mixture random vector $\boldsymbol{M}$ with coefficient matrix A and factors $\boldsymbol{Z}_{J_1}^{(1)}, \ldots, \boldsymbol{Z}_{J_r}^{(r)}$. Using Theorem 2.4.4, the random vector $\boldsymbol{M}$ follows an mev distribution $G_{\boldsymbol{M}}$ with unit-Fréchet margins and exponent measure $\tilde{\mu}$ satisfying for $\boldsymbol{x} \in \mathbb{E}$,

$$\tilde{\mu}\left([\boldsymbol{0}, \boldsymbol{x}]^c\right) = \sum_{k \in R} \mu_{J_k}^{(k)} \left(\left[0, \left(x_j / a_{jk}\right)_{j \in J_k}\right]^c \right) = \sum_{k \in R} \mu_{J_k}^{(k)} \left(\left(1/a_{jk}\right)_{j \in J_k} \cdot \left[0, \boldsymbol{x}_{J_k}\right]^c \right)$$
$$= \sum_{k \in R} \mu \left(\pi_{J_k}^{-1} \left(\left(a_{jk}\right)_{j \in J_k} \cdot \left(1/a_{jk}\right)_{j \in J_k} \cdot \left[0, \boldsymbol{x}_{J_k}\right]^c \right) \right) = \sum_{k \in R} \mu \left(\pi_{J_k}^{-1} \left([\boldsymbol{0}, \boldsymbol{x}_{J_k}]^c \right) \right) = \mu\left([\boldsymbol{0}, \boldsymbol{x}]^c\right).$$

The last equality follows from $\pi_{J_k}^{-1}\left([\boldsymbol{0}, \boldsymbol{x}_{J_k}]^c\right) = \mathbb{E}_{J_k} \setminus [\boldsymbol{0}, \boldsymbol{x}]$ since the sets $\mathbb{E}_{J_k}$ are all disjoint and μ has no mass outside $\bigcup_{k \in R} \mathbb{E}_{J_k}$. Hence G and $G_{\boldsymbol{M}}$ are both mev distributions with unit-Fréchet margins and the same exponent measure. Thus $G_{\boldsymbol{M}} = G$. □

Proof of Proposition 2.4.11. The distribution of $\boldsymbol{U}$ in (2.4.7) is a mixture of the distributions of the random vectors $\pi_{J_k}^{-1}\left(\boldsymbol{U}^{(k)} + \left(\ln\left(a_{jk}/m_k\right)\right)_{j \in J_k}\right)$ over $k \in R$. The latter random vectors are absolutely continuous with respect to υ in (2.3.10) with density

$$f^{(k)}(\boldsymbol{u}) = f_{\boldsymbol{U}^{(k)}}\left(\boldsymbol{u}_{J_k} - \left(\ln\left(a_{jk}/m_k\right)\right)_{j \in J_k}\right) \mathbb{1}\left\{\boldsymbol{u} \in \mathbb{A}_{J_k}\right\}, \qquad \boldsymbol{u} \in [-\infty, \infty)^d,$$

where $\mathbb{A}_{J_k}$ as in (2.2.1). Hence $\boldsymbol{U}$ is absolutely continuous with respect to υ with density

$$f_{\boldsymbol{U}}(\boldsymbol{u}) = \sum_{k \in R} m_k f_{\boldsymbol{U}^{(k)}}\left(\boldsymbol{u}_{J_k} - \left(\ln\left(a_{jk}/m_k\right)\right)_{j \in J_k}\right) \mathbb{1}\left\{\boldsymbol{u} \in \mathbb{A}_{J_k}\right\}, \qquad \boldsymbol{u} \in [-\infty, \infty)^d.$$

Finally, Theorem 2.3.9 yields that the mgp random vector $\boldsymbol{Y}$ is absolutely continuous with respect to υ with density $h_{\boldsymbol{Y}}$ in (2.4.8). □

Proof of Proposition 2.4.14. For $k \in R$, the random vector $\boldsymbol{U}^{(k)}$ in (2.4.11) satisfies (2.2.14) and (2.2.15) for the logistic stdf with parameter α_k [43, Proposition 1.2.1]. Finally, Proposition 2.4.7 finishes the proof. □

Proof of Proposition 2.4.15. For $k \in R$, the random vector $\boldsymbol{U}^{(k)}$ in (2.4.11) has independent Gumbel components. Thus $\boldsymbol{U}^{(k)}$ is absolutely continuous with respect to $\lambda_{|J_k|}$. Proposition 2.4.11 yields that the mgp random vector $\boldsymbol{Y}$ associated with the mixture logistic model is absolutely continuous with respect to υ in (2.3.10) with density $h_{\boldsymbol{Y}}$ given by (2.4.8). We use Kiriliouk et al. [66, Subsection 7.2] to calculate the integrals in the latter equation. This yields the density of $\boldsymbol{Y}$ in (2.4.12). □

Proof of Proposition 2.4.17. For $k \in R$, the random vector $\boldsymbol{U}^{(k)}$ in (2.4.14) satisfies (2.2.14) and (2.2.15) for the Hüsler–Reiss stdf with variogram matrix $\Gamma^{(k)}$; see, e.g., Wadsworth and Tawn [99]. Finally, Proposition 2.4.7 finishes the proof. □

Proof of Proposition 2.4.18. For $k \in R$, the random vector $\boldsymbol{U}^{(k)}$ in (2.4.14) is a $|J_k|$-variate normal random vector with positive definite covariance matrix. Thus $\boldsymbol{U}^{(k)}$ is absolutely continuous w.r.t. $\lambda_{|J_k|}$. Proposition 2.4.11 yields that the mgp random vector $\boldsymbol{Y}$ associated with the mixture Hüsler–Reiss model is absolutely continuous with respect to υ with density $h_{\boldsymbol{Y}}$ given by (2.4.8). We use Kiriliouk et al. [66, Section 7.2] to calculate the integrals in the latter equation. This yields the density of $\boldsymbol{Y}$ given by (2.4.15). □

2.C Proofs of Section 2.5

Proof of Proposition 2.5.1. For $k \in R$, let the random vector $\boldsymbol{W}^{(k)} = \left(W_1^{(k)}, \ldots, W_d^{(k)}\right)$ with values in $\mathbb{A}_{J_k}$ be defined by

$$W_j^{(k)} = \begin{cases} P_j^{(k)} & \text{if } j \in J_k, \\ -\infty & \text{if } j \in D \setminus J_k. \end{cases}$$

Then Eq. (2.4.7) is the same as

$$\mathsf{P}\left[\boldsymbol{U} \in \cdot\,\right] = \sum_{k \in R} m_k \, \mathsf{P}\left[\boldsymbol{W}^{(k)} \in \cdot\,\right] \tag{2.C.1}$$

Further, we have, by the relation between the stdf $\ell^{(k)}$ and the generator $\boldsymbol{U}^{(k)}$,

$$\begin{aligned} \mathsf{E}\left[\exp\left\{\max\left(\boldsymbol{W}^{(k)}\right)\right\}\right] &= \mathsf{E}\left[\exp\left\{\max\left(\boldsymbol{P}^{(k)}\right)\right\}\right] \\ &= \mathsf{E}\left[\max_{j \in J_k}\left\{\frac{a_{jk}}{m_k} \mathrm{e}^{U_j^{(k)}}\right\}\right] = \frac{1}{m_k} \ell^{(k)}\left((a_{jk})_{j \in J_k}\right). \end{aligned}$$

Since also $\ell(\mathbf{1}) = \mathsf{E}\left[\mathrm{e}^{\max(\boldsymbol{U})}\right]$, we find

$$\frac{m_k \, \mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{W}^{(k)}}\right]}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{U}}\right]} = \frac{\ell^{(k)}\left((a_{jk})_{j \in J_k}\right)}{\ell(\mathbf{1})} = w_k \tag{2.C.2}$$

as defined in (2.5.1). By Equations (2.2.17), (2.C.1) and (2.C.2), we find, Borel sets $B \subseteq [-\infty, \infty)^d$,

$$\begin{aligned}
\mathsf{P}\left[\boldsymbol{T} \in B\right] &= \frac{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{U}}\, \mathbb{1}\left\{\boldsymbol{U} \in B\right\}\right]}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{U}}\right]} = \sum_{k \in R} \frac{m_k}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{U}}\right]} \mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{W}^{(k)}}\, \mathbb{1}\left\{\boldsymbol{W}^{(k)} \in B\right\}\right] \\
&= \sum_{k \in R} w_k \frac{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{W}^{(k)}}\, \mathbb{1}\left\{\boldsymbol{W}^{(k)} \in B\right\}\right]}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{W}^{(k)}}\right]} \\
&= \sum_{k \in R} w_k \frac{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{P}^{(k)}}\, \mathbb{1}\left\{\boldsymbol{P}^{(k)} \in \pi_{J_k}\left(\mathbb{A}_{J_k} \cap B\right)\right\}\right]}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{P}^{(k)}}\right]} \\
&= \sum_{k \in R} w_k\, \mathsf{P}\left[\boldsymbol{T}^{(k)} \in \pi_{J_k}\left(\mathbb{A}_{J_k} \cap B\right)\right],
\end{aligned}$$

where $\boldsymbol{T}^{(k)}$ is a random vector on $\mathbb{R}^{J_k}$ which satisfies (2.5.1).

The expression for the density of $\boldsymbol{T}^{(k)}$ follows from Lemma 2.C.1 below applied to $\boldsymbol{P} = \boldsymbol{P}^{(k)}$ upon noting that

$$\mathsf{E}\left[\exp\left(P_j^{(k)}\right)\right] = \mathsf{E}\left[\exp\left\{U_j^{(k)} + \ln(a_{jk}/m_k)\right\}\right] = a_{jk}/m_k, \qquad j \in J_k,$$

as well as $\mathsf{E}\left[\exp\left\{\max\left(\boldsymbol{P}^{(k)}\right)\right\}\right] = \ell^{(k)}\left((a_{jk})_{j \in J_k}\right)/m_k$, which was shown above. □

Lemma 2.C.1. *Let J be a positive integer. If $\boldsymbol{P}$ is a random vector on $\mathbb{R}^J$ with Lebesgue density $f_{\boldsymbol{P}}$ such that $p_j := \mathsf{E}\left[\exp(P_j)\right] < \infty$ for all $j \in \{1, \ldots, J\}$, then the random vector $\boldsymbol{T}$ with distribution $\mathsf{P}\left[\boldsymbol{T} \in \cdot\,\right] = \mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{P}}\, \mathbb{1}\left\{\boldsymbol{P} \in \cdot\right\}\right] / \mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{P}}\right]$ has Lebesgue density*

$$f_{\boldsymbol{T}}(\boldsymbol{t}) = c\, g(\boldsymbol{t}) \sum_{j=1}^{J} n_j\, q_j(\boldsymbol{t}), \qquad \boldsymbol{t} \in \mathbb{R}^J,$$

with $c = \sum_{j=1}^J p_j / \mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{P}}\right]$, $g(\boldsymbol{t}) = \mathrm{e}^{\max \boldsymbol{t}} / \sum_{j=1}^J \mathrm{e}^{t_j}$, mixture weights $n_j = p_j / \sum_{i=1}^J p_i$ and probability densities $q_j(\boldsymbol{t}) = \mathrm{e}^{t_j} f_{\boldsymbol{P}}(\boldsymbol{t})/p_j$.

Proof of Lemma 2.C.1. The density of $\boldsymbol{T}$ evaluated in $\boldsymbol{t} \in \mathbb{R}^J$ is

$$f_{\boldsymbol{T}}(\boldsymbol{t}) = \frac{\mathrm{e}^{\max \boldsymbol{t}}}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{P}}\right]} f_{\boldsymbol{P}}(\boldsymbol{t}) = \frac{\sum_{i=1}^J p_i}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{P}}\right]} \frac{\mathrm{e}^{\max \boldsymbol{t}}}{\sum_{j=1}^J \mathrm{e}^{t_j}} \sum_{j=1}^{J} \frac{p_j}{\sum_{i=1}^J p_i} \frac{\mathrm{e}^{t_j}}{p_j} f_{\boldsymbol{P}}(\boldsymbol{t}),$$

which is the expression stated in the lemma. □

Proof of Proposition 2.5.2. Equation (2.2.16) yields that for measurable, nonnegative functions g, the distributions of $\boldsymbol{S}$ and $\boldsymbol{U}$ are related via

$$\mathsf{E}\left[g(\boldsymbol{S})\right] = \frac{\mathsf{E}\left[g\left(\boldsymbol{U} - \max \boldsymbol{U}\right) \mathrm{e}^{\max \boldsymbol{U}}\right]}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{U}}\right]}.$$

Fix $\boldsymbol{y} \in [0,\infty)^d$ with $y_j \geqslant 1$. For $g(\boldsymbol{s}) = \mathbb{1}\left\{\boldsymbol{s} - s_j \leqslant \ln(\boldsymbol{y})\right\} \mathrm{e}^{s_j}$, we get

$$\begin{aligned} g(\boldsymbol{U} - \max \boldsymbol{U}) &= \mathbb{1}\left\{(\boldsymbol{U} - \max \boldsymbol{U}) - \left(U_j - \max \boldsymbol{U}\right) \leqslant \ln(\boldsymbol{y})\right\} \mathrm{e}^{U_j - \max \boldsymbol{U}} \\ &= \mathbb{1}\left\{\boldsymbol{U} - U_j \leqslant \ln(\boldsymbol{y})\right\} \mathrm{e}^{U_j - \max \boldsymbol{U}} \end{aligned}$$

and thus

$$\begin{aligned} \mathsf{E}\left[\mathbb{1}\left\{\boldsymbol{S} - S_j \leqslant \ln(\boldsymbol{y})\right\} \mathrm{e}^{S_j}\right] &= \frac{\mathsf{E}\left[\mathbb{1}\left\{\boldsymbol{U} - U_j \leqslant \ln(\boldsymbol{y})\right\} \mathrm{e}^{U_j - \max \boldsymbol{U}} \mathrm{e}^{\max \boldsymbol{U}}\right]}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{U}}\right]} \\ &= \frac{\mathsf{E}\left[\mathbb{1}\left\{\boldsymbol{U} - U_j \leqslant \ln(\boldsymbol{y})\right\} \mathrm{e}^{U_j}\right]}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{U}}\right]}. \end{aligned}$$

Further, we deduce from (2.2.6), (2.2.8), (2.2.14) and (2.2.15) that

$$\mathsf{P}\left[Y_j > 0\right] = \frac{1}{\mu\left(\{\boldsymbol{x} \geqslant 0 :\ \boldsymbol{x} \not\leqslant \mathbf{1}\}\right)} = \frac{1}{\ell(\mathbf{1})} = \frac{1}{\mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{U}}\right]}.$$

Recall representation $\boldsymbol{Y} \overset{\mathsf{d}}{=} E + \boldsymbol{S}$ with E a unit exponential random variable independent from $\boldsymbol{S}$. Since, by definition, $\mathsf{P}\left[\boldsymbol{Q}^{(j)} \in \mathrm{d}\boldsymbol{t}\right] = \mathrm{e}^{t_j} \mathsf{P}\left[\boldsymbol{U} \in \mathrm{d}\boldsymbol{t}\right]$, we conclude that

$$\begin{aligned} \mathsf{P}\left[\mathrm{e}^{\boldsymbol{Q}^{(j)} - Q_j^{(j)}} \leqslant \boldsymbol{y}\right] &= \mathsf{E}\left[\mathbb{1}\left\{\mathrm{e}^{\boldsymbol{U} - U_j} \leqslant \boldsymbol{y}\right\} \mathrm{e}^{U_j}\right] \\ &= \mathsf{E}\left[\mathbb{1}\left\{\boldsymbol{S} - S_j \leqslant \ln(\boldsymbol{y})\right\} \mathrm{e}^{S_j}\right] \cdot \mathsf{E}\left[\mathrm{e}^{\max \boldsymbol{U}}\right] \\ &= \mathsf{P}\left[(E + \boldsymbol{S}) - (E + S_j) \leqslant \ln(\boldsymbol{y}),\, E > -S_j\right] / \mathsf{P}\left[Y_j > 0\right] \\ &= \mathsf{P}\left[\mathrm{e}^{\boldsymbol{Y}} / \mathrm{e}^{Y_j} \leqslant \boldsymbol{y},\, Y_j > 0\right] / \mathsf{P}\left[Y_j > 0\right] \\ &= \mathsf{P}\left[\boldsymbol{\mathcal{U}}^{(j)} \leqslant \boldsymbol{y}\right], \end{aligned}$$

where the last equality is proved in Engelke and Hitz [34, Lemma 2]. □

Proof of Proposition 2.5.3. Fix $j \in J_k$. If $i \in J_k \setminus \{j\}$, then the density of ψ_i is clearly $t \mapsto f_{U_i^{(k)}}\left(t - \ln\left(a_{ik}/m_k\right)\right)$. Otherwise if $i = j$, by straightforward calculus,

we get

$$\frac{\partial}{\partial t}\,\mathsf{P}\left[-\alpha_k \ln N + \ln\left(a_{jk}/\left(m_k\Gamma(1-\alpha_k)\right)\right) \leqslant t\right] = \frac{\mathrm{e}^t f_{U_j^{(k)}}\left(t - \ln\left(a_{jk}/m_k\right)\right)}{\int_{s\in\mathbb{R}} \mathrm{e}^s f_{U_j^{(k)}}\left(s - \ln\left(a_{jk}/m_k\right)\right)\,\mathrm{d}s}.$$

□

Proof of Proposition 2.5.4. The proof is by direct calculations and is omitted for brevity. □

Funding The research of Anas Mourahib was supported financially by the *Fonds de la Recherche Scientifique – FNRS*, Belgium (grant number T.0203.21).

Acknowledgments We are grateful to the anonymous reviewers for their suggestions which greatly helped to improve the manuscript.

Chapter 3

Estimation procedure and extreme directions identification algorithm

Anna Kiriliouk[1] Anas Mourahib[2] Johan Segers[3]

Contents

Abstract

Estimating the parameters of max-stable parametric models poses significant challenges, particularly when some parameters lie on the boundary of the parameter space. This situation arises when a subset of variables exhibits extreme values simultaneously, while the remaining variables do not—a phenomenon referred to as an extreme direction in the literature. In this paper, we propose a novel estimator for the parameters of a general parametric mixture model, incorporating a threshold

[1]UNamur, NaXys, anna.kiriliouk@unamur.be.
[2]UCLouvain, LIDAM/ISBA, anas.mourahib@uclouvain.be.
[3]KU Leuven, Department of Mathematics, jjjsegers@kuleuven.be.

exceedances approach based on a pseudo-norm penalization. The latter plays a crucial role in accurately identifying parameters at the boundary of the parameter space. Additionally, our estimator comes with a data-driven algorithm to detect groups of variables corresponding to extreme directions. We assess the performance of our estimator in terms of both parameter estimation and the identification of extreme directions through extensive simulation studies. Finally, we apply our method to two real-world datasets: first, to discharge measurements at stations along the Danube River, and second, to financial portfolio losses from stocks listed on the NYSE, AMEX, and NASDAQ. In both applications, we identify the sets of variables that can become large simultaneously.

3.1 Introduction

Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ of risk factors and suppose the focus is on rare events. In this regime, neither parametric nor nonparametric methods are satisfactory: parametric models rely on distributional assumptions that can be dominated by the bulk of the data, and thus tend to underestimate extreme outcomes, while nonparametric techniques demand large sample sizes—often unavailable when studying extremes. Extreme value theory offers a rigorous framework for extrapolating beyond observed levels and has found applications in finance, hydrology, weather forecasting, and other fields.

Univariate extreme value theory, that is, when $d = 1$, has been extensively studied [24, 4, 16] and modeling rare events in this case is done via a finite-dimensional parametric family of distributions. Multivariate modeling of $d \geqslant 2$ risk factors is a more challenging topic, since risks may exhibit some sort of dependence. The goal is not only to model rare events of each risk factor independently but to capture the extremal dependence between these risk factors. Two main approaches appear in this context: the block-maxima approach and the threshold exceedances approach. In the first approach, standardized componentwise maxima of $\boldsymbol{X}$ are asymptotically modeled using a max-stable random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$. In the second approach, asymptotically, observations where at least one risk factor is extreme are modeled using a multivariate (generalized) Pareto random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ [83].

For both approaches, many parametric families have been proposed; popular examples include the Hüsler–Reiss model [58] and the logistic model [96], the latter being a special case of the scaled extremal Dirichlet model [7]. Estimation methods for such models are numerous, for example, through censored likelihood [78], weighted least squares [33], or based on neural networks [70, 82].

A drawback of most parametric models is that they force all components of $\boldsymbol{X}$ to be extreme simultaneously. Figure 3.1 (left) shows a log-transformed simulation

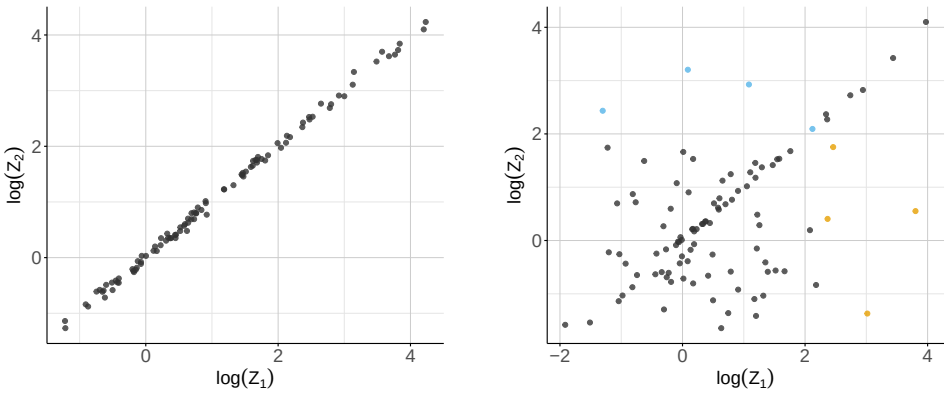


Figure 3.1
A sample of size $n = 100$ from a max-stable Hüsler–Reiss model (left) and a max-stable mixture Hüsler–Reiss model (right). Points in the extreme direction $\{1, 2\}$ are shown in black, those in $\{1\}$ in orange, and those in $\{2\}$ in blue.

from a max-stable Hüsler–Reiss model. Observations are close to the diagonal line $x = y$, suggesting that both components are large simultaneously. However, in practice, it may happen that only certain subsets $J \subset \{1, \dots, d\}$ of variables are jointly extreme while the others are not. Such a subset J is called an *extreme direction* [75, 90].

Both the max-linear model [46, 32] and the asymmetric logistic model [96] can accommodate multiple extreme directions. However, the first model has a singular distribution (not discrete nor absolutely continuous), thus making statistical inference complicated. The second one is too specific in the sense that it cannot represent a large variety of max-stable distributions.

To capture multiple extreme directions, Mourahib et al. [75] introduce the *mixture model* that extends the classic max-linear model and can represent *any* max-stable distribution. It is identified via two sets of parameters:

- a $(d \times r)$ coefficient matrix A with entries in $[0, 1]$, whose columns correspond to distinct extreme directions of $\boldsymbol{X}$. Zero entries in A indicate which variables are not contained in a given extreme direction [75, Proposition 4.5];
- a second set of parameters that governs the dependence structure *within* each extreme direction.

A simulation example of this model is given in Figure 3.1 (right), where we can see that the model, in fact, captures not only the extreme direction $\{1, 2\}$, represented by the points in black, but also $\{1\}$ and $\{2\}$, represented by the points in orange and blue, respectively.

A natural question is how to estimate the model parameters. The weighted least-squares estimator of Einmahl et al. [33] requires the true parameters to be in

the interior of the parameter space. In contrast, the mixture model's reliance on zero entries in A pushes parameters to the boundary of the space, violating this condition.

Contribution 1. We propose an estimator based on threshold exceedances for the parameters of max-stable distributions that accommodate arbitrary collections of extreme directions, building on the penalized least squares estimator introduced in [33]. A pseudo-norm L_p with $p \leqslant 1$ (See Section 3.3) is used to penalize the coefficient matrix A of the mixture model, which leads to zero entries on A and thus multiple extreme directions, in particular those different from $\{1, \ldots, d\}$. The proposed penalized least squares estimator can simultaneously identify extreme directions *and* estimate the mixture model parameters. This sets our approach apart from the existing literature, which focuses exclusively on identifying extreme directions [50, 90, 74].

Contribution 2. A common criticism of extreme-value models with asymptotic dependence is their difficulty in handling situations where not all variables are extreme, that is, when the extreme directions differ from the full set $\{1, \ldots, d\}$. This issue has motivated alternative approaches such as conditional extremes [52] and geometric extremes [98]. We address this limitation by developing an algorithm based on asymptotically dependent mixture models [75]. Our method can detect multiple extreme directions, including those involving only subsets of variables. We compare our approach with the *DAMEX* algorithm from [50], which considers a subset $J \subset \{1, \ldots, d\}$ to define an extreme direction if the corresponding cone $\mathcal{E}_J = \{\mathbf{x} \geqslant 0 : x_j > 0 \text{ iff } j \in J\}$ has positive mass under an empirical estimate of the exponent measure [50, Equation 20]. However, as noted in [98], at finite levels, data rarely lie exactly on such cones when $J \neq \{1, \ldots, d\}$, which can lead to missed directions. In contrast, our algorithm uses tuning parameters that can be selected via cross-validation, improving robustness and practical applicability.

Outline of the paper. The paper is structured as follows. Section 3.2 provides essential background on extreme value theory, focusing in particular on the notions of extreme directions [90], the mixture model [75] and the least-squares estimator [33]. In Section 3.3, we introduce a penalized estimator for the mixture model and subsequently develop a data-driven algorithm to identify the extreme directions present in a given sample. In Section 3.4, using a simulation study case, we show that the choice of the tuning parameters in our estimation procedure effectively reduces to the choice of a single penalization parameter, which we determine via cross-validation. Moreover, with this choice of tuning parameter, we show that our estimation procedure gives good results both in terms of estimation of the

mixture model and extreme directions identification. Finally, in Section 3.5, we illustrate the practical relevance of our approach by applying the algorithm, first to river discharge data across stations from the Danube basin, and second to financial industry portfolios constructed from stocks listed on the NYSE, AMEX, and NASDAQ. Finally, we compare our results with those obtained by applying algorithms from the literature, in particular the DAMEX algorithm introduced by Goix et al. [50].

Notation. Let d be a positive integer and write $D = \{1, \dots, d\}$. Throughout, bold symbols will refer to multivariate quantities. Let $\mathbf{0}$ and $\mathbf{1}$ denote vectors of zeroes and ones, respectively, of a dimension clear from the context. Let $\boldsymbol{x} = (x_1, \dots, x_d)$ denote a d-dimensional vector and, for a non-empty set $J \subseteq \{1, \dots, d\}$, write $\boldsymbol{x}_J = (x_j)_{j \in J}$. Mathematical operations on vectors such as addition, multiplication and comparison are considered component-wise. For two vectors $\boldsymbol{a},\ \boldsymbol{b}$, the set $[\boldsymbol{a}, \boldsymbol{b}]$ will denote the Cartesian product $\prod_{j=1}^{d}[a_j, b_j]$ and so forth for $(\boldsymbol{a}, \boldsymbol{b})$, $[\boldsymbol{a}, \boldsymbol{b})$, $(\boldsymbol{a}, \boldsymbol{b}]$. For two integers $d_1,\ d_2 > 0$ and a set $\mathbb{S} \subset \mathbb{R}^+ = [0, \infty)$, let $\mathcal{M}_{d_1, d_2}(\mathbb{S})$ denote the set of $d_1 \times d_2$ matrices with entries in $\mathbb{S}$. The arrow $\rightsquigarrow$ denotes convergence in distribution (weak convergence). For two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^d$, define the *lexicographic order* $\leqslant_{\text{lex}}$ as

$$\mathbf{x} \leqslant_{\text{lex}} \mathbf{y} \quad \text{if and only if} \quad x_j < y_j \text{ for the first index } j \in D \text{ for which } x_j \neq y_j. \tag{3.1.1}$$

The lexicographic order is a total order in $\mathbb{R}^d$, meaning that either $\boldsymbol{x} \leqslant_{\text{lex}} \boldsymbol{y}$ or $\boldsymbol{y} \leqslant_{\text{lex}} \boldsymbol{x}$. For a set S, let $S^n = S \times \cdots \times S$ (n times). Finally, let $\mathbb{E} := [0, \infty)^d \setminus \{\mathbf{0}\}$.

3.2 Background

3.2.1 Multivariate extreme value distributions

For $i = 1, \dots, n$, let $\boldsymbol{X}_i = (X_{i1}, \dots, X_{id})$ be i.i.d. copies of $\boldsymbol{X}$, a random vector on $\mathbb{R}^d$ with joint cumulative distribution function (cdf) F and margins $F_1, \dots, F_d$. We say that F is in the (max-)domain of attraction of a *max-stable* distribution G if there exist sequences $\boldsymbol{a}_n \in (0, \infty)^d$ and $\boldsymbol{b}_n \in \mathbb{R}^d$ such that

$$\frac{\max_{i=1,\dots,n} \boldsymbol{X}_i - \boldsymbol{b}_n}{\boldsymbol{a}_n} \rightsquigarrow G, \qquad n \to \infty.$$

The margins of G, denoted $G_1, \dots, G_d$, are univariate extreme-value distributions themselves, and any max-stable distribution G admits the representation

$$G(\boldsymbol{x}) = \exp\left\{-\ell\left(-\ln G_1(x_1), \dots, -\ln G_d(x_d)\right)\right\},$$

where $\ell : [0,\infty)^d \to [0,\infty)$ is the *stable tail dependence function (stdf)*. Then the distribution of $\boldsymbol{X}^* = (X_1^*, \ldots, X_d^*)$ with $X_j^* = 1/\left\{1 - F_j(X_j)\right\}$ for $j \in D$ is in the domain of attraction of $G^*(\boldsymbol{x}) = \exp\{-\ell(1/x_1, \ldots, 1/x_d)\}$, a max-stable distribution with unit-Fréchet margins.

The stable tail dependence function can be retrieved via

$$\ell(\boldsymbol{x}) = \lim_{t \downarrow 0} t^{-1} \, \mathbb{P}\left[F_1(X_1) \geqslant 1 - tx_1 \text{ or } \ldots \text{ or } F_d(X_d) \geqslant 1 - tx_d\right]. \tag{3.2.1}$$

The function ℓ is convex and homogeneous in the sense that for $c \geqslant 0$ and $\boldsymbol{x} \in [0,\infty)^d$, we have $\ell(c\boldsymbol{x}) = c\ell(\boldsymbol{x})$. Moreover, unit-Fréchet margins of G^* imply that $\ell(0, \ldots, 0, x_j, 0, \ldots, 0) = x_j$ for $x_j \geqslant 0$ and $j \in D$. Finally, ℓ can be bounded by

$$\max(x_1, \ldots, x_d) \leqslant \ell(x) \leqslant x_1 + \cdots + x_d, \tag{3.2.2}$$

where the lower bound corresponds to perfect tail dependence and the upper bound to asymptotic independence.

Suppose that $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$ follows a max-stable distribution with unit-Fréchet margins. For non-empty $J \subseteq D$, let ℓ_J be the stdf associated with $\boldsymbol{Z}_J$, that is,

$$\ell_J(\boldsymbol{x}_J) = \ell\left(\tilde{\boldsymbol{x}}\right), \qquad \boldsymbol{x}_J \in [0,\infty)^{|J|},$$

with $\tilde{\boldsymbol{x}} = \sum_{j \in J} x_j \boldsymbol{e}_j$ and $\boldsymbol{e}_j$ the j-th canonical unit vector in $\mathbb{R}^d$. The *extremal coefficients* [86] are defined via $\zeta_J := \ell(\sum_{j \in J} \boldsymbol{e}_j)$, for set $J \subseteq D$ with $|J| \geqslant 2$. From (3.2.2), we get that $1 \leqslant \zeta_J \leqslant |J|$. The value ζ_J can be interpreted as the effective number of tail independent variables among $\boldsymbol{Z}_J$.

The *exponent measure* μ of G^* is the unique Borel measure concentrated on $\mathbb{E} = [0,\infty)^d \setminus \{\boldsymbol{0}\}$ such that

$$\mu\left(\mathbb{E} \setminus [\boldsymbol{0}, 1/\boldsymbol{x}]\right) = \ell(\boldsymbol{x}), \qquad \boldsymbol{x} \in [\boldsymbol{0}, \boldsymbol{\infty})^d.$$

The measure μ is finite on Borel sets that are bounded away from the origin. Moreover, $\mu\left(\mathbb{E} \setminus [\boldsymbol{0}, \boldsymbol{y}]\right) = \infty$ as soon as $y_j = 0$ for some $j \in D$.

3.2.2 Extreme directions and the mixture model

For non-empty $J \subseteq D$, define $\mathbb{E}_J := \left\{\boldsymbol{x} \in \mathbb{E} : \; x_j > 0 \text{ iff } j \in J\right\}$. Then $\{\mathbb{E}_J : \; \varnothing \neq J \subseteq D\}$ forms a partition of $\mathbb{E}$. Let $\boldsymbol{X}^*$ be again a d-dimensional random vector in the domain of attraction of a max-stable distribution with unit-Fréchet margins and exponent measure μ. The following definition was introduced in Simpson et al. [90] in order to identify *extreme directions* of $\boldsymbol{X}^*$. Intuitively, an extreme direction is a group of components of $\boldsymbol{X}^*$ that may take on large values simultaneously while the other components are of smaller order.

Definition 3.2.1. The set J is said to be an extreme direction of $\boldsymbol{X}^*$ or μ if $\mu(\mathbb{E}_J) > 0$.

Note that μ has no mass outside $\bigcup_{J\in\mathcal{R}(\mu)} \mathbb{E}_J$, where $\mathcal{R}(\mu)$ is the collection of extreme directions of μ. Further, even if a certain set $J \subseteq D$ is not an extreme direction of $\boldsymbol{X}^*$, it is still possible that J is a superset or a subset of an extreme direction.

Many standard max-stable models are only appropriate for data exhibiting a single extreme direction $J = D$. Notable exceptions are the asymmetric logistic [95] and the max-linear [100, 32] models. [75] introduce the *mixture model*, a smoothed version of the max-linear model, which allows any collection of non-empty subsets of D that together covers D, so that no element of D is left out, to be the extreme directions of μ. The mixture model is key to modeling data with non-standard dependence structures as it covers *all* max-stable distributions.

Definition 3.2.2 (Mourahib et al. [75])**.** Let r be a positive integer and write $R = \{1, \dots, r\}$. Let $\boldsymbol{Z}_{\bullet 1}, \dots, \boldsymbol{Z}_{\bullet r}$ be independent random vectors in $\mathbb{R}^d$. For all $s \in R$, suppose that $\boldsymbol{Z}_{\bullet s} = (Z_{1s}, \dots, Z_{ds})$ are max-stable random vectors with unit-Fréchet margins, stdf $\ell^{(s)}$ and a single extreme direction D. Let $A = (a_{js})_{j\in D; s\in R}$ be a $(d \times r)$ coefficient matrix with $a_{js} \in [0, 1]$ for all $j \in D$ and $s \in R$, unit row sums $\sum_{s=1}^{r} a_{js} = 1$ for all $j \in D$ and positive column sums $\sum_{j=1}^{d} a_{js} > 0$ for all $s \in R$. The d-variate random vector

$$\boldsymbol{M} = (M_1, \dots, M_d) = \left(\max_{s\in R}\{a_{1s}Z_{1s}\}, \dots, \max_{s\in R}\{a_{ds}Z_{ds}\} \right), \tag{3.2.3}$$

will be called the $(d \times r)$ *mixture model* with coefficient matrix A and factors $\boldsymbol{Z}_{\bullet 1}, \dots, \boldsymbol{Z}_{\bullet r}$.

Remark 3.2.3. Note that two different coefficient matrices, containing the same columns but in different orders, will lead to the same mixture model $\boldsymbol{M}$ if the distributions of $\boldsymbol{Z}_{\bullet 1}, \dots, \boldsymbol{Z}_{\bullet r}$ are identical. To ensure identifiability, we will always assume that the columns of A are decreasingly ordered according to the lexicographic order (3.1.1).

In (3.2.3), for a given $j \in D$, only those $s \in R$ that satisfy $a_{js} > 0$ matter for the value of the maximum. Let

$$J_s := \left\{ j \in D : \ a_{js} > 0 \right\}, \tag{3.2.4}$$

denote the s-th *signature* of the coefficient matrix A.

Theorem 3.2.4 (Stephenson [94], Mourahib et al. [75])**.** *The random vector* $\boldsymbol{M}$ *in* (3.2.3) *follows a max-stable distribution* $G_{\boldsymbol{M}}$ *with unit-Fréchet margins and stdf*

$$\ell(\boldsymbol{x}) = \sum_{s\in R} \ell^{(s)}_{J_s} \left(\left(a_{js} x_j \right)_{j\in J_s} \right), \qquad \boldsymbol{x} \in [0, \infty)^d. \tag{3.2.5}$$

Zero entries of the coefficient matrix A allow some groups of components of the random vector $\boldsymbol{M}$ to constitute an extreme direction. Mourahib et al. [75, Proposition 4.5] show that the set of extreme directions of the mixture model is given by the signatures of its coefficient matrix A. The model is general in the sense that any max-stable distribution with unit-Fréchet margins is the cdf of a mixture model whose coefficient matrix has mutually different signatures [75, Theorem 4.8].

In the rest of the paper, we will assume that an extreme direction only corresponds to a single column of the coefficient matrix A, i.e., we will assume that the signatures of A are mutually different.

Example 3.2.5 (Mixture logistic model)**.** For $s \in R$, suppose that the stdf associated with $\boldsymbol{Z}_{J_s,s}$ is

$$\ell^{(s)}_{J_s}(x_1, \ldots, x_{|J_s|}) = \left(x_1^{1/\alpha_s} + \cdots + x_{|J_s|}^{1/\alpha_s}\right)^{\alpha_s},$$

for $\boldsymbol{x} \in [0,\infty)^{|J_s|}$, where $0 < \alpha_s < 1$. The stdf of the *mixture logistic model* $\boldsymbol{M}$ is

$$\ell(\boldsymbol{x}) = \sum_{s\in R} \left\{ \sum_{j\in J_s} \left(a_{js} x_j\right)^{1/\alpha_s} \right\}^{\alpha_s}, \qquad \boldsymbol{x} \in [0,\infty)^d.$$

Example 3.2.6 (Mixture Hüsler–Reiss model)**.** Let $s \in R$. For $|J_s| > 1$, suppose that $\ell^{(s)}_{J_s}$ is the Hüsler–Reiss stdf associated with the $(|J_s| \times |J_s|)$ variogram matrix[1] $\Gamma^{(s)}$; see, e.g., Huser and Davison [57]. That is,

$$\ell^{(s)}_{J_s}(\boldsymbol{x}) = \sum_{j\in J_s} x_j \, \Phi_{|J_s|-1}\left(\boldsymbol{\eta}^{j,(s)}(\boldsymbol{x}) \, ; \Sigma^{j,(s)}\right), \qquad \boldsymbol{x} \in [0,\infty)^{J_s},$$

where for $j \in J_s$, $\boldsymbol{\eta}^{j,(s)}(\boldsymbol{x}) := \left\{\ln(x_j/x_t) + \Gamma^{(s)}_{jj}/2\right\}_{t\in J_s; t\neq j}$ where by convention, $0/0$ is set to ∞, and where the $(|J_s \setminus \{j\}| \times |J_s \setminus \{j\}|)$ matrix $\Sigma^{j,(s)}$ is defined by

$$\Sigma^{j,(s)} := \frac{1}{2}\left\{\Gamma^{(s)}_{jt} + \Gamma^{(s)}_{jt'} - \Gamma^{(s)}_{tt'}\right\}_{t,t'\in J_s\setminus\{j\}},$$

and finally, where $\Phi_m(\,\cdot\,;\Sigma)$ is the cdf of the centered m-variate normal distribution with covariance matrix Σ. The stdf of the mixture model $\boldsymbol{M}$ is

$$\ell(\boldsymbol{x}) = \sum_{s\in R} \left\{ \sum_{j\in J_s} a_{js} x_j \, \Phi_{|J_s|-1}\left(\boldsymbol{\eta}^{j,(s)}\left(\left(a_{js}x_j\right)_{j\in J_s}\right); \Sigma^{j,(s)}\right) \right\}, \qquad \boldsymbol{x} \in [0,\infty)^d.$$

[1] A matrix $\Gamma \in \mathbb{R}^{d\times d}$ is a variogram matrix if it is symmetric, has zero diagonal, and satisfies $\boldsymbol{v}^\top \Gamma \boldsymbol{v} < 0$ for all $\boldsymbol{0} \neq \boldsymbol{v}^\top \perp \boldsymbol{1}$. To each positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is associated a variogram matrix $\Gamma \in \mathbb{R}^{d\times d}$ via $\Gamma_{st} = \Sigma_{ss} + \Sigma_{tt} - 2\Sigma_{st}$ for $s,t \in D$.

3.2.3 Least squares estimator of tail dependence

Let again $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote i.i.d. random vectors in the domain of attraction of a max-stable distribution with stdf ℓ. We will first recall how to estimate ℓ non-parametrically. Let $k = k_n \in (0, n]$ be an intermediate sequence such that

$$k \to \infty \quad \text{and} \quad k/n \to 0, \quad \text{as } n \to \infty. \tag{3.2.6}$$

The k determines the fraction of the data that is considered as extreme and that will be used for inference. We will refer to the quantity k/n as the *tail fraction*. A straightforward non-parametric estimator of the stdf is

$$\hat{\ell}_{n,k}(\boldsymbol{x}) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}\left\{R_{i1}^n > n + 1/2 - kx_1 \text{ or } \ldots \text{ or } R_{id}^n > n + 1/2 - kx_d\right\}, \tag{3.2.7}$$

where R_{ij}^n denotes the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$ for $j = 1, \ldots, d$. Note that (3.2.7) is a minor modification of the standard empirical stdf [30]. Alternatives are the bias-corrected estimator proposed in Beirlant et al. [5] or a smooth estimator based on the empirical beta copula [67].

Suppose now that ℓ belongs to a parametric family $\{\ell(\,\cdot\,;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, with $\Theta \subset \mathbb{R}^p$ for some integer $p \geqslant 1$. Furthermore, assume the existence of a unique parameter $\boldsymbol{\theta}_0 \in \Theta$ such that $\ell(\boldsymbol{x}) = \ell(\boldsymbol{x}; \boldsymbol{\theta}_0)$ for all $\boldsymbol{x} \in [0, \infty)^d$. Einmahl et al. [33] propose a continuous updating weighted least-squares estimator for $\boldsymbol{\theta}_0$, which, in its simplest form, reduces to the ordinary least-squares estimator

$$\hat{\boldsymbol{\theta}}_{n,k} := \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \sum_{m=1}^{q} \left(\hat{\ell}_{n,k}(\boldsymbol{c}_m) - \ell(\boldsymbol{c}_m; \boldsymbol{\theta})\right)^2. \tag{3.2.8}$$

Here, $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0, \infty)^d$, with $\boldsymbol{c}_m = (c_{m1}, \ldots, c_{md})$ for $m = 1, \ldots, q$ are the q points in which ℓ and $\hat{\ell}_{n,k}$ are evaluated. The estimator (3.2.8) is easy to compute and shows good performance for many well-known max-stable models [33, Section 3].

3.3 Estimation of the mixture model

In Section 3.3.1, we will start by introducing our estimation procedure for the simplified setting where the number of extreme directions, r, is assumed to be known. The case where r is unknown will be discussed further in Section 3.3.2.

3.3.1 Known number of extreme directions

As in Definition 3.2.2, let $\boldsymbol{M}$ be the $d \times r$ mixture model with coefficient matrix A and factors $\boldsymbol{Z}_{\cdot 1}, \ldots, \boldsymbol{Z}_{\cdot r}$ with stdfs $\ell^{(1)}, \ldots, \ell^{(r)}$. Recall that $R = \{1, \ldots, r\}$. From

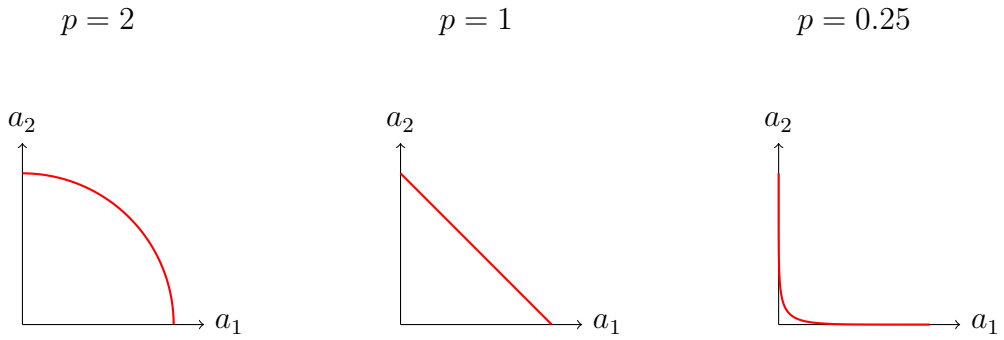


Figure 3.2
Bivariate contours of $(\Sigma_s a_s^p)^{1/p} = 1$ for given values of p.

now on, suppose that the stdfs $\ell^{(1)}, \ldots, \ell^{(r)}$ belong to a common parametric family $\{\ell(\,\cdot\,;\boldsymbol{\theta}_Z) : \boldsymbol{\theta}_Z \in \Theta_Z\}$, where $\Theta_Z \subset \mathbb{R}^v$ for some integer $v \geqslant 1$, and that for each $s \in R$, there exists a unique parameter $\boldsymbol{\theta}_Z^{(s)} \in \Theta_Z$ such that $\ell^{(s)} = \ell(\,\cdot\,;\boldsymbol{\theta}_Z^{(s)})$. In what follows, for simplicity, we assume that the dependence parameters across the columns are identical, meaning that all the $\boldsymbol{\theta}_Z^{(s)}$ are equal for $s \in R$. However, all the results remain valid in the general case. Let $\ell(\cdot\,;\boldsymbol{\theta})$ be the stdf associated with $\boldsymbol{M}$ as in Theorem 3.2.4. More specifically, let $\boldsymbol{\theta} = (A, \boldsymbol{\theta}_Z)$, where we recall $A \in \Theta_A := \mathcal{M}_{d,r}([0,1])$ and $\boldsymbol{\theta}_Z = \boldsymbol{\theta}_Z^{(1)} \in \Theta_Z \subset \mathbb{R}^v$. Note that A can equivalently be represented as a vector $\boldsymbol{\theta}_A \in [0,1]^{d\times r}$ by stacking its rows sequentially. For simplicity, we will use A and $\boldsymbol{\theta}_A$ interchangeably throughout the paper.

The *preliminary non-standardized penalized least-squares (PNPLS) estimator* is defined as

$$\hat{\boldsymbol{\theta}}_{\mathrm{pr}} = \left(\hat{A}_{\mathrm{pr}}\,,\, \hat{\boldsymbol{\theta}}_{Z,\mathrm{pr}}\right) = \underset{\boldsymbol{\theta}=(A,\boldsymbol{\theta}_Z)\in(\Theta_A,\Theta_Z)}{\arg\min} \left\{ \sum_{m=1}^{q} \left(\hat{\ell}_{n,k}(\boldsymbol{c}_m) - \ell(\boldsymbol{c}_m;\boldsymbol{\theta})\right)^2 + \lambda \mathcal{P}(A) \right\}, \tag{3.3.1}$$

where $\lambda > 0$ is called the *penalization parameter* and $\mathcal{P}(A) = \sum_{j\in D} \left(\sum_{s\in R} a_{js}^p\right)^{1/p}$, for some *penalization exponent* $p \in (0,1)$.

For $\lambda = 0$, the PNPLS estimator is equal to the ordinary least-squares estimator defined in (3.2.8). The penalization term $\lambda P(A)$ in (3.3.1) is needed to find zero entries in the coefficient matrix A. We restrict our attention to $p \leqslant 1$. Geometrically, using a penalization $\mathcal{P}(A)$ with $p \leqslant 1$ promotes sparsity because its level sets are non-round and tend to intersect the loss contours at sparse solutions—similar to the well-known Lasso case when $p = 1$. This is in contrast to the case of $p > 1$, as illustrated in Figure 3.2. In Remark 3.3.3, we present some additional arguments for the choice $p \leqslant 1$ related to the minimization algorithm.

The minimization problem in (3.3.1) is non-convex and is not guaranteed to give a global minimum. The parameter (vector) $\boldsymbol{\theta}_Z$, corresponding to the latent factors

$\boldsymbol{Z}_{\cdot 1}, \ldots, \boldsymbol{Z}_{\cdot r}$, may be particularly difficult to estimate. A straightforward way to improve the PNPLS estimator is as follows. After obtaining $\hat{\boldsymbol{\theta}}_{\text{pr}}$, we follow-up with a two-step estimation procedure, where each step minimizes (3.3.1) with respect to either $\boldsymbol{\theta}_Z \in \Theta_Z$ or $A \in \Theta_A$. More precisely,

1. let $\hat{A}_{\text{new}}$ be the partial solution of (3.3.1) w.r.t. A, where $\boldsymbol{\theta}_Z$ is fixed to $\hat{\boldsymbol{\theta}}_{Z,\text{pr}}$;
2. write $\hat{A}_{\text{new}} = (\hat{a}_{11,\text{new}}, \ldots, \hat{a}_{d1,\text{new}}, \ldots, \hat{a}_{1r,\text{new}}, \ldots, \hat{a}_{dr,\text{new}})$ and standardize to ensure unit row sums,

$$\hat{a}_{js} = \hat{a}_{js,\text{new}} \Big/ \sum_{l=1,\ldots,r} \hat{a}_{jl,\text{new}}, \qquad j \in D, s \in R. \tag{3.3.2}$$

   This yields $\hat{A} \in \mathcal{M}_{d,r}([0,1])$ with unit row sums;
3. let $\hat{\boldsymbol{\theta}}_Z$ be the partial solution of (3.3.1) w.r.t. $\boldsymbol{\theta}_Z$, where A is fixed to $\hat{A}_{\text{new}}$.

The resulting estimate $\hat{\boldsymbol{\theta}} = (\hat{A}, \hat{\boldsymbol{\theta}}_Z)$ is called the *penalized least-squares (PLS) estimator*.

The two steps of the estimation procedure can also be iteratively alternated until a predefined stopping criterion is met. However, as we will demonstrate in the simulation studies in Section 3.4, performing these steps only once is typically sufficient to achieve satisfactory results in practice. The full estimation procedure is explained in Algorithm 2.

Remark 3.3.1. Since A has unit row sums by definition, one can argue that the minimization problem (3.3.1) can be reduced to a lower-dimensional formulation where $\Theta_A = [0,1]^{d\times(r-1)}$ by eliminating the column $\boldsymbol{a}_{\cdot r}$ and imposing the constraint $a_{j1} + \cdots + a_{jr-1} \leqslant 1$, for each $j \in D$. However, such an approach penalizes only the first $r-1$ columns of the matrix A, leading to an asymmetric treatment of the parameters.

Remark 3.3.2. By Theorem 3.2.4, the extreme directions of a mixture model correspond to the set of signatures of its coefficient matrix A. For $s \in R$, the sets $\hat{J}_s = \left\{ j \in D : \hat{a}_{js} > 0 \right\}$ thus explicitly provide estimates of the extreme directions associated with the data.

Minimization algorithm and choice of p. To solve the minimization problem (3.3.1), we use the bounded limited-memory modification of the BFGS quasi-Newton method (`L-BFGS-B` algorithm in `R`), see Byrd et al. [10]. This iterative method has the update step

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m - \alpha_m H_m^{-1} \nabla L\left(\boldsymbol{\theta}_m\right), \qquad m \geqslant 0, \tag{3.3.4}$$

where L is the loss function in (3.3.1), $\alpha_m > 0$ is the step-size, H_m is a $(d \times r + v) \times (d \times r + v)$ approximate curvature positive definite matrix capturing second-order

information based on the gradients observed during optimization, and $\nabla L(\boldsymbol{\theta}_m)$ is the gradient of L evaluated at $\boldsymbol{\theta}_m$. At each step, the gradient is approximated using finite differences method.

Remark 3.3.3. Another motivation for choosing the penalization exponent $p \in (0, 1]$ is that the update direction in (3.3.4) is opposite to the gradient. For a fixed parameter a_{js}, the penalization term in (3.3.1) then contributes to the gradient proportionally to

$$T(p) = a_{js}^{p-1} \left(\sum_{t \in R} a_{jt}^{p} \right)^{1/p-1} . \tag{3.3.5}$$

Recall that $p < 1$ and thus $p - 1 < 0$. Consequently, if the true value of a_{js} is zero, then as the corresponding coefficient approaches zero at some step of the minimization algorithm L-BFGS-B, the term $T(p)$ in Equation (3.3.5) diverges to $+\infty$. As a result, the gradient also diverges to $+\infty$, and since the update direction in (3.3.4) is opposite to the gradient, this leads to a solution where $\hat{a}_{js} = 0$.

Stopping criterion. The stopping criterion in Step 9 of Algorithm 2 plays a crucial role in ensuring convergence while preventing unnecessary computations. Here, we adopt a criterion based on parameter improvement, terminating the iterations when the change between successive parameter estimates becomes sufficiently small:

$$\|\hat{\boldsymbol{\theta}}_{\text{new}} - \hat{\boldsymbol{\theta}}_{\text{old}}\|_{d \times r + v} < \epsilon,$$

where ϵ is a predefined small threshold, $\hat{\boldsymbol{\theta}}_{\text{new}}$ and $\hat{\boldsymbol{\theta}}_{\text{old}}$ are as defined in Algorithm 2, and $\|\cdot\|_{d \times r + v}$ is the Euclidean norm. Additionally, to prevent infinite loops, we impose a hard limit on the number of iterations. Specifically, the algorithm is forced to stop if it reaches a predefined upper bound max.

Choice of tuning parameters. There are three tuning parameters for the estimation procedure in Algorithm 2: the tail fraction k/n, the penalization exponent p and the penalization parameter λ. First, the choice of the tail fraction is a classical trade-off between bias and variance; a smaller k/n or larger k/n introduces more variance or more bias, respectively, for the empirical stdf in (3.2.7). Second, for $0 < p_1 \leqslant p_2 < 1$, we have $T(p_1)/T(p_2)$ explodes to ∞ when a_{js} is closer to zero, with T defined in (3.3.5). The same reasoning after (3.3.5) shows that a smaller penalization exponent yields stronger penalization. Different choices of p will be illustrated in Section 3.4. Finally, the penalization parameter λ will be chosen using the K-fold cross validation in Algorithm 3.

3.3.2 Identifying the extreme directions

In practice, the number of extreme directions r is unknown. Using the same assumptions and notation as in Section 3.3.1, we now estimate $\boldsymbol{\theta}$ without taking any prior knowledge of the number of columns, r, of A into account. Our proposed method is iterative: at each step $t \geqslant 1$, we fix the number of columns to t and compute the standardized penalized least-squares estimator $\hat{\boldsymbol{\theta}} = \left(\hat{A}, \hat{\boldsymbol{\theta}}_Z\right) \in \Theta_A \times \Theta_Z$ using Algorithm 2. We then iterate to $t+1$ as long as all sub-vectors $(a_{1s}, \ldots, a_{ds})$ of $\boldsymbol{\theta}_A = (a_{11}, \ldots, a_{d1}, \ldots, a_{1t}, \ldots, a_{dt})$ remain non-null for all $s = 1, \ldots, t$. The method is presented in Algorithm 4.

The *DAMEX* algorithm introduced by Goix et al. [50, Algorithm 1] can also be used to identify extreme directions. It relies on three key parameters: a threshold beyond which data are considered extreme, analogous to the tail fraction k/n used in Algorithm 2; a tolerance parameter that heuristically defines the region associated with an extreme direction; and a mass threshold, which determines whether a subset $J \subset V$ with sufficiently large mass is considered an extreme direction. As explained in Goix et al. [50], in a supervised setting, these parameters are selected via cross-validation. As we will detail in the simulation study in Section 3.4, this is also the case for Algorithm 2: the choice of the tail fraction involves a trade-off between bias and variance. In contrast, the choice of the penalization exponent p is less critical, provided that the penalization parameter λ is appropriately tuned—this being achieved, as outlined in Algorithm 3, through K-fold cross validation.

Our procedure is tailored to max-stable mixture models, since Mourahib et al. [75, Theorem 4.8] show that every max-stable distribution with unit-Fréchet margins can in fact be represented as a max-stable mixture model. Of course, in practice, we limit ourselves to a limited number of parametric models for the factors $\boldsymbol{Z}_{\bullet 1}, \ldots, \boldsymbol{Z}_{\bullet r}$.

As in Section 3.3.1, at each step $t \geqslant 1$ of Algorithm 4, the factors $\boldsymbol{Z}_{\bullet 1}, \ldots, \boldsymbol{Z}_{\bullet t}$ are assumed to follow a parametric model ensuring that the only extreme direction of each factor is D. This assumption could be seen as a limitation of Algorithm 4, as it requires selecting and imposing a specific model for the dependence structure of the factors. However, for the purpose of identifying extreme directions, the choice of a parametric model for the factors is less critical, since the extreme directions of a mixture model $\boldsymbol{M}$ depend only on the matrix A and not on the specific parametric model imposed on the factors; see Theorem 3.2.4. This will also be validated through simulations in Section 3.4.

3.4 Simulation study

In this section, we conduct a simulation study to evaluate the estimation procedure introduced in Section 3.3. We begin by outlining the setup used throughout the simulations in Section 3.4.1. First we consider the setting where the number of extreme directions is known, as described in Algorithm 2 and detailed in Section 3.4.2. We then turn to the more general case in which the number of extreme directions is unknown, corresponding to Algorithm 4, and discussed in Section 3.4.3.

3.4.1 Preliminaries

Coefficient matrix. Throughout the simulation study, we fix the dimension $d = 4$ and the coefficient matrix

$$A = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 3/4 & 1/4 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{pmatrix}. \tag{3.4.1}$$

Domain of attraction. We simulate samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of size n in the domain of attraction of a mixture model $\boldsymbol{M}$ as defined in (3.2.3). To do so, we simply perturb $\boldsymbol{M}$ by adding a lighter-tailed noise. More precisely,

$$\boldsymbol{X}_i = \boldsymbol{M}_i + \boldsymbol{N}_i, \qquad i = 1, \ldots, n, \tag{3.4.2}$$

where for $i = 1, \ldots, n$, $\boldsymbol{M}_i \in \mathbb{R}^d$ follows a mixture model and $\boldsymbol{N}_i \in \mathbb{R}^d$ has independent centered Gaussian margins with variance $\sigma^2 > 0$. Algorithm 5 outlines the procedure for simulating from a mixture model, relying on the generation of the max-stable random vectors $\boldsymbol{Z}_{\cdot 1}, \ldots, \boldsymbol{Z}_{\cdot r}$. This is feasible for most popular models; see Dombry et al. [28, Algorithm 1] or the `R` package `mev` [6]. Furthermore, code for Algorithm 5 is freely accessible for both the mixture logistic model and the mixture Hüsler–Reiss model.[2]

Starting points in the minimization algorithm. To calculate the PNPLS estimator using the `L-BFGS-B` algorithm, we need to provide good starting values for $\boldsymbol{\theta} = (A, \boldsymbol{\theta}_Z)$. Suppose that the number of extreme directions is fixed to t, and therefore $A \in \mathcal{M}_{d,t}([0,1])$. Consider a d-variate sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and write $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})$, for $i = 1, \ldots, n$. We determine a starting value for A

[2] `https://github.com/AnasMourahib/Penalized_least-squares_estimator/blob/main/Functions/Simulation_mixture_model.R`.

by applying the k-means clustering algorithm with t clusters to the unit-Pareto transformed data,

$$\left(\frac{n}{n+1-R_{i1}}, \ldots, \frac{n}{n+1-R_{id}}\right), \qquad i = 1, \ldots, n,$$

where for $i = 1, \ldots, n$ and $j = 1, \ldots, d$, R_{ij} is the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$. We retain only the top 10% of observations with the highest sum of transformed values.

Starting values for the dependence parameters $\boldsymbol{\theta}_Z$ will be assigned randomly, depending on the selected model. As in Section 3.3, recall that for simplicity, we suppose that the dependence parameters across the columns of A are the same.

Grid points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$. Both the parametric and the empirical stdf are evaluated in $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$ during the minimization (3.3.1). The choice of the number q of points involves a trade-off: selecting a sufficiently large q ensures reliable estimation, while keeping q reasonably small reduces computational expense. In our case, using the `R` package `tailDepFun` [63], we select points that can be generated in a 4-dimensional space:

- **for the mixture logistic fit:** using values $\{1/4, 1/3, 1/2, 3/4, 1\}$ where each point satisfies the condition of having either two or three non-zero components. This gives $q = 650$ points;
- **for the mixture Hüsler–Reiss fit:** using values $\{1/6, 1/8, 1/4, 1/3, 2/3, 3/4, 1\}$, where each point satisfies the condition of having only two non-zero components. This gives $q = 384$ points.

Performance assessment. We simulate $N = 100$ samples as in (3.4.2) and for each sample, we obtain the PLS estimator $\hat{\boldsymbol{\theta}}^{(l)} = (\hat{A}^{(l)}, \hat{\boldsymbol{\theta}}_Z^{(l)})$, $l = 1, \ldots, N$, using Algorithm 2. We focus on the mixture logistic and the mixture Hüsler–Reiss models (see Section 3.2.2). Therefore, $\boldsymbol{\theta}_Z \in \mathbb{R}$ for the mixture logistic model and $\boldsymbol{\theta}_Z \in \mathbb{R}^{d(d-1)/2}$ for the mixture Hüsler–Reiss model.

First, we start by assessing the performance of our estimator in terms of the identification of extreme directions. We first introduce the score, denoted "ED-S", which represents the Jaccard distance and computes the fraction of misidentified extreme directions,

$$\text{ED-S} = \frac{1}{N}\sum_{l=1}^{N} \mathrm{D}(\hat{\mathcal{J}}_l, \mathcal{J}), \qquad \mathrm{D}(\hat{\mathcal{J}}_l, \mathcal{J}) = 1 - \frac{|\hat{\mathcal{J}}_l \cap \mathcal{J}|}{|\hat{\mathcal{J}}_l \cup \mathcal{J}|} = \frac{|\hat{\mathcal{J}}_l \triangle \mathcal{J}|}{|\hat{\mathcal{J}}_l \cup \mathcal{J}|} \tag{3.4.3}$$

where $\mathcal{J}$ and $\hat{\mathcal{J}}_l$ denote the signatures of A and $\hat{A}^{(l)}$, respectively, as defined in (3.2.4).

Second, we assess the performance of our estimator by measuring its proximity to the true parameter $\boldsymbol{\theta} = (A, \boldsymbol{\theta}_Z)$ of the mixture model. In the following, for $l = 1, \ldots, N$, we reorder the columns of $\hat{A}^{(l)}$ so as to minimize the difference $\operatorname{diff}\left(\hat{A}^{(l)}, A\right)$, as defined in (3.4.5), but without the denominator term. For simplicity, we continue to denote the resulting matrix by $\hat{A}^{(l)}$. To construct a robust performance metric and ensure that the marginal components of our estimator are equally scaled, we mitigate the disproportionate influence of their magnitudes. This is achieved by normalizing each coefficient of $\hat{\boldsymbol{\theta}}$ using the maximum value computed over the N estimates corresponding to that component. More precisely, let $\hat{m}_A = (\hat{m}_{A,js})_{j\in D,\, s\in R}$ denote the $(d \times r)$ matrix where the (j,s)-th entry represents the maximum value calculated based on the coefficients $\hat{a}_{js}^{(l)}$ obtained from the matrices $\hat{A}^{(l)}$, for $l = 1, \ldots, N$. Similarly, recall from Section 3.3 the dimension v of the dependence parameter vectors $\hat{\boldsymbol{\theta}}_Z$. Let $\hat{\boldsymbol{m}}_Z = (\hat{m}_{Z,t})_{t=1,\ldots,v}$ denote the v-dimensional vector where the t-th entry represents the maximum value calculated based on the coefficients $\hat{\theta}_t^{(l)}$ obtained from the vectors $\hat{\boldsymbol{\theta}}_Z^{(l)}$, for $l = 1, \ldots, N$. We define a standardized version of the mean-squared error as

$$\text{SMSE} = \frac{1}{N}\sum_{l=1}^{N}\left[\operatorname{diff}_{m_A}\left(\hat{A}^{(l)},\, A\right) + \operatorname{diff}_{m_Z}\left(\hat{\boldsymbol{\theta}}_Z^{(l)}, \boldsymbol{\theta}_Z\right)\right], \quad \text{where} \tag{3.4.4}$$

$$\operatorname{diff}_{m_A}\left(\hat{A}^{(l)},\, A\right) = \sum_{j\in D}\sum_{s\in R}\frac{\left(\hat{a}_{js}^{(l)} - a_{js}\right)^2}{\hat{m}_{A,js}^2} \quad \text{and} \quad \operatorname{diff}_{m_Z}\left(\hat{\boldsymbol{\theta}}_Z^{(l)}, \boldsymbol{\theta}_Z\right) = \sum_{t=1}^{v}\frac{\left(\hat{\theta}_{Z,t}^{(l)} - \theta_{Z,t}\right)^2}{\hat{m}_{Z,t}^2}. \tag{3.4.5}$$

Note that the maximum of a coefficient across N estimates is zero if and only if all estimates of that coefficient are equal to zero. This occurs only for the null coefficients of A. In such cases, by convention, we define $0/0 = 0$.

3.4.2 Known number of columns

We consider the mixture logistic model in Example 3.2.5 and the mixture Hüsler–Reiss model in Example 3.2.6, both with coefficient matrix A in (3.4.1) and

1. dependence parameter $\alpha := \alpha_1 = \ldots = \alpha_4 = 0.25$ for the mixture logistic model,
2. variogram matrix $\Gamma := \Gamma^{(1)} = \ldots = \Gamma^{(4)} \in \mathcal{M}_{4,4}([0,\infty))$, with all off-diagonal elements equal to 1 for the mixture Hüsler–Reiss model.

Note that for the mixture Hüsler–Reiss model, we do not suppose that all the off-diagonal elements of the variogram matrix are equal in the estimation procedure, meaning that $\Theta_Z \subset (0,\infty)^6$ in (3.3.1).

In this section, we suppose that the number of extreme directions, $r = 4$, is known. We generate $N = 100$ samples of size $n \in \{1\,000, 2\,000, 3\,000\}$ in the domain of attraction of the mixture logistic model and of the mixture Hüsler–Reiss model as in (3.4.2), both with a noise variance set to $\sigma^2 = 9$.

Effect of the penalization exponent. We start by studying the effect of the penalization exponent p used in the minimization problem (3.3.1) on the scores ED-S in (3.4.3) and SMSE in (3.4.4). For the tail fraction, we set $k/n = n^c/n$ with $c = \{0.4, 0.5, 0.6\}$. With sample sizes $n \in \{1\,000, 2\,000, 3\,000\}$, this corresponds approximately to $k/n \in \{0.01, 0.02, 0.05\}$. Recall that we choose the penalization parameter λ in (3.3.1) based on a 10-fold cross validation (Algorithm 3) over a well chosen grid. We use an adaptive penalization parameter grid, meaning that for each value of the penalization exponent p, we have a distinct grid of penalization parameters λ.

Figure 3.3 displays the scores ED-S and SMSE as functions of the penalization exponent p. Results are shown for the mixture logistic fit. Qualitatively similar results arise under the mixture Hüsler–Reiss fit. The effect of p is not important as long as the penalization parameter λ in (3.3.1) is well chosen. In fact, for $0 < p_1 < p_2 < 1$ and a parameter a of $\boldsymbol{\theta}_A$, the term T in (3.3.5), which governs the opposite direction step in the minimization algorithm, satisfies $T(p_1)/T(p_2) \to \infty$, whenever the parameter $a \to 0$. It follows that the penalization in (3.3.1) is stronger with smaller penalization exponent. This explains the use of an adaptive grid of penalization parameters, where the values of the grid associated with p_1 are smaller than the ones used for p_2.

Effect of the tail fraction. We study the effect of the tail fraction k/n used in (3.3.1) on the scores ED-S in (3.4.3) and SMSE in (3.4.4). Figure 3.3 shows that the effect of the penalization exponent p is not important as long as the penalization parameter λ is well chosen using 10-fold cross validation. Therefore, we set $p = 0.4$ and perform a 10-fold cross validation to choose λ, where for each value p, we use an adapted grid.

Figure 3.4 illustrates the behavior of the scores ED-S and SMSE as functions of the tail fraction k/n for a mixture logistic and a mixture Hüsler–Reiss fit, respectively.

For the mixture logistic model, we observe that both scores tend to improve near the middle of the tail fraction grid. This pattern reflects a bias–variance trade-off: a smaller tail fraction k/n results in higher variance but lower bias, as fewer points are treated as extreme observations. In contrast, a larger k/n leads to lower variance but increased bias, since more points—possibly not truly extreme—are included in the estimation. We also observe that as the sample

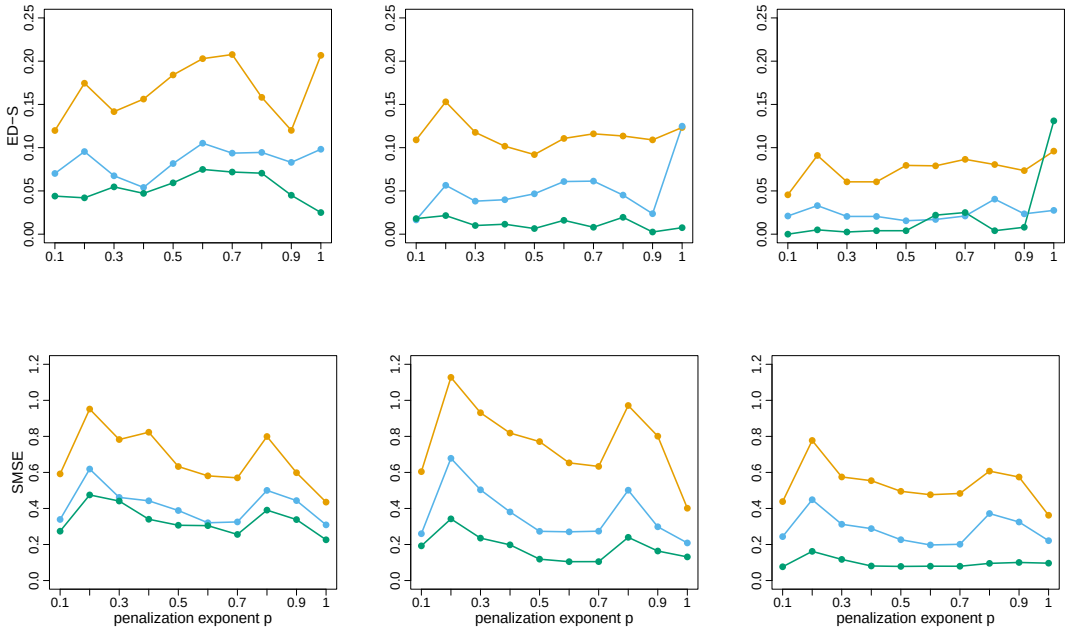


Figure 3.3
Scores ED-S (top) and SMSE (bottom) defined in (3.4.3) and (3.4.4), respectively, as functions of the penalization exponent p, for the mixture logistic model perturbed as in (3.4.2) and sample sizes $n = 1\,000$ (left), $n = 2\,000$ (center), and $n = 3\,000$ (right). Curves correspond to varying tail fractions, with approximately $k/n = 0.01$ (orange), $k/n = 0.02$ (blue), and $k/n = 0.05$ (green).

size increases, the optimal choice of the tail fraction k/n—that is, the optimal proportion of points considered as extreme—tends to decrease. This is intuitive: with a larger sample size, a smaller proportion of extreme observations is sufficient to achieve accurate estimation.

The score ED-S for the mixture Hüsler–Reiss fit behaves similarly to that of the mixture logistic fit. In contrast, the score SMSE does not reflect the bias-variance trade-off effect observed for the mixture logistic fit. We can also see that this score is higher for the mixture Hüsler–Reiss model. This makes sense, since the dependence coefficient Γ for the mixture Hüsler–Reiss model is of dimension $d = 6$ (in our example), while the dependence coefficient for the mixture logistic model is of dimension $d = 1$. This makes the estimation of the dependence parameter for the mixture Hüsler–Reiss model more complicated.

We now discuss the estimation of the variogram matrix Γ of the mixture Hüsler–Reiss model in Point 2. Figure 3.5 shows the boxplots of the upper off-diagonal entries of Γ, except for the entry Γ_{23}; since no signature of the matrix A in (3.4.1)

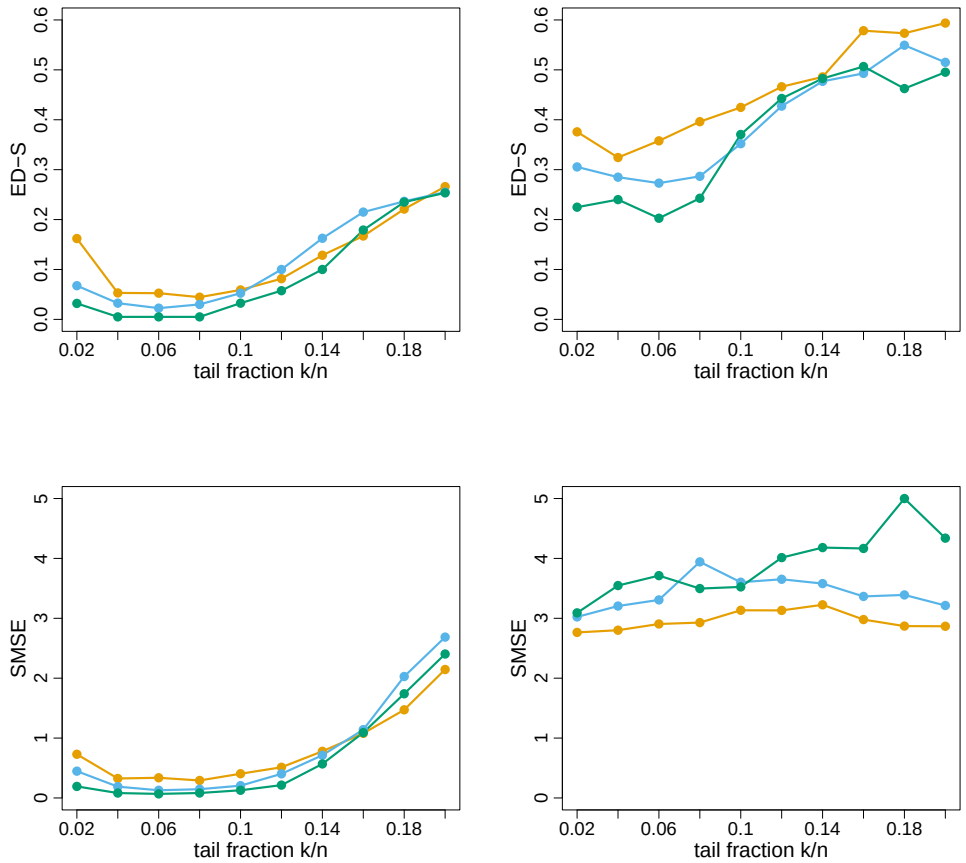


Figure 3.4
Scores ED-S (top) and SMSE (bottom) defined in (3.4.3) and (3.4.4), respectively, as functions of the tail fraction k/n, for the mixture logistic model (left) and the mixture Hüsler–Reiss model (right), both perturbed as in (3.4.2), with penalization exponent $p = 0.4$. Curves correspond to $n = 1\,000$ (orange), $n = 2\,000$ (blue) and $n = 3\,000$ (green).

.

contains the pair $(2, 3)$, the coefficient Γ_{23} is not identifiable. More precisely, Γ_{23} does not contribute to the stdf in (3.2.5) associated to the corresponding mixture model.

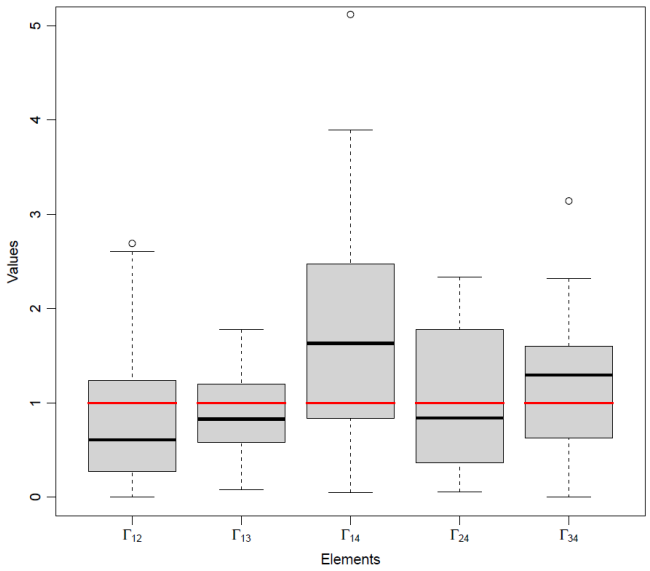


Figure 3.5
Boxplots of the estimated coefficients Γ_{ij} from the variogram matrix Γ for all pairs (i, j) included in some signature J of A. The true value 1 for each coefficient is presented with a horizontal red line. The tail fraction k/n is set to 0.06, the penalization exponent to $p = 0.4$ and the sample size to $n = 1\,000$.

3.4.3 Unknown number of columns

We now suppose that the number of extreme directions is unknown and identify the extreme directions using Algorithm 4. Following Figure 3.3, the choice of the penalization parameter p is not important as long as the penalization parameter λ is well chosen using K-fold cross validation as explained in Algorithm 3. Therefore, we simply choose $p = 0.4$ and use a 10-fold cross validation procedure to choose λ from the grid associated with $p = 0.4$. Following Figure 3.4, we choose $k/n \in \{0.04, 0.06, 0.08\}$. We simulate samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the mixture logistic model in Point 1 and samples $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ from the mixture Hüsler–Reiss model in Point 2, both perturbed as in (3.4.2). We apply Algorithm 4 with the following two settings:

1. **Correct model specification.** Fit a mixture logistic model to $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and a mixture Hüsler–Reiss model to $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$;

2. **Misspecified model.** Fit a mixture logistic model to $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ and a mixture Hüsler–Reiss model to $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

We investigate the impact of the sample size n on the score ED-S defined in (3.4.3). The results for the mixture logistic fit are shown in Figure 3.6. Qualitatively similar results arise under the mixture Hüsler–Reiss fit. Both the fit under the correct model specification and the fit under the misspecified model yield satisfactory performance even for relatively small sample sizes, with good results observed from $n = 500$ onward.

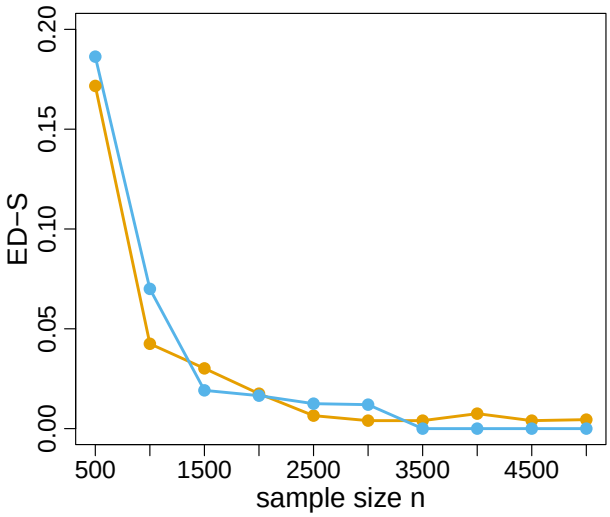


Figure 3.6
The score ED-S defined in (3.4.3), for simulations from the mixture logistic model (blue) and the mixture Hüsler–Reiss model (orange), both perturbed as in (3.4.2). In both cases, a mixture logistic model is fitted using the estimation procedure in Algorithm 4. The penalization exponent is fixed to $p = 0.4$ and the tail fraction to $k/n = 0.08$.

3.5 Applications

3.5.1 River discharge data

We illustrate our method using river discharge data from 31 stations along the Danube River. The dataset consists of daily river discharge measurements from the

summer months between 1960 and 2010. To mitigate temporal dependence, the data were declustered by Engelke et al. [40], resulting in approximately seven to ten observations per year and yielding $n = 428$ observations in total. The declustered dataset is available in the `GraphicalExtremes` package in `R`; see Engelke et al. [40].

To reduce the dimensionality and thereby lower the computational cost of identifying extreme directions, we apply a spatial clustering technique. Specifically, we partition the stations into K clusters using the Partitioning Around Medoids (PAM) algorithm [62, 8], adapted to threshold exceedances as in Kiriliouk and Naveau [65]. As a pseudo-distance between two stations s and t, we use

$$\hat{d}_{st} = \frac{1 - \hat{\chi}_{st}}{2 \cdot (3 - \hat{\chi}_{st})}, \qquad \hat{\chi}_{st} := 2 - \hat{\ell}_{n,k,st}(1,1), \tag{3.5.1}$$

where k/n represents the tail fraction and $\hat{\ell}_{n,k,st}$ the pairwise empirical stdf in (3.2.7) for the margin $\{s, t\}$.

On the one hand, the number K of clusters balances model simplicity (fewer clusters) and intra-cluster homogeneity. Various methods exist for selecting an appropriate value of K, such as minimizing within-cluster inertia [62]. On the other hand, as explained in Section 3.4, the choice of the tail fraction k/n is a trade-off between bias and variance. Here, we simply choose $K = 5$ and $q = 0.05$ which gives spatially coherent results as displayed in Figure 3.7. The cluster centers are given then by the five stations $\{7, 18, 24, 27, 30\}$.

Consider the $d = K = 5$ stations of interest to be the cluster centers. The data $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})$ representing the time series of river discharge over the five stations are assumed to be independent and identically distributed for $i = 1, \ldots, n$ and we will identify their associated extreme directions using Algorithm 4.

Choice of tuning parameters. Before estimating the extreme directions associated with the five central stations, we first address the choice of the tuning parameters: the tail fraction k/n, the penalization exponent p, the penalization coefficient λ and the grid points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0, \infty)^5$. As noted in the simulation study, the choice of the penalization exponent p has limited impact provided that the penalization parameter λ is appropriately selected. We simply choose $p = 0.4$, although, in terms of results, other values of the penalization exponent yield similar findings. Guided by the results in Section 3.4 and given the sample size, $n = 428$, we set the tail fraction to $k/n = 0.1$. The penalization parameter λ is selected via 5-fold cross-validation procedure as described in Algorithm 3 over a grid of values. This grid is constructed in a greedy iterative way such that the optimal value λ^* lies roughly in the middle of the range: we start with a grid $\{0.1, 0.2, \ldots, 10\}$, ranging from 0.1 to 10 in steps of 0.1, we then find the optimal value with respect to the 5-fold cross-validation, and construct a new grid centered

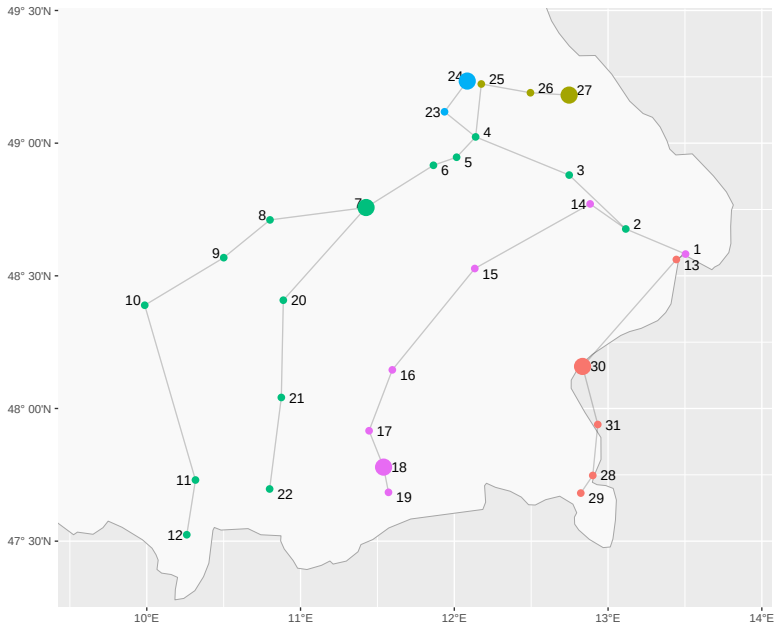


Figure 3.7
31 stations from the Danube dataset clustered into $K = 5$ groups using an adapted version of the PAM algorithm with the pseudo-distance $\hat{d}$ in (3.5.1) with tail fraction $k/n = 0.1$. Stations with the same color belong to the same cluster, and within each cluster, the station represented by a larger-sized shape indicates the cluster center.

around this value. We repeat this process until results are stable. This yields a grid of values, $\{0.01, 0.02, \ldots, 1\}$, ranging from 0.01 to 1 in steps of 0.01. Note that given the limited sample size, $n = 428$, we perform 5-fold cross-validation, instead of 10-fold cross-validation, as in Section 3.4, where the sample size was larger. Finally, we select points that can be generated in a 5-dimensional space using values $\{1/4, 1/3, 1/2, 3/4, 1\}$ where each point satisfies the condition of having either two or three non-zero components. This yields $q = 1500$ evaluation points.

Results. At each step of Algorithm 4, we fit a mixture logistic model to the five-dimensional river discharge dataset. While in the simulation study, the initial dependence parameter α of the mixture logistic model was randomly initialized within the interval $[0, 1]$, here we set a fixed initial value $\alpha = 0.5$. The results are summarized in Table 3.1 (left). The algorithm identifies five extreme directions, which appear spatially consistent with the stations indicated on the map in Figure 3.7.

As indicated in Table 3.1 (left), each extreme direction is associated with a weight, see [75, Proposition 5.1]. Specifically, for a mixture logistic model with column coefficients $(\boldsymbol{a}_{\bullet J} : J \in \mathcal{J})$, where $\mathcal{J}$ is the set of extreme directions (equivalently, the signature of the model) and a dependence parameter $\alpha \in [0, 1]$, the scalar

$$w_{\tilde{J}} = \frac{\left(\sum_{j \in \tilde{J}} a_{j\tilde{J}}^{1/\alpha}\right)^{\alpha}}{\sum_{J \in \mathcal{J}} \left(\sum_{j \in J} a_{jJ}^{1/\alpha}\right)^{\alpha}} \tag{3.5.2}$$

can be interpreted as the weight corresponding to the extreme direction $\tilde{J} \in \mathcal{J}$.

Results make sense given the hydrological nature of the dataset. The standard extreme direction $\{7, 18, 24, 27, 30\}$ can be explained by the fact that extreme river discharges are often caused by large-scale precipitation during the summer months; see Böhm and Wetzel [9]. The other non-standard extreme directions can be divided into two clusters. The first cluster corresponds to the extreme directions $\{24\}$, $\{24, 27\}$, and $\{7, 24\}$, with stations 7, 24, and 27 located along Bavarian rivers. Stations 24 and 27 are geographically close and often experience the same weather systems. However, the presence of the extreme direction $\{7, 24\}$ is surprising, since the two stations receive water from different sources; see [41, Figure 9]. This might explain why the extreme direction $\{7, 24\}$ is detected with the lowest weight. The second cluster corresponds to the extreme direction $\{7, 18, 30\}$, with stations 7, 18, and 30 located along Alpine tributaries (southern Germany and western Austria). This can be explained by the fact that these three stations receive water from the same region, flowing from south to north; see [41, Figure 9].

Comparison to other algorithms from literature. The *DAMEX* algorithm introduced in Goix et al. [50, Algorithm 1] identifies only the extreme direction $\{7, 18, 24, 27, 30\}$, even when varying its tuning parameters. This extreme direction was also identified by our algorithm with the largest weight (see Table 3.1 (left)). Our extreme direction identification method thus appears capable of detecting sparser extreme-direction structures. However, it is computationally more expensive than the DAMEX algorithm.

Performance assessment. To assess the fit of the mixture logistic model to the Danube dataset, we compare the empirical extremal correlation in (3.5.1) with the estimated extremal correlation for each pair of stations. For a mixture logistic model with estimated matrix $\hat{A} \in \mathcal{M}_{d,r}([0,1])$ and dependence coefficient $\hat{\alpha}$, the extremal correlation between two margins s and t is

$$\hat{\chi}_{st,\,\mathrm{est}} = \sum_{k=1}^{r} \left(\hat{a}_{sk}^{1/\hat{\alpha}} + \hat{a}_{tk}^{1/\hat{\alpha}} \right)^{\hat{\alpha}}. \tag{3.5.3}$$

Figure 3.8 (left) compares the empirical and fitted extremal correlations. Here we use $p = 0.4$ and $k/n = 0.1$, noting that other choices of tuning parameters yield qualitatively similar results. The agreement is strong, as most points lie close to the line $x = y$. Note that the evaluation process is entirely separate from the model fitting.

3.5.2 Financial industry portfolios

We consider the value-weighted returns for $d = 10$ industry portfolios, downloaded from the Kenneth French Data Library[3]: "1 = Nondurables", "2 = Durables", "3 = Manufacturing", "4 = Energy", "5 = HiTech", "6 = Telecom", "7 = Shops", "8 = Health", "9 = Utilities", "10 = Other". The individual stocks that make up the five industry portfolios are taken from all listed firms on the NYSE, AMEX, and NASDAQ. Data is available between July 1926 and March 2025, leading to a sample of size $n = 51\,922$.

Data pre-processing. We convert the standard returns to log-returns and multiply by (-1) since we are interested in extreme *losses*. Let $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{td})$ denote the negative log-returns of the five portfolios at time $t = 1, \ldots, n$. In order to obtain a time series without temporal dependence, we fit a GARCH$(1,1)$ model marginally. More precisely, for each portfolio $j = 1, \ldots, 10$,

$$X_{tj} = \mu_{tj} + \sigma_{tj}\varepsilon_{tj}, \qquad t = 1, \ldots, n, \tag{3.5.4}$$

[3] See `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_10_ind_port.html` for more details about the data and the definition of each portfolio.

where σ_t^2 is the conditional variance (volatility), and $\varepsilon_{jt} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. For more details about the GARCH model and its application to financial data, see for example Francq and Zakoian [47]. We continue with the fitted residuals

$$\hat{\varepsilon}_{tj} = \frac{X_{tj} - \hat{\mu}_{tj}}{\hat{\sigma}_{tj}}, \qquad t = 1, \ldots, n.$$

We apply the Ljung–Box test for serial correlation to each of the $d = 10$ marginal residual series $\hat{\varepsilon}_{1j}, \ldots, \hat{\varepsilon}_{nj}$, using lags 1 through 10. In every case, the test fails to reject the null hypothesis of no autocorrelation, suggesting that the residuals are serially uncorrelated.

Choice of tuning parameters. Similarly to Section 3.5.1, we set the penalization exponent to $p = 0.4$. Other values of the penalization exponent yielded similar findings. The sample size ($n = 51\,922$) of the industry portfolios dataset is much larger than the one of the Danube dataset ($n = 428$) in Section 3.5.1. Therefore, in this section, we set the tail fraction k/n to a smaller value $k/n = 0.05$. The penalization parameter λ is selected via the 10-fold cross-validation procedure described in Algorithm 3 over a grid of values. As explained in Section 3.5.1, this grid is chosen in a greedy iterative way until results stabilize. This yields a grid of values, $\{0.001, 0.002, \ldots, 0.02\}$, ranging from 0.001 to 0.02 in steps of 0.001. Finally, we select points that can be generated in a 10-dimensional space using values $\{0, 1/3, 1/2, 1\}$ where each point has two non-zero components which yields $q = 405$ evaluation points.

Results. As in Section 3.5.1, at each step of Algorithm 4 we fit a mixture logistic model (Example 3.2.5) to the five industry portfolio and set the initial value for the dependence parameter to $\alpha = 0.5$. The results are summarized in Table 3.1 (right). Unlike the DAMEX algorithm, our approach identifies multiple extreme directions, including non-standard ones. Among these, the standard extreme direction $\{1, \ldots, 10\}$ receives the highest weight, corresponding to the unique extreme direction detected by DAMEX, even when varying its tuning parameters. Again, this emphasizes that our extreme direction identification method is capable of detecting sparser extreme-direction structures. The standard extreme direction $\{1, \ldots, 10\}$ corresponds to the one assigned the highest weight, which is intuitive given that the portfolios are composed of stocks from U.S.-based exchanges (NYSE, AMEX, and NASDAQ). As such, losses are not confined to a single sector but rather spread across the entire U.S. economy. Some of the other extreme directions can also be meaningfully interpreted. For instance, "1 = NonDurables", includes stocks from sectors such as Food and Tobacco, while "10 = Other", comprises sectors like Transportation and Entertainment; these two belong to the same extreme direction,

Extreme directions	Weights
$\{7, 18, 24, 27, 30\}$	0.49
$\{7, 18, 30\}$	0.17
$\{24\}$	0.15
$\{24, 27\}$	0.14
$\{7, 24\}$	0.06

Extreme directions	Weights
$\{1, \ldots, 10\}$	0.52
$\{7, 8\}$	0.10
$\{6\}$	0.09
$\{1, 9, 10\}$	0.09
$\{4\}$	0.08
$\{1, 2, 3, 6, 7\}$	0.07
$\{2\}$	0.02

Table 3.1
Extreme directions identified by Algorithm 4 for the Danube river discharge data at five stations, discussed in Section 3.5.1 (left), and the ten industry portfolios: "1 = Nondurables", "2 = Durables", "3 = Manufacturing", "4 = Energy", "5 = HiTech", "6 = Telecom", "7 = Shops", "8 = Health", "9 = Utilities", "10 = Other", discussed in Section 3.5.2 (right), along with their associated weights as in (3.5.2).

suggesting similar tail-risk behavior. Likewise, "6 = Telecom" (including Telephone and Television Transmission), and "7 = Shops" (including Repair Shops), also fall within the same extreme direction. However, some extreme directions are more challenging to interpret without deeper economic insight. This includes singleton directions such as "2 = Durables", 4 = Energy", "6 = Telecom", which may reflect more nuanced or isolated sector-specific behaviors. Finally, the standard extreme direction $\{1, \ldots, 10\}$ prevents the existence of a negative dependence structure between every pair of distinct sectors.

Performance assessment. To assess the fit of the mixture logistic model to the industry portfolios dataset, we compute for each pair (s, t) both the empirical extremal correlation in (3.5.1) and the fitted extremal correlation in (3.5.3). Figure 3.8 (right) compares the empirical and fitted extremal correlations, computed both with $p = 0.4$ and $k/n = 0.05$. Other choices of tuning parameters yielded qualitatively similar results. The agreement is strong, as most points lie close to the line $x = y$.

3.6 Conclusion

Given a vector of d risk factors, extreme directions refer to groups of factors $J \subset \{1, \ldots, d\}$ that are large simultaneously while those outside this group are not.

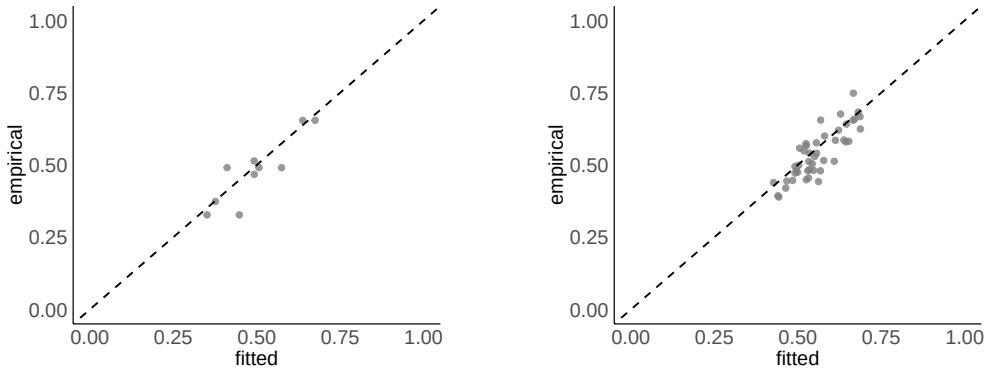


Figure 3.8
Empirical (3.5.1) versus fitted (3.5.3) extremal correlations for the results from the Danube dataset in Section 3.5.1 (left) and industry portfolios in Section 3.5.2 (right).

Mourahib et al. [75] propose a construction method of multivariate generalized Pareto distributions, or equivalently max-stable distributions, with multiple extreme directions. Extreme directions are a consequence of parameters of the mixture model being on the border of the parameter space [75, Proposition 4.5]. Parameter estimation is therefore a challenging topic. We proposed a penalized least-squares estimator (3.3.1) based on the estimator from [33]. The penalization term in this estimator is based on a "quasi-norm" L_p with $p \leqslant 1$. We motivated the choice of $p \leqslant 1$ both geometrically and heuristically. Next, we use this estimator to propose a data-driven algorithm for extreme-directions identification. Using a simulation study, we show that although the algorithm relies on a parametric assumption for the mixture model (e.g., mixture Hüsler–Reiss graphical model or mixture logistic model), the specific choice of the model is not crucial. In addition, the algorithm involves three tuning parameters; again through simulations, we demonstrate that the tuning essentially reduces to selecting two parameters: one that controls the proportion of data considered as extreme observations and the other that controls the penalization strength, chosen using K-fold cross-validation. We illustrated our algorithm on two real datasets, determining the sets of variables that form extreme directions. We compare our Algorithm with the *DAMEX* algorithm in Goix et al. [50] and show that our algorithm appears capable of detecting non-standard extreme directions.

We obtain good simulation results for our penalized least-squares estimator proposed in (3.3.1). However, no results on consistency or asymptotic normality are provided. Two main difficulties prevent us from establishing asymptotic theoretical results:

1. when the penalization exponent of the penalization term is such that $p \leqslant 1$, our PLS estimator is the solution of a non-convex minimization problem;
2. some of the true parameters lie on the boundary of the parameter space. Under certain conditions, this is treated in Kiriliouk [64] for the unpenalized version of our estimator, i.e., the weighted least-squares estimator in Einmahl et al. [33]. However, these conditions seem difficult to verify for the parametric mixture models.

In the simplest case where the true parameters are in the interior of the parameter space and no penalization is needed, asymptotic results are proved in Einmahl et al. [33] under certain conditions. However, as pointed out in the same paper, these conditions are either difficult to verify, for instance with the Hüsler–Reiss parametric model, or are not verified at all, for instance with the max-linear model.

Acknowledgment

The authors thank Sebastian Engelke for insightful discussions on the interpretation of the river discharge data analysis in Section 3.5.1, Michel Denuit for his input on the interpretation of the portfolio data in Section 3.5.2, and Anne Sabourin for generously sharing her code and providing a clear explanation of the DAMEX algorithm.

3.A Algorithms

Algorithm 5 Simulation of a mixture model

Input: integers $d, r \geqslant 1$, matrix $A = \left(a_{jk}\right)_{j\leqslant d, k\leqslant r} \in \mathcal{M}_{d,r}([0,1])$ and max-stable random vectors $\boldsymbol{Z}_{\bullet 1}, \ldots, \boldsymbol{Z}_{\bullet r} \in \mathbb{R}^d$ as in Definition 3.2.2. ▷ For simplicity, we consider the vector $\boldsymbol{Z}_{\cdot k} = (Z_{1k}, \ldots, Z_{dk}) \in \mathbb{R}^d$ for each $k = 1, \ldots, r$. However, as discussed after Definition 3.2.2, only the sub-vector $\boldsymbol{Z}_{J_k,k}$ is relevant.

Output: mixture logistic model $\boldsymbol{M} \in \mathbb{R}^d$ in (3.2.3) with coefficient matrix A and factors $\boldsymbol{Z}_{\bullet 1}, \ldots, \boldsymbol{Z}_{\bullet r}$.

1: Initialize a vector $\boldsymbol{M} = (0, \ldots, 0) \in \mathbb{R}^d$
2: Initialize a list $\tilde{Z} = \left\{\tilde{\boldsymbol{Z}}_1, \ldots, \tilde{\boldsymbol{Z}}_r\right\}$, where each $\tilde{\boldsymbol{Z}}_k \in \mathbb{R}^d$ is initialized as a zero vector.
3: **for** each $k \in \{1, \ldots, r\}$ **do**
4: Define J_k as in (3.2.4), the k-th signature of the coefficient matrix A.
5: Set $\tilde{Z}_{k,j}$, the j-th component of $\tilde{\boldsymbol{Z}}_k$, to $a_{jk} \cdot Z_{jk}$, for every $j \in J_k$.
6: **end for**
7: Compute $M_j = \max\left\{\tilde{Z}_{j1}, \ldots, \tilde{Z}_{jr}\right\}$, for each $j \in \{1, \ldots, d\}$.
8: **return** $\boldsymbol{M}$

Algorithm 2 Estimation of the parameters of a mixture model for known r

Input: integers $d, r \geqslant 1$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0,\infty)^d$, sample $\mathcal{S} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0,\infty)^d$, sample size $n \in \mathbb{N}$, integer $k \leqslant n$, reals $0 < p \leqslant 1$ and $\lambda > 0$.

Output: penalized least-squares estimator $\hat{\boldsymbol{\theta}}$.

1: Compute the empirical stdf estimates $\left(\hat{\ell}_{n,\,k}\left(\boldsymbol{c}_m\right)\right)_{m=1,\ldots,q}$ based on the sample $\mathcal{S}$.

2: **Preliminary non-standardized penalized least-squares estimator**: Compute the parameter estimate $\hat{\boldsymbol{\theta}}_{\mathrm{pr}} = \left(\hat{A}_{\mathrm{pr}},\, \hat{\boldsymbol{\theta}}_{Z,\mathrm{pr}}\right) \in \Theta_A \times \Theta_z$ as the solution of (3.3.1) based on empirical stdf estimates calculated in Step 1.

3: Set $\hat{\boldsymbol{\theta}}_{\mathrm{old}} := (\hat{A}_{\mathrm{old}}, \hat{\boldsymbol{\theta}}_{Z,\mathrm{old}})$ to $\hat{\boldsymbol{\theta}}_{\mathrm{pr}}$.

4: Initialize stopping criterion to **FALSE**.

5: **while** stopping criterion is **FALSE do**

6: **Step 1. Update the coefficient matrix**: Compute $\hat{A}_{\mathrm{new}}$ as the solution of (3.3.1) with $\boldsymbol{\theta}_Z$ fixed at $\boldsymbol{\theta}_{Z,\mathrm{old}}$ and using the starting value $\hat{A}_{\mathrm{old}}$.

7: **Step 2. Standardization**: Apply the transformation (3.3.2) to $\hat{\boldsymbol{\theta}}_{A,\mathrm{new}}$ to ensure unit row sums.

8: **Step 3. Update the dependence parameter**: Compute $\hat{\boldsymbol{\theta}}_{Z,\mathrm{new}}$ as the solution of (3.3.1) with A fixed at $\hat{A}_{\mathrm{new}}$ and using the starting value $\hat{\boldsymbol{\theta}}_{Z,\mathrm{old}}$.

9: Check the stopping criterion; see the discussion on stopping criteria immediately following the algorithm. If verified, set stopping criterion to **TRUE**. Else, set $\hat{\boldsymbol{\theta}}_{\mathrm{old}}$ to $\hat{\boldsymbol{\theta}}_{\mathrm{new}} = (\hat{A}_{\mathrm{new}}, \hat{\boldsymbol{\theta}}_{Z,\mathrm{new}})$.

10: **end while**

11: **Standardized penalized least-squares estimator:** Define

$$\hat{\boldsymbol{\theta}} := \left(\hat{A}, \hat{\boldsymbol{\theta}}_Z\right) = \left(\hat{A}_{\mathrm{new}}, \hat{\boldsymbol{\theta}}_{Z,\mathrm{new}}\right). \tag{3.3.3}$$

12: **return** $(\hat{\boldsymbol{\theta}})$.

Algorithm 3 K-fold cross validation score for a penalization parameter λ

Input: integers $d, r \geqslant 1$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0, \infty)^d$, sample $\mathcal{S} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0, \infty)^d$, sample size $n \in \mathbb{N}$, integer $k \leqslant n$, reals $0 < p \leqslant 1$ and $\lambda > 0$.

Output: cross validation score associated to the penalization parameter λ.

1: Split the sample $\mathcal{S}$ into K (approximately) equal-sized folds $\mathcal{S}_1, \ldots, \mathcal{S}_K$.
2: **for** each fold $i \in \{1, \ldots, K\}$ **do**
3: **Validation set:** set $\mathcal{S}_i$ as the validation set.
4: **Training set:** set $\mathcal{S} \setminus \mathcal{S}_i$ (all other folds) as the training set.
5: Let $\hat{\boldsymbol{\theta}}^{(-i)}$ be the output of Algorithm 2 with parameters d, r, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$, sample $\mathcal{S} \setminus \mathcal{S}_i$, sample size $\lfloor n(K-1)/k \rfloor$, integer $\lfloor k(K-1)/K \rfloor$, reals p and λ.
6: Calculate the empirical stdf estimates $\left(\hat{\ell}^{(i)}_{\lfloor n/K \rfloor, \lfloor k/K \rfloor}(\boldsymbol{c}_m)\right)_{m=1,\ldots,q}$ using the the validation set.
7: Evaluate the trained model on the validation set using the i-th score

$$\mathrm{CV}^{(-i)}(\lambda) = \sum_{m=1}^{q} \left(\hat{\ell}^{(i)}_{\lfloor n/K \rfloor, \lfloor k/K \rfloor}(\boldsymbol{c}_m) - \ell\left(\boldsymbol{c}_m; \hat{\boldsymbol{\theta}}^{(-i)}\right) \right)^2.$$

8: **end for**
9: Compute the average cross validation score along all folds via

$$\mathrm{CV}(\lambda) = \frac{1}{K} \sum_{i=1}^{K} \mathrm{CV}^{(-i)}(\lambda).$$

10: **return** $\mathrm{CV}(\lambda)$.

Algorithm 4 Estimation of the extreme directions associated with a data sample

Input: integer $d \geqslant 1$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0, \infty)^d$, sample $\mathcal{S} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0, \infty)^d$, sample size $n \in \mathbb{N}$, integer $k \leqslant n$, reals $0 < p \leqslant 1$ and $\lambda > 0$.

Output: extreme directions associated with the sample $\mathcal{S}$.

1: Set $t = 1$.
2: Let $\hat{\boldsymbol{\theta}}_t = \left(\hat{A}, \hat{\theta}_Z\right)$ be the output of Algorithm 2 with parameters d, t, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$, sample $\mathcal{S}$, sample size n, integer $k \leqslant n$, reals p and λ.
3: Compute the signatures $\hat{J}_s = \left\{j \in D : \hat{a}_{js} > 0\right\}$ for $s = 1, \ldots, t$ and let $\hat{\mathcal{J}} = \{\hat{J}_1, \ldots, \hat{J}_t\}$.
4: **while** $\varnothing \notin \hat{\mathcal{J}}$ **do**
5: Iterate $t \leftarrow t + 1$.
6: Repeat Steps 2 and 3.
7: **end while**
8: Remove $\{\varnothing\}$ from $\hat{\mathcal{J}}$, that is, update $\hat{\mathcal{J}} \leftarrow \hat{\mathcal{J}} \setminus \{\varnothing\}$.
9: **return** $\hat{\mathcal{J}}$.

Chapter 4

Extremal graphical models with non-standard extreme directions and extremal independence

Anas Mourahib[1] Sebastian Engelke[2] Johan Segers[3]

Contents

[1]UCLouvain, LIDAM/ISBA, anas.mourahib@uclouvain.be.
[2]Research Center for Statistics, University of Geneva, sebastian.engelke@unige.ch.
[3]KU Leuven, Department of Mathematics, jjjsegers@kuleuven.be.

Abstract

Extremal graphical models offer a powerful framework for encoding conditional independencies among risk factors in the tail of their joint distribution. However, most existing constructions assume that all variables become large simultaneously—a restriction that fails when some risks remain essentially independent even in extreme events. In this paper, we propose a novel methodology for building extremal graphical models that seamlessly accommodates both (i) scenarios in which a subset of variables is extreme while others are not (such a group of variables will be called an *extreme direction*), and (ii) genuine extremal independence between certain components. As a concrete instance, we introduce the *mixture Hüsler–Reiss graphical model*, which extends the Hüsler–Reiss model. We then leverage this construction to develop an efficient algorithm for learning forest-structured extremal graphical models and for identifying extreme directions in multivariate risk data.

4.1 Introduction

Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ of risk factors and suppose our focus is on rare events in the tail of its distribution. In this regime, neither parametric nor nonparametric methods are satisfactory: parametric models rely on distributional assumptions that can be dominated by the bulk of the data, and thus tend to underestimate extreme outcomes, while nonparametric techniques demand large sample sizes—often unavailable when studying extremes. Extreme value theory (EVT) offers a rigorous framework for extrapolating beyond observed data and has found applications in finance, hydrology, weather forecasting, and other fields.

Univariate extreme value theory, that is, when the dimension d of $\boldsymbol{X}$ is $d = 1$, is a well studied topic [24, 4, 16] and modeling in this case is done via a finite-dimensional parametric family of distributions. Multivariate modeling of $d \geqslant 2$ risk factors is a more challenging topic, since risks may exhibit some sort of dependence. The goal is not only to model rare events of each risk factor independently but to capture the extremal dependence between these risk factors. Two main approaches appear in this context: the block-maxima approach and the threshold exceedances approach. In the first approach, standardized componentwise maxima of $\boldsymbol{X}$ are asymptotically modeled using a max-stable random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$ [44]. In the second approach, asymptotically, observations where at least one risk factor

is extreme are modeled using a multivariate (generalized) Pareto random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ [83].

As in classical statistics, one challenge in modeling multivariate extremal dependence is the dimension d. As the number of risk factors becomes larger, inference becomes more complicated given the limited number of observations in the tail. One solution is to look for sparsity among the dependence structure underlying the risk factors by studying the concept of extremal conditional independence.

Studying conditional independence in the context of extreme value theory is not straightforward. Given risk factors X_1, X_2, X_3, we first need to define "X_1 is extremely independent of X_3 given X_2". A straightforward first idea is to use max-stable distributions. However, it turns out that this is not so useful. In fact, under certain conditions, Papastathopoulos and Strokorb [79] show that if (Z_1, Z_2, Z_3) is a max-stable random vector, then $Z_1 \perp\!\!\!\perp Z_3 \mid Z_2$ reduces simply to $Z_1 \perp\!\!\!\perp Z_3$.

Engelke and Hitz [34] propose to define extremal conditional independence using a conditioned version of the multivariate Pareto random vector (Y_1, Y_2, Y_3) denoted by $Y_1 \perp\!\!\!\perp_e Y_3 \mid Y_2$. Once this is defined, conditional extremal independencies are encoded using a graph $\mathcal{G} = (\{1, \ldots, d\}, E)$ [68]. A d-variate Pareto random vector $\boldsymbol{Y}$ is then said to be an *extremal graphical model with respect to the graph* $\mathcal{G}$ if

$$Y_i \perp\!\!\!\perp_e Y_j \mid \boldsymbol{Y}_{\{1,\ldots,d\}\setminus\{i,j\}}, \qquad (i,j) \notin E.$$

Many parametric extremal graphical models exist in the literature, for instance, the chain-logistic model [92, 91, 17], and the Hüsler–Reiss graphical model [34]. However, most of these models suppose that all risk factors are large simultaneously. This condition is not realistic for example when some of the risk factors are nearly independent and thus a group $J \subset \{1, \ldots, d\}$ of risk factors may be large simultaneously without the remaining group of factors $\{1, \ldots, d\} \setminus J$. Such a group J of variables is called an *extreme direction* [75, 90]. To our knowledge, the only extremal graphical model that covers multiple extreme directions is the one constructed in Engelke et al. [38, Example 8.8] where the underlying graph structure is a tree.

The theoretical background of extremal graphical models with multiple extreme directions is discussed in Engelke et al. [38]. However, statistical methods and parametric constructions of such models are still to be treated.

The goal of this paper is to give a construction method of parametric extremal graphical models that on the one hand covers multiple extreme directions and on the other hand allows for extremal independence between some group of risk factors.

Contribution and outline of the paper. The paper is organized as follows. Section 4.2 introduces notation and presents the necessary background. In Sec-

tion 4.3, we state our main result, Theorem 4.3.1, which provides a construction of extremal graphical models with multiple extreme directions and possible extremal independence among groups of variables, along with the required elements. Section 4.4 applies this construction to define the *mixture Hüsler–Reiss graphical model* with respect to a graph and details its parameterization; in the special case where the graph is a forest, inference reduces to bivariate procedures. In Section 4.5, we propose a data-driven algorithm (Algorithm 6) to learn both the extreme directions and the forest dependence structure. Section 4.6 presents a simulation study evaluating the algorithm, demonstrating good finite-sample performance under appropriate choices of tuning parameters. Finally, in Section 4.7, we employ Algorithm 6 in two applications: first to characterize the extremal dependence of river-discharge measurements recorded at 31 stations from the Danube river, and then to capture the extremal dependence among financial portfolios across eight different sectors. The Appendix contains proofs, supplementary results, algorithms and figures.

4.2 Background

Notation. Let d be a positive integer and write $V = \{1, \ldots, d\}$. Throughout, bold symbols will refer to multivariate quantities. A d-dimensional vector $\boldsymbol{x}$ will be written as $\boldsymbol{x} = (x_1, \ldots, x_d)$ and for a non-empty set $I \subset \{1, \ldots, d\}$, write $\boldsymbol{x}_I = (x_i)_{i \in I}$. Let $\mathbf{0}_{|I|}$ and $\mathbf{1}_{|I|}$ denote vectors of zeroes and ones, respectively, of dimension $|I|$, the cardinality of I.

We also fix notation for some sets that will be used throughout the paper. Let $I \subset V$ be a non-empty set. For two $|I|$-dimensional vectors $\boldsymbol{a}$, $\boldsymbol{b}$, the set $[\boldsymbol{a}, \boldsymbol{b}]$ will denote the Cartesian product $\prod_{i \in I}[a_i, b_i]$ and so forth for $(\boldsymbol{a}, \boldsymbol{b})$, $[\boldsymbol{a}, \boldsymbol{b})$, $(\boldsymbol{a}, \boldsymbol{b}]$. Write $\mathbb{E}_I = [0, \infty)^I$. Let $\mathcal{B}(F)$ denote the Borel sets over a topological space F. For a set E and $A \subset E$, let $A^c = E \setminus A$. Let $\mathbb{R}^+ = [0, \infty)$ and $\bar{\mathbb{R}}^+ = \mathbb{R}^+ \cup \{+\infty\}$. For an integer $p > 0$ and a set $\mathbb{E} \subset \mathbb{R}^p$, let $\mathbb{E}^* = \mathbb{E} \setminus \left\{\mathbf{0}_p\right\}$. For a collection of sets $\mathcal{S}$, let $\mathcal{S}^*$ denote the non-empty elements of $\mathcal{S}$. For two integers d_1, $d_2 > 0$ and a set $\mathbb{S} \subset \bar{\mathbb{R}}^+$, let $\mathcal{M}_{d_1, d_2}(\mathbb{S})$ denote the set of $d_1 \times d_2$ matrices with entries in $\mathbb{S}$. For two non-empty subsets $I_1 \subset \{1, \ldots, d_1\}$ and $I_2 \subset \{1, \ldots, d_2\}$ and a matrix $M \in \mathcal{M}_{d_1, d_2}(\mathbb{S})$, let $M_{I_1, I_2} \in \mathcal{M}_{i_1, i_2}(\mathcal{S})$ be the submatrix of M obtained by selecting only the rows from I_1 and the columns from I_2, where $i_1 = |I_1|$ and $i_2 = |I_2|$. We will also use the notation M_I to refer to the matrix $M_{I,I}$. For two non-empty sets $J \subset I \subset V$, consider the sub-face

$$\mathbb{A}_{I,J} := \left\{\boldsymbol{x} \in \mathbb{E}_I : \; x_j > 0 \text{ iff } j \in J\right\}. \tag{4.2.1}$$

For fixed I, the collection $\{\mathbb{A}_{I,J} : \varnothing \neq J \subset I\}$ is a partition of $\mathbb{E}_I^*$. Let $\pi_{I,J} : \mathbb{E}_I \to \mathbb{R}^J$ be the projection $\boldsymbol{x} \mapsto \boldsymbol{x}_J$, with $\mathbb{R}^J$ the set of real vectors $(x_j)_{j \in J}$ indexed by J.

Throughout the paper, whenever the notation becomes intricate, we write π_J in place of $\pi_{I,J}$, leaving I implicit when it is clear from context. Let E be an arbitrary set and $S_1, S_2 \subset E$ with $S_1 \cap S_2 = \varnothing$. We write $S_1 \dot{\cup} S_2$ to denote the disjoint union $S_1 \cup S_2$. For two σ-finite measures μ, λ on $\mathbb{R}$, let $\mu \otimes \lambda$ denote the product measure of μ and λ. Let δ_0 be the unit-point mass at 0 and let μ_V be the product measure on $\mathbb{E}_V$ given by

$$\mu_V(\,\mathrm{d}\boldsymbol{x}_V) := \bigotimes_{j\in V} \mu_0(\,\mathrm{d}x_j) := \bigotimes_{j\in V} \Big(\,\mathrm{d}x_j + \delta_0(\,\mathrm{d}x_j)\Big)\,. \tag{4.2.2}$$

4.2.1 Multivariate extreme value theory

The tail behavior of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ with values in $\mathbb{R}^d$ can be analyzed using two primary methods: the block-maxima and the threshold exceedances methods [16, 4]. These two approaches employ various functions and measures to characterize the extremal dependence structure of $\boldsymbol{X}$. In this section, we introduce these techniques and focus on a key measure used in both—the exponent measure.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent and identically distributed (i.i.d.) copies of $\boldsymbol{X}$. Suppose that the marginal distributions of X are continuous. Typically, in extremal dependence structure analysis, each marginal distribution F_j of X_j (for $j = 1, \ldots, d$) is estimated first, followed by standardization to a unit-Pareto distribution via $1/(1 - F_j(X_j))$. For simplicity, we assume throughout that the margins of $\boldsymbol{X}$ have been transformed accordingly. In practice, during data analysis, this transformation must be based on the estimated marginal distributions. We say that $\boldsymbol{X}$ belongs to the max-domain of attraction of a random vector $\boldsymbol{M} \in \mathbb{R}^d$, denoted as $\boldsymbol{X} \in \mathrm{DA}(\boldsymbol{M})$, if for every $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{E}_V$, the following holds:

$$\lim_{n\to\infty} \mathsf{P}\left(\max_{i=1,\ldots,n} X_{i1} \leqslant nx_1, \ldots, \max_{i=1,\ldots,n} X_{id} \leqslant nx_d\right) = \mathsf{P}\left(M_1 \leqslant x_1, \ldots, M_d \leqslant x_d\right). \tag{4.2.3}$$

Here, $\boldsymbol{M}$ is referred to as a *max-stable random vector* and it has unit-Fréchet margins, meaning $\mathsf{P}(M_j \leqslant x) = \exp(-1/x)$ for $x > 0$ [4, Table 2.1].

A central concept in extreme value theory is the *exponent measure* Λ, which is a certain Borel measure associated with the max-stable random vector $\boldsymbol{M}$. It is concentrated on $\mathbb{E}_V^*$ and satisfies:

$$\mathsf{P}\left(\boldsymbol{M} \leqslant \boldsymbol{x}\right) = \exp\left\{-\Lambda\left(\boldsymbol{x}\right)\right\}, \qquad \boldsymbol{x} \in \mathbb{E}_V^*, \tag{4.2.4}$$

where $\Lambda\left(\boldsymbol{x}\right) := \Lambda\left([\mathbf{0}_d, \boldsymbol{x}]^c\right)$.

An alternative approach in multivariate extreme value theory uses threshold exceedances. Resnick [81, Proposition 5.17] states that convergence in (4.2.3) is

equivalent to

$$u\left(1-\mathsf{P}\left(\boldsymbol{X}\leqslant u\boldsymbol{x}\right)\right)\to\Lambda\left(\boldsymbol{x}\right),\qquad u\to\infty,\qquad\forall\boldsymbol{x}>\mathbf{0}_d.$$

From this, it follows that the threshold exceedances $\boldsymbol{Y}$ of $\boldsymbol{X}$ satisfy

$$\mathsf{P}\left(\boldsymbol{Y}\leqslant\boldsymbol{y}\right):=\lim_{u\to\infty}\mathsf{P}\left(\boldsymbol{X}/u\leqslant\boldsymbol{y}\mid\max(\boldsymbol{X})>u\right)=\frac{\Lambda\left(\boldsymbol{y}\wedge\mathbf{1}\right)-\Lambda\left(\boldsymbol{y}\right)}{\Lambda\left(\mathbf{1}_d\right)},\qquad\boldsymbol{y}>\mathbf{0}_d. \tag{4.2.5}$$

The limiting vector $\boldsymbol{Y}$, known as the *multivariate Pareto random (mpr) vector* associated to Λ, has support contained in the L-shaped set $\mathcal{L}=\{\boldsymbol{z}\in\mathbb{E}_V:\max(\boldsymbol{z})>1\}$. Additionally, setting $\boldsymbol{Z}=\log(\boldsymbol{Y})$ yields the standard multivariate generalized Pareto random vector [84].

Under the condition that the exponent measure Λ admits a density λ with respect to the dominating measure μ_V in (4.2.2), Equation (4.2.5) implies that the mpr vector $\boldsymbol{Y}$ admits also a density $h_{\boldsymbol{Y}}$ with respect to the restriction of μ_V to $\mathcal{L}$, given by

$$h_{\boldsymbol{Y}}(\boldsymbol{y})=\frac{\lambda(\boldsymbol{y})}{\Lambda(\mathbf{1}_d)},\qquad\boldsymbol{y}\in\mathcal{L}. \tag{4.2.6}$$

Further details are given in Mourahib et al. [75, Section 3.2]. The density $h_{\boldsymbol{Y}}$ is proportional to the exponent measure density λ and the normalizing constant $\Lambda([\mathbf{0}_d,\mathbf{1}_d]^c)\in[1,d]$ coincides with the extremal coefficient [4, Chapter 8]. In the following, we will call λ the exponent measure density associated to the mpr vector $\boldsymbol{Y}$. In this work, we construct models for exponent measures, thereby implicitly modeling mpr vectors.

4.2.2 Extreme directions and mixture model

Let $J\subseteq V$ be a fixed non-empty set and recall the sub-face $\mathbb{A}_{V,J}$ in (4.2.1). Recall the d-variate random vector $\boldsymbol{X}$ with unit-Pareto margins and suppose that $\boldsymbol{X}$ is in the max-domain of attraction of a max-stable random vector $\boldsymbol{M}$ with unit-Fréchet margins as defined in (4.2.3) and associated exponent measure Λ as in (4.2.4). The following definition was introduced in Simpson et al. [90] in order to identify extreme directions of the random vector $\boldsymbol{X}$.

Definition 4.2.1. The set J is an *extreme direction* of $\boldsymbol{X}$ or Λ if $\Lambda(\mathbb{A}_{V,J})>0$.

In the following, let $\mathcal{J}(\Lambda)$ denote the set of extreme directions of the exponent measure Λ. An extreme direction $J\in\mathcal{J}(\Lambda)$ is said to be a *non-standard extreme direction* if $J\neq V$. The extreme direction $J=V$ is the "standard" situation encountered in popular parametric models, such as the Hüsler–Reiss or logistic models. Hence, the first use of the term "non-standard extreme direction" in

Mourahib et al. [75] to refer to the other extreme directions. Mourahib et al. [75] introduced a smoothed variant of the max-linear model [32], known as the *mixture model*, which generates max-stable random vectors or equivalently mpr vectors exhibiting non-standard extreme directions.

Definition 4.2.2 (Mourahib et al. [75])**.** Let r be a positive integer and write $R = \{1, \ldots, r\}$. Let $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(r)}$ be independent random vectors in $\mathbb{R}^d$. For all $k \in R$, suppose that $\boldsymbol{Z}^{(k)} = (Z_1^{(k)}, \ldots, Z_d^{(k)})$ is a max-stable random vector with unit-Fréchet margins, exponent measure $\Lambda^{(k)}$ and a single extreme direction V, meaning that its exponent measure is concentrated on $(0, \infty)^d$. Let $A = (a_{jk})_{j \in V;\, k \in R} \in \mathcal{M}_{d,r}([0,1])$ be a coefficient matrix with unit row sums $\sum_{k=1}^{r} a_{jk} = 1$ for all $j \in V$ and positive column sums $\sum_{j=1}^{d} a_{jk} > 0$ for all $k \in R$. The d-variate random vector

$$\boldsymbol{M} = (M_1, \ldots, M_d) = \left(\max_{k \in R} \{a_{1k} Z_1^{(k)}\}, \ldots, \max_{k \in R} \{a_{dk} Z_d^{(k)}\} \right), \tag{4.2.7}$$

will be called the $(d \times r)$ *mixture model* with coefficient matrix A and factors $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(r)}$.

Mourahib et al. [75, Theorem 4.4] established that $\boldsymbol{M}$ in (4.2.7) is a max-stable random vector with unit-Fréchet margins. Furthermore, zero entries in the coefficient matrix A lead to non-standard extreme directions for the exponent measure Λ of $\boldsymbol{M}$. Specifically, for $k \in R$, let J_k denote the set of row indices in the k-th column of A that contain nonzero entries, i.e.,

$$J_k := \left\{ j \in V : \; a_{jk} > 0 \right\}. \tag{4.2.8}$$

We refer to J_k as the k-th signature of A. Mourahib et al. [75, Proposition 4.5] further showed that the set of extreme directions $\mathcal{J}(\Lambda)$ of Λ coincides with the set of signatures of A. Remark that in (4.2.7), for a given $j \in V$, only those $k \in R$ that satisfy $a_{jk} > 0$ matter for the value of the maximum. Therefore, only the sub-vectors $\boldsymbol{Z}_{J_k}^{(k)}$ need to be given. Hence, we adapt notation and let $\Lambda^{(k)}$ denote the exponent measure associated with $\boldsymbol{Z}_{J_k}^{(k)}$. Mourahib et al. [75, proof of Proposition 4.5] express the exponent measure Λ in terms of $\Lambda^{(1)}, \ldots, \Lambda^{(r)}$. From this expression, it follows that if $\Lambda^{(k)}$ admits a density $\lambda^{(k)}$ with respect to the $|J_k|$-dimensional Lebesgue measure, then Λ admits a density λ with respect to the measure μ_V defined in (4.2.2), given for μ_V-almost all $\boldsymbol{y} \in \mathbb{E}_V^*$ by

$$\lambda(\boldsymbol{y}) = \sum_{k \in R} \mathbb{1}\left\{ \boldsymbol{y} \in \mathbb{A}_{V,J_k} \right\} \lambda^{(k)} \left(\left(y_j / a_{jk} \right)_{j \in J_k} \right) \prod_{j \in J_k} (1/a_{jk}), \tag{4.2.9}$$

where we recall the sub-face $\mathbb{A}_{V,J}$ as in (4.2.1). The density λ in (4.2.9) is called a *mixture exponent measure density* with matrix A and factors $\left\{ \boldsymbol{Z}_{J_k}^{(k)} : \; k \in R \right\}$. The

associated mpr vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ with corresponding density $h_{\boldsymbol{Y}}$ in (4.2.6) is called the *mixture mpr vector*. Note that since $\boldsymbol{M}$ has unit-Fréchet margins, then $\lambda(y) = y^{-2}$, for $y > 0$, in case $d = 1$.

For a non-empty set $I \subset V$, we define the marginal density $\lambda_{V \to I}$ for $\boldsymbol{y}_I \in \mathbb{E}_I^*$ by

$$\lambda_{V \to I}(\boldsymbol{y}_I) = \int_{\mathbb{E}_{V \setminus I}} \lambda_V(\boldsymbol{y}_V) \, \mathrm{d}\mu_{V \setminus I}(\boldsymbol{y}_{V \setminus I}), \tag{4.2.10}$$

with $\mu_{V \setminus I}$ as in (4.2.2). Let $K_{A,I}$ denote the indices of the columns of A corresponding to signatures that intersect I. Formally, let

$$K_{A,I} = \{k \in R : \, J_k \cap I \neq \varnothing\}, \tag{4.2.11}$$

where J_k denotes the k-th signature of A as in (4.2.8). Straightforward calculus show that $\lambda_{V \to I}$ is an $|I|$-dimensional mixture exponent measure density with coefficient matrix $A_{I,K_{A,I}}$ and factors $\left\{\boldsymbol{Z}_{J_k}^{(k)} : \, k \in K_{A,I}\right\}$.

4.2.3 Conditional independence

Let $\boldsymbol{Y}$ be the mpr vector in (4.2.5) with support $\mathcal{L} = \{\boldsymbol{z} \in \mathbb{E}_V : \max(\boldsymbol{z}) > 1\}$. Since $\mathcal{L}$ is not a product space, the definition of extremal conditional independence directly through $\boldsymbol{Y}$ is impractical. Instead, Engelke and Hitz [34] considered the random vector $\boldsymbol{Y}^{(j)} = \boldsymbol{Y} \mid Y_j > 1$, defined as the mpr vector $\boldsymbol{Y}$ conditioned on the event $\left\{Y_j > 1\right\}$, for $j \in V$. The advantage is that for every $j \in V$, the random vector $\boldsymbol{Y}^{(j)}$ lives in the product space $\mathcal{L}^{(j)} = \left\{\boldsymbol{z} \in \mathcal{L} : \, z_j > 1\right\}$. For $A, B, C \subset V$ non-empty disjoint subsets whose union is $V = \{1, \ldots, d\}$, Engelke and Hitz [34] defined extremal conditional independence $\boldsymbol{Y}_A \perp\!\!\!\perp_{\mathrm{e}} \boldsymbol{Y}_C \mid \boldsymbol{Y}_B$ as

$$\boldsymbol{Y}_A^{(j)} \perp\!\!\!\perp \boldsymbol{Y}_C^{(j)} \mid \boldsymbol{Y}_B^{(j)}, \qquad \forall j \in V. \tag{4.2.12}$$

This concept of extremal conditional independence leads to a natural definition of graphical models for threshold exceedances. Let $\mathcal{G} = (V, E)$ be an undirected graph with nodes $V = \{1, \ldots, d\}$ and edge set E; see Section 4.A for more details about graph theory. Similarly to classical probabilistic graphical models [68], we will say that $\boldsymbol{Y}$ satisfies the global Markov property with respect to $\mathcal{G}$ if $\boldsymbol{Y}_A \perp\!\!\!\perp_{\mathrm{e}} \boldsymbol{Y}_C \mid \boldsymbol{Y}_B$ in the sense of (4.2.12) for every partition (A, B, C) of V, where C (possibly empty) separates A from B in $\mathcal{G}$; see Engelke et al. [38, Definition 6.1]. We will call an mpr vector satisfying the global Markov property with respect to $\mathcal{G}$ an extremal graphical model with respect to $\mathcal{G}$. Note that a definition of pairwise Markov property where A and C are singletons is given in Engelke and Hitz [34] too and proved to be equivalent to the global Markov property.

From now on, we suppose that $\mathcal{G}$ is a decomposable graph as in Section 4.A.1 and let $\mathcal{C}$ and $\mathcal{D}$ denote the set of its cliques and separators, respectively. Let also $\mathcal{D}^*$ denote the collection of non-empty separators of $\mathcal{G}$. For classical probabilistic graphical models, the Hammersley–Clifford theorem [68, Theorem 3.9] states that a random vector $\boldsymbol{P}$ that admits a positive and continuous density $f_{\boldsymbol{P}}$ with respect to a product measure satisfies the global Markov property with respect to the graph $\mathcal{G}$ if and only if $f_{\boldsymbol{P}}$ factorizes in terms of the marginal densities of $\boldsymbol{P}$ on the cliques $\mathcal{C}$ and the separators $\mathcal{D}$ of $\mathcal{G}$. Engelke et al. [38, Theorem 6.4] showed a similar result for exponent measures. Suppose that the mpr vector $\boldsymbol{Y}$ admits an exponent measure density λ (Eq. (4.2.6)) with respect to the dominating measure μ_V defined in (4.2.2). For $I \in \mathcal{C} \cup \mathcal{D}^*$, consider the marginal density, $\lambda_I := \lambda_{V \to I}$, defined on $\mathbb{E}_I^*$ as in (4.2.10). Moreover, let $\bar{\lambda}_I$ denote the extension of λ_I to $\mathbb{E}_I$, defined as

$$\bar{\lambda}_I(\boldsymbol{y}_I) = \begin{cases} \lambda_I(\boldsymbol{y}_I) & \text{if } \boldsymbol{y}_I \in \mathbb{E}_I^*, \\ 1 & \text{if } \boldsymbol{y}_I = \boldsymbol{0}_I. \end{cases} \tag{4.2.13}$$

By convention, we set $\bar{\lambda}_\varnothing = 1$. Engelke et al. [38, Theorem 6.4] show that the mpr vector $\boldsymbol{Y}$ satisfies the global Markov property with respect to $\mathcal{G}$ if and only if

$$\bar{\lambda}(\boldsymbol{y}) \prod_{D \in \mathcal{D}} \bar{\lambda}_D(\boldsymbol{y}_D) = \prod_{C \in \mathcal{C}} \bar{\lambda}_C(\boldsymbol{y}_C), \qquad \boldsymbol{y} \notin \mathcal{Z}(\mathcal{G}), \text{ with} \tag{4.2.14}$$

$$\mathcal{Z}(\mathcal{G}) = \bigcup_{\substack{s,\, t \in V,\ D \in \mathcal{D}: \\ D \text{ separates } s \text{ and } t}} \{\boldsymbol{y} \in \mathbb{E}_V : y_s > 0,\, y_t > 0,\, \boldsymbol{y}_D = \boldsymbol{0}_D\}. \tag{4.2.15}$$

Note that if $s, t \in V$ are not in the same connected component of $\mathcal{G}$ (see Section 4.A.1), $\varnothing$ separates s and t and the corresponding set in (4.2.15) becomes simply $\{\boldsymbol{y} \in \mathbb{E}_V : y_s > 0,\, y_t > 0\}$. In the following sections, we will call $\mathcal{Z}(\mathcal{G})$ the zero set of $\mathcal{G}$. Finally, throughout the paper, we will need to define the *safe set* $\mathcal{S}(\mathcal{G}, \lambda)$ of λ associated with the graph $\mathcal{G}$. This set includes all points that do not violate the positivity conditions imposed by the separators of the graph $\mathcal{G}$. Formally,

$$\mathcal{S}(\mathcal{G}, \lambda) = \mathbb{E}_V \setminus \bigcup_{D \in \mathcal{D}^*} \{\boldsymbol{y} \in \mathbb{E}_V : \boldsymbol{y}_D \neq \boldsymbol{0}_D, \lambda_D(\boldsymbol{y}_D) = 0\}. \tag{4.2.16}$$

4.3 Model construction

Outline of the section. In Section 4.3.1, we present our first main result, Theorem 4.3.1, which provides a construction of an extremal graphical model on a given graph from mutually consistent exponent measure densities on its cliques, accommodating non-standard extreme directions. Section 4.3.2 then shows

how to obtain such a consistent family of clique-wise exponent measure densities by marginalizing a global exponent measure density—assumed to have only the standard extreme direction—using an appropriate choice of clique-specific coefficient matrices.

4.3.1 Construction

In this section, we present our main result, Theorem 4.3.1, which provides a construction method for extremal graphical models that exhibit both lower-dimensional extreme directions and extremal independence. This construction is based on the exponent measure density.

From now on, let $\mathcal{G} = (V, E)$ be a decomposable undirected graph (not necessarily connected) with cliques $\mathcal{C}$, separators $\mathcal{D}$ and let $\mathcal{D}^* \subset \mathcal{D}$ denote the non-empty separators of $\mathcal{G}$. Let $\{\lambda_I : I \in \mathcal{C} \cup \mathcal{D}\}$ be a set of mixture exponent densities as in (4.2.9), with $\lambda_\varnothing = 1$, satisfying the *clique-to-separator marginalization consistency constraints*:

$$\lambda_D(\boldsymbol{y}_D) = \lambda_{C\to D}(\boldsymbol{y}_D), \quad \text{for every } D \subset C, \quad C \in \mathcal{C}, \quad D \in \mathcal{D}^*, \tag{4.3.1}$$

where we recall the marginal density $\lambda_{C\to D}$ in (4.2.10). Equation (4.3.1) means that for every separator D and any clique C that contains D, the marginalization of the density associated with C onto D must be equal to the density assigned directly to D. Note that if $\mathcal{G}$ is a decomposable disconnected graph, then its connected components are also decomposable; see Section 4.A.1. Consequently, condition (4.3.1) holds for the entire graph $\mathcal{G}$ if and only if it holds for each of its connected components separately. Also, the mixture model in (4.2.7) having unit-Fréchet margins yields that for a singleton minimal separator $D = \{i\} \in \mathcal{D}$, the density $\lambda_{C\to D}$ defined in (4.3.1) takes the simple form $\lambda_{C\to D}(y_i) = y_i^{-2}$ for every $y_i > 0$ and for every clique $C \in \mathcal{C}$ such that $C \supset D$. Consequently, the marginalization constraint (4.3.1) associated with singleton D is automatically satisfied. In particular, note that for a block graph [51], every separator is a singleton and thus (4.3.1) is satisfied.

For each clique or non-empty separator $I \in \mathcal{C} \cup \mathcal{D}^*$, let Λ_I denote the exponent measure with density λ_I with respect to the dominating measure μ_I in (4.2.2) and thereafter. Recall the collection of extreme directions $\mathcal{J}(\Lambda_I)$ as in Definition 4.2.1. Throughout the whole paper, we assume

$$\forall I \in \mathcal{C} \cup \mathcal{D}^* : \; J \in \mathcal{J}(\Lambda_I) \implies \lambda_I(\boldsymbol{y}) > 0, \, \forall \boldsymbol{y} \in \mathbb{A}_{I,J}, \tag{4.3.2}$$

where $\mathbb{A}_{I,J}$ is defined as in (4.2.1). Condition (4.3.2) guarantees that the density λ_I remains strictly positive on each sub-face $\mathbb{A}_{I,J}$ whenever $J \in \mathcal{J}(\Lambda_I)$. This property

holds for most known parametric exponent measure density models, including the mixture Hüsler–Reiss model or the mixture logistic model [75, Section 4].

Given the densities $\{\lambda_I : I \in \mathcal{C} \cup \mathcal{D}\}$ satisfying (4.3.1) and (4.3.2), our goal is to construct a new exponent measure density λ_V on the whole space $\mathbb{E}_V$ such that the exponent measure Λ_V with density λ_V with respect to μ_V in (4.2.2) satisfies the global Markov property with respect to $\mathcal{G}$. Engelke et al. [38, Theorem 6.4] show that it is sufficient to prove the following two properties:

(P1) **factorization property with respect to** $\mathcal{G}$, meaning that $\lambda_V(\boldsymbol{y}) = 0$ for μ_V-almost all $\boldsymbol{y} \in \mathcal{Z}(\mathcal{G})$, and for μ_V-almost all $\boldsymbol{y} \notin \mathcal{Z}(\mathcal{G})$, we have

$$\bar{\lambda}(\boldsymbol{y}) \prod_{D \in \mathcal{D}} \bar{\lambda}_D(\boldsymbol{y}_D) = \prod_{C \in \mathcal{C}} \bar{\lambda}_C(\boldsymbol{y}_C); \tag{4.3.3}$$

(P2) **whole-to-clique marginalization property**, meaning that for each clique $C \in \mathcal{C}$,

$$\lambda_C(\boldsymbol{y}_C) = \lambda_{V \to C}(\boldsymbol{y}_C) := \int_{\mathbb{E}_{V \setminus C}} \lambda_V(\boldsymbol{y}_V) \, \mathrm{d}\mu_{V \setminus C}(\boldsymbol{y}_{V \setminus C}), \quad \boldsymbol{y}_C \neq \mathbf{0}, \tag{4.3.4}$$

where we recall the zero set $\mathcal{Z}(G)$ in (4.2.15) and for every $I \in \mathcal{C} \cup \mathcal{D}^*$, the extension $\bar{\lambda}_I$ of the exponent measure density λ_I in (4.2.13). Engelke and Hitz [34, Corollary 1] have already addressed this for the specific case of a connected decomposable graph $\mathcal{G}$ and clique densities $\{\lambda_C : C \in \mathcal{C}\}$ that do not exhibit non-standard extreme directions $J \neq V$. The main contribution of this paper is to generalize these results by incorporating non-standard extreme directions within the cliques and to allow for extremal independence between some group of variables.

In the following, suppose the connected components of $\mathcal{G}$ are induced by the partition $V = V_1 \dot{\cup} \ldots \dot{\cup} V_p$. Let $\mathcal{C}_i$ and $\mathcal{D}_i$ be the cliques and separators of the connected component $\mathcal{G}_{V_i}$ for every $i = 1, \ldots, p$. Define $\lambda_{V_i} : \mathbb{E}^*_{V_i} \to [0, \infty)$ by

$$\lambda_{V_i}(\boldsymbol{y}) = \begin{cases} \displaystyle\prod_{C \in \mathcal{C}_i} \bar{\lambda}_C(\boldsymbol{y}_C) \Big/ \prod_{D \in \mathcal{D}_i} \bar{\lambda}_D(\boldsymbol{y}_D), & \text{if } \boldsymbol{y} \in \mathcal{S}(\mathcal{G}_{V_i}, \lambda) \setminus \mathcal{Z}(\mathcal{G}_{V_i}), \\ 0, & \text{otherwise,} \end{cases}$$

where the sets $\mathcal{Z}(\mathcal{G})$ and $\mathcal{S}(\mathcal{G}, \lambda)$ are defined in (4.2.15) and (4.2.16), respectively. Finally, define $\lambda_V : \mathbb{E}_V \to [0, \infty)$ by $\lambda_V(\mathbf{0}_d) = 0$ and

$$\lambda_V(\boldsymbol{y}) = \begin{cases} \lambda_{V_i}(\boldsymbol{y}_{V_i}) & \text{if there exists a (unique) } i \text{ such that } y_s = 0 \text{ for all } s \in V \setminus V_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3.5}$$

Theorem 4.3.1 (Extension of Corollary 1 in Engelke and Hitz [34])**.** *The density λ_V in* (4.3.5) *is a valid d-dimensional exponent measure density as explained in Section 4.C.3. Moreover, the mpr vector $\boldsymbol{Y}$ with exponent measure density λ_V has the following properties:*

- *$\boldsymbol{Y}$ is an extremal graphical model w.r.t. $\mathcal{G}$, meaning that it satisfies the global Markov property w.r.t. $\mathcal{G}$ as discussed in Section 4.2.3;*
- *The set of extreme directions $\mathcal{J}(\Lambda_V)$ associated to $\boldsymbol{Y}$ or equivalently to the associated exponent measure Λ_V is equal to the collection of non-empty sets $J \subset V$ with the following two properties: (1) the subgraph $\mathcal{G}_J$ is connected and (2) for every clique C intersecting J, the set $J \cap C$ is an extreme direction of the exponent measure Λ_C.*

Remark 4.3.2. Let V_1, $V_2 \subset V$ be the indices associated to two different connected components of $\mathcal{G}$; therefore, $\varnothing$ separates V_1 and V_2. Since $\boldsymbol{Y}$ is an extremal graphical model with respect to $\mathcal{G}$, we have $\boldsymbol{Y}_{V_1} \perp\!\!\!\perp_e \boldsymbol{Y}_{V_2}$.

As stated in Theorem 4.3.1, the exponent measure Λ_V has support on the disjoint union of sub-faces $\left\{\mathbb{A}_{V,J} : J \in \mathcal{J}(\Lambda_V)\right\}$, with $\mathbb{A}_{V,J}$ as in (4.2.1). For every $J \in \mathcal{J}(\Lambda_V)$, let $\Lambda_{V,J}$ be the measure Λ_V projected into $\mathbb{A}_{V,J}$, that is,

$$\Lambda_{V,J}(B) = \Lambda\left(\pi_{V,J}^{-1}(B) \cap \mathbb{A}_{V,J}\right), \qquad B \in \mathcal{B}(\mathbb{E}_J), \tag{4.3.6}$$

where we recall the projection $\pi_{V,J}$ from the beginning of Section 4.2.

Proposition 4.3.3. *For every $J \in \mathcal{J}(\Lambda_V)$, the measure $\Lambda_{V,J}$ defined in* (4.3.6) *admits a density h_J with respect to the $|J|$-dimensional Lebesgue measure given by*

$$h_J(\boldsymbol{y}) = \frac{\prod_{C \in \mathcal{C}_J} \lambda_C\left(\boldsymbol{y}_{C\cap J}, \mathbf{0}_{C\setminus J}\right)}{\prod_{D \in \mathcal{D}_J} \lambda_D\left(\boldsymbol{y}_{D\cap J}, \mathbf{0}_{D\setminus J}\right)}, \qquad \boldsymbol{y} \in \mathbb{E}_J, \tag{4.3.7}$$

where $\mathcal{C}_J$ and $\mathcal{D}_J$ denote the sets of cliques and separators intersecting J, respectively.

4.3.2 Clique-to-separator marginalization consistency constraints

The construction of extremal graphical models in Section 4.3.1 requires a collection of exponent measure densities $\{\lambda_I : I \in \mathcal{C} \cup \mathcal{D}\}$ that satisfy the clique-to-separator marginalization constraints in (4.3.1). In this section, we provide sufficient conditions to construct such a collection of densities.

Simplification. As noted in Section 4.3.1, when the underlying graph is a block graph—such as a forest—the separators reduce to singletons or the empty set. In this case, the constraints in (4.3.1) are automatically satisfied. Consequently, Section 4.3.2 can be bypassed, and one may proceed directly to Section 4.4.

Recall $d \geqslant 1$, $V = \{1, \dots, d\}$. Let $\mathcal{G} = (V, E)$ be a decomposable undirected graph with cliques $\mathcal{C}$, separators $\mathcal{D}$, and let $\mathcal{D}^*$ denote the set of non-empty separators of $\mathcal{G}$. In the following, for a matrix A possibly having zero columns, let $\tilde{A}$ denote the submatrix obtained by removing all zero columns from A, preserving the original order of the non-zero columns.

Recall that a $\underline{d} \times \underline{r}$ mixture model, as in Definition 4.2.2, requires a coefficient matrix $A \in \mathcal{M}_{\underline{d},\underline{r}}([0,1])$ with unit row sums and strictly positive column sums, along with max-stable random vectors $\boldsymbol{Z}^{(1)}, \dots, \boldsymbol{Z}^{(\underline{r})}$ whose only extreme direction is $\{1, \dots, \underline{d}\}$. For brevity, we will refer to these as the conditions of Definition 4.2.2.

Let $\mathcal{A} = \{A^C : C \in \mathcal{C}\}$ be a collection of coefficient matrices, one for each clique in $\mathcal{G}$. For each $C \in \mathcal{C}$, assume that the matrix $A^C = (a^C_{jk})_{j \in C,\, k \in \{1,\dots,r_C\}}$, for some integer $r_C \geqslant 1$, satisfies the conditions of Definition 4.2.2. We further assume that these coefficient matrices are *compatible* in the following sense: for every two cliques whose intersection is neither empty nor a singleton, the corresponding submatrices on the common variables coincide up to removal of zero columns and a permutation of the remaining columns. Specifically,

$$A^{C'}_{C' \cap C'', \cdot} \sim A^{C''}_{C' \cap C'', \cdot}, \qquad \forall C', C'' \in \mathcal{C}, \qquad C' \cap C'' \neq \varnothing, \tag{4.3.8}$$

where, for a clique $C \in \mathcal{C}$ and a subset $J \subset C$, we write $A^C_{J,\cdot} = (a^C_{jk})_{j \in J,\, k \in \{1,\dots,r_C\}}$; for two matrices A, B having the same number of rows, we write $A \sim B$ if (i) the two matrices have the same number of non-zero columns and (ii) the submatrices $\tilde{A}$ and $\tilde{B}$ are equal up to a permutation of columns.

Example 4.3.4. Consider the 4-dimensional induced subgraph $\mathcal{G}_S$ with $S = \{1, 2, 3, 4\}$ from Figure 4.9. This is a decomposable graph with cliques $C' = \{1, 2, 3\}$ and $C'' = \{2, 3, 4\}$. Consider the two matrices

$$A^{C'} = \begin{pmatrix} a^{C'}_{\cdot 1} & 0 & a^{C'}_{\cdot 3} & a^{C'}_{\cdot 4} \\ a^{C'}_{\cdot 1} & 0 & a^{C'}_{\cdot 3} & 0 \\ a^{C'}_{\cdot 1} & a^{C'}_{\cdot 2} & 0 & 0 \end{pmatrix}, \qquad A^{C''} = \begin{pmatrix} 0 & a^{C''}_{\cdot 2} & a^{C''}_{\cdot 3} \\ a^{C''}_{\cdot 1} & a^{C''}_{\cdot 2} & 0 \\ 0 & a^{C''}_{\cdot 2} & a^{C''}_{\cdot 3} \end{pmatrix},$$

where, for simplicity, we assume that within each column, the entries are either zero or a fixed coefficient specific to that column.

The choices $a^{C'}_{\cdot 1} = a^{C''}_{\cdot 2}$, $a^{C'}_{\cdot 2} = a^{C''}_{\cdot 1}$ and $a^{C'}_{\cdot 3} = a^{C''}_{\cdot 3}$ ensure that $A^{C'}_{C' \cap C'', \cdot} \sim$

$A^{C''}_{C'\cap C'',\cdot}$, where

$$A^{C'}_{C'\cap C'',\cdot} = \begin{pmatrix} a^{C'}_{\cdot 1} & 0 & a^{C'}_{\cdot 3} & 0 \\ a^{C'}_{\cdot 1} & a^{C'}_{\cdot 2} & 0 & 0 \end{pmatrix}, \qquad A^{C''}_{C'\cap C'',\cdot} = \begin{pmatrix} 0 & a^{C''}_{\cdot 2} & a^{C''}_{\cdot 3} \\ a^{C''}_{\cdot 1} & a^{C''}_{\cdot 2} & 0 \end{pmatrix}.$$

For a collection $\mathcal{A} = \{A^C : C \in \mathcal{C}\}$ of matrices satisfying (4.3.8) and for a non-empty separator $D \in \mathcal{D}^*$, let $A^D := \tilde{A}^C_{D,\cdot}$, for any choice of $C \in \mathcal{C}$ such that $D \subset C$.

In the following, recall the partition $V = V_1 \dot\cup \ldots \dot\cup V_p$ of V inducing the connected components of $\mathcal{G}$.

Proposition 4.3.5. *Let $\{\tilde\lambda_{V_i} : i = 1, \ldots, p\}$ be a collection of exponent measure densities and suppose that $\tilde\lambda_{V_i}$ is concentrated on $(0,\infty)^{|V_i|}$, for every $i \in \{1,\ldots,p\}$. Let $\mathcal{A}$ be a collection of matrices over the cliques of $\mathcal{G}$ satisfying* (4.3.8) *and the conditions of Definition 4.2.2. For every $I \in \mathcal{C} \cup \mathcal{D}^*$, where $I \subset V_i$ for some $i \in \{1,\ldots,p\}$, consider the mixture exponent measure density λ_I as in* (4.2.9) *with coefficient matrix A^I and each factor being a max-stable random vector with exponent measure density $\tilde\lambda_I$, the I-th margin of $\tilde\lambda_{V_i}$. The exponent measure densities $\{\lambda_I : I \in \mathcal{C} \cup \mathcal{D}^*\}$ satisfy the clique-to-separator marginalization consistency constraints* (4.3.1).

4.4 Mixture Hüsler–Reiss graphical model

Similarly as for Hüsler–Reiss graphical models [53], the main result of this section (Theorem 4.4.8) states that if all cliques have mixture Hüsler–Reiss densities, then so has the full d-dimensional graphical model constructed in Theorem 4.3.1.

Recall $d \geqslant 1$, $V = \{1,\ldots,d\}$, and let $\mathcal{G} = (V, E)$ be a decomposable undirected graph (not necessarily connected) with cliques $\mathcal{C}$ and separators $\mathcal{D}$. Recall the partition $V = V_1 \dot\cup \cdots \dot\cup V_p$ inducing the connected components of $\mathcal{G}$; see Section 4.A.1. Given the increasing interest in Hüsler–Reiss parametric models in recent literature, as highlighted in Engelke and Hitz [34], Engelke et al. [39], Hentschel et al. [53], we adopt the same approach and assume that the exponent measure densities on the cliques and separators of $\mathcal{G}$ used for the construction in Theorem 4.3.1 are mixture Hüsler–Reiss exponent measure densities as defined in (4.B.3). In terms of Proposition 4.3.5, this is the case when the original exponent measure density $\tilde\lambda_{V_i}$ on each connected component is a Hüsler–Reiss exponent measure density as in (4.B.2) with some variogram matrix $\Gamma \in \mathcal{V}_d$.

Outline of the section. In Section 4.4.1, we introduce the *generalized variogram matrix*, which extends the variogram matrix in (4.B.1) to account for

extremal independence between components. Section 4.4.2 presents the *mixture Hüsler–Reiss forest model* for the case where the underlying graph $\mathcal{G}$ is a forest, demonstrating that parameter inference simplifies to a collection of bivariate problems. In Section 4.4.3, we generalize this framework to *decomposable graphs* and define the *mixture Hüsler–Reiss graphical model*.

4.4.1 Generalized variogram matrix

As pointed out in Remark 4.3.2, if the graph $\mathcal{G}$ is disconnected, then the constructed model exhibits extremal independence between the connected components of this graph. In the following, we provide a natural extension of the notions of variogram matrix and variogram matrix completion studied for connected graphs in Hentschel et al. [53] to also account for extremal independence.

Definition 4.4.1. A matrix $\Gamma \in \mathcal{M}_{d,d}(\bar{\mathbb{R}}^+)$ is a *generalized variogram matrix with respect to* $\mathcal{G}$ if for every $i = 1, \ldots, p$, the submatrix Γ_{V_i} is a variogram matrix in the sense of (4.B.1), and $\Gamma_{st} = \infty$ whenever s and t belong to two different connected components of $\mathcal{G}$.

Write $\bar{\mathbb{R}}^{+,?} = \bar{\mathbb{R}}^+ \cup \{?\}$ and let $\Gamma^? \in \mathcal{M}_{d,d}(\bar{\mathbb{R}}^{+,?})$. The question mark refers to an element from $\Gamma^?$ whose value is still to be specified. For every $i = 1, \ldots, p$, suppose that the submatrix $\Gamma^?_{V_i}$ is partially conditionally negative definite, and specified w.r.t. the connected component $\mathcal{G}_{V_i}$; see Hentschel et al. [53, Section 4]. We complete $\Gamma^?$ w.r.t. $\mathcal{G}$ as follows:

1. Solve the matrix completion on each of the components $\mathcal{G}_{V_i}$ as in [53].
2. Fill up the remaining elements of $\Gamma^?$ with ∞.

Since the solution to Step 1 is unique and, by construction, yields a variogram matrix for each connected component $\mathcal{G}_{V_i}$ [53, Proposition 4.4], it follows that the resulting matrix Γ is likewise unique and constitutes a generalized variogram matrix with respect to $\mathcal{G}$, as per Definition 4.4.1. We refer to Γ as the *generalized completion of* $\Gamma^?$ with respect to $\mathcal{G}$.

Example 4.4.2. Consider the 5-dimensional forest structure $\mathcal{F}$ in Figure 4.10, consisting of two connected components, $\mathcal{T}_1 := \mathcal{F}_{V_1}$ and $\mathcal{T}_2 := \mathcal{F}_{V_2}$, where $V_1 = \{1, 2\}$ and $V_2 = \{3, 4, 5\}$. The generalized completion Γ, shown on the right in (4.4.1), of the matrix $\Gamma^?$, shown on the left in (4.4.1), with respect to $\mathcal{F}$ is obtained by completing the submatrix Γ_{V_2} with respect to $\mathcal{T}_2$:

$$\Gamma^? = \begin{pmatrix} 0 & 0.5 & \infty & \infty & \infty \\ 0.5 & 0 & \infty & \infty & \infty \\ \infty & \infty & 0 & 0.5 & ? \\ \infty & \infty & 0.5 & 0 & 1 \\ \infty & \infty & ? & 1 & 0 \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} 0 & 0.5 & \infty & \infty & \infty \\ 0.5 & 0 & \infty & \infty & \infty \\ \infty & \infty & 0 & 0.5 & 1.5 \\ \infty & \infty & 0.5 & 0 & 1 \\ \infty & \infty & 1.5 & 1 & 0 \end{pmatrix}. \tag{4.4.1}$$

Note that the completion with respect to a tree is $\Gamma_{st} = \sum_{(i,j)\in \mathrm{path}(s,t)} \Gamma^{?}_{ij}$, where $\mathrm{path}(s,t)$ denotes the unique path between s and t in the tree [2, 36]. The completion problem for general connected decomposable graphs is discussed in Hentschel et al. [53, Section 4].

4.4.2 Forests

In this section, we use the construction in Theorem 4.3.1 to the special case when the underlying graph is a forest $\mathcal{F}$ and densities on the cliques are bivariate mixture Hüsler–Reiss densities as in (4.B.3). We present our main result, Proposition 4.4.3, which establishes that the parameters of the d-variate model—namely, the associated coefficient and variogram matrices, and thus the corresponding extreme directions and dependence structure—can be completely characterized by covering the bivariate densities on the edges of the forest.

Let $\mathcal{F} = (V, E)$ be an undirected forest. A forest is a graph in which all connected components are trees. Note that in this case, cliques correspond to edges and each separator is either a singleton or the empty set. For every edge $(s,t) \in E$, consider a bivariate mixture Hüsler–Reiss exponent measure density $\lambda_{st} = \lambda_{\{s,t\}}$ as in (4.B.3) with coefficient matrix $A^{\{s,t\}}$ as in (4.4.6) and variogram coefficient $\Gamma_{st} > 0$. Note that an edge (s,t) is always represented such that $s < t$. Therefore, $A^{\{s,t\}}$ can be identified with a coefficient $a_{st} > 0$ and we have

$$\lambda_{st}(y_s, y_t) = \begin{cases} (1 - a_{st}) y_s^{-2} & \text{if } y_s > 0 = y_t, \\ (1 - a_{st}) y_t^{-2} & \text{if } y_s = 0 < y_t, \\ a_{st} \lambda((y_s, y_t); \Gamma_{st}) & \text{if } y_s > 0 \text{ and } y_t > 0, \end{cases} \tag{4.4.2}$$

where $\lambda(\,\cdot\,;\Gamma)$ is simply the bivariate Hüsler–Reiss exponent measure density as in (4.B.2) with coefficient Γ. Note that in the special case where $a_{st} = 1$, the measure λ_{st} simplifies to the Hüsler–Reiss density $\lambda(\,\cdot\,;\Gamma)$, and consequently, the associated exponent measure places mass only on the extreme direction $\{s,t\}$. In contrast, if $0 < a_{st} < 1$, the associated exponent measure distributes mass across all the three possible extreme directions $\{s,t\}, \{s\}, \{t\}$; see Figure 4.1 for a simulation from the associated mpr vector.

Hereafter, consider the density λ_V as in (4.3.5) constructed via $\{\lambda_{st} : (s,t) \in E\}$.

Proposition 4.4.3. *The density λ_V constructed above is a mixture Hüsler–Reiss exponent measure density as in* (4.B.3) *with the following ingredients:*

- ***generalized variogram matrix*** $\Gamma = (\Gamma_{st})_{(s,t)\in V\times V} \in \mathcal{M}_{d,d}(\bar{\mathbb{R}}^{+})$, *where*

$$\Gamma_{st} = \begin{cases} \sum_{(i,j)\in \mathrm{path}(s,t)} \Gamma_{ij}, & \text{if } s \text{ and } t \text{ are in the same connected component,} \\ \infty, & \text{otherwise,} \end{cases} \tag{4.4.3}$$

and $\mathrm{path}(s,t)$ *denotes the unique path connecting* s *and* t *in the forest* $\mathcal{F}$*;*

- ***extreme directions*** $\mathcal{J}(\Lambda_V)$*, defined as all non-empty subsets* $J \subset V$ *for which the subgraph* $\mathcal{F}_J$ *is connected and*

$$\forall\,(s,t) \in E \text{ such that } \{s,t\} \cap J \neq \varnothing : \qquad a_{st} = 1 \implies \{s,t\} \subset J. \tag{4.4.4}$$

- ***coefficient matrix*** $A \in \mathcal{M}_{d,r}([0,1])$*, where* $r = |\mathcal{J}(\Lambda_V)|$*. For each extreme direction* $J \in \mathcal{J}(\Lambda_V)$*, the corresponding column* $a_{\cdot J}$ *of* A *is given by* $a_{jJ} = 0$ *if* $j \notin J$*, and*

$$a_{\cdot J} = a_{jJ} = \prod_{\substack{(s,t)\in E \\ |\{s,t\}\cap J|=2}} a_{st} \cdot \prod_{\substack{(s,t)\in E \\ |\{s,t\}\cap J|=1}} (1 - a_{st}), \quad j \in J. \tag{4.4.5}$$

Let $\boldsymbol{Y}$ be the mpr vector with density λ_V. Theorem 4.3.1 shows that $\boldsymbol{Y}$ is an extremal graphical model with respect to $\mathcal{F}$. We introduce the following definition.

Definition 4.4.4. The mpr $\boldsymbol{Y}$ will be called a mixture Hüsler–Reiss forest model with respect to $\mathcal{F}$ with coefficient matrix A and generalized variogram matrix Γ.

Remark 4.4.5. In the case where the graph consists of a single tree, Hu et al. [56, Section 3] construct a multivariate exponent measure on the entire tree by assembling bivariate exponent measures along its edges, a model which arises as the extremal attractor of a Markov tree [89]. Proposition 4.4.3 corresponds to a particular instance of this framework, specifically when the bivariate exponent measures are mixture Hüsler–Reiss exponent measures. Note that the exponent measure density of a mixture Hüsler–Reiss forest model as in Proposition 4.4.3 is also presented in Engelke et al. [38, Example 8.8].

Remark 4.4.6. Equations (4.4.3) and (4.4.5) demonstrate that identifying the parameters A—and hence the extreme directions—and the matrix Γ of the mixture Hüsler–Reiss forest model $\boldsymbol{Y}$ requires only the estimation of (a_{st}, Γ_{st}) for each edge (s,t) in the forest $\mathcal{F}$. Moreover, for each edge, this reduces to fitting a bivariate dependence model with just two parameters.

Example 4.4.7. Suppose $d = 5$ and consider the forest $\mathcal{F}$ shown in Figure 4.10 with two sub-trees $\{1,2\}$ and $\{3,4,5\}$. For each edge (s,t) from $\mathcal{F}$, consider a bivariate mixture Hüsler–Reiss density λ_{st} as in (4.4.2) with coefficient matrix $A^{\{s,t\}}$ and variogram coefficient $\Gamma^{\{s,t\}}$ given respectively by

$$A^{\{s,t\}} = \begin{pmatrix} a_{st} & 1 - a_{st} & 0 \\ a_{st} & 0 & 1 - a_{st} \end{pmatrix}, \qquad \Gamma^{\{s,t\}} = \begin{pmatrix} 0 & \Gamma_{st} \\ \Gamma_{st} & 0 \end{pmatrix}. \tag{4.4.6}$$

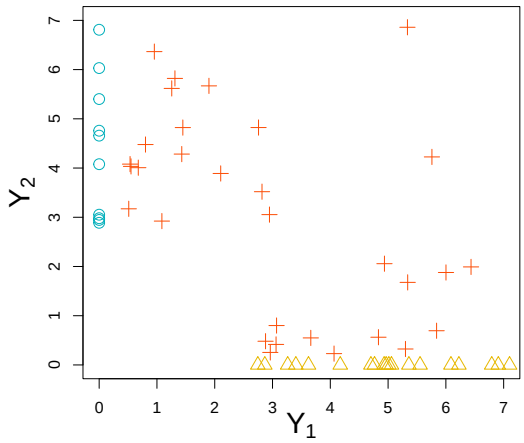


Figure 4.1
Sample of size 50 from a bivariate mixture Hüsler–Reiss mpr vector with matrices as defined in (4.4.6), setting $a_{st} = \Gamma_{st} = 0.5$. Red plus signs correspond to the extreme direction $\{1, 2\}$. Open yellow triangles correspond to the extreme direction $\{1\}$. Open blue circles correspond to the extreme direction $\{2\}$. These symbols correspond respectively to the first, second, and third signatures of the matrix $A^{\{1,2\}}$ given in (4.4.6).

Let $\boldsymbol{Y}$ be the resulting 5-dimensional mixture Hüsler–Reiss forest model as in Proposition 4.4.3. Suppose $0 < a_{12}, a_{34} < 1$ and $a_{45} = 1$. Proposition 4.4.3 shows that the set of extreme directions $\mathcal{J}$ associated to $\boldsymbol{Y}$ or equivalently to the corresponding exponent measure is

$$\mathcal{J} = \{\{1,2\}, \{1\}, \{2\}, \{3,4,5\}, \{4,5\}, \{3\}\}$$

Moreover, the coefficient matrix $A \in \mathcal{M}_{5,6}([0,1])$ and the generalized variogram matrix $\Gamma \in \mathcal{M}_{5,5}(\bar{\mathbb{R}}^+)$ are given respectively by

$$A = \begin{pmatrix} a_{12} & 1-a_{12} & 0 & 0 & 0 & 0 \\ a_{12} & 0 & 1-a_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & a_{34} & 0 & 1-a_{34} \\ 0 & 0 & 0 & a_{34} & (1-a_{34}) & 0 \\ 0 & 0 & 0 & a_{34} & (1-a_{34}) & 0 \end{pmatrix},$$

$$\Gamma = \begin{pmatrix} 0 & \Gamma_{12} & \infty & \infty & \infty \\ \Gamma_{12} & 0 & \infty & \infty & \infty \\ \infty & \infty & 0 & \Gamma_{34} & \Gamma_{34}+\Gamma_{45} \\ \infty & \infty & \Gamma_{34} & 0 & \Gamma_{45} \\ \infty & \infty & \Gamma_{34}+\Gamma_{45} & \Gamma_{45} & 0 \end{pmatrix}.$$

4.4.3 Decomposable graphs

We are now ready to state the main theorem of this section. In the following, we fix a matrix $\Gamma^? \in \mathcal{M}_{d,d}(\bar{\mathbb{R}}^{+,?})$ as in Section 4.4.1. For every $i = 1, \ldots, p$, let $\tilde{\lambda}_{V_i}$ be the Hüsler–Reiss density in (4.B.2) with variogram matrix $\Gamma^?_{V_i}$. Note that the latter has entries in $(0, \infty)$ since $\mathcal{G}_{V_i}$ is connected. Consider a collection of coefficient matrices $\mathcal{A} = \{A^C : C \in \mathcal{C}\}$ satisfying (4.3.8) and suppose that each matrix satisfies the conditions in Definition 4.2.2. Consider the resulting mixture Hüsler–Reiss exponent measure densities $\{\lambda_I : I \in \mathcal{C} \cup \mathcal{D}\}$ defined in Proposition 4.3.5 and the resulting d-dimensional exponent measure density λ_V in (4.3.5) factorizing w.r.t. $\mathcal{G}$. Suppose finally that for every matrix A^C from $\mathcal{A}$ and for every column index $k \in \{1, \ldots, r_C\}$, all entries in that column are either a fixed coefficient or zero, more precisely,

$$\forall C \in \mathcal{C}, \quad \forall k \in \{1, \ldots, r_C\}, \quad \exists \tilde{a}^C_{J_k} \in (0,1] \text{ such that } a^C_{jk} \in \{0, \tilde{a}^C_{J_k}\}, \quad \forall j \in C. \tag{4.4.7}$$

All in all, we impose the following conditions on the matrices A^C: Definition 4.2.2 and Equations (4.3.8) and (4.4.7).

Theorem 4.4.8. *The density λ_V constructed above is a mixture Hüsler–Reiss exponent measure density as in* (4.B.3) *with the following ingredients:*

- *set of extreme directions* $\mathcal{J}(\Lambda_V)$ *of the associated exponent measure* Λ_V *consisting precisely of the subsets* $J \subset V$ *such that (1) the subgraph* $\mathcal{G}_J$ *is connected, (2) for every clique* $C \in \mathcal{C}$ *such that* $C \cap J \neq \varnothing$*, the set* $C \cap J$ *is a signature of the matrix* A^C*, as defined in* (4.2.8)*;*

- *generalized variogram matrix* Γ*, the generalized completion as in Steps 1 and 2 of* $\Gamma^?$ *with respect to* $\mathcal{G}$*;*

- *coefficient matrix* $A \in \mathcal{M}_{d,r}([0,1])$*, where* $r = |\mathcal{J}(\Lambda_V)|$ *with signatures* $\mathcal{J}(\Lambda_V)$*. Moreover, let* $\boldsymbol{a}_{\cdot J}$ *denote the column of* A *associated with an extreme direction* $J \in \mathcal{J}(\Lambda_V)$*. Then* $a_{jJ} = 0$ *if* $j \notin J$ *and*

$$a_{jJ} = a_J = \prod_{n=1}^{f_J} a_{J,n}, \quad j \in J, \tag{4.4.8}$$

  where f_J *and* $(a_{J,1}, \ldots, a_{J,f_J})$ *are given in Lemma 4.F.2.*

Note that the coefficient matrix A and the generalized variogram matrix Γ from Theorem 4.4.8 are compatible as defined in the last paragraph of Section 4.B.2. Later in Proposition 4.4.3, we will express (4.4.8) for the case where the graph $\mathcal{G}$ is a forest; see Section 4.4.2.

Definition 4.4.9. An extremal graphical model $\boldsymbol{Y}$ with density λ_V as in Theorem 4.4.8 will be called a *mixture Hüsler–Reiss graphical model with respect to* $\mathcal{G}$ with coefficient matrix A and generalized variogram matrix Γ.

4.5 Structure learning and extreme directions identification of a mixture Hüsler–Reiss forest model

The goal of this section is to jointly learn the graph structure and the extreme directions of the mixture Hüsler–Reiss forest model $\boldsymbol{Y}$ introduced in Section 4.4.2.

Let Λ_V denote the exponent measure associated with $\boldsymbol{Y}$. Suppose we observe independent copies $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of a d-dimensional random vector $\boldsymbol{X}$ that lies in the max-domain of attraction of an mev distribution G with unit-Fréchet margins and exponent measure Λ_V. Our estimation procedure is as follows:

1. Estimate $\hat{\mathcal{T}} = (V, \hat{E})$, the minimum spanning tree, following the method of Engelke and Volgushev [36].

2. For each edge $(s,t) \in \hat{E}$, fit a bivariate mixture Hüsler–Reiss exponent measure density with parameters (a_{st}, Γ_{st}) as in (4.4.2), allowing for the possibility that $a_{st} = 0$, and estimate these parameters.

3. Define the estimated forest $\hat{\mathcal{F}}$ by removing from $\hat{E}$ any edge (s,t) for which the estimated mixture coefficient $\hat{a}_{st}$ is zero and compute the estimated extreme directions $\hat{\mathcal{J}}$ as described in (4.4.4).

For detailed information on the minimum spanning tree procedure in Step 1, see the final paragraph of Section 4.A.1; for the construction of the extreme directions $\hat{\mathcal{J}}$ in Step 3, refer to Algorithm 14.

Estimation of a bivariate mixture Hüsler–Reiss model. In the sequel, suppose that we have estimated the minimum spanning tree $\hat{\mathcal{T}} = (V, \hat{E})$ as in Step 1 above and fix an edge $(s,t) \in \hat{E}$. Let $\boldsymbol{X}_{i,st}$ denote the bivariate $\{s,t\}$-margin of $\boldsymbol{X}_i$ for $i = 1, \ldots, n$, and define the penalized least-squares estimator (PLS) proposed in Section 3.3:

$$(\widehat{a}_{st}, \widehat{\Gamma}_{st}) = \underset{0 \leqslant a \leqslant 1, \Gamma > 0}{\arg\min} \sum_{m=1}^{q} \Big(\widehat{\ell}_{n,k,st}(\boldsymbol{c}_m) - \ell_{st}(\boldsymbol{c}_m; a, \Gamma) \Big)^2 + \beta \cdot \{a^\alpha + (1-a)^\alpha\}^{1/\alpha}, \tag{4.5.1}$$

where k/n is the *tail fraction* in Section 3.3, $\ell_{st}(\,\cdot\,; a, \Gamma)$ is the bivariate mixture Hüsler–Reiss stable tail dependence function (stdf) in Section 4.B.3 with coefficients a and Γ, and $\widehat{\ell}_{n,k,st}(\,\cdot\,)$ is the empirical stdf of the pair of variables (s,t) based on $\boldsymbol{X}_{1,st}, \ldots, \boldsymbol{X}_{n,st}$ as defined in Einmahl et al. [33, Section 2.5.1]. The points $\boldsymbol{c}_m = (c_{m1}, c_{m2}) \in [0,\infty)^2$ for $m = 1, \ldots, q$ are the evaluation points for $\widehat{\ell}_{n,k,st}(\,\cdot\,)$ and $\ell_{st}(\,\cdot\,; a, \Gamma)$, while $0 < \alpha < 1$ is the *penalization exponent* and $\beta > 0$ is the *penalization coefficient*. For more intuition and details about the tuning parameters of the estimator (4.5.1) see Section 3.3.

Steps 1 and 2 involve methodological choices and selection of the tuning parameters. These choices are summarized in Table 4.1. The overall estimation procedure—covering both the forest structure $\mathcal{F}$ and the extreme directions $\mathcal{J}(\Lambda_V)$—is outlined in Algorithm 6. Sub-algorithms of Algorithm 6 are presented in Section 4.G.1.

Minimum spanning tree estimation [36]	**Bivariate mixture Hüsler–Reiss estimation in Section 3.3**
• Method used to fit the minimum spanning tree: – empirical correlation; – extremal variogram. • Exceedance proportion $p \in [0,1]$.	• tail fraction $k/n \in [0,1]$; • grid of bivariate points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m$; • penalization parameter $\beta > 0$; • penalization exponent $0 < \alpha < 1$.

Table 4.1
Methods and tuning parameters of minimum spanning tree estimation and bivariate mixture Hüsler–Reiss model fitting.

Algorithm 6 Estimation of the mixture Hüsler–Reiss forest model

Input: integers n, $K \in \mathbb{N}$, $k \leqslant n$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0,\infty)^2$, sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0,\infty)^d$ as defined above, reals $0 < p$, $\alpha < 1$, $\left\{\beta_1, \ldots, \beta_g\right\} \in [0,\infty)^g$ and starting point $(a_0, \Gamma_0) \in [0,1] \times \mathbb{R}^+_{>0}$.

Output: extremal forest estimation $\hat{\mathcal{F}}$.

1: Define the graph structure $\hat{\mathcal{F}} = (V, \hat{E})$ being the minimum spanning tree obtained as in Engelke and Volgushev [36], with an associated probability p used as a threshold. ▷ See the `emst()` function from the `R` package `graphicalExtremes` [40].
2: **for** each edge (s,t) from $\hat{E}$ **do**
3: Define the estimator $(\hat{a}_{st}, \hat{\Gamma}_{st})$ as the output of Algorithm 9 with inputs: sample size $n \in \mathbb{N}$, integers $k \leqslant n$ and $K \in \mathbb{N}$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0,\infty)^2$, sample $\boldsymbol{X}_{1,st}, \ldots, \boldsymbol{X}_{n,st} \in [0,\infty)^2$, reals $0 < \alpha < 1$, $\left\{\beta_1, \ldots, \beta_g\right\} \in [0,\infty)^g$ and starting point $(a_0, \Gamma_0) \in [0,1] \times \mathbb{R}^+_{>0}$.
4: **if** $\hat{a}_{st} = 0$ **then**
5: Update the forest $\hat{\mathcal{F}}$ by removing the edge (s,t), that is

$$\hat{E} = \hat{E} \setminus \{(s,t)\}.$$

6: **end if**
7: **end for**
8: Let $\hat{\mathcal{T}}_1, \ldots, \hat{\mathcal{T}}_p$ be the subtrees of $\mathcal{F}$. ▷ See the function `components()` from the `R` package `igraph` [22].
9: **for** $i \in \{1, \ldots, p\}$ **do**

$$\hat{\mathcal{J}}_i = \mathrm{EDC}\left(\mathcal{T}_i, (\hat{a}_{st} : (s,t) \in \hat{E}_i)\right) \text{ (see Algorithm 14)}$$

10: **end for**
11: Define $\hat{\mathcal{J}} = \bigcup_{i=1}^{p} \hat{\mathcal{J}}_i$.
12: **return** $(\hat{\mathcal{F}}, \hat{\mathcal{J}})$.

4.6 Simulation study

In this section, we conduct a simulation study to evaluate the estimation procedure introduced in Algorithm 6. We begin by outlining the setup used throughout the simulations in Section 4.6.1. We also introduce a simulation procedure algorithm from a max-stable random vector with mixture Hüsler–Reiss forest exponent measure density as in Section 4.4.2. We finally introduce two scores to asses the performance of our estimator in Algorithm 6 both in terms of graph structure recovery and extreme directions identification. These two scores will be studied in Section 4.6.2 as a function of the tuning parameters of our estimator and also as a function of the sample size.

4.6.1 Setup

We first establish the setup that will be used throughout the estimation procedure.

Forest structure. In this simulation study, we fix the 20-dimensional forest structure $\mathcal{F} = (V, E)$ given in Figure 4.11.

Simulation. We simulate a d-variate sample $\boldsymbol{M}_1, \dots, \boldsymbol{M}_n$ following a max-stable distribution with unit-Fréchet margins and mixture Hüsler–Reiss forest exponent measure density λ_V as in Section 4.4.2 with some coefficient matrix $A \in \mathcal{M}_{d,r}([0,1])$ and generalized variogram matrix $\Gamma \in \mathcal{M}_{d,d}(\mathbb{R}^{+,?})$. The simulation algorithm is presented in Section 4.G.2. Similarly to Remark 4.4.6, the algorithm requires only bivariate parameters (a_{st}, Γ_{st}) for every edge (s,t) from the forest $\mathcal{F}$.

To introduce a light-tailed perturbation and make the simulation more realistic, we model

$$(X_{i1}, \dots, X_{id}) = (M_{i1}, \dots, M_{id}) + (N_{i1}, \dots, N_{id}), \qquad i = 1, \dots, n, \tag{4.6.1}$$

where $\boldsymbol{N}_i$ is a d-variate Gaussian random vector for every $i = 1, \dots, n$. Note that we will present results for a d-dimensional centered Gaussian noise with independent margins and standard deviation $\sigma = 5$. Qualitatively similar results arise under other noise-dependence scenarios, such as centered Gaussian noise with correlated margins.

Minimization algorithm. As in Section 3.4, in order to solve the bivariate minimization problem (4.5.1), we use the bounded limited-memory modification of the BFGS quasi-Newton method (L-BFGS-B algorithm from the function `optim()` in R), see Byrd et al. [10].

Starting points. In order to solve the minimization problem (4.5.1) using the L-BFGS-B algorithm, we sample randomly (a_0, Γ_0) from the set $[0.1, 0.9] \times [0.2, 2]$.

Performance assessment. We first evaluate the estimation procedure in Algorithm 6 in terms of graph structure recovery. To do so, we introduce an error measure between two forests that computes the proportion of edges that are not shared. More precisely, for two graphs $\mathcal{F} = (V, E)$ and $\hat{\mathcal{F}} = (V, \hat{E})$, we define

$$\mathrm{D}(\hat{E}, E) = 1 - \frac{|\hat{E} \cap E|}{|\hat{E} \cup E|} = \frac{|\hat{E} \triangle E|}{|\hat{E} \cup E|} \tag{4.6.2}$$

with $A \triangle B = (A \cup B) \setminus (A \cap B)$ the symmetric difference between sets A and B.

Second, in order to evaluate the performance of Algorithm 6 in terms of extreme directions identification, we define a similarity distance between two sets that computes the proportion of non-shared elements. More precisely, for two collections $J, \hat{J} \subset \mathcal{P}(V)$, we define

$$\text{ED-S}(\mathcal{J}, \hat{\mathcal{J}}) = 1 - \frac{|J \cap \hat{J}|}{|J \cup \hat{J}|} = \frac{|J \triangle \hat{J}|}{|J \cup \hat{J}|} \tag{4.6.3}$$

Grid points, $\boldsymbol{c_1}, \ldots, \boldsymbol{c_q}$. Recall the bivariate points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$ at which both the stdf and the empirical stdf are evaluated during the estimation procedure within each edge; see Step 5 of Algorithm 8. The choice of the number q involves a trade-off: selecting a sufficiently large q ensures reliable estimation, while keeping q reasonably small reduces computational expense. In our case, we set $q = 25$ and sample q points uniformly from $[0, 1]^2$.

Empirical estimates. Each experiment of the form (4.6.1), with sample size n, is repeated 100 times to empirically estimate the error measures (4.6.2), (4.6.3) and (4.6.5) below.

4.6.2 Effect of tuning parameters

Recall that our estimation procedure summarized in Algorithm 6 is based on four tuning parameters: the exceedance proportion p used for the minimum spanning tree fit in Step 1 on page 100, the tail fraction k/n and the penalization parameters α, β in Algorithm 7.

Exceedance proportion. The choice of the exceedance proportion p is the topic of the simulation study in Engelke and Volgushev [36, Section 5]. The authors state that a value $p = n^{0.8}/n$ gives good results in practice. In Step 1, we need the

minimum spanning tree $\hat{\mathcal{T}} = (V, \hat{E}_{\mathcal{T}})$ to contain all edges of the forest $\mathcal{F} = (V, E)$. More precisely, we seek a theoretical guarantee of the form

$$\lim_{n\to\infty} \mathsf{P}\Big(E \subset \hat{E}_{\mathcal{T}}\Big) \;=\; 1. \tag{4.6.4}$$

A result of the form (4.6.4) is not yet shown in this paper. In the simulation study, we compute the proportion of missing edges in $\hat{\mathcal{T}}$ from $\mathcal{F}$, that is,

$$\mathrm{MER}_E(\hat{E}_{\mathcal{T}}) = 1 - \frac{|\hat{E}_{\mathcal{T}} \cap E|}{|E|} = \frac{|E \setminus \hat{E}_{\mathcal{T}}|}{|E|}. \tag{4.6.5}$$

Results for the missing edge rate. Figure 4.2 shows the resulting proportion of missing edges in (4.6.5) as a function of the sample size n. In each run, the minimum spanning tree is fitted via the empirical extremal variogram method (blue) and via the empirical correlation method (orange). The variogram-based estimator consistently outperforms the correlation-based one—mirroring the findings of Engelke and Volgushev [36, Section 5] in a tree-recovery framework.

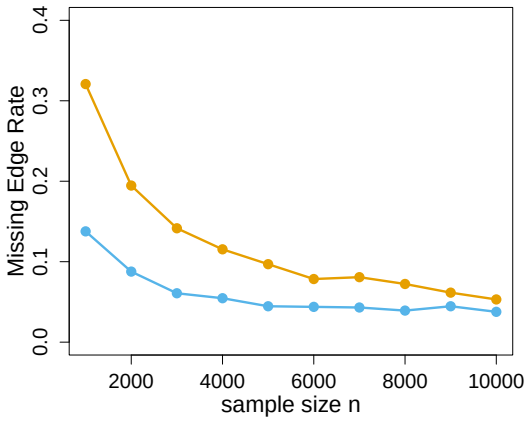


Figure 4.2
Empirical estimates of the proportion of missing edges in (4.6.5) for the forest $\mathcal{F}$ of Figure 4.11, after fitting a minimum spanning tree (Step 1 in Algorithm 6). The estimates are plotted against the sample size n. The exceedance proportion is set to $p = n^{0.8}/n$. Estimates from the extremal variogram method are shown in blue, and those from the empirical correlation method in orange.

Penalization exponent. As discussed in Mourahib et al. [76, Section 4], the choice of the penalization exponent α is not important as long as the penalization

parameter β is well chosen using a K-fold cross validation procedure (Algorithm 8). For simplicity, we fix $\alpha = 0.2$.

Tail fraction. As noted in Section 3.4, choosing the tail fraction k/n in the bivariate estimation (Step 3 of Algorithm 6) involves a classic bias–variance trade-off: a smaller k/n yields lower bias but higher variance, since fewer observations are deemed extreme, whereas a larger k/n reduces variance at the cost of increased bias by including points that are not large "enough" (or small "enough", depending on the case) to be considered extreme. Figure 4.3 plots the empirical structure error from (4.6.2) against the tail fraction, using the perturbed model in (4.6.1) with independent centered Gaussian noise. We compare two fitting methods—the empirical extremal variogram (blue) and the empirical correlation (orange). The U-shaped curves clearly reflect the bias–variance trade-off, with an optimal k/n roughly in $[0.05, 0.1]$. Moreover, as the sample size grows, the structure-recovery error systematically decreases.

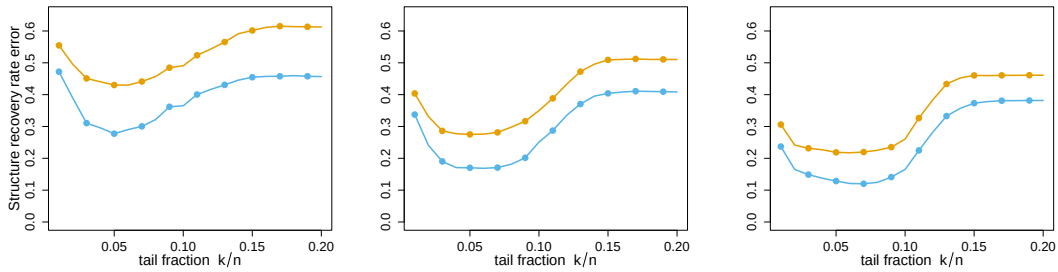


Figure 4.3
Empirical estimates of the recovery error (4.6.2) of the forest $\mathcal{F}$ in Figure 4.11 as a function of the tail fraction k/n. In blue, the extremal variogram method and in orange the empirical correlation method. The exceedance proportion is set to $p = n^{0.8}/n$. The penalization exponent is fixed to $\alpha = 0.2$ and the grid used for the 10-fold cross-validation is $\beta \in \{0.01, 0.03, \ldots, 1\}$.

Bivariate error share. Note that the error recovery rate D in (4.6.2) is a compound error introduced by both the minimum spanning tree estimation (Step 1) and the bivariate error estimation within each edge (Step 3). We compute the proportion of the bivariate error in the compound error D, that is

$$\mathrm{BES}(\hat{E}, \hat{E}_{\mathcal{T}}, E) = 1 \; - \; \frac{\mathrm{MER}_E(\hat{E}_{\mathcal{T}})}{\mathrm{D}(\hat{E}, E)}, \tag{4.6.6}$$

where we recall that $\hat{E}_{\mathcal{T}}$ and $\hat{E}$ are the edges of the minimum spanning tree and the final forest from Algorithm 6, and that MER and D are defined in (4.6.5) and (4.6.2), respectively.

Figure 4.4 shows the empirical estimator of this proportion as a function of the tail fraction k/n with sample size $n = 3000$. Unsurprisingly, the error share of the bivariate estimation is lowest for k/n around $[0.05, 0.1]$, which matches the dip in the compound D in that interval (Figure 4.3). We also see that for $k/n \in [0.05, 0.1]$, less than 50% of the total error D comes from the bivariate step. Finally, observe that using the extremal empirical correlation for tree recovery results in a lower bivariate-error share than the empirical extremal variogram approach—consistent with the superior performance of the variogram-based method reported in Engelke and Volgushev [36, Section 5] and illustrated in Figure 4.2.

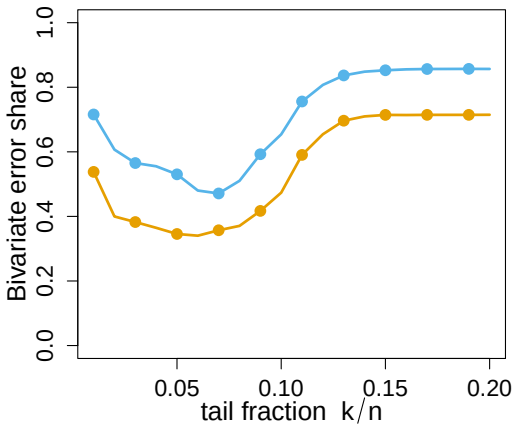


Figure 4.4
Empirical estimates of the bivariate error share (4.6.6) for the forest $\mathcal{F}$ in Figure 4.11 against the tail fraction k/n. The exceedance proportion is set to $p = n^{0.8}/n$ and sample size to $n = 3000$. Estimates from the extremal variogram method are shown in blue, and those from the empirical extremal correlation method in orange. The penalization exponent is fixed to $\alpha = 0.2$ and the grid used for the 10-fold cross-validation is The penalization exponent is fixed to $\alpha = 0.2$ and the grid used for the 10-fold cross-validation is $\beta \in \{0.01, 0.03, \ldots, 1\}$.

Sample size. Finally, we assess the performance of our estimator as the sample size grows over $n \in \{1000, 2000, \ldots, 10000\}$. We again fix the exceedance proportion at $p = n^{0.8}/n$ and set the penalization exponent to $\alpha = 0.2$. Consistent with earlier experiments, we use a tail fraction $k/n = 0.1$. As shown in Figure 4.5, the structure-

recovery error defined in (4.6.2) converges to zero for both the extremal–variogram method (blue) and the extremal–correlation method (orange) employed in Step 1 of Algorithm 6. Unsurprisingly, the variogram-based approach yields slightly lower error, reflecting its superior edge-identification performance in the underlying forest $\mathcal{F}$ as presnted in Figure 4.2.

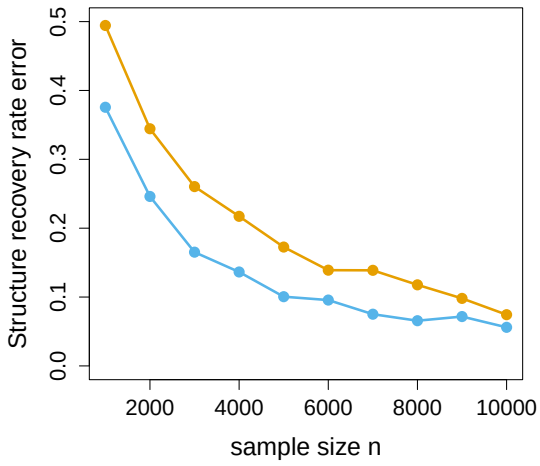


Figure 4.5
Empirical estimates of the recovery error (4.6.2) of the forest $\mathcal{F}$ in Figure 4.11 as a function of the sample size n. In blue, the extremal variogram method and in orange the empirical correlation method. The exceedance proportion is set to $p = n^{0.8}/n$. The penalization exponent is fixed to $\alpha = 0.2$ and the grid used for the 10-fold cross-validation is $\beta \in \{0.01, 0.03, \ldots, 1\}$.

4.7 Application

4.7.1 River discharge data

We illustrate Algorithm 6 using river discharge data from $d = 31$ stations along the Danube River (see Figure 1.1). The dataset consists of daily river discharge measurements from the summer months between 1960 and 2010. To mitigate temporal dependence, the data are declustered, resulting in approximately seven to ten observations per year and yielding $\boldsymbol{X}_i$, $i = 1, \ldots, n$, with $n = 428$ observations in total. The dataset is available in the `GraphicalExtremes` package in `R`; see Engelke et al. [40].

The data $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{in})$ are approximately independent and identically

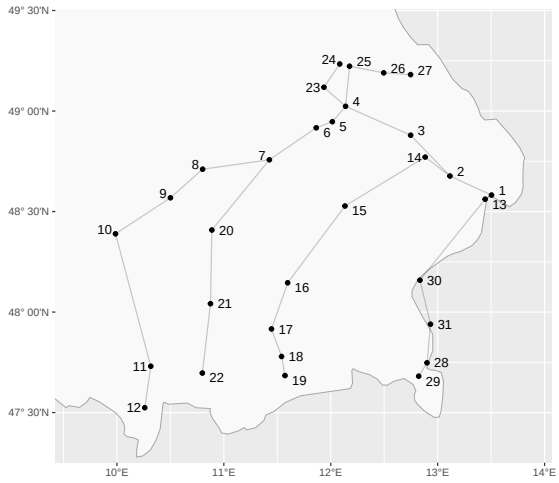


Figure 4.6
Stations from the Danube river dataset available in the `R` package `graphicalExtremes`; see Engelke et al. [40].

distributed for $i = 1, \dots, n$ and we will model their tail dependence using a mixture Hüsler–Reiss forest model as in Section 4.4.2. Before estimating this dataset, we first discuss the choice of the tuning parameters.

Tuning parameters. As discussed in the simulation study in Section 4.6, the exceedance proportion used to estimate the minimum spanning tree is the topic of Engelke and Volgushev [36]. The author suggest to choose the exceedance proportion as to minimize the deviation of the empirical variogram matrix to form a variogram matrix over a tree. However, this choice does not guarantee an optimal p that gives the "best" minimum spanning tree to model the extremal behavior of the dataset. As in Section 4.6, we simply fix $p = n^{0.8}/n \approx 0.3$.

The choice of the tail fraction is a bias–variance trade-off. In Section 4.6, the sample size was proportional to 10^3 and the optimal tail fraction was between 0.05 and 0.1. Given the limited sample size of only $n = 428$, we will need a slightly higher tail fraction to have enough data for inference. We set $k/n \in \{0.1, 0.15\}$.

As discussed in Section 4.6, the choice of the penalization exponent α is not important as long as the penalization parameter α is well chosen. Here, we simply choose $\alpha \in \{0.2, 0.5\}$.

As explained in Algorithm 8, we use a K-fold cross validation to choose the penalization parameter β. Here, we set $K = 10$ and we choose the penalization parameter β using a K-fold cross validation as in Algorithm 8 from a grid of values starting at 10^{-5} and increasing by 2×10^{-5} up to 10^{-3}.

Results. Figure 4.13 shows the resulting forest structure from Algorithm 6. Unsurprisingly, the estimated forest is simply a tree for all values $k/n \in \{0.1, 0.15\}$ and $\alpha \in \{0.2, 0.5\}$. This is explained by the spatial structure of the stations where the stations are not geographically far from each other.

In terms of extreme directions, as shown in Figure 4.13, some mixture coefficients $\hat{a}_{st}$ take values strictly less than 1. Proposition 4.4.3 then implies the existence of non-standard extreme directions $J \subset \{1, \dots, d\}$. Determining the full set of extreme directions in dimension $d = 31$ remains computationally expensive, since Algorithm 12 must enumerate all subsets $J \subset V$ which results in a complexity $\mathcal{O}(2^d)$. Note that Station s belongs to an extreme direction $J \subset \{1, \dots, d\}$ excluding Station t if and only if the unique path between s and t contains an edge (i, j) with $\hat{a}_{ij} < 1$. For example, Station 1 can lie in an extreme direction that excludes Station 12.

As a validation, we fit a bivariate mixture Hüsler–Reiss model as in Section 4.4.2 for the station pair $j_1 = 1$ and $j_2 = 12$, estimating the parameters $a_{j_1 j_2}$ and $\Gamma_{j_1 j_2}$ via (4.5.1) with the same tuning parameters used above. We obtain $\hat{a}_{j_1 j_2} = 0.74 < 1$, which confirms that one station can lie in an extreme direction that excludes the other.

Note also that the same results are obtained with different values of the tuning parameters α and k/n which means that our model is stable under a change of these parameters.

To assess the quality of our fit, we compare the model-based fitted extremal correlations to their empirical counterparts in (4.A.3). Under the mixture Hüsler–Reiss forest model on a tree $\mathcal{F} = (V, E)$ with edge-specific parameters $\{(a_{st}, \Gamma_{st}) : (s, t) \in E\}$, Engelke et al. [38, Example 8.8] show that the theoretical extremal correlation between two stations $j_1, j_2 \in V$ is

$$\chi_{j_1 j_2} = \Big[2 - 2\Phi\Big(\sqrt{\Gamma_{j_1 j_2}/2}\Big)\Big] \cdot \prod_{(s,t) \in \mathrm{path}(j_1, j_2)} a_{st}, \tag{4.7.1}$$

where $\Gamma_{j_1 j_2} = \sum_{(s,t) \in \mathrm{path}(j_1, j_2)} \Gamma_{st}$, Φ is the standard normal cdf, and $\mathrm{path}(j_1, j_2)$ denotes the unique path in $\mathcal{F}$ between j_1 and j_2 when it exists. We plug in our estimates $\{(\hat{a}_{st}, \hat{\Gamma}_{st}) : (s, t) \in E\}$ from Step 3 of Algorithm 6 to obtain the fitted correlations $\{\hat{\chi}_{j_1 j_2} : j_1, j_2 \in V\}$.

Figure 4.7 compares the empirical and fitted extremal correlations. Here we use $\alpha = 0.2$ and $k/n = 0.05$, noting that other choices of tuning parameters yield qualitatively similar results. The agreement is strong, as most points lie close to the line $x = y$.

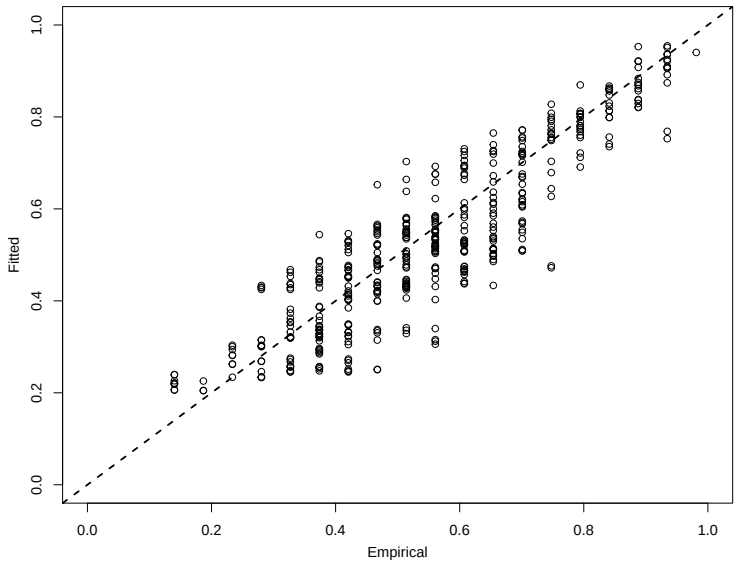


Figure 4.7
Scatterplot of empirical versus fitted extremal correlations for all station pairs, computed with $\alpha = 0.2$ and $k/n = 0.05$. The 45° line $x = y$ indicates relatively good agreement; the clustering of points near this line reflects the strong correspondence between the empirical estimates and the model fit.

4.7.2 Financial portfolios data

In this section, T will denote the sample size and t time. We focus on the daily negative value-averaged returns of $d = 8$ industry portfolios downloaded from the Kenneth French Data Library[1]. The portfolios are from the following sectors: "Food", "Beer", "Smoke", "Healthcare", "Utilities", "Electrical equipment", "Books" and "Games". Data are available for the period from January 1970 until December 2019, leading to a sample of size $T = 12\,613$.

Data pre-processing. Let $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{td})$ denote the negative value-averaged returns of the portfolios at time $t = 1, \ldots, T$. In order to obtain a time series without temporal dependence, we fit a vector autoregressive model [71] of order p:

$$\mathbf{X}_t \;=\; \mathbf{c} \;+\; \sum_{i=1}^{p} A_i\, \mathbf{X}_{t-i} \;+\; \boldsymbol{\varepsilon}_t, \qquad t = p+1, \ldots, T,$$

[1] `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

where $\mathbf{c}$ is a d-vector of intercepts, each A_i is a $d \times d$ coefficient matrix on the lagged values, and $\boldsymbol{\varepsilon}_t$ is a d-vector of innovations assumed to be serially uncorrelated. We select the lag order p by minimizing the AIC and find $p = 5$. Let $\hat{\mathbf{X}}_t$ be the fitted value at time t. We define the residuals

$$\boldsymbol{\varepsilon}_t \;=\; \mathbf{X}_t \;-\; \hat{\mathbf{X}}_t, \qquad t = 1, \ldots, T.$$

We then apply the Box–Ljung test for serial correlation to each of the $d = 8$ marginal series residual series $\{\boldsymbol{\varepsilon}_t\}_{t=1}^T$, using lags 1 through 10. In every case, the test fails to reject the null hypothesis of no autocorrelation, indicating that the residuals are now serially uncorrelated. We therefore adopt $\{\boldsymbol{\varepsilon}_t\}_{t=1}^T$ as our new observational series.

Results. We use the same tuning parameters as in Section 4.7.1. Figure 4.14 shows the resulting forest structure $\hat{\mathcal{F}}$ from Algorithm 6. Results are shown for penalization exponent $\alpha = 0.5$, tail fraction $k/n = 0.1$, and penalization parameter β chosen via 10-fold cross-validation as in Algorithm 8 from a grid of values starting at 10^{-5} and increasing by 2×10^{-5} up to 10^{-3}. Other values of the tuning parameters give quantitatively similar results. We obtain a simple tree structure, which indicates that no sectors exhibit extremal independence.

In terms of extreme directions, as presented in Figure 4.14, we have $0 < \hat{a}_{st} < 1$ for every edge (s, t). Proposition 4.4.3 implies that every group of sectors $J \subset \{1, \ldots, d\}$ such that $\hat{\mathcal{F}}_J$ is connected is an extreme direction. The results seem to make sense given the types of sectors.

To assess the quality of our fit, Figure 4.8 compares the fitted and empirical extremal correlations in (4.7.1) and (4.A.3) (with $q = 0.95$), respectively. The results are relatively good; however, we observe that the fitted model underestimates the extremal correlation for most sector pairs, except for seven pairs that correspond exactly to the edges (s, t) of the estimated forest. Intuitively, given (4.7.1), the fitted extremal correlation between two sectors j_1, j_2 is decreasing as a function of the length of the path that connects them in the estimated graph. Equivalently, if our model systematically underestimates $\chi_{j_1 j_2}$, this suggests that in the true underlying graph those two sectors are actually joined by a shorter path than the one recovered by the estimate. Another explanation would be that the model assumption of extremal conditional independence along a forest is not correct.

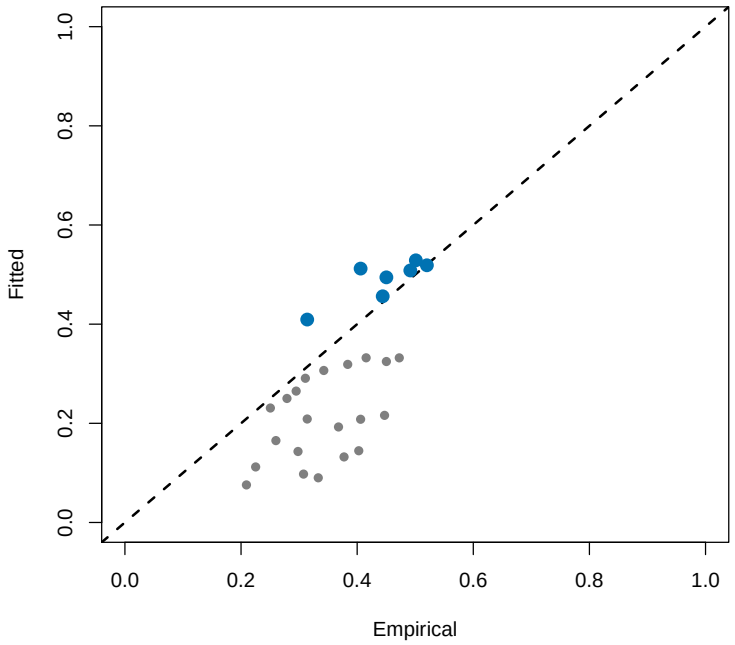


Figure 4.8
Scatterplot of empirical extremal correlations in (4.A.3) with threshold $q = 0.95$ versus fitted extremal correlations in (4.7.1), using tail fraction $k/n = 0.1$ and penalization exponent $\alpha = 0.5$, for all sector pairs (s, t). Blue points highlight those pairs (s, t) that form the edges of the estimated forest $\hat{\mathcal{F}}$ in 4.14.

4.8 Conclusion

Given a vector of d risk factors, extremal graphical models proposed in the literature, for instance the Hüsler–Reiss graphical model [53, 34], cover the case where the only possible extreme direction is $\{1, \ldots, d\}$, meaning that all risk factors are large simultaneously. However, in many fields, some group of risk factors $J \subsetneq \{1, \ldots, d\}$ can be large simultaneously while those outside the group are not. Such a case is illustrated in Section 4.7.1 with the example of Danube river discharge, and in Section 4.7.2 with the example of financial portfolios.

In this paper, we proposed a construction method of extremal graphical models with non-standard extreme directions, i.e., extreme directions $J \neq \{1, \ldots, d\}$. This construction also allows some group of risk factors to exhibit extremal independence. As an example, we generalized the Hüsler–Reiss graphical model to what we have

called the mixture Hüsler–Reiss graphical model, expressed the parameters of this model, and therefore identified the set of its extreme directions. In a more specific example, we discussed the mixture Hüsler–Reiss graphical model when the underlying structure is a forest, and showed that parameter inference of this model in dimension d reduces to bivariate inference. We proposed a data-driven algorithm for forest structure learning, generalizing the one proposed in Engelke and Volgushev [36] that treats the case of extremal tree models. Finally, we performed a simulation study of this algorithm and a data application showing that choosing good tuning parameters for our algorithm leads to good results.

Asymptotic consistency results of the forest structure learning procedure in Algorithm 6 still need to be established. This goes through two points:

1. Show that the minimum spanning tree $\hat{T}$ fitted in Step 1 contains all edges of the underlying true forest structure, that is, Equation (4.6.4).

2. Consider a bivariate mixture Hüsler–Reiss graphical model with coefficient matrix A as in (4.4.6), or equivalently a mixture coefficient $0 \leqslant a \leqslant 1$. We need to show that the PLS estimator $\hat{a}_n$ in (4.5.1) is consistent in the case of extremal independence, in the sense that, when $a = 0$, we have

$$\lim_{n\to\infty} \mathsf{P}(\hat{a}_n = 0) = 1.$$

Another idea is to slightly change the parameterization of the coefficient matrix A in (4.4.6) to

$$A = \begin{pmatrix} a & b & 0 \\ a & 0 & c \end{pmatrix}, \qquad 0 \leqslant a, b, c \leqslant 1. \tag{4.8.1}$$

The idea is to estimate the parameters a, b, c via the PLS estimator and then standardize to obtain unit row sums of A. Note that this parametrization allows having the extreme direction $\{1\}$ without having the extreme direction $\{2\}$ (case $b \neq 0$, $c = 0$) or having the extreme direction $\{2\}$ without having the extreme direction $\{1\}$ (case $b = 0$, $c \neq 0$). Conversely, the parametrization of A in (4.4.6) implies that the existence of one of these two extreme directions entails the existence of the other. Note that a drawback of the new parametrization in (4.8.1) is that the resulting constructed density λ_V in Equation (4.3.5) is no longer a mixture Hüsler–Reiss forest density as in Proposition 4.4.3. This is because Condition (4.4.7) is no longer valid for A.

4.A Decomposable graphs

4.A.1 Background

We provide a brief overview of graph theory concepts used in the paper. For more detailed information, refer to Cowell et al. [21], Lauritzen [68]. Recall the integer $d \geqslant 1$. A graph $G = (V, E)$ consists of a set of nodes $V = \{1, \ldots, d\}$ and a set of edges $E \subset V \times V$, where each edge is a pair of distinct nodes. The graph $\mathcal{G}$ is called undirected if, for every two nodes $s, t \in V$, the presence of an edge $(s, t) \in E$ implies that $(t, s) \in E$ as well. For notational convenience, edges (s, t) and (t, s) will be identified and represented in increasing order of node indices, meaning (s, t) if $t > s$ and (t, s) if $s > t$. When counting the number of edges, each edge is considered only once. A path between two nodes $s, t \in V$ in a graph, denoted as $\text{path}(s, t)$, is a sequence of edges that allows to travel from one node to the other without interruption. It consists of a series of connected nodes, where each step moves along an edge linking two nodes directly. The graph $\mathcal{G}$ is called connected if there exists a $\text{path}(s, t)$ between s and t, for every pair of nodes $s, t \in V$. Otherwise, $\mathcal{G}$ is said to be disconnected. For a disconnected graph, nodes can be partitioned into subsets, where each subset induces a subgraph called a connected component, in which every two nodes are connected by a path, but there is no path between nodes in different components. Finally the graph $\mathcal{G}$ is called complete if it is fully connected, meaning that there exists an edge $(s, t) \in E$ for every pair of distinct nodes s and t from V.

For a non-empty set $C \subset V$, the subgraph $\mathcal{G}_C = (C, E_C)$ is the graph whose node set is C and whose edge set E_C consists of all the edges in E that have both endpoints in C, that is, $E_C = E \cap (C \times C)$. A set C is said to be a clique if the subgraph $\mathcal{G}_C$ is complete and C is maximal in the sense that it cannot be properly contained within any other non-empty $\bar{C} \subset V$ such that $\mathcal{G}_{\bar{C}}$ is complete. A subset $D \subset V$ separates $\mathcal{G}$ if at least two nodes in the same connected component of $\mathcal{G}$ are in two distinct connected components of $\mathcal{G}_{V \setminus D}$. Note that in the case where $\mathcal{G}$ is connected, a subset D separates $\mathcal{G}$ if and only if $\mathcal{G}_{V \setminus D}$ is no longer connected.

The class of *decomposable graphs* [68] holds significant utility within the realm of extremal graphical models, as elucidated by Engelke and Hitz [34] and Engelke et al. [38]. These graphs serve as a foundational framework for constructing extreme value models across the entirety of a graph $\mathcal{G}$ by leveraging lower-dimensional extreme value models, as detailed in Engelke and Hitz [34, Section 5]. Next, we will formally define the class of decomposable graphs and present the relevant results that are essential for our discussion in this paper. A partition (A, B, C) of V is said to form a decomposition of the graph $\mathcal{G}$ if B separates A from C, meaning that all paths from elements from A to elements from C intersect B, and the subgraph $\mathcal{G}_B$ is complete. This decomposition is proper if A and C are both non-empty. When

this is the case, we say that (A, B, C) separates $\mathcal{G}$ into the subgraphs $\mathcal{G}_{A\cup B}$ and $\mathcal{G}_{B\cup C}$. A graph is *prime* if no proper decomposition exists.

The graph $\mathcal{G}$ is called *decomposable* if it is complete or if there exists a proper decomposition (A, B, C) into decomposable subgraphs $\mathcal{G}_{A\cup B}$ and $\mathcal{G}_{B\cup C}$. This definition is recursive; in particular, $\mathcal{G}$ is decomposable if and only if every maximal prime subgraph of $\mathcal{G}$ forms a complete graph. Here, a "maximal prime subgraph" refers to a prime subgraph of $\mathcal{G}$ that is not strictly contained within any larger prime subgraph of $\mathcal{G}$. In other words, it is as large as possible while still maintaining its property of being prime. Finally, remark that a graph $\mathcal{G}$ is decomposable if and only if its connected components are also decomposable. See Appendix 4.A.3 for an example of a decomposable graph.

We introduce now some results from the literature about decomposable graphs that will be useful throughout the paper. For more clarity, all the concepts are illustrated through examples in Section 4.A.3 from the appendix. In the following, let $\mathcal{G} = (V, E)$ be a fixed decomposable graph with cliques $\mathcal{C}$.

A *junction tree* associated to a decomposable graph $\mathcal{G}$ is a tree[2] $\mathcal{T} = (\mathcal{C}, E_{\mathcal{T}})$ whose node set is $\mathcal{C}$ and satisfies the *path intersection property*:

$$\forall C_s,\, C_t \in \mathcal{C}, \quad \forall C \in \text{path}\,(C_s, C_t): \qquad C_s \cap C_t \subset C, \tag{4.A.1}$$

where $\text{path}\,(C_s, C_t)$ here should be seen as the node-cliques that appear in the unique path between C_s and C_t in $\mathcal{T}$. Lauritzen [68, Proposition 2.27] establishes that the existence of such a tree is both a necessary and sufficient condition for a graph to be decomposable. However, a junction tree is not unique.

An ordering $\{C_1, \ldots, C_m\}$ of $\mathcal{C}$ satisfies the *running intersection property* if, for every $i = 2, \ldots, m$, there exists an index $k < i$ such that

$$D_i := C_i \cap \bigcup_{j=1}^{i-1} C_j \subset C_k. \tag{4.A.2}$$

Lauritzen [68, Proposition 2.17] shows that such an ordering always exists for decomposable graphs, although it is not unique. Note that the multiset $\mathcal{D} = \{D_2, \ldots, D_m\}$ may contain duplicate elements. Dirac [26, Theorem 1] established that the multiset $\mathcal{D}$ is uniquely determined up to ordering. Henceforth, we refer to $\mathcal{D}$ as the separators of $\mathcal{G}$. Finally, observe that $\mathcal{D}$ contains the empty set if and only if $\mathcal{G}$ is disconnected.

A *depth-first traversal* (DFT) of an oriented tree [42, Chapter 3] is a tree traversal algorithm that explores each branch of the tree as deeply as possible before backtracking to explore the next branch.

[2]A tree is an undirected graph in which every two nodes are connected by exactly one path.

Minimum spanning tree. The minimum spanning tree is one of the most known approaches in tree structure learning. Given a candidate tree $\mathcal{T} = (V, E)$, each edge $(s, t) \in E$ is assigned a cost ρ_{st}. The idea is to choose the tree $\mathcal{T}_{\text{mst}}$ that minimizes the sum of these weights along its edges. More precisely

$$\mathcal{T}_{\text{mst}} = \underset{\mathcal{T}=(V,E)}{\arg\min} \sum_{(s,t)\in E} \rho_{st}.$$

Given a d-variate sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with underlying graph structure $\mathcal{T}$. Let $0 \leqslant p_n \leqslant 1$ be the *exceedance proportion* depending on the sample size n. Consider the empirical extremal correlation in (4.A.3) and the empirical extremal variogram (for any fixed $m \in V$) in (4.A.3)

$$\hat{\chi}_{st} = \frac{1}{p_n} \sum_{i=1}^{n} \mathbb{1}\left\{\hat{F}_s(X_{is}) > 1 - p_n,\ \hat{F}_s(X_{it}) > 1 - p_n\right\}, \tag{4.A.3}$$

$$\hat{\Gamma}_{st} = \widehat{\text{Var}}\left[\log\left(1 - \hat{F}_s(X_{is})\right) - \log\left(1 - \hat{F}_s(X_{is})\right) :\ \hat{F}_m(X_{im}) \geqslant 1 - p_n\right], \tag{4.A.4}$$

where $\hat{F}_j$ denotes the empirical cumulative distribution function associated to $X_{1j}, \ldots, X_{nj}$. Under certain conditions, the choice of the cost

$$\rho_{st} = \hat{\Gamma}_{st}, \qquad \text{or,} \qquad \rho_{st} = -\log(\hat{\chi}_{st})$$

guarantees consistency results on the associated minimum spanning tree $\widehat{\mathcal{T}}_{\text{mst}}$, meaning that

$$\mathsf{P}(\widehat{\mathcal{T}}_{\text{mst}} = \mathcal{T}) \to 1, \qquad n \to \infty.$$

For more details about the empirical extremal correlation, the empirical extremal variogram and the consistency result of the minimum spanning tree, see Engelke and Volgushev [36], Hu et al. [56].

4.A.2 Results

Recall the definitions from Section 4.A.1. Throughout this section, let $V = \{1, \ldots, d\}$ and $\mathcal{G} = (V, E)$ be a decomposable, undirected, non-complete graph with cliques $\mathcal{C}$, separators $\mathcal{D}$, and an associated junction tree $\mathcal{T} = (\mathcal{C}, E_{\mathcal{T}})$.

Lemma 4.A.1. *Select a clique $C^* \in \mathcal{C}$ as the root of $\mathcal{T}$ and establish the natural orientation of $\mathcal{T}$ emanating from this root. Consider the ordering $\{C_1 = C^*, C_2, \ldots, C_m\}$ of the cliques in $\mathcal{C}$ induced by any DFT of $\mathcal{T}$. This ordering satisfies the running intersection property in* (4.A.2)*. Moreover, the separators $\mathcal{D} = \{D_2, \ldots, D_m\}$ of $\mathcal{G}$ as defined in* (4.A.2) *satisfy the* separator property*,*

$$\mathcal{D} = \{C_s \cap C_t : (C_s, C_t) \in E_{\mathcal{T}}\}. \tag{4.A.5}$$

Proof of Lemma 4.A.1. Let $i \in \{2, \ldots, m\}$ and let C_i be the i-th clique from $\{C_1, \ldots, C_m\}$. Denote by $\mathrm{pa}(C_i)$ the parent of the clique C_i in $\mathcal{T}$ with respect to its natural orientation. The path intersection property in (4.A.1) implies that $C_i \cap C_j \subset \mathrm{pa}(C_i)$ for every $j < i$. Consequently, we have $C_i \cap \bigcup_{j=1}^{i-1} C_j \subset C_i \cap \mathrm{pa}(C_i) \subset \mathrm{pa}(C_i)$, which establishes the running intersection property in (4.A.2) with $C_k = \mathrm{pa}(C_i)$.

Applying the same reasoning and noting that also $C_i \cap \mathrm{pa}(C_i) \subset C_i \cap \bigcup_{j=1}^{i-1} C_j$ and thus $C_i \cap \mathrm{pa}(C_i) = C_i \cap \bigcup_{j=1}^{i-1} C_j$, we obtain

$$\mathcal{D} = \left\{ C_i \cap \bigcup_{j=1}^{i-1} C_j : i \in \{2, \ldots, m\} \right\} = \{ C_s \cap C_t : (C_s, C_t) \in E_{\mathcal{T}} \}, \qquad (4.A.6)$$

where the first equality in (4.A.6) follows from the fact that the ordering $\{C_1, \ldots, C_m\}$ satisfies the running intersection property. □

For the following lemma, consider two adjacent cliques C^* and C^{**} in the junction tree $\mathcal{T}$. Note that removing the edge (C^*, C^{**}) partitions $\mathcal{T}$ into two disjoint subtrees, denoted by $\mathcal{T}^*$ and $\mathcal{T}^{**}$.

Lemma 4.A.2. *Let V^* and V^{**} be the sets of nodes corresponding to the union of the cliques in $\mathcal{T}^*$ and $\mathcal{T}^{**}$, respectively. Then, the separator $D = C^* \cap C^{**}$ separates $V^* \setminus D$ and $V^{**} \setminus D$ in $\mathcal{G}$.*

Proof. See Cowell et al. [21, Proof of Theorem 4.6]. □

Lemma 4.A.3. *Let $J \subset V$ be a non-empty set such that $\mathcal{G}_J$ is connected. Let $\mathcal{G}_J = \{C \in \mathcal{C} : C \cap J \neq \varnothing\}$. Then the subgraph $\mathcal{T}_{\mathcal{C}_J}$ is connected. Moreover, $C^* \cap C^{**} \cap J \neq \varnothing$ for every edge (C^*, C^{**}) of $\mathcal{T}_{\mathcal{C}_J}$.*

Proof. Suppose that $\mathcal{T}_{\mathcal{C}_J}$ is disconnected. Then, there exist two cliques $C', C'' \in \mathcal{C}_J$ and a clique $C^* \notin \mathcal{C}_J$ such that C^* separates C' and C'' in $\mathcal{T}$. Let C^{**} be any adjacent clique to C^* in $\mathcal{T}$ (possibly C' or C''). Removing the edge (C^*, C^{**}) partitions $\mathcal{T}$ into two disjoint subtrees and C', C'' are not in the same subtree. Consider $D^* = C^* \cap C^{**}$. On the one hand, $D^* \cap J = \varnothing$, and on the other hand, Lemma 4.A.2 shows that D^* separates $C' \cap J$ and $C'' \cap J$ in $\mathcal{G}$. This contradicts the assumption that $\mathcal{G}_J$ is connected.

Let (C^*, C^{**}) be an edge in $\mathcal{T}_{\mathcal{C}_J}$ such that $(C^* \cap C^{**}) \cap J = \varnothing$. It follows that $C^* \cap J$ and $C^{**} \cap J$ are nonempty and disjoint. Moreover, the same reasoning as above shows that $C^* \cap C^{**}$ separates $C^* \cap J$ and $C^{**} \cap J$. Hence there exist two elements of J separated by the set $I = C^* \cap C^{**}$ with $I \cap J = \varnothing$, contradicting the connectivity of $\mathcal{G}_J$. □

For two different cliques $C', C'' \in \mathcal{C}$, let $\mathrm{path}(C', C'')$ denote the cliques and $\mathcal{D}_{C'C''}$ the separators in the unique path between C' and C'' in $\mathcal{T}$. More precisely,

$$\mathrm{path}(C', C'') = \left\{C', C''\right\} \cup \left\{C \in \mathcal{C} : C \text{ separates } C' \text{ and } C'' \text{ in } \mathcal{T}\right\}, \tag{4.A.7}$$

$$\mathcal{D}_{C'C''} = \left\{C^* \cap C^{**} : (C^*, C^{**}) \in E_{\mathcal{T}},\, C^*, C^{**} \in \mathrm{path}(C', C'')\right\}. \tag{4.A.8}$$

Lemma 4.A.4. *For a separator $D^* \in \mathcal{D}$ and two different cliques C', $C'' \in \mathcal{C}$ such that $D^* \subset C' \cap C''$, we assert that $D^* \subset D$ for all $D \in \mathcal{D}_{C'C''}$.*

Proof of Lemma 4.A.4. The path intersection property in (4.A.1) yields that $D^* \subset C' \cap C'' \subset C$ for every clique $C \in \mathrm{path}(C', C'')$, and therefore also $D^* \subset C^* \cap C^{**}$ for every $C^*, C^{**} \in \mathrm{path}(C', C'')$, and thus $D^* \subset D$ for all $D \in \mathcal{D}_{C'C''}$. □

In the following, for a non-empty set $J \subset V$, let $\mathcal{D}_J = \{D \in \mathcal{D} : D \cap J \neq \varnothing\}$ and recall $\mathcal{C}_J$ as defined in Lemma 4.A.3.

Lemma 4.A.5. *Let $J \subset V$ such that $\mathcal{G}_J$ is connected. We have*

$$|J| = \sum_{C \in \mathcal{C}_J} |J \cap C| - \sum_{D \in \mathcal{D}_J} |J \cap D|. \tag{4.A.9}$$

Proof. Lemma 4.A.3 shows that $\mathcal{T}_{\mathcal{C}_J}$ is connected and

$$\mathcal{D}_J = \{C_s \cap C_t : (C_s, C_t) \in E_{\mathcal{T}},\, C_s, C_t \in \mathcal{C}_J\}. \tag{4.A.10}$$

Let $\{C_1, \ldots, C_m\}$ be the order induced by a DFT order of $\mathcal{T}_{\mathcal{C}_J}$. Note that this is a continuous traversal of $\mathcal{T}_{\mathcal{C}_J}$ in the sense that for $k \leqslant m$, the graph $\mathcal{T}_k$ induced by the node-cliques $C_1, \ldots, C_k$ is connected. The inclusion-exclusion principle shows

$$|J \cap (C_1 \cup C_2)| = |J \cap C_1| + |J \cap C_1| - |J \cap (C_1 \cap C_2)|.$$

Let $k \leqslant m$ and denote by E_{k-1} the edges of $\mathcal{T}_{k-1}$. Suppose

$$\Big|J \cap \bigcup_{j=1}^{k-1} C_j\Big| = \sum_{j=1}^{k-1} |J \cap C_j| - \sum_{\substack{1 \leqslant j_1 < j_2 \leqslant k-1 \\ (C_{j_1}, C_{j_2}) \in E_{k-1}}} |J \cap (C_{j_1} \cap C_{j_2})|. \tag{4.A.11}$$

Consider now the sub-tree $\mathcal{T}_k$ and let $\mathrm{N}(C_k) \in \{C_1, \ldots, C_{k-1}\}$ be the unique adjacent node-clique of C_k in $\mathcal{T}_k$. The path intersection property in (4.A.1) shows that $C_i \cap C_k \subset \mathrm{N}(C_k)$, for every $i \leqslant k-1$. The inclusion-exclusion principle,

followed by (4.A.11) yield

$$
\begin{aligned}
|J \cap \bigcup_{j=1}^{k} C_j| &= |J \cap \bigcup_{j=1}^{k-1} C_j| + |J \cap C_k| - \left| \left(J \cap \bigcup_{j=1}^{k-1} C_j \right) \cap (J \cap C_k) \right| \\
&= \sum_{j=1}^{k} |J \cap C_j| - \sum_{\substack{1 \leqslant j_1 < j_2 \leqslant k-1 \\ (C_{j_1}, C_{j_2}) \in E_{k-1}}} |J \cap (C_{j_1} \cap C_{j_2})| - |J \cap \mathrm{N}(C_k) \cap C_k| \\
&= \sum_{j=1}^{k} |J \cap C_j| - \sum_{\substack{1 \leqslant j_1 < j_2 \leqslant k \\ (C_{j_1}, C_{j_2}) \in E_k}} |J \cap (C_{j_1} \cap C_{j_2})|. \qquad (4.A.12)
\end{aligned}
$$

In particular, Equation (4.A.12) for $k = m$, followed by (4.A.10) show (4.A.9). □

4.A.3 Examples and Figures

Consider the decomposable graph in Figure 4.9. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ denote the connected components associated to the node sets $N_1 = \{1, \ldots, 6\}$ and $N_2 = \{7, \ldots, 10\}$, respectively. We illustrate the background of graph theory in Section 4.A.1 with examples.

- The subgraphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are both decomposable, with cliques $\mathcal{C}_1, \mathcal{C}_2$ and separators $\mathcal{D}_1, \mathcal{D}_2$, given by:

$$
\begin{aligned}
\mathcal{C}_1 &= \Big\{\{1,2,3\}, \{2,3,4\}, \{3,4,5,6\}\Big\}, \qquad \mathcal{C}_2 = \Big\{\{7,8\}, \{8,9\}, \{9,10\}\Big\}, \\
\mathcal{D}_1 &= \Big\{\{2,3\}, \{3,4\}\Big\}, \qquad\qquad\qquad\quad \mathcal{D}_2 = \Big\{\{8\}, \{8\}\Big\}.
\end{aligned}
$$

- For the subgraph $\mathcal{G}_1$, the order $C_1 = \{2,3,4\}$, $C_2 = \{1,2,3\}$, $C_3 = \{3,4,5,6\}$ satisfies the running intersection property in (4.A.2). However, this order does not satisfy the separator property (4.A.5). The order $C_1 = \{1,2,3\}$, $C_2 = \{2,3,4\}$, $C_3 = \{3,4,5,6\}$ satisfies both the running intersection property and the separator property (4.A.5).
- For the subgraph $\mathcal{G}_2$, the element $\{8\}$ appears twice in $\mathcal{D}_2$ as a duplicate.
- Let $\mathcal{D}$ denote the set of separators of $\mathcal{G}$. Note that $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \{\varnothing\}$.

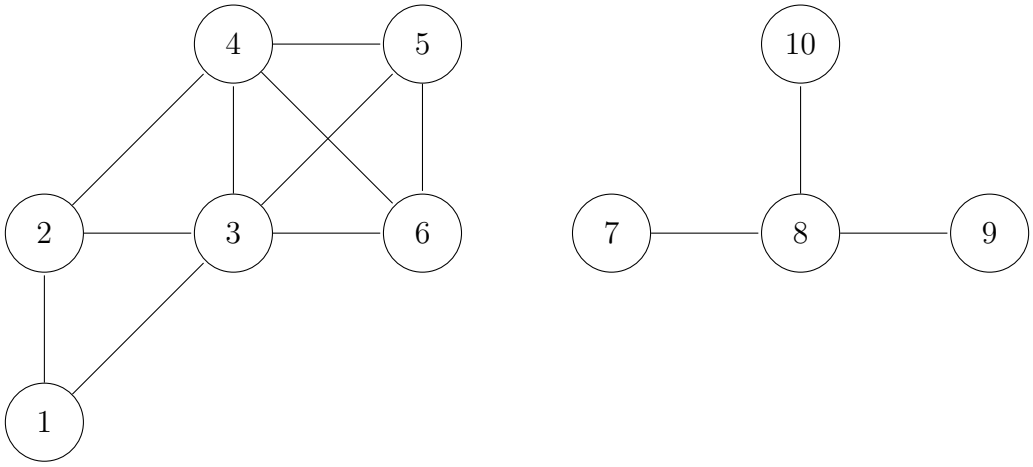


Figure 4.9
A decomposable graph with two connected components, $\mathcal{G}_1$ and $\mathcal{G}_2$, which are the induced subgraphs of the node sets $N_1 = \{1, \ldots, 6\}$ and $N_2 = \{7, \ldots, 10\}$, respectively.

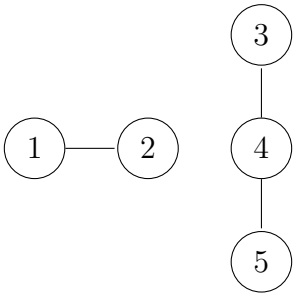


Figure 4.10
(a) A 5-dimensional forest with two connected sub-trees $\{1, 2\}$ and $\{3, 4, 5\}$.

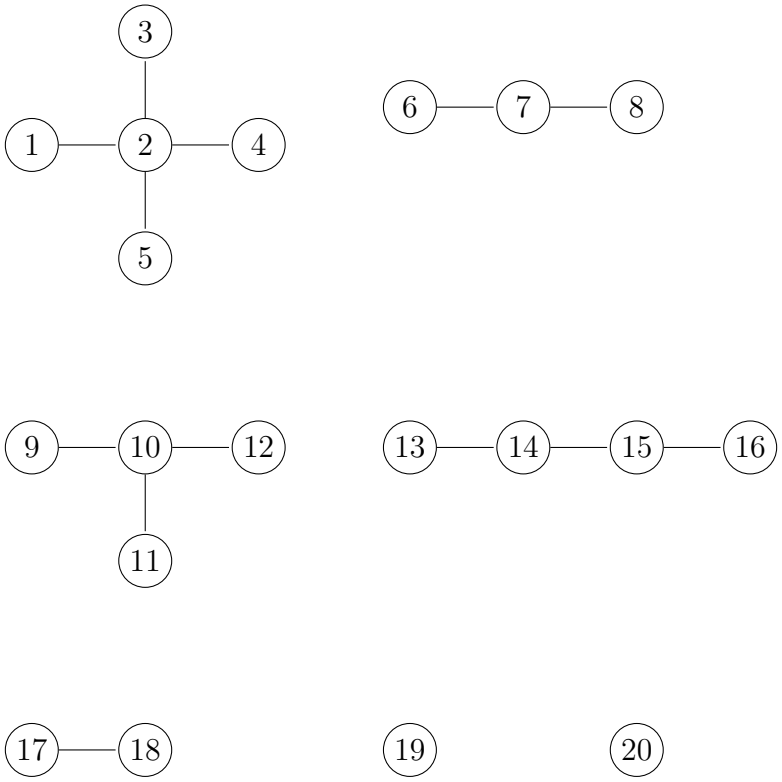


Figure 4.11
A $d = 20$ forest with components $\{1, 2, 3, 4, 5\}$ and $\{6, 7, 8\}$ on row 1; $\{9, 10, 11, 12\}$ and $\{13, 14, 15, 16\}$ on row 2; $\{17, 18\}$, $\{19\}$, and $\{20\}$ on row 3.

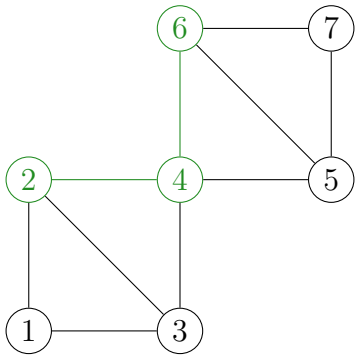


Figure 4.12
A 7-dimensional decomposable graph with cliques $C_1 = \{1, 2, 3\}$, $C_2 = \{2, 3, 4\}$, $C_3 = \{4, 5, 6\}$, $C_4 = \{5, 6, 7\}$.

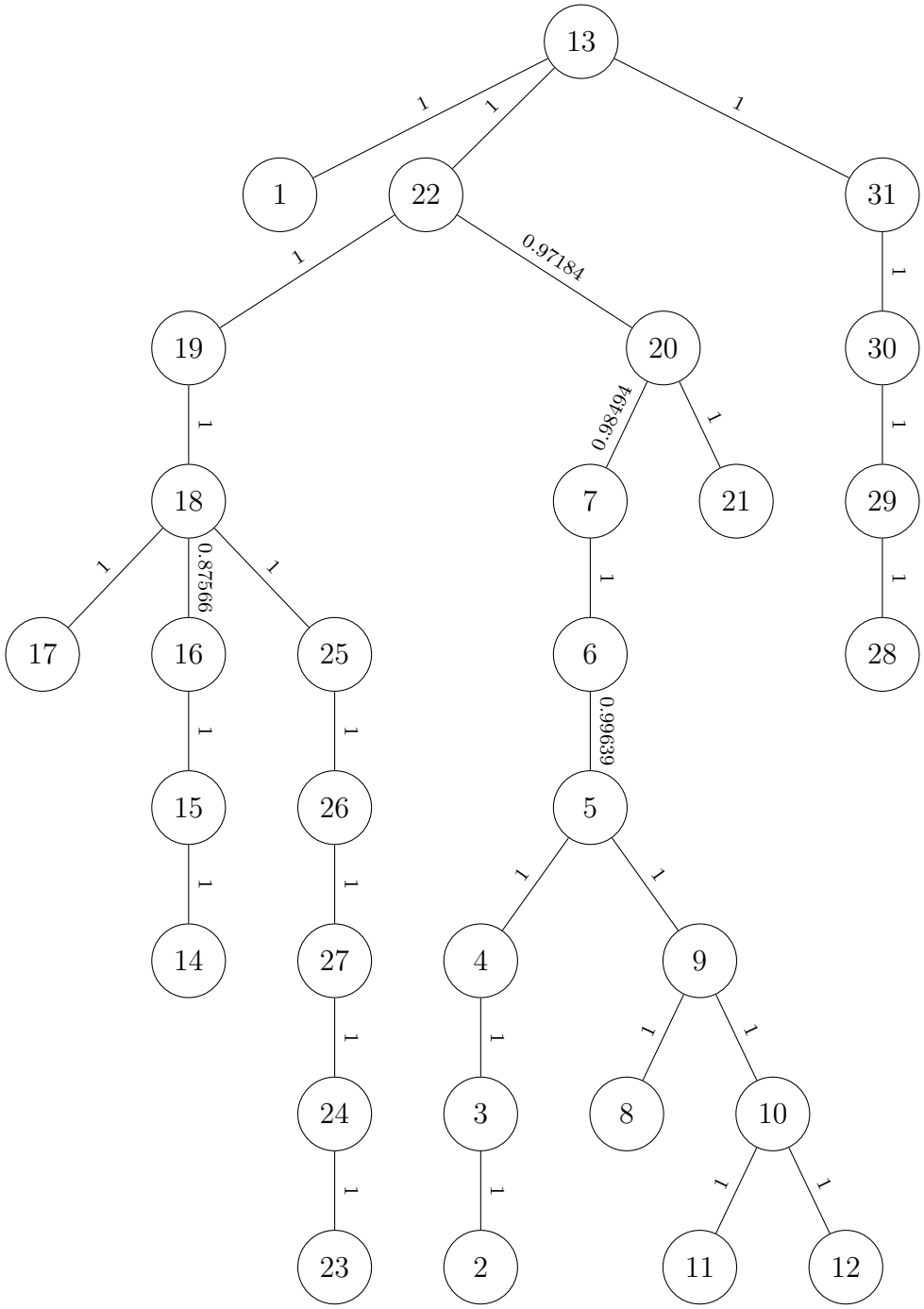


Figure 4.13
Estimated forest from Algorithm 6 on the Danube dataset of $d = 31$ stations; see Section 4.7.1. For each pair (s, t), we report the mixture coefficient $\hat{a}_{st}$ estimated in Step 3 of Algorithm 6. Station s can lie in an extreme direction $J \subset \{1, \ldots, d\}$ that excludes station t if along the unique path between s and t there exists an edge (i, j) with mixture coefficient $\hat{a}_{ij} < 1$.

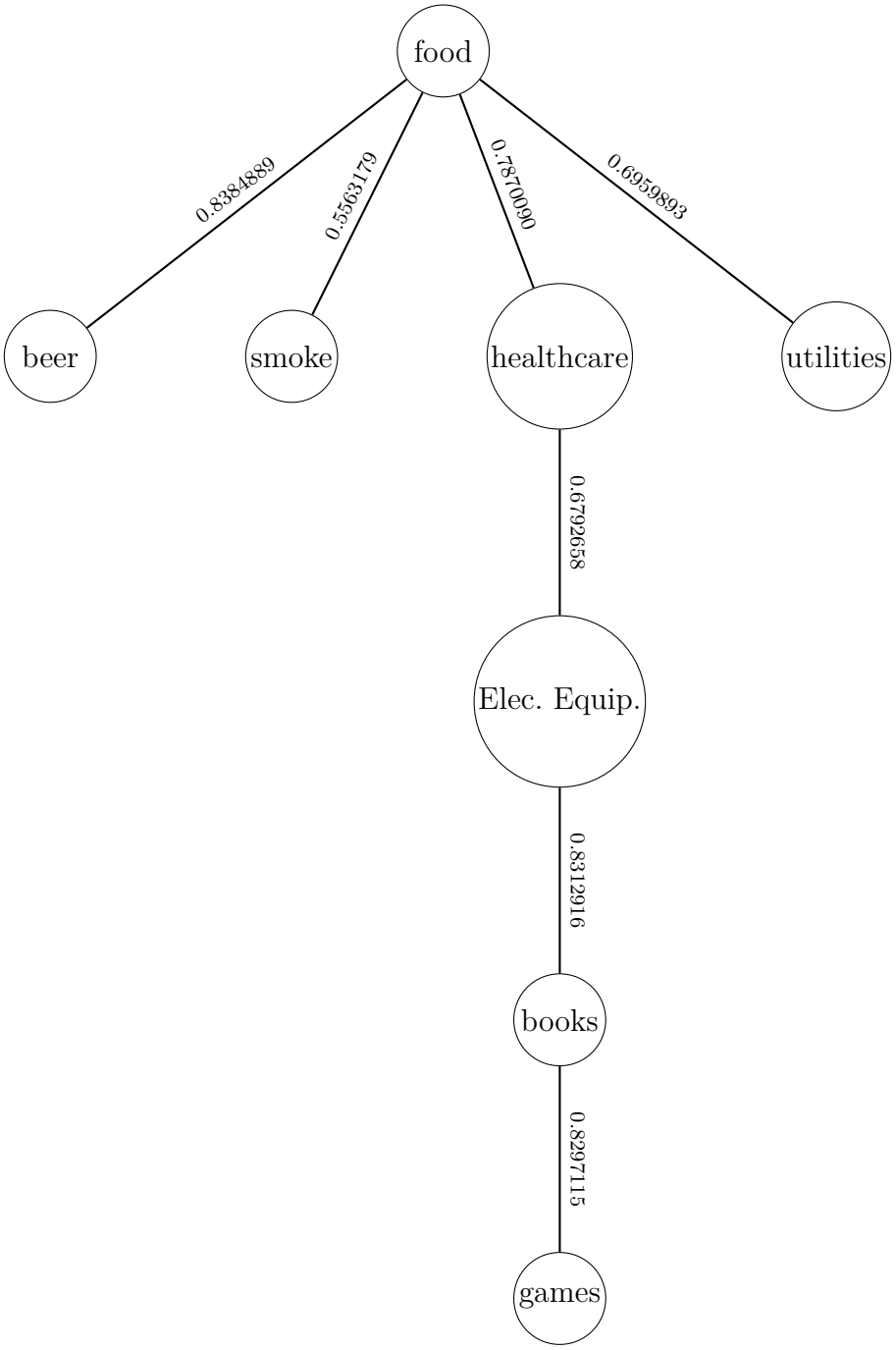


Figure 4.14
Estimated forest $\hat{\mathcal{F}}$ from Algorithm 6 on the portfolios dataset of $d = 8$ sectors (see Section 4.7.2). For each pair (s, t), we show the mixture coefficient $\hat{a}_{st}$ estimated in Step 3 of Algorithm 6. Since $0 < \hat{a}_{st} < 1$ for every edge (s, t), then any subset $J \subset \{1, \ldots, d\}$ of sectors inducing a connected sub-forest $\hat{\mathcal{F}}_J$ constitutes an extreme direction.

4.B Hüsler–Reiss and mixture Hüsler–Reiss models

4.B.1 Hüsler–Reiss model

Hüsler–Reiss parametric models received increasing interest in recent literature [34, 53]. In the following, we briefly give some background about this model. The well-known family of Hüsler–Reiss exponent measure densities is parametrized by symmetric conditionally negative definite matrices, denoted by Γ in this paper and called variogram matrices:

$$\mathcal{V}_d = \left\{\Gamma \in \mathcal{M}_{d,d}(\mathbb{R}^+) : \Gamma = \Gamma^\top, \operatorname{diag}(\Gamma) = 0, \boldsymbol{v}^\top \Gamma \boldsymbol{v} < 0, \forall\, \mathbf{0}_d \neq \boldsymbol{v} \perp \mathbf{1}\right\}. \quad (4.B.1)$$

The Hüsler–Reiss model is closely related to the Gaussian model. Notably, given a variogram matrix $\Gamma \in \mathcal{V}_d$ and $j \in V$, the matrix defined in Engelke and Hitz [34] by

$$\tilde{\varphi}^{(j)}(\Gamma) := \tilde{\Sigma}^{(j)} := \tfrac{1}{2}\left(\Gamma_{sj} + \Gamma_{tj} - \Gamma_{st}\right)_{s,t\in V} \in \mathcal{M}_{d,d}(\mathbb{R}),$$

is a covariance matrix, i.e., symmetric and positive semi-definite. By construction, the j-th row and column of $\tilde{\Sigma}^{(j)}$ are identically zero. Let $\varphi^{(j)}(\Gamma) \in \mathcal{M}_{d-1,d-1}(\mathbb{R})$ denote the matrix with these entries removed.

Definition 4.B.1. Let $\Gamma \in \mathcal{V}_d$ be a variogram matrix, $j \in V$, and $\Sigma^{(j)} = \varphi^{(j)}(\Gamma)$. The d-variate *Hüsler–Reiss exponent measure density* with variogram matrix Γ is given by

$$\lambda(\boldsymbol{y}, \Gamma) = y_j^{-2} \prod_{i\neq j} y_i^{-1} \phi_{d-1}\left(\tilde{\boldsymbol{y}}_{V\setminus\{j\}}; \Sigma^{(j)}\right), \qquad \boldsymbol{y} \in \mathbb{E}_V^*, \quad (4.B.2)$$

where $\phi_p(\,\cdot\,;\Sigma)$ is the density of the centered p-dimensional normal distribution with covariance Σ and $\tilde{\boldsymbol{y}} = \left\{\log(y_i/y_j) + \Gamma_{ij}/2\right\}_{i=1,\dots,d}$.

The representation of the density in (4.B.2) appears to depend on the choice of j, but in reality, the value of the right-hand side of this equation is independent of j [37]. Finally, note that $\lambda(\boldsymbol{y}, \Gamma) > 0$ for every $\boldsymbol{y} \in \mathbb{E}_V^*$. This remark will be useful later in the paper, for instance to verify condition (4.3.2).

4.B.2 Mixture Hüsler–Reiss model

Recall the mixture epxonent measure density λ in (4.2.9). Setting the density $\lambda^{(k)}$ to a Hüsler–Reiss exponent measure density as in (4.B.2) with some variogram matrix $\Gamma^{(k)} \in \mathcal{M}_{|J_k|,|J_k|}(\mathbb{R}^+)$, for every $k \in R$, yields what we call a *mixture Hüsler–Reiss exponent measure density*. In the following sections, we will consider a simplified case where all matrices $\Gamma^{(k)}$ are derived from the same $d \times d$ variogram

matrix. Specifically, let A be a matrix as in Definition 4.2.2 and let $\Gamma \in \mathcal{V}_d$ be a variogram matrix as defined in (4.B.1) (See the last paragraph of Section 4.B.2 for the case where $\Gamma \in \mathcal{M}_{d,d}(\overline{\mathbb{R}}^+)$). By construction, $\Gamma_J \in \mathcal{V}_{|J|}$, for every non-empty set $J \subset V$. Therefore, for every $k \in R$, set $\Gamma^{(k)} = \Gamma_{J_k}$ and define the *mixture Hüsler–Reiss exponent measure density* with variogram matrix Γ and coefficient matrix A, given by

$$\lambda_V(\boldsymbol{y}) = \sum_{k \in R} \mathbb{1}\left\{\boldsymbol{y} \in \mathbb{A}_{V,J_k}\right\} \lambda\left(\left(y_j/a_{jk}\right)_{j \in J_k}; \Gamma_{J_k}\right) \prod_{j \in J_k} (1/a_{jk}). \tag{4.B.3}$$

This assumption ensures that any two fixed components from different factors, $\boldsymbol{Z}^{(k)}$ and $\boldsymbol{Z}^{(k')}$, share the same variogram coefficient. This property will be particularly useful in Section 4.4, where it helps satisfy certain marginalization constraints. Moreover, for every non-empty subset $I \subset V$, the marginal density $\lambda_{V \to I}$, as defined in (4.2.10), remains a mixture Hüsler–Reiss exponent measure density with coefficient matrix $A_{I,K_{A,I}}$, with $K_{A,I}$ as in (4.2.11) and variogram matrix Γ_I.

Throughout this paper, we have introduced the concept of a generalized variogram matrix $\Gamma \in \mathcal{M}_{d,d}(\overline{\mathbb{R}}^+)$ (Definition 4.4.1). Such a matrix induces lower-dimensional variogram matrices in the sense of (4.B.1). Define $\mathcal{J}_\Gamma = \{J \subset V : \Gamma_J \in \mathcal{V}_{|J|}\}$. If every signature J_k of the coefficient matrix A lies in $\mathcal{J}_\Gamma$, we say that A and Γ are *compatible*. Under this compatibility assumption, the density λ_V from (4.B.3) remains well-defined. We then refer to λ_V as the *mixture Hüsler–Reiss exponent measure density* with generalized variogram matrix Γ and coefficient matrix A.

4.B.3 Bivariate mixture Hüsler–Reiss stable tail dependence function

The stable tail dependence function (stdf) associated to a d-variate max-stable random vector $\boldsymbol{M}$ with unit Fréchet margins and exponent measure Λ_V is given by

$$\ell(\boldsymbol{x}) = \Lambda_V\left(\mathbb{E}_V \setminus [\mathbf{0}_d, \mathbf{1}/\boldsymbol{x}]\right), \qquad \boldsymbol{x} \in [\mathbf{0}_d, \boldsymbol{\infty}). \tag{4.B.4}$$

See Beirlant et al. [4], Coles et al. [16] for more details about the stdf. Mourahib et al. [75, Section 4] express the stdf when Λ_V is a mixture Hüsler–Reiss exponent measure. In the bivariate case, Λ_V has two parameters: the mixture coefficient $0 \leqslant a \leqslant 1$ and the variogram coefficient $\Gamma > 0$. The bivariate mixture Hüsler–Reiss stdf is then given for $\boldsymbol{x} = (x_1, x_2) \geqslant \mathbf{0}_2$ by

$$\begin{aligned}\ell(\boldsymbol{x}\,;\, a, \Gamma) = {} & a x_1 \Phi\left(\frac{\log(x_1/x_2)}{\sqrt{\Gamma}} + \sqrt{\Gamma}/2\right) + a x_2 \Phi\left(\frac{\log(x_2/x_1)}{\sqrt{\Gamma}} + \sqrt{\Gamma}/2\right) \\ & + (1-a)(x_1 + x_2),\end{aligned} \tag{4.B.5}$$

where Φ is the standard Gaussian cumulative distribution function. Equation (4.B.5) will be useful for the estimation procedure in Section 4.5.

4.C Proof of Theorem 4.3.1

The proof of Theorem 4.3.1 is lengthy and relies on several auxiliary results. For simplification, we split the argument into three main parts:

1. prove that the associated multivariate Pareto vector $\boldsymbol{Y}$ satisfies the global Markov property with respect to the graph $\mathcal{G}$, that is (P1) and (P2);
2. characterize the extreme directions $\mathcal{J}$ of $\boldsymbol{Y}$ as stated in Theorem 4.3.1;
3. verify that λ_V defines a valid exponent-measure density.

The lemmas and propositions needed for Part 1 appear in Section 4.C.1, those for Part 2 in Section 4.C.2, and the argument for Part 3 is carried out in Section 4.C.3. In the following, we will use the same notation as in Section 4.3.

4.C.1 Markov property

Factorization with respect to $\mathcal{G}$

Recall the partition $V = V_1 \dot\cup \ldots \dot\cup V_p$ inducing the connected components of $\mathcal{G}$. First note that the zero set in (4.2.15) can be written as

$$\mathcal{Z}(\mathcal{G}) = \left\{\boldsymbol{y} \in \mathbb{E}_V : \boldsymbol{y}_{V_{i_1}} \neq \mathbf{0}_{V_{i_1}}, \boldsymbol{y}_{V_{i_2}} \neq \mathbf{0}_{V_{i_2}}, i_1 \neq i_2\right\} \cup \bigcup_{i=1}^{p} \left\{\boldsymbol{y} \in \mathbb{E}_V : \boldsymbol{y}_{V_i} \in \mathcal{Z}(\mathcal{G}_{V_i})\right\}. \quad (4.C.1)$$

By construction of λ_V in (4.3.5), it follows that $\lambda_V(\boldsymbol{y}) = 0$, for every $\boldsymbol{y} \in \mathcal{Z}(\mathcal{G})$.

Next, let $\boldsymbol{y} \in \mathbb{E}_V \setminus \mathcal{Z}(\mathcal{G})$. Equation (4.C.1) shows that there exists a unique $i \leqslant p$ such that $\boldsymbol{y}_{V_i} \neq \mathbf{0}_{V_i}$. If $\boldsymbol{y}_{V_i} \in \mathcal{S}(\mathcal{G}_{V_i}, \lambda)$, then (4.3.3) is straightforward. Otherwise, if $\boldsymbol{y}_{V_i} \notin \mathcal{S}(\mathcal{G}_{V_i}, \lambda)$, then there exists a separator $D \subset V_i$ such that $\boldsymbol{y}_D \neq \mathbf{0}_D$ and $\lambda_D(\boldsymbol{y}_D) = 0$. Consider a clique $C \subset V_i$ that contains D. The positivity condition in (4.3.2) and the clique-to-separator property (4.3.1) show that $\lambda_C(\boldsymbol{y}_C) = 0$. Therefore, both the left- and right-hand sides of (4.3.3) are equal to zero.

Whole-to-clique constraints

For a fixed clique $C^* \in \mathcal{C}$, our goal is to prove the constraint (P2). The idea of the proof is as follows: at each step, we start by selecting a clique $\tilde{C} \neq C^*$ that includes nodes (at least one) not present in any other clique. We then perform integration with respect to these nodes.

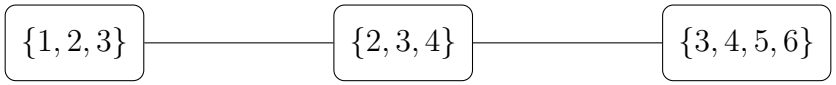
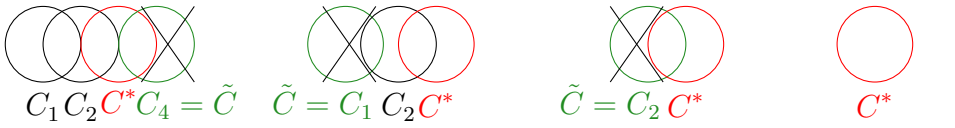


Figure 4.15
A junction tree associated to the 6-dimensional graph in Figure 4.9 (left).



Given the form of λ_V in (4.3.5), it is sufficient to show Condition (P2) for each connected component of the graph $\mathcal{G}$. Thus, we suppose that the graph $\mathcal{G}$ is connected. In the following, suppose that $\mathcal{G}$ is not complete and let $m \geqslant 2$ be its number of cliques (otherwise, there is nothing to prove in (P2)). Let $C^* \in \mathcal{C}$ be a fixed clique for which we aim to prove the marginalization condition in (P2). Let $\mathcal{T}$ be a fixed junction tree of $\mathcal{G}$. Nodes of $\mathcal{T}$ will be referred to as node-cliques. From basic graph theory, we know that every non-trivial tree has at least two nodes of degree 1. In fact, among all paths S of a tree, we can choose a path P with a maximum number of edges. The initial and the final nodes of the path P must be nodes of degree 1. It yields that the junction tree $\mathcal{T}$ admits at least two node-cliques of degree 1; see also Cowell et al. [21, Corollary 4.7].

Definition 4.C.1 (Exterior clique, history, separator, residual)**.** An exterior clique $\tilde{C}$ of $\mathcal{G}$ with respect to $\mathcal{T}$ is a node-clique of order 1 in $\mathcal{T}$. For such a clique, we define the history $\mathrm{H}(\tilde{C})$, the separator $\mathrm{S}(\tilde{C})$, and the residual $\mathrm{R}(\tilde{C})$ of $\tilde{C}$ as

$$\mathrm{H}(\tilde{C}) = \bigcup_{C \neq \tilde{C}} C, \qquad \mathrm{S}(\tilde{C}) = \tilde{C} \cap \mathrm{H}(\tilde{C}), \qquad \mathrm{R}(\tilde{C}) = \tilde{C} \setminus \mathrm{S}(\tilde{C}). \tag{4.C.2}$$

Example 4.C.2. Consider the 6-dimensional decomposable graph illustrated in Figure 4.9 (left). A possible junction tree $\mathcal{T}$ is given in Figure 4.15.

The clique $\tilde{C} = \{1, 2, 3\}$ is an exterior clique of $\mathcal{G}$ with respect to $\mathcal{T}$. We have

$$\mathrm{H}(\tilde{C}) = \{2, \ldots, 6\}\,, \qquad \mathrm{S}(\tilde{C}) = \{2, 3\}\,, \qquad \mathrm{R}(\tilde{C}) = \{1\}\,.$$

From now on, let $\tilde{C} \neq C^*$ be a fixed exterior clique of $\mathcal{G}$ with respect to $\mathcal{T}$. For simplicity, since the graph $\mathcal{G}$ and the junction tree $\mathcal{T}$ are fixed, we will call $\tilde{C}$ simply an exterior clique.

Lemma 4.C.3. *We have* $\mathrm{S}(\tilde{C}) \in \mathcal{D}$ *and* $\mathrm{R}(\tilde{C}) \neq \varnothing$*. Moreover, the subgraph* $\mathcal{G}_{V_{m-1}}$*, where* $V_{m-1} = \mathrm{H}(\tilde{C})$*, is a decomposable connected graph with cliques* $\mathcal{C}_{m-1} = \mathcal{C} \setminus \{\tilde{C}\}$ *and separators* $\mathcal{D}_{m-1} = \mathcal{D} \setminus \{\mathrm{S}(\tilde{C})\}$*, where the latter notation should be understood as the removal of one instance of* $\mathrm{S}(\tilde{C})$ *from the multiset* $\mathcal{D}$*.*

Proof of Lemma 4.C.3. The path intersection in (4.A.1) shows that $\mathrm{S}(\tilde{C}) = \tilde{C} \cap \mathrm{N}(\tilde{C})$, where $\mathrm{N}(\tilde{C})$ is the unique adjacent clique of $\tilde{C}$ in $\mathcal{T}$. The separator property (4.A.5) shows that $\mathrm{S}(\tilde{C}) \in \mathcal{D}$. Moreover, $\mathrm{R}(\tilde{C}) \neq \varnothing$, otherwise $\tilde{C} \subset \mathrm{N}(\tilde{C})$.

Consider now the induced subgraph $\mathcal{G}_{V_{m-1}}$, where $V_{m-1} = \mathrm{H}(\tilde{C})$. Let $C \in \mathcal{C}_{m-1}$ and a, b from C. On the one hand, $(a, b) \in E_{V_{m-1}} := E \cap (V_{m-1} \times V_{m-1})$, implying that C is complete in $\mathcal{G}_{V_{m-1}}$. On the other hand, $E_{V_{m-1}} \subset E$ and thus C is maximal complete in $\mathcal{G}_{V_{m-1}}$, that is, a clique in $\mathcal{G}_{V_{m-1}}$. By definition, the node-clique $\tilde{C}$ is of degree 1 in the junction tree $\mathcal{T}$, it follows that the subtree $\mathcal{T}_{\mathcal{C}_{m-1}}$ with the clique $\tilde{C}$ removed is still a tree of cliques $\mathcal{C}_{m-1}$ for $\mathcal{G}_{V_{m-1}}$. The path intersection property clearly holds for $\mathcal{T}_{\mathcal{C}_{m-1}}$, so that $\mathcal{T}_{\mathcal{C}_{m-1}}$ is a junction tree of $\mathcal{G}_{V_{m-1}}$. Lauritzen [68, Proposition 2.27] shows that $\mathcal{G}_{V_{m-1}}$ is decomposable with cliques $\mathcal{C}_{m-1}$. Finally, the separator property in (4.A.5) shows that the set of separators of $\mathcal{G}_{V_{m-1}}$ is $\mathcal{D}_{m-1}$, which implies that $\mathcal{G}_{V_{m-1}}$ is still connected since $\varnothing \notin \mathcal{D}_{m-1}$. □

Lemma 4.C.4. *The set* $\mathrm{S}(\tilde{C})$ *separates* $\mathrm{R}(\tilde{C})$ *and* $\mathrm{H}(\tilde{C}) \setminus \mathrm{S}(\tilde{C})$ *in* $\mathcal{G}$.

Lauritzen [68, Lemma 2.11] shows a similar result for the last clique C_m in an order $\{C_1, \ldots, C_m\}$ of $\mathcal{C}$ that satisfies the running intersection property in (4.A.2).

Proof of Lemma 4.C.4. Suppose there exists a path between $\mathrm{R}(\tilde{C})$ and $\mathrm{H}(\tilde{C}) \setminus \mathrm{S}(\tilde{C})$ that avoids $\mathrm{S}(\tilde{C})$. Then, there exists necessarily an edge (a, b) between $a \in \mathrm{R}(\tilde{C})$ and $b \in C \setminus \mathrm{S}(\tilde{C})$ for some $C \neq \tilde{C}$. Then (a, b) must be an edge from some clique of $\mathcal{G}$. This cannot be $\tilde{C}$, since $b \notin \tilde{C}$ and it cannot be C for some $C \neq \tilde{C}$, since $a \notin C$. Such an edge can therefore not exist and thus $\mathrm{S}(\tilde{C})$ separates $\mathrm{R}(\tilde{C})$ and $\mathrm{H}(\tilde{C}) \setminus \mathrm{S}(\tilde{C})$. □

Lemma 4.C.5. *Let* $a, b \in V$. *If* $S \subset V$ *separates* a *and* b, *then there exists a separator* $D \in \mathcal{D}$ *such that* $D \subset S$.

Proof. Let C_a and C_b be cliques of $\mathcal{G}$ containing a and b, respectively. Note that $C_a \neq C_b$. Recall $\mathrm{path}(C_a, C_b)$ the path between C_a and C_b in the junction tree $\mathcal{T}$ as in (4.A.7) and the intermediate separators $\mathcal{D}_{C_a C_b}$ as in (4.A.8). Suppose that $D \setminus S \neq \varnothing$ for every $D \in \mathcal{D}_{C_a C_b}$ (Note that here $\mathcal{G}$ is connected and thus every $D \in \mathcal{D}$ is non-empty). Construct the path between a and b in $\mathcal{G}$ starting from a in C_a and going to the next clique in $\mathrm{path}(C_a, C_b)$ by visiting $D \setminus S$. We have therefore constructed a path from a to b that avoids S which contradicts the separation between a and b by S. □

In the following, we define the admissible set $\mathcal{A}(\mathcal{G}) = \mathbb{E}_V \setminus \mathcal{Z}(\mathcal{G})$.

Definition 4.C.6. The admissible and safe sets with respect to the exterior clique $\tilde{C}$ are defined, respectively, by

$$\mathcal{A}_{\tilde{C}} = \left\{ \boldsymbol{y} \in [0,\infty)^{|\tilde{C}|} : \ \boldsymbol{y}_{\mathrm{S}(\tilde{C})} = \boldsymbol{0}_{\mathrm{S}(\tilde{C})} \implies \boldsymbol{y}_{\mathrm{R}(\tilde{C})} = \boldsymbol{0}_{\mathrm{R}(\tilde{C})} \right\},$$
$$\mathcal{S}_{\tilde{C}} = \left\{ \boldsymbol{y} \in [0,\infty)^{|\mathrm{S}(\tilde{C})|} : \ \boldsymbol{y}_{\mathrm{S}(\tilde{C})} \neq \boldsymbol{0}_{\mathrm{S}(\tilde{C})} \implies \lambda_{\mathrm{S}(\tilde{C})}(\boldsymbol{y}_{\mathrm{S}(\tilde{C})}) > 0 \right\}.$$

The following lemma expresses the admissible and safe sets of the subgraph induced by removing $\mathrm{R}(\tilde{C})$ in terms of the admissible and safe sets of the original graph. In the following, define $\mathbb{E}_{V,\{\boldsymbol{y}_{C^*} \neq \boldsymbol{0}_{C^*}\}} = \mathbb{E}_V \setminus \{\boldsymbol{y}_{C^*} = \boldsymbol{0}_{C^*}\}$.

Lemma 4.C.7. *We have*

$$\mathcal{A}(\mathcal{G}) \cap \mathbb{E}_{V,\{\boldsymbol{y}_{C^*} \neq \boldsymbol{0}_{C^*}\}} = \left\{ \boldsymbol{y} \in \mathbb{E}_{V,\{\boldsymbol{y}_{C^*} \neq \boldsymbol{0}_{C^*}\}} : \ \boldsymbol{y}_{\mathrm{H}(\tilde{C})} \in \mathcal{A}(\mathcal{G}_{\mathrm{H}(\tilde{C})}),\ \boldsymbol{y}_{\tilde{C}} \in \mathcal{A}_{\tilde{C}} \right\}, \tag{4.C.3}$$

$$\mathcal{S}(\mathcal{G}) \cap \mathbb{E}_{V,\{\boldsymbol{y}_{C^*} \neq \boldsymbol{0}_{C^*}\}} = \left\{ \boldsymbol{y} \in \mathbb{E}_{V,\{\boldsymbol{y}_{C^*} \neq \boldsymbol{0}_{C^*}\}} : \ \boldsymbol{y}_{\mathrm{H}(\tilde{C})} \in \mathcal{S}(\mathcal{G}_{\mathrm{H}(\tilde{C})}),\ \boldsymbol{y}_{\mathrm{S}(\tilde{C})} \in \mathcal{S}_{\tilde{C}} \right\}. \tag{4.C.4}$$

Recall that Lemma 4.C.3 shows that $\mathcal{G}_{\mathrm{H}(\tilde{C})}$ is decomposable. Let $\mathcal{D}_{m-1}$ denote the set of its separators.

Proof. Equality (4.C.4) is straightforward. Let us show Equality (4.C.3).

$\rightarrow$ Let $\boldsymbol{y}$ be in the left-hand side set of (4.C.3). Separation in Lemma 4.C.4 and $\boldsymbol{y}_{C^*} \neq \boldsymbol{0}_{C^*}$ shows that $\boldsymbol{y}_{\tilde{C}} \in \mathcal{A}_{\tilde{C}}$. Next, let a, b from $\mathrm{H}(\tilde{C})$ and let $D \in \mathcal{D}_{m-1}$ separate a and b in $\mathcal{G}_{\mathrm{H}(\tilde{C})}$.

Separation in $\mathcal{G}_{\mathrm{H}(\tilde{C})}$ implies separation in $\mathcal{G}$. Suppose D does not separate a and b in $\mathcal{G}$. There exists then a path $\tau = (n_1 = a, \ldots, n_l = b)$ that does not intersect D, and a node $n_k \in \mathrm{R}(\tilde{C})$ from τ. Lemma 4.C.4 yields that τ intersects $\mathrm{S}(\tilde{C})$ in at least two nodes. Let n_i and n_j be the first and last nodes, respectively, where τ intersects $\mathrm{S}(\tilde{C})$. The set $\mathrm{S}(\tilde{C})$ is complete and we consider the sub-path $\eta_{\setminus \mathrm{S}(\tilde{C})} = (n_1 = a, \ldots, n_i, n_j, \ldots, n_l = b)$. Therefore, $\eta_{\setminus \mathrm{S}(\tilde{C})}$ is a path in $\mathcal{G}_{\mathrm{H}(\tilde{C})}$ that does not intersect D, leading to absurdity. Therefore, we deduce that D separates a and b in $\mathcal{G}$, implying $\boldsymbol{y}_D = 0 \implies (y_a = 0 \text{ or } y_b = 0)$. It follows that $\boldsymbol{y}_{\mathrm{H}(\tilde{C})} \in \mathcal{A}(\mathcal{G}_{\mathrm{H}(\tilde{C})})$.

$\rightarrow$ Conversely, let now $\boldsymbol{y}$ be in the right-hand side set of (4.C.3). Let a, b from V and let $S \in \mathcal{D}$ separate a and b in $\mathcal{G}$. We need to show that $\boldsymbol{y}_S = \boldsymbol{0}_S$ implies $y_a = 0$ or $y_b = 0$.

1. **Case 1.** Let a, b from $\mathrm{H}(\tilde{C})$. On the one hand, we know that $\mathcal{D}_{m-1}$ consists of all separators of $\mathcal{G}_{\mathrm{H}(\tilde{C})}$. On the other hand, $S \subset \mathrm{H}(\tilde{C})$ separates a and b in

$\mathcal{G}_{\mathrm{H}(\tilde{C})}$. Lemma 4.C.5 implies that there exists $D \in \mathcal{D}_{m-1}$ such that $D \subset S$. Therefore, $\boldsymbol{y}_{\mathrm{H}(\tilde{C})} \in \mathcal{A}(\mathcal{G}_{\mathrm{H}(\tilde{C})})$ yields that $\boldsymbol{y}_S = \mathbf{0}_S \implies \boldsymbol{y}_D = \mathbf{0}_D \implies (y_a = 0 \text{ or } y_b = 0)$.

2. **Case 2.** Let $b \in \mathrm{R}(\tilde{C})$ and thus $a \in \mathrm{H}(\tilde{C}) \setminus \tilde{C}$.

   (a) **Case 2.1.** If $\mathrm{S}(\tilde{C}) \subset S$, then $\boldsymbol{y}_S = \mathbf{0}_S \implies \boldsymbol{y}_{\mathrm{S}(\tilde{C})} = \mathbf{0}_{\mathrm{S}(\tilde{C})} \implies y_b = 0$.

   (b) **Case 2.2.** If $\mathrm{S}(\tilde{C}) \setminus S \neq \varnothing$, then S separates $\mathrm{S}(\tilde{C}) \setminus S$ and $\{a\}$; otherwise, there exists a path $\tau = (n_1 = a, \ldots, n_l = e)$ for some $e \in \mathrm{S}(\tilde{C}) \setminus S$ that does not intersect S. Let (e, b) be the edge in $\tilde{C}$ (the latter is complete by definition), then $\tau_+ = (n_1 = a, \ldots, n_l = e, n_{l+1} = b)$ does not intersect S, leading to an absurdity. Thus, $\boldsymbol{y}_S = \mathbf{0}_S \implies (y_a = 0 \text{ or } \boldsymbol{y}_{\mathrm{S}(\tilde{C})} = \mathbf{0}_{\mathrm{S}(\tilde{C})}) \implies (y_a = 0 \text{ or } y_b = 0)$. □

We are now ready to prove the marginalization constraint (P2) for the clique C^*. Suppose first that $m > 1$. Let $\boldsymbol{y}_{C^*} \neq \mathbf{0}_{C^*}$. Lemma 4.C.7 shows that

$$I = \int_{[0,\infty)^{|V \setminus C^*|}} \lambda_V(\boldsymbol{y}) \, \mathrm{d}\mu_{V \setminus C^*}(\boldsymbol{y}_{V \setminus C^*}) = \int_{[0,\infty)^{|V_{m-1} \setminus C^*|}} \frac{\prod_{C \in \mathcal{C}_{m-1}} \bar{\lambda}_C(\boldsymbol{y}_C)}{\prod_{D \in \mathcal{D}_{m-1}} \bar{\lambda}_D(\boldsymbol{y}_D)} \times$$
$$\mathbb{1}\left\{\boldsymbol{y}_{V_{m-1}} \in \mathcal{A}\left(\mathcal{G}_{V_{m-1}}\right) \cap \mathcal{S}(\mathcal{G}_{V_{m-1}})\right\} \times \mathbb{1}\left\{\boldsymbol{y}_{\mathrm{S}(\tilde{C})} \in \mathcal{S}_{\tilde{C}}\right\} \times \frac{1}{\bar{\lambda}_{\mathrm{S}(\tilde{C})}(\boldsymbol{y}_{\mathrm{S}(\tilde{C}_1)})} \times$$
$$\underbrace{\int_{[0,\infty)^{|\mathrm{R}(\tilde{C})|}} \bar{\lambda}_{\tilde{C}}(\boldsymbol{y}_{\tilde{C}}) \, \mathbb{1}\{\boldsymbol{y}_{\tilde{C}} \in \mathcal{A}_{\tilde{C}}\} \, \mathrm{d}\mu_{\mathrm{R}(\tilde{C})}(\boldsymbol{y}_{\mathrm{R}(\tilde{C})})}_{I(\mathrm{R}(\tilde{C}))} \, \mathrm{d}\mu_{V_{m-1} \setminus C^*}(\boldsymbol{y}_{V_{m-1} \setminus C^*}). \quad (4.C.5)$$

Let $\boldsymbol{y}_{V_{m-1} \setminus C^*} \in [0,\infty)^{|V_{m-1} \setminus C^*|}$. First, the clique-to-separator marginalization consistency constraints in (4.3.1) show that $I(\mathrm{R}(\tilde{C})) = \bar{\lambda}_{\mathrm{S}(\tilde{C})}(\boldsymbol{y}_{\mathrm{S}(\tilde{C})})$; remark that $I(\mathrm{R}(\tilde{C})) = \bar{\lambda}_{\mathrm{S}(\tilde{C})}(\boldsymbol{y}_{\mathrm{S}(\tilde{C})}) = 1$ for the case where $\boldsymbol{y}_{\mathrm{S}(\tilde{C})} = \mathbf{0}_{\mathrm{S}(\tilde{C})}$. Next, if $\boldsymbol{y}_{\mathrm{S}(\tilde{C})} \in \mathcal{S}_{\tilde{C}}$, then $\mathbb{1}\left\{\boldsymbol{y}_{\mathrm{S}(\tilde{C})} \in \mathcal{S}_{\tilde{C}}\right\} = 1$. Otherwise, there exists a clique $C \in \mathcal{C}_{m-1}$ such that $\mathrm{S}(\tilde{C}) \subset C$ and the clique-to-separator constraint in (4.3.1) from C to $\mathrm{S}(\tilde{C})$ shows that $\bar{\lambda}_C(\boldsymbol{y}_C) = 0$. It follows that

$$I = \int_{[0,\infty)^{|V_{m-1} \setminus C^*|}} \frac{\prod_{C \in \mathcal{C}_{m-1}} \bar{\lambda}_C(\boldsymbol{y}_C)}{\prod_{D \in \mathcal{D}_{m-1}} \bar{\lambda}_D(\boldsymbol{y}_D)}$$
$$\times \mathbb{1}\left\{\boldsymbol{y}_{V_{m-1}} \in \mathcal{A}\left(\mathcal{G}_{V_{m-1}}\right) \cap \mathcal{S}(\mathcal{G}_{V_{m-1}})\right\} \mathrm{d}\mu_{V_{m-1} \setminus C^*}(\boldsymbol{y}_{V_{m-1} \setminus C^*}). \quad (4.C.6)$$

The integrand in (4.C.6) factorizes with respect to the cliques, separators, admissible and safe set of the graph $\mathcal{G}_{V_{m-1}}$. By induction, we repeat the same reasoning $m-1$

times, where each time we obtain a decomposable connected graph (Lemma 4.C.3) and choose an exterior clique different from C^*. We finally obtain $I = \lambda_{C^*}(\boldsymbol{y}_{C^*})$.

In the case where $m = 2$, the proof is slightly similar. There are only two cliques, C^* and $\tilde{C}$, and one separator, $\mathrm{S}(\tilde{C}) = C^* \cap \tilde{C}$. Since C^* is complete, we have $\mathcal{Z}(\mathcal{G}_{C^*}) = \varnothing$, so

$$\begin{aligned} I &= \int_{[0,\infty)^{|V \setminus C^*|}} \lambda_V(\boldsymbol{y}) \, \mathrm{d}\mu_{V \setminus C^*}(\boldsymbol{y}_{V \setminus C^*}) \\ &= \mathbb{1}\left\{\boldsymbol{y}_{C^*} \in \mathcal{S}(\mathcal{G}_{C^*})\right\} \times \mathbb{1}\left\{\boldsymbol{y}_{\mathrm{S}(\tilde{C})} \in \mathcal{S}_{\tilde{C}}\right\} \times \frac{\bar{\lambda}_{C^*}(\boldsymbol{y}_{C^*})}{\bar{\lambda}_{\mathrm{S}(\tilde{C})}(\boldsymbol{y}_{\mathrm{S}(\tilde{C})})} \times \\ &\quad \underbrace{\int_{[0,\infty)^{|\tilde{C} \setminus C^*|}} \bar{\lambda}_{\tilde{C}}(\boldsymbol{y}_{\tilde{C}}) \, \mathbb{1}\left\{\boldsymbol{y}_{\tilde{C}} \in \mathcal{A}_{\tilde{C}}\right\} \, \mathrm{d}\mu_{\tilde{C} \setminus C^*}(\boldsymbol{y}_{\tilde{C} \setminus C^*})}_{I(\mathrm{R}(\tilde{C}))}. \end{aligned}$$

The indicator $\mathbb{1}\left\{\boldsymbol{y}_{C^*} \in \mathcal{S}(\mathcal{G}_{C^*})\right\}$ is equal to 1 since $\mathcal{S}(\mathcal{G}_{C^*}) = \mathbb{E}_{C^*}$. Next, if $\boldsymbol{y}_{\mathrm{S}(\tilde{C})} \notin \mathcal{S}_{\tilde{C}}$, then $\boldsymbol{y}_{\mathrm{S}(\tilde{C})} \neq \mathbf{0}_{\mathrm{S}(\tilde{C})}$ and $\lambda_{\mathrm{S}(\tilde{C})}(\boldsymbol{y}_{\mathrm{S}(\tilde{C})}) = 0$. The clique-to-separator constraint in (4.3.1) from C^* to $\mathrm{S}(\tilde{C})$ shows that $\lambda_{C^*}(\boldsymbol{y}_{C^*}) = 0$. It follows that the indicator $\mathbb{1}\left\{\boldsymbol{y}_{\mathrm{S}(\tilde{C})} \in \mathcal{S}_{\tilde{C}}\right\}$ does not matter. Finally, similarly to above, the clique-to-separator constraint from $\tilde{C}$ to $\mathrm{S}(\tilde{C})$ yields that $I(\mathrm{R}(\tilde{C}))$ is equal to $\bar{\lambda}_{\mathrm{S}(\tilde{C})}(\boldsymbol{y}_{\mathrm{S}(\tilde{C})})$ where we note that $I(\mathrm{R}(\tilde{C})) = \bar{\lambda}_{\mathrm{S}(\tilde{C})}(\boldsymbol{y}_{\mathrm{S}(\tilde{C})}) = 1$ for the case where $\boldsymbol{y}_{\mathrm{S}(\tilde{C})} = \mathbf{0}_{\mathrm{S}(\tilde{C})}$.

4.C.2 Extreme directions

In the following, let $\mathcal{J}_{\mathrm{adm}}$ be the collection of non-empty sets $J \subset V$ satisfying (1) and (2) from Theorem 4.3.1. Recall that Λ_V is the exponent measure associated to λ_V and $\mathcal{J}(\Lambda_V)$ is the set of extreme directions of Λ_V. In this section, we suppose that λ_V satisfies (P1) and (P2). For a non-empty set $J \subset V$, recall that $\mathcal{C}_J$ and $\mathcal{D}_J$ denote the cliques and separators, respectively, intersecting J.

Lemma 4.C.8. *If $J \in \mathcal{J}(\Lambda_V)$, then $J \cap I \in \mathcal{J}(\Lambda_I)$ for every $I \in \mathcal{C}_J \cup \mathcal{D}_J$.*

Proof of Lemma 4.C.8. Suppose there exists $I \in \mathcal{C}_J \cup \mathcal{D}_J$ such that $I \cap J \notin \mathcal{J}(\Lambda_I)$. If $\boldsymbol{y} \in \mathbb{A}_{V,J}$, then $\boldsymbol{y}_I \in \mathbb{A}_{I,I \cap J}$. It follows

$$\begin{aligned} \Lambda_V\left(\mathbb{A}_{V,J}\right) &= \int_{\mathbb{E}_V} \lambda_V(\boldsymbol{y}) \, \mathbb{1}\left\{\boldsymbol{y} \in \mathbb{A}_{V,J}\right\} \, \mathrm{d}\mu_V(\boldsymbol{y}) \\ &\leqslant \int_{\mathbb{E}_I} \mathbb{1}\left\{\boldsymbol{y}_I \in \mathbb{A}_{I,I \cap J}\right\} \cdot \left(\int_{\mathbb{E}_{V \setminus I}} \lambda_V\left(\boldsymbol{y}\right) \, \mathrm{d}\mu_{V \setminus I}(\boldsymbol{y}_{V \setminus I}) \right) \, \mathrm{d}\mu_I(\boldsymbol{y}_I) \\ &= \int_{\mathbb{E}_I} \mathbb{1}\left\{\boldsymbol{y}_I \in \mathbb{A}_{I,I \cap J}\right\} \lambda_I(\boldsymbol{y}_I) \, \mathrm{d}\mu_I(\boldsymbol{y}_I). \end{aligned} \tag{4.C.7}$$

Since $I \cap J \notin \mathcal{J}(\Lambda_I)$, the integrand in (4.C.7) is equal to zero for μ_I-almost all $\boldsymbol{y}_I \in \mathbb{E}_I^*$. This contradicts $J \in \mathcal{J}(\Lambda_V)$. □

Lemma 4.C.9. *We have* $\mathcal{J}(\Lambda_V) \subset \mathcal{J}_{\mathrm{adm}}$.

Proof of Lemma 4.C.9. Let $J \in \mathcal{J}(\Lambda_V)$. We verify conditions (1) and (2) in Theorem 4.3.1, and thus $J \in \mathcal{J}_{\mathrm{adm}}$ by definition.

(1) Suppose that $\mathcal{G}_J$ is disconnected and let $\boldsymbol{y} \in \mathbb{A}_{V,J}$. Note that in this case, $\boldsymbol{y} \in \mathcal{Z}(\mathcal{G})$ and thus, $\lambda_V(\boldsymbol{y}) = 0$, by construction of λ_V. It follows that $\Lambda_V(\mathbb{A}_{V,J}) = 0$ which contradicts $J \in \mathcal{J}(\Lambda_V)$.

(2) Lemma 4.C.8 shows (2) of Theorem 4.3.1. □

Lemma 4.C.10. *We have* $\mathcal{J}_{\mathrm{adm}} \subset \mathcal{J}$.

Proof. Recall the partition $V = V_1 \dot\cup \cdots \dot\cup V_p$ inducing the connected components of $\mathcal{G}$. Let $J \in \mathcal{J}_{\mathrm{adm}}$. Since $\mathcal{G}_J$ is connected, then there exists $i(J) \leqslant p$ such that $J \subset V_J := V_{i(J)}$. Thus,

$$\begin{aligned}\Lambda_V(\mathcal{A}_{V,J}) &= \int_{\boldsymbol{y}\in\mathbb{A}_{V,J}} \lambda_V(\boldsymbol{y})\ \mathrm{d}\mu_V(\boldsymbol{y}) \\ &= \int_{\boldsymbol{y}\in\mathbb{A}_{V,J}} \frac{\prod_{C\in\mathcal{C}_J} \lambda_C(\boldsymbol{y}_{C\cap J}, \mathbf{0}_{C\setminus J})}{\prod_{D\in\mathcal{D}_J} \lambda_D(\boldsymbol{y}_{D\cap J}, \mathbf{0}_{D\setminus J})}\, \mathbb{1}\left\{\boldsymbol{y}_{V_J} \in \mathcal{S}(\mathcal{G}_{V_J}, \lambda) \setminus \mathcal{Z}(\mathcal{G}_{V_J})\right\}\ \mathrm{d}\mu_V(\boldsymbol{y}_V).\end{aligned} \tag{4.C.8}$$

Ley $\boldsymbol{y} \in \mathbb{A}_{V,J}$. First, since $\mathcal{G}_J$ is connected, then $\boldsymbol{y} \notin \mathcal{Z}(\mathcal{G}_{V_J})$. Next, suppose $\boldsymbol{y} \notin \mathcal{S}(\mathcal{G}_{V_J}, \lambda)$. It follows that there exists $D \in \mathcal{D}_J$ such that $\lambda_D(\boldsymbol{y}_D) = 0$. Hence the clique-to-separator constraint (4.3.1) and the positivity condition (4.3.2) show that there exists $C \in \mathcal{C}_J$ such that $C \cap J \notin \mathcal{J}(\Lambda_C)$ which contradicts J being admissible. Finally, the integrand in (4.C.8) is strictly positive and thus $J \in \mathcal{J}(\Lambda_V)$. □

4.C.3 Homogeneity

By definition, the density λ_V is a valid exponent measure density with respect to μ_V if it satisfies the two following conditions:

- normalized margins, meaning that

$$\int_{y_j>1} \lambda_V(\boldsymbol{y})\ \mathrm{d}\mu_V(\boldsymbol{y}) = 1, \qquad j \in V; \tag{4.C.9}$$

- The measure Λ_V associated to λ_V with respect to μ_V is (-1)-homogeneous, meaning $\Lambda_V(rB) = r^{-1}\Lambda_V(B)$, for every $B \in \mathcal{B}(\mathbb{E}_V)$.

In this section, we suppose that λ_V satisfies (P1) and (P2).

Lemma 4.C.11. *The density* λ_V *satisfies Equation* (4.C.9).

Proof. Let $j \in V$ and let $C \in \mathcal{C}$ such that $j \in C$. Marginalization constraint (P2) for the clique C yields

$$\begin{aligned}\int_{y_j>1} \lambda_V(\boldsymbol{y})\ \mathrm{d}\mu_V(\boldsymbol{y}) &= \int_{\{\boldsymbol{y}_C \in \mathbb{E}_C : y_j>1\}} \int_{\boldsymbol{y}_{V\setminus C} \in \mathbb{E}_{V\setminus C}} \lambda_V(\boldsymbol{y})\ \mathrm{d}\mu_{V\setminus C}(\boldsymbol{y}_{V\setminus C})\ \mathrm{d}\mu_C(\boldsymbol{y}_C) \\ &= \int_{y_j>1} \lambda_C(\boldsymbol{y}_C)\ \mathrm{d}\mu_C(\boldsymbol{y}_C). \end{aligned} \tag{4.C.10}$$

Finally, λ_C is the exponent measure density of the random vector $\mathbf{M}$ (Definition 4.2.2) with respect to the measure μ_C, and by construction $\mathbf{M}$ has unit Fréchet margins. It follows that (4.C.10) is equal to 1. □

Lemma 4.C.12. *If the density h_J in* (4.3.7) *is $-(|J|+1)$-homogeneous for every $J \in \mathcal{J}(\Lambda_V)$, then Λ_V is -1 homogeneous.*

Proof. Let $r > 0$ and $B \in \mathcal{B}(\mathbb{E}_V^*)$. The change of variable $\boldsymbol{v} = \boldsymbol{y}/r$ on the appropriate dimensions, and the $-(|J|+1)$-homogeneity of h_J for each $J \in \mathcal{J}(\Lambda_V)$ yield that

$$\begin{aligned}&\Lambda_V(rB) = \sum_{J\in\mathcal{J}} \Lambda_V(rB \cap \mathbb{A}_{V,J}) = \sum_{J\in\mathcal{J}} \Lambda_{V,J}(\pi_J(rB \cap \mathbb{A}_{V,J})) \\ &= \sum_{J\in\mathcal{J}} \int_{\boldsymbol{y}\in r\pi_J(B\cap\mathbb{A}_{V,J})} h_J(\boldsymbol{y})\ \mathrm{d}\boldsymbol{y} = \sum_{J\in\mathcal{J}} \int_{\boldsymbol{v}\in\pi_J(B\cap\mathbb{A}_{V,J})} h_J(r\boldsymbol{v})\, r^{|J|}\ \mathrm{d}\boldsymbol{v} \\ &= r^{-1} \sum_{J\in\mathcal{J}} \int_{\boldsymbol{v}\in\pi(B\cap\mathbb{A}_{V,J})} h_J(\boldsymbol{v})\ \mathrm{d}\boldsymbol{v} = r^{-1}\Lambda_V(B). \end{aligned}$$ □

Lemma 4.C.13. *The density h_J in* (4.3.7) *is $-(|J|+1)$-homogeneous for every $J \in \mathcal{J}(\Lambda_V)$.*

Proof. Let $J \in \mathcal{J}$ and $I \in \mathcal{C}_J \cup \mathcal{D}_J$. On the one hand, Lemma 4.C.8 shows that $I \cap J$ is an extreme direction of Λ_I. On the other hand λ_I is a mixture exponent measure density and has the form (4.2.9). It follows that $\lambda_{I\cap J} : \mathbb{E}_{I\cap J} \to \mathbb{R}^+,\ \boldsymbol{z} \mapsto \lambda_I\left(\boldsymbol{z}, \mathbf{0}_{I\setminus J}\right)$ is homogeneous of order $-(|I\cap J|+1)$. We get, $\lambda_I(r\boldsymbol{y}_{I\cap J}, \mathbf{0}_{I\setminus J}) = r^{-(|I\cap J|+1)}\lambda_I(\boldsymbol{y}_{I\cap J}, \mathbf{0}_{I\setminus J})$, for every $\boldsymbol{y} \in \mathbb{A}_{I,I\cap J}$ and thus Lemma 4.A.5 shows

$$h_J(r\boldsymbol{y}_J) = \frac{\prod_{C\in\mathcal{C}_J} r^{-(|C\cap J|+1)}}{\prod_{D\in\mathcal{D}_J} r^{-(|D\cap J|+1)}} h_J(\boldsymbol{y}_J) = r^{-(|J|+1)} h_J(\boldsymbol{y}_J), \qquad r > 0. \quad \square$$

4.D Proof of Proposition 4.3.3

Proof of Proposition 4.3.3. Recall the partition $V = V_1 \dot\cup \cdots \dot\cup V_p$ inducing the connected components of $\mathcal{G}$. Let $J \in \mathcal{J}(\Lambda_V)$ and consider $B \in \mathcal{B}(\mathbb{E}_J)$. Lemma 4.C.9

shows that there exists $i(J) \leqslant p$ such that $J \subset V_J := V_{i(J)}$. Hence, we obtain $\left\{j \in V : \, y_j > 0\right\} \subset V_J$ for every $\boldsymbol{y} \in \mathbb{A}_{V,J}$. We get

$$\begin{aligned}\Lambda_{V,J}(B) &= \int_{\boldsymbol{y} \in \pi_J^{-1}(B) \cap \mathbb{A}_{V,J}} \lambda_V(\boldsymbol{y}) \, \mathrm{d}\mu_V(\boldsymbol{y}) \\ &= \int_{\boldsymbol{y} \in \pi_J^{-1}(B) \cap \mathbb{A}_{V,J}} \frac{\prod_{C \in \mathcal{C}_J} \bar{\lambda}_C(\boldsymbol{y}_C)}{\prod_{D \in \mathcal{D}_J} \bar{\lambda}_D(\boldsymbol{y}_D)} \, \mathbb{1}\left\{\boldsymbol{y}_{V_J} \in \mathcal{S}(\mathcal{G}_{V_J}, \lambda) \setminus \mathcal{Z}(\mathcal{G}_{V_J})\right\} \, \mathrm{d}\mu_V(\boldsymbol{y}_V).\end{aligned} \tag{4.D.1}$$

Let $\boldsymbol{y} \in \mathbb{A}_{V,J}$ and $D \in \mathcal{D}_J$ such that $\boldsymbol{y}_D \neq \mathbf{0}_D$. On the one hand, $\boldsymbol{y} \in \mathbb{A}_{V,J}$ yields that $\boldsymbol{y}_D \in \mathbb{A}_{D,J \cap D}$. On the other hand, Lemma 4.C.8 shows that $D \cap J \in \mathcal{J}(\Lambda_D)$. Condition (4.3.2) shows that $\lambda_D(\boldsymbol{y}_D) > 0$. Hence $\boldsymbol{y}_{V_J} \in \mathcal{S}(\mathcal{G}_{V_J}, \lambda)$. Next, note that $\boldsymbol{y} \in \mathbb{A}_{V,J}$ and $\mathcal{G}_J$ is connected. It follows that $\boldsymbol{y} \notin \mathcal{Z}(\mathcal{G}_{V_J})$. Hence, the indicator function on the integrand in (4.D.1) can be removed.

Consider the class $\mathcal{A}$ of product-form sets in $\mathbb{E}_V$, that is

$$\mathcal{A} = \left\{ B = \bigtimes_{j \in J} B_j : \; B_j \in \mathcal{B}\left(\mathbb{R}^+\right) \right\}.$$

Let $B \in \mathcal{A}$ and suppose $B_j \neq \{0\}$ for every $j \in J$ (else $\pi_J^{-1}(B) \cap \mathbb{A}_{V,J} = \varnothing$ and there is nothing to prove). We get $\pi_J^{-1}(B) \cap \mathbb{A}_{V,J} = \times_{j \in J} B_j^* \times \left\{\mathbf{0}_{V \setminus J}\right\}$. It follows that

$$\begin{aligned}&\int_{\boldsymbol{y} \in \pi_J^{-1}(B) \cap \mathbb{A}_{V,J}} \lambda_V(\boldsymbol{y}) \, \mathrm{d}\mu_V(\boldsymbol{y}) \\ &= \int_{\boldsymbol{y}_J \in \times_{j \in J} B_j^*} \int_{\boldsymbol{y}_{V \setminus J} = \mathbf{0}_{V \setminus J}} \frac{\prod_{C \in \mathcal{C}_J} \bar{\lambda}_C\left(\boldsymbol{y}_C\right)}{\prod_{D \in \mathcal{D}_J} \bar{\lambda}_D\left(\boldsymbol{y}_D\right)} \, \mathrm{d}\mu_{V \setminus J}(\boldsymbol{y}_{V \setminus J}) \, \mathrm{d}\mu_J(\boldsymbol{y}_J) \\ &= \int_{\boldsymbol{y}_J \in \times_{j \in J} B_j} \frac{\prod_{C \in \mathcal{C}_J} \lambda_C(\boldsymbol{y}_{C \cap J}, \mathbf{0}_{C \setminus (C \cap J)})}{\prod_{D \in \mathcal{D}_J} \lambda_D(\boldsymbol{y}_{D \cap J}, \mathbf{0}_{D \setminus (D \cap J)})} \, \mathrm{d}\boldsymbol{y}_J.\end{aligned} \tag{4.D.2}$$

It follows that (4.D.2) is satisfied for every $B \in \mathcal{A}$. Finally, the sigma-algebra generated by the class of product sets on one dimension generates $\mathcal{B}(\mathbb{E}_J)$, the π-λ theorem yields that (4.D.2) holds for every $B \in \mathcal{B}(\mathbb{E}_J)$. □

4.E Proof of Proposition 4.3.5

Let $D \subset C$, with $C \in \mathcal{C}$ and $D \in \mathcal{D}^*$. Recall the mixture exponent measure density λ_C from (4.2.9), with coefficient matrix $A^C \in \mathcal{M}_{d,r_C}([0,1])$ and exponent measure density $\tilde{\lambda}_C$ on each factor. For each $k \in \{1, \ldots, r_C\}$, let J_k denote the k-th signature of A^C as in (2.4.2). For $I \in \{C \setminus D, D\}$, define

$$J_{k,I} := I \cap J_k, \qquad \mathbb{A}_{I,J_k} := \mathbb{A}_{I, I \cap J_k},$$

where for $J \subset I$, $\mathbb{A}_{I,J}$ is given in (4.2.1). Fix $\boldsymbol{y}_D \neq \mathbf{0}_D$ and $k \in \{1, \dots, r_C\}$. Assume $D \cap J_k \neq \varnothing$.

- **Case $\boldsymbol{y}_D \in \mathbb{A}_{D,J_k}$.** Note that in this case, $(\boldsymbol{y}_D, \boldsymbol{y}_{C\setminus D}) \in \mathbb{A}_{C,J_k}$ if and only if $\boldsymbol{y}_{C\setminus D} \in \mathbb{A}_{C\setminus D,J_k}$. Therefore, we obtain

$$\begin{aligned}
&\int_{\boldsymbol{y}\in\mathbb{E}_{C\setminus D}} \mathbb{1}\left\{\boldsymbol{y}_C \in \mathbb{A}_{C,J_k}\right\} \tilde{\lambda}_{J_k}\left((y_j/a_{jk})_{j\in J_k}\right) \prod_{j\in J_k} (a_{jk})^{-1}\, \mathrm{d}\mu_{C\setminus D}(\boldsymbol{y}_{C\setminus D}) \\
&= \prod_{j\in J_k} (a_{jk})^{-1} \cdot \int_{\boldsymbol{y}\in\mathbb{E}_{C\setminus D}} \mathbb{1}\left\{\boldsymbol{y}_{C\setminus D} \in \mathbb{A}_{C\setminus D,J_k}\right\} \tilde{\lambda}_{J_k}\left((y_j/a_{jk})_{j\in J_k}\right) \mathrm{d}\mu_{C\setminus D}(\boldsymbol{y}_{C\setminus D}) \\
&= \prod_{j\in J_k} (a_{jk})^{-1} \cdot \int_{\boldsymbol{y}_{J_{k,C\setminus D}}>\mathbf{0}} \tilde{\lambda}_{J_k}\left((y_j/a_{jk})_{j\in J_k}\right) \otimes_{j\in J_{k,C\setminus D}} \mathrm{d}y_j \times \\
&\int_{\boldsymbol{y}_{(C\setminus D)\setminus J_k}=\mathbf{0}} \mathbb{1}\left\{\boldsymbol{y}_{C\setminus D} \in \mathbb{A}_{C\setminus D,J_k}\right\} \otimes_{j\in(C\setminus D)\setminus J_k} \mathrm{d}\delta_0(y_j) \\
&= \prod_{j\in J_k} (a_{jk})^{-1} \cdot \int_{\boldsymbol{y}_{J_{k,C\setminus D}}>\mathbf{0}} \tilde{\lambda}_{J_k}\left((y_j/a_{jk})_{j\in J_k}\right) \otimes_{j\in J_{k,C\setminus D}} \mathrm{d}y_j \\
&= \prod_{j\in J_{k,D}} (a_{jk})^{-1} \tilde{\lambda}_{J_{k,D}}\left((y_j/a_{jk})_{j\in J_{k,D}}\right),
\end{aligned}$$

  where the last equality follows by change of variables, and for simplicity, we write $\mathbf{0}$ for a zero vector of appropriate dimension, and $A^C = (a_{jk})_{j\in C,\, k\in\{1,\dots,r_C\}}$.

- **Case $\boldsymbol{y}_D \notin \mathbb{A}_{D,J_k}$.** In this case, $(\boldsymbol{y}_D, \boldsymbol{y}_{C\setminus D}) \notin \mathbb{A}_{C,J_k}$ for all $\boldsymbol{y}_{C\setminus D} \in \mathbb{E}_{C\setminus D}$. Thus, the integral above evaluates to zero.

Additionally, observe that when $D \cap J_k = \varnothing$, the integral is zero. Summing over all $k \in \{1, \dots, r_C\}$, we conclude that $\lambda_{C\to D}$ as defined in (4.2.10) coincides exactly with the mixture exponent measure density λ_D with coefficient matrix A^D and exponent measure density $\tilde{\lambda}_D$ in each factor.

Finally, note that the choice of the clique $C \supset D$ is not essential. Indeed, if $D \subset C' \cap C''$, then Equation (4.3.8) guarantees that $A^{C'}_{D,\cdot} \sim A^{C''}_{D,\cdot}$. Hence, the submatrices $\tilde{A}^{C'}_{D,\cdot}$ and $\tilde{A}^{C''}_{D,\cdot}$ differ at most by a permutation of their columns. Since mixture models are invariant under permutations of columns in the associated coefficient matrix, the choice between $\tilde{A}^{C'}_{D,\cdot}$ or $\tilde{A}^{C''}_{D,\cdot}$ as A^D does not matter.

4.F Proofs of Section 4.4

Note that Proposition 4.4.3 arises as a special case of Theorem 4.4.8 when the underlying graph is a forest. Accordingly, we proceed directly to the proof of

Theorem 4.4.8 in this section. The most technical part is to compute the coefficient a_J in (4.4.8) associated to an extreme direction $J \subset V$. Let $J \subset V$ be an extreme direction, and set $\mathcal{C}_J = \{C \in \mathcal{C} : C \cap J \neq \varnothing\}$. The key idea is to partition $\mathcal{C}_J$ into f_J "maximal" collections of cliques such that the nodes appearing within each collection induce a connected subgraph of $\mathcal{G}$ and such that all cliques C in the n-th collection share the same coefficient $a_{J,n}$ corresponding to the extreme direction $C \cap J$.

For a collection $\tilde{\mathcal{C}} \subset \mathcal{C}$ of cliques, let $\tilde{S} = \bigcup_{C \in \tilde{\mathcal{C}}} C$. We will say that $\tilde{\mathcal{C}}$ is a 2-*complete collection of cliques* if

1. the subgraph $\mathcal{G}_{\tilde{S}}$ is connected;
2. for two cliques $C', C'' \in \tilde{\mathcal{C}}$, either $C' \cap C'' = \varnothing$ or $|C' \cap C''| \geqslant 2$.

A 2-complete collection of cliques is *maximal* if it is not strictly included in another complete collection of cliques. We illustrate this concept of complete collection of cliques in Example 4.F.3.

Lemma 4.F.1. *There exists a disjoint partition $\mathcal{C}_J = \dot{\bigcup}_{n=1}^{f_J} \mathcal{C}_{J,n}$ such that $\mathcal{C}_{J,n}$ is a maximal 2-complete collection of cliques for every $n = 1, \ldots, f_J$.*

Proof of Lemma 4.F.1. Let $\mathcal{T}$ be a junction tree of $\mathcal{G}$ as in (4.A.1). Since J is an extreme direction, the induced subgraph $\mathcal{G}_J$ is connected. By Lemma 4.A.3, the induced subtree $\mathcal{T}_{\mathcal{C}_J}$ is connected and $C^* \cap C^{**} \neq \varnothing$ for every edge (C^*, C^{**}) of $\mathcal{T}_{\mathcal{C}_J}$. Decompose $\mathcal{T}_{\mathcal{C}_J}$ into maximal connected subtrees $\mathcal{T}_{J,n} = (\mathcal{C}_{J,n}, E_{J,n})$, $n = 1, \ldots, f_J$, such that adjacent cliques within each sub-tree $\mathcal{T}_{J,n}$ do not form singleton intersections. Since each $\mathcal{T}_{J,n}$ is connected, the union of its vertex set, that is, $S_{J,n} = \cup_{C \in \mathcal{C}_{J,n}} C$, induces a connected subgraph $\mathcal{G}_{S_{J,n}}$. Moreover, by construction $\mathcal{C}_{J,n}$ satisfies Condition (2), hence it is a 2-complete collection of cliques. Finally, maximality of each $\mathcal{C}_{J,n}$ follows from the maximality of the corresponding subtree in the above decomposition. □

For the following lemma, recall $\mathcal{A}$ the collection of matrices as in Section 4.3.2 satisfying (4.3.8) and (4.4.7).

Lemma 4.F.2. *For every $n = 1, \ldots, f_J$, the collection $\mathcal{C}_{J,n}$ from Lemma 4.F.1 satisfies*

1. *$J \cap C$ is a signature of A^C, for every clique $C \in \mathcal{C}_{J,n}$;*
2. *there exists a unique $a_{J,n}$ such that*

$$a^C_{J \cap C} = a_{J,n}, \qquad \forall C \in \mathcal{C}_{J,n},$$

where $a^C_{J \cap C}$ is the coefficient associated to the signature $J \cap C$ on the matrix A^C.

Proof of Lemma 4.F.2. Let C', C'' two different cliques of $\mathcal{C}_{J,n}$. Consider $C' = C_1$ as the root of the tree $\mathcal{T}_{\mathcal{C}_{J,n}}$ and define the order induced by naturally following

path(C', C'') through $\mathcal{T}_{\mathcal{C}_{J,n}}$ until $C'' = C_m$. Lemma 4.A.3 shows that $C_i \cap C_{i+1} \cap J \neq \varnothing$ for every $i \leqslant m-1$. Therefore (4.3.8) and (4.4.7) ensure that

$$a^{C_i}_{C_i \cap J} = a^{C_{i+1}}_{C_{i+1} \cap J}, \qquad i \leqslant m-1.$$

Simple induction shows that $a^{C'}_{C' \cap J} = a^{C''}_{C'' \cap J}$. □

Example 4.F.3. Consider the decomposable graph $\mathcal{G}$ given in Figure 4.12 with cliques $\mathcal{C}$. Consider matrices $\mathcal{A}$ over $\mathcal{C}$ given by

$$A^{C_1} = \begin{pmatrix} 0 & a^{C_1}_{\cdot 2} & a^{C_1}_{\cdot 3} \\ a^{C_1}_{\cdot 1} & a^{C_1}_{\cdot 2} & 0 \\ 0 & a^{C_1}_{\cdot 2} & a^{C_1}_{\cdot 3} \end{pmatrix}, \quad A^{C_2} = \begin{pmatrix} a^{C_2}_{\cdot 1} & a^{C_2}_{\cdot 2} & 0 \\ 0 & a^{C_2}_{\cdot 2} & a^{C_2}_{\cdot 3} \\ a^{C_2}_{\cdot 1} & 0 & a^{C_2}_{\cdot 3} \end{pmatrix},$$

$$A^{C_3} = \begin{pmatrix} a^{C_3}_{\cdot 1} & a^{C_3}_{\cdot 2} & 0 \\ 0 & a^{C_3}_{\cdot 2} & a^{C_3}_{\cdot 3} \\ a^{C_3}_{\cdot 1} & a^{C_3}_{\cdot 2} & 0 \end{pmatrix}, \quad A^{C_4} = \begin{pmatrix} 0 & a^{C_4}_{\cdot 2} & a^{C_4}_{\cdot 3} \\ a^{C_4}_{\cdot 1} & a^{C_4}_{\cdot 2} & 0 \\ 0 & a^{C_4}_{\cdot 2} & a^{C_4}_{\cdot 3} \end{pmatrix}.$$

where $0 < a^{C_i}_{\cdot k} \leqslant 1$ for every $k \leqslant 3, i \leqslant 4$. Suppose that $\mathcal{A}$ satisfies (4.3.8) and (4.4.7). Let $J = \{2, 4, 6\}$. We have $\mathcal{C}_J = \mathcal{C}$ and $C_i \cap J$ is a signature of A^{C_i} for every $i = 1, \ldots, 4$; see columns in green. Theorem 4.4.8 shows that J is an extreme direction of Λ_V. We have

(i) $\mathcal{C}_J = \mathcal{C}_{J,1} \dot\cup \mathcal{C}_{J,2}$, with $\mathcal{C}_{J,1} = \{C_1, C_2\}$ and $\mathcal{C}_{J,2} = \{C_3, C_4\}$ are maximal 2-complete collections of cliques;

(ii) Equation (4.3.8) and (4.4.7) ensure that $a^{C_1}_{C_1 \cap J} = a^{C_2}_{C_2 \cap J} = a^{C_1}_{\cdot 1}$ and $a^{C_3}_{C_3 \cap J} = a^{C_4}_{C_4 \cap J} = a^{C_3}_{\cdot 1}$.

Remark 4.F.4. Note that we can always construct the partition $\mathcal{C}_J = \mathcal{C}_{J,1} \dot\cup \ldots \dot\cup \mathcal{C}_{J,f_J}$ in a "continuous way" such that for every $n < f_J$, the graph $\mathcal{T}_{\mathcal{C}_{J,1} \dot\cup \ldots \dot\cup \mathcal{C}_{J,n}}$ is connected. This choice guarantees that $S_n \cap S_{n+1}$ is a singleton that separates S_n and S_{n+1}, where $S_n = \bigcup_{C \in \mathcal{C}_{J,n}} C$.

Lemma 4.F.5. *Recall the set of separators $\mathcal{D}$ as in* (4.A.2) *and consider $\mathcal{D}_J = \{D \in \mathcal{D} : D \cap J \neq \varnothing\}$. There exist $d_{1,2}, \ldots, d_{f_{J-1}, f_J} \in V$ such that*

$$\mathcal{D}_J = \bigcup_{n=1}^{f_J} \left\{ C' \cap C'' : (C', C'') \in E_{\mathcal{T}_{J,n}} \right\} \cup \bigcup_{n=1}^{f_J - 1} \left\{ \left\{ d_{n,n+1} \right\} \right\},$$

where $\mathcal{C}_J = \mathcal{C}_{J,1} \dot\cup \ldots \dot\cup \mathcal{C}_{J,n}$ is the partition given in Lemma 4.F.1.

Proof of Lemma 4.F.5. Recall the subtrees $\mathcal{T}_{J,1}, \ldots, \mathcal{T}_{J,n}$ of $\mathcal{T}_{\mathcal{C}_J}$ induced by the maximal 2-complete collections $\mathcal{C}_{J,1}, \ldots, \mathcal{C}_{J,n}$ as in the proof of Lemma 4.F.1. Equation (4.A.10) followed by connectivity of $\mathcal{T}_{J,n}$ for every $n = 1, \ldots, f_J$ and Remark 4.F.4 yield that

$$\begin{aligned}
\mathcal{D}_J &= \left\{ C' \cap C'' : \ (C', C'') \in E_{\mathcal{T}},\, C', C'' \in \mathcal{C}_J \right\} \\
&= \left\{ C' \cap C'' : (C', C'') \in E_{\mathcal{T}},\, C', C'' \in \bigcup_{n=1}^{f_J} \mathcal{C}_{J,n} \right\} \\
&= \bigcup_{n=1}^{f_J} \left\{ C' \cap C'' : (C', C'') \in E_{\mathcal{T}_{J,n}} \right\} \cup \bigcup_{n=1}^{f_J - 1} \{ S_n \cap S_{n+1} \} \\
&= \bigcup_{n=1}^{f_J} \left\{ C' \cap C'' : (C', C'') \in E_{\mathcal{T}_{J,n}} \right\} \cup \bigcup_{n=1}^{f_J - 1} \left\{ \left\{ d_{n,n+1} \right\} \right\}. \qquad \square
\end{aligned}$$

Proof of Theorem 4.4.8. Let $J \subset V$ be an extreme direction and consider the partition $\mathcal{C}_J = \mathcal{C}_{J,1} \dot\cup \ldots \dot\cup \mathcal{C}_{J,f_J}$ as introduced in Lemma 4.F.1 and the associated subtrees $\mathcal{T}_{J,1}, \ldots, \mathcal{T}_{J,n}$. Let $\boldsymbol{y} \in \mathcal{A}_{V,J}$ as in (4.2.1). It follows that $\boldsymbol{y}_I \in \mathbb{A}_{I, I \cap J}$, for every $I \in \mathcal{C}_J \cup \mathcal{D}_J$. Lemma 4.F.5 shows that

$$\begin{aligned}
\lambda_V(\boldsymbol{y}) &= \frac{\prod_{C \in \mathcal{C}} \bar\lambda_C(\boldsymbol{y}_C)}{\prod_{D \in \mathcal{D}} \bar\lambda_D(\boldsymbol{y}_D)} = \frac{\prod_{C \in \mathcal{C}_J} \lambda_C(\boldsymbol{y}_C)}{\prod_{D \in \mathcal{D}_J} \lambda_D(\boldsymbol{y}_D)} \\
&= \frac{\prod_{n=1}^{f_J} \prod_{C \in \mathcal{C}_{J,n}} \lambda_C(\boldsymbol{y}_C)}{\prod_{n=1}^{f_J} \prod_{(C', C'') \in E_{J,n}} \lambda_{C' \cap C''}(\boldsymbol{y}_{C' \cap C''})} \times \prod_{n=1}^{f_J} \frac{1}{y_{d_{n,n+1}}^{-2}}.
\end{aligned}$$

Inside a maximal 2-complete collection of cliques. Fix $n \leqslant f_J$ and write $\mathcal{T}_{J,n} = (\mathcal{C}_{J,n}, E_{J,n})$. Let $(C',\, C'')$ be two cliques from $\mathcal{C}_{J,n}$. Lemma 4.F.2 shows that $a^{C'}_{J \cap C'} = a^{C''}_{J \cap C'} = a_{J,n}$. Let $J' = J \cap C'$, $J'' = J \cap C''$ and $j' = |J'|$, $j'' = |J''|$. A d-variate Hüsler–Reiss exponent measure density being $-(d+1)$ homogeneous, completion of a variogram matrix in Hentschel et al. [53], $C' \cap C''$ separating C' and C'' and calculus in 1 just right after the proof yield

$$\frac{\lambda_C(\boldsymbol{y}_C) \times \lambda_C(\boldsymbol{y}_C)}{\lambda_{C' \cap C''}(\boldsymbol{y}_{C' \cap C''})} = a_{J,n}^{-|J \cap (C' \cup C'')|} \lambda_{\Gamma_{J \cap (C' \cup C'')}}(a_{J,n}^{-1} \cdot \boldsymbol{y}_{J \cap (C' \cup C'')}). \tag{4.F.1}$$

Write $S_n = \bigcup_{C \in \mathcal{C}_{J,n}} C$. It follows that

$$\lambda_{J,n}(\boldsymbol{y}_{J \cap S_n}) := \frac{\prod_{C \in \mathcal{C}_{J,n}} \lambda_C(\boldsymbol{y}_C)}{\prod_{(C', C'') \in E_{J,n}} \lambda_{C' \cap C''}(\boldsymbol{y}_{C' \cap C''})} = a_{J,n}^{-|J \cap S_n|} \lambda_{\Gamma_{J \cap S_n}} \left(a_{J,n}^{-1} \boldsymbol{y}_{J \cap S_n} \right). \tag{4.F.2}$$

Between two maximal 2-complete collections of cliques. Now, let $n < f_J$ and recall from Remark 4.F.4 that $\left\{d_{n,n+1}\right\}$ separates S_n and S_{n+1}. A d-variate Hüsler–Reiss exponent measure density being $-(d+1)$ homogeneous, completion of a variogram matrix in Hentschel et al. [53], Equation (4.F.2) and calculus in 2 just right after the proof yield

$$\begin{aligned}
&\frac{\lambda_{J,n}(\boldsymbol{y}_{J\cap S_n}) \times \lambda_{J,n+1}(\boldsymbol{y}_{J\cap S_{n+1}})}{y_{d_{n,n+1}}^{-2}} \\
&= \left(a_{J,n} \cdot a_{J,n+1}\right)^{-|J\cap(S_n\cup S_{n+1})|} \lambda_{\Gamma_{J\cap(S_n\cup S_{n+1})}}\left((a_{J,n}a_{J,n+1})\boldsymbol{y}_{J\cap(S_n\cup S_{n+1})}\right).
\end{aligned}$$

Let $\tilde{\lambda}_V$ be the mixture Hüsler–Reiss exponent measure density as in (4.B.3) with coefficient matrix A and variogram matrix Γ as in Theorem 4.4.8. It follows that the exponent measures Λ_V and $\tilde{\Lambda}_V$, with densities λ_V and $\tilde{\lambda}_V$, have the same extreme directions, and that λ_V and $\tilde{\lambda}_V$ coincide on these extreme directions. Consequently, we obtain $\lambda_V = \tilde{\lambda}_V$. □

Calculus for the proof of Theorem 4.4.8.

1. For two cliques in the same 2-complete collection:

$$\begin{aligned}
\frac{\lambda_{C'}(\boldsymbol{y}_{C'}) \times \lambda_{C''}(\boldsymbol{y}_{C''})}{\lambda_{C'\cap C''}(\boldsymbol{y}_{C'\cap C''})} &= \frac{a_{J,n}^{-j'}\lambda_{\Gamma_{J'}}(a_{J,n}^{-1}\cdot\boldsymbol{y}_{J'}) \times a_{J,n}^{-j''}\lambda_{\Gamma_{J''}}(a_{J,n}^{-1}\cdot\boldsymbol{y}_{J''})}{a_{J,n}^{-|J'\cap J''|}\lambda_{\Gamma_{J'\cap J''}}(a_{J,n}^{-1}\cdot\boldsymbol{y}_{J'\cap J''})} \\
&= \frac{a_{J,n}^{-j'}a_{J,n}^{j'-1}\lambda_{\Gamma_{J'}}(\boldsymbol{y}_{J'}) \times a_{J,n}^{-j''}a_{J,n}^{j''-1}\lambda_{\Gamma_{J''}}(\boldsymbol{y}_{J''})}{a_{J,n}^{-|J'\cap J''|}a_{J,n}^{|J'\cap J''|-1}\lambda_{\Gamma_{J'\cap J''}}(\boldsymbol{y}_{J'\cap J''})} \\
&= a_{J,n}^{-1}\lambda_{\Gamma_{J\cap(C'\cup C'')}}(\boldsymbol{y}_{J\cap(C'\cup C'')}) \\
&= a_{J,n}^{-|J\cap(C'\cup C'')|}\lambda_{\Gamma_{J\cap(C'\cup C'')}}(a_{J,n}^{-1}\cdot\boldsymbol{y}_{J\cap(C'\cup C'')}).
\end{aligned}$$

2. Between two 2-complete collections $\mathcal{C}_{J,n}$ and $\mathcal{C}_{J,n+1}$

$$\begin{aligned}
&\frac{\lambda_{J,n}(\boldsymbol{y}_{J\cap S_n}) \times \lambda_{J,n+1}(\boldsymbol{y}_{J\cap S_{n+1}})}{y_{d_{n,n+1}}^{-2}} \\
&= \frac{a_{J,n}^{-|J\cap S_n|}\lambda_{\Gamma_{J\cap S_n}}\left(a_{J,n}^{-1}\boldsymbol{y}_{J\cap S_n}\right) \times a_{J,n+1}^{-|J\cap S_{n+1}|}\lambda_{\Gamma_{J\cap S_{n+1}}}\left(a_{J,n+1}^{-1}\boldsymbol{y}_{J\cap S_{n+1}}\right)}{y_{d_{n,n+1}}^{-2}}. \\
&= \frac{a_{J,n}^{-|J\cap S_n|}a_{J,n}^{|J\cap S_n|-1}\lambda_{\Gamma_{J\cap S_n}}(\boldsymbol{y}_{J\cap S_n})a_{J,n+1}^{-|J\cap S_{n+1}|}a_{J,n+1}^{|J\cap S_{n+1}|-1}\lambda_{\Gamma_{J\cap S_{n+1}}}(\boldsymbol{y}_{J\cap S_{n+1}})}{y_{d_{n,n+1}}^{-2}} \\
&= \left(a_{J,n} \cdot a_{J,n+1}\right)^{-|J\cap(S_n\cup S_{n+1})|} \lambda_{\Gamma_{J\cap(S_n\cup S_{n+1})}}\left((a_{J,n}a_{J,n+1})\boldsymbol{y}_{J\cap(S_n\cup S_{n+1})}\right).
\end{aligned}$$

4.G Algorithms

4.G.1 Sub-algorithms for Algorithm 6

Algorithm 7 Estimation of the bivariate mixture Hüsler–Reiss model for a single penalization parameter $\beta > 0$

Input: sample size $n \in \mathbb{N}$, integer $k \leqslant n$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0, \infty)^2$, bivariate sample $\mathcal{S} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0, \infty)^2$, reals $0 < \alpha < 1$, $\beta > 0$ and starting point $(a_0, \Gamma_0) \in [0, 1] \times \mathbb{R}^+_{>0}$.

Output: estimation $(\hat{a}, \hat{\Gamma}) \in [0, 1] \times \mathbb{R}^+$ of the bivariate mixture Hüsler–Reiss model.

1: Compute the bivariate empirical stdf estimates $\left(\hat{\ell}_{n,\,k}\left(\boldsymbol{c}_m\right)\right)_{m=1,\ldots,q}$ based on the sample $\mathcal{S}$. ▷ See Einmahl et al. [33, Section 2].

2: Compute $(\hat{a},\, \hat{\Gamma}) \in [0, 1] \times \mathbb{R}^+$ as the solution of the minimization problem (4.5.1) with penalization coefficient β, penalization exponent α and starting point (a_0, Γ_0).

3: **return** $(\hat{a}, \hat{\Gamma})$.

Algorithm 8 Bivariate K-fold cross validation score for a penalization parameter $\beta > 0$

Input: integers n, $k \leqslant n$ and $K \in \mathbb{N}$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0, \infty)^2$, bivariate sample $\mathcal{S} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0, \infty)^2$, reals $0 < \alpha < 1$, $\beta > 0$ and starting point $(a_0, \Gamma_0) \in [0, 1] \times \mathbb{R}^+_{>0}$.

Output: cross validation score associated to the penalization parameter β.

1: Split the sample $\mathcal{S}$ into K approximately equal-sized folds $\mathcal{S}_1, \ldots, \mathcal{S}_K$.
2: **for** each fold $i \in \{1, \ldots, K\}$ **do**
3: **Validation set:** set $\mathcal{S}_i$ as the validation set.
4: **Training set:** set $\mathcal{S} \setminus \mathcal{S}_i$ (all other folds) as the training set.
5: Compute the empirical stdf estimates $\left(\widehat{\ell}^{(i)}_{\lfloor n/K \rfloor, \lfloor k/K \rfloor}(\boldsymbol{c}_m)\right)_{m=1,\ldots,q}$ based on the sample $\mathcal{S}_i$. ▷ See Einmahl et al. [33, Section 2].
6: Compute $\left(\hat{a}^{(-i)}, \hat{\Gamma}^{(-i)}\right)$ as the output of Algorithm 7 with parameters: sample size $\lfloor n(K-1)/K \rfloor$, integer $\lfloor k(K-1)/K \rfloor$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$, sample $\mathcal{S} \setminus \mathcal{S}_i$, reals α and β and starting point (a_0, Γ_0).
7: Evaluate the trained model on the validation set via the i-th score

$$\mathrm{CV}^{(-i)}(\beta) = \sum_{m=1}^{q} \left\{ \widehat{\ell}^{(i)}_{\lfloor n/K \rfloor, \lfloor k/K \rfloor}(\boldsymbol{c}_m) - \ell\left(\boldsymbol{c}_m; \left(\widehat{a}^{(-i)}, \widehat{\Gamma}^{(-i)}\right)\right) \right\}^2.$$

8: **end for**
9: Compute the average cross validation score along all folds via

$$\mathrm{CV}(\beta) = \sum_{i=1}^{K} \mathrm{CV}^{(-i)}(\beta)/K.$$

10: **return** $(\mathrm{CV}(\beta))$.

Algorithm 9 Estimation of the bivariate mixture Hüsler–Reiss model for a grid $\left\{\beta_1, \ldots, \beta_g\right\}$ of penalization parameter

Input: sample size $n \in \mathbb{N}$, integers $k \leqslant n$ and $K \in \mathbb{N}$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0,\infty)^2$, bivariate sample $\mathcal{S} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0,\infty)^2$, reals $0 < \alpha < 1$, $\left\{\beta_1, \ldots, \beta_g\right\} \in [0,\infty)^g$ and starting point $(a_0, \Gamma_0) \in [0,1] \times \mathbb{R}^+_{>0}$.

Output: estimation $(\hat{a}, \hat{\Gamma}) \in [0,1] \times \mathbb{R}^+$ of the bivariate mixture Hüsler–Reiss model.

1: Define $\mathrm{CV}(\beta)$ the cross validation score for $\beta \in \{\beta_1, \ldots, \beta_r\}$ as defined in Algorithm 8.
2: Define $\beta^* = \arg\min_{\beta \in \left\{\beta_1, \ldots, \beta_g\right\}} \mathrm{CV}(\beta)$.
3: Compute $(\hat{a}, \hat{\Gamma})$ as the output of Algorithm 7 with parameters: sample size n, integer k, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$, bivariate sample $\mathcal{S}$, reals α, β^* and starting point (a_0, Γ_0).
4: **return** $(\hat{a}, \hat{\Gamma})$.

4.G.2 Algorithm to simulate from a mixture Hüsler–Reiss forest max-stable random vector

Let $\mathcal{F} = (V, E)$ be a forest and let $\{(a_{st}, \Gamma_{st}) : (s,t) \in E\}$ be the bivariate coefficients as in Section 4.4.2. Denote by $\boldsymbol{Y}$ the mixture Hüsler–Reiss forest graphical model on $\mathcal{F}$ with coefficient matrix A and variogram matrix Γ constructed as in Theorem 4.4.8 and let Λ_V be its exponent measure. In this section, we will simulate a max-stable random vector $\boldsymbol{M}$ with unit Fréchet margins and exponent measure Λ_V.

Write $V = V_1 \dot\cup V_2 \dot\cup \cdots \dot\cup V_p$ for the partition of V inducing the connected-component subtrees of $\mathcal{F}$. Since $\boldsymbol{Y}$ satisfies the global Markov property w.r.t. $\mathcal{F}$, one has $\Lambda_V\{\boldsymbol{y} \in \mathbb{E}_V : \boldsymbol{y}_{V_i} \neq \boldsymbol{0}_{V_i}, \boldsymbol{y}_{V_j} \neq \boldsymbol{0}_{V_j}\} = 0$ if $i \neq j$, and hence by Engelke et al. [38, Proposition 5.1 and Corollary 8.2] the subvectors $\boldsymbol{M}_{V_1}, \ldots, \boldsymbol{M}_{V_p}$ are independent. Consequently, we may simulate each $\boldsymbol{M}_{V_i}$ separately, reducing to the case where $\mathcal{F}$ is a single tree. From here on assume $\mathcal{F}$ is a tree.

Simulation of $\boldsymbol{M}$ proceeds in two stages:

1. construct the full coefficient matrix A and variogram matrix Γ from the edge-wise pairs (a_{st}, Γ_{st}). The variogram is completed by the function `completeGamma()` in the `graphicalExtremes` package [40], and A is assembled by Algorithm 13;
2. given A and Γ, simulate a mixture Hüsler–Reiss max-stable vector via Algorithm 10.

Algorithm 10 Simulation of a max-stable random vector with mixture Hüsler–Reiss exponent measure

Input: matrices $A = \left(a_{jk}\right)_{j \leqslant d,\, k \leqslant r} \in \mathcal{M}_{d,r}([0,1])$ and $\Gamma \in \mathcal{M}_{d,d}([0,\infty))$.

Output: max-stable random vector with unit-Fréchet margins and mixture Hüsler–Reiss exponent measure Λ_V with matrices A and Γ.

Define the symmetric matrix $\Sigma \in \mathcal{M}_{d,d}(\mathbb{R})$, given by $\Sigma_{jk} = \Gamma_{jk}/4$, for each $j \leqslant d$, $k \leqslant r$.
Initialize a list $\tilde{Z} = \left\{\tilde{\boldsymbol{Z}}_1, \ldots, \tilde{\boldsymbol{Z}}_r\right\}$, where each $\tilde{\boldsymbol{Z}}_k \in \mathbb{R}^d$ is initialized as a zero vector.
for each $k \in \{1, \ldots, r\}$ **do**
    Define $J_k \leftarrow \{j \in V : a_{jk} > 0\}$, the signature of the k-th column of A.
    Let $l_k = |J_k|$ be the size of J_k
    if $l_k = 1$, that is, $J_k = \{j\}$ is a singleton **then**
        Set $\tilde{Z}_{k,j}$, the j-th component of $\tilde{\boldsymbol{Z}}_k$, to $-a_{jk}/\log(U)$, where $U \sim \mathcal{U}(0,1)$.
    else if $l_k > 1$ **then**
        Extract the submatrix Γ_{J_k} from Γ.
        Define a vector $\boldsymbol{Z}_k \in \mathbb{R}^d$ and initialize it as the zero vector
        Set $\boldsymbol{Z}_{J_k}$ as a max-stable Hüsler–Reiss random vector with unit-Fréchet margins and variogram matrix Γ_{J_k}. ▷ **Refer to Algorithms 1 and 2 in Dombry et al. [27], as well as the `rmev()` function from the `mev` R package, where the squared coefficient matrix λ^2 is defined as $\lambda^2 = \Gamma/4$.**
        Set $\tilde{Z}_{jk} = Z_{jk} \cdot a_{jk}$, for each $j \in J_k$
    end if
end for
Compute $M_j = \max\left\{\tilde{Z}_{j1}, \ldots, \tilde{Z}_{jr}\right\}$, for each $j \in \{1, \ldots, d\}$.
return $\boldsymbol{M}$.

Algorithm 11 CoeffConstruction(CS) - Computation of an extreme direction coefficient

Input: tree $\mathcal{T} = (V, E)$, coefficient vector $\boldsymbol{a} = (a_{st} : (s,t) \in E) \in (0,1]^{|E|}$, a non-empty set $J \subset V$ such that $\mathcal{T}_J$ is connected.

Output: coefficient $a_{\cdot J}$ as in (4.4.5).

Initialize $a_{\cdot J} \leftarrow 1$.
Extract edge set $E \subset V \times V$ from $\mathcal{T}$. ▷ See for instance the function `E()` from the `R` package `igraph` [22].
for each edge $(s,t) \in E$ **do**
    if $s \in J$ or $t \in J$ **then**
        if $s, t \in J$ **then**
            $a_{\cdot J} \leftarrow a_{\cdot J} \cdot a_{st}$.
        else
            $a_{\cdot J} \leftarrow a_{\cdot J} \cdot (1 - a_{st})$.
        end if
    end if
end for
return $a_{\cdot J}$.

Algorithm 12 GetAllConnectedSubgraphs (GACS) - Extract the index sets of all connected subgraphs

Input: graph $\mathcal{G}$.

Output: index sets of all connected subgraphs of $\mathcal{G}$.

Initialize an empty list $\mathcal{S}$.
Let $V \leftarrow V(\mathcal{G})$ and $E \leftarrow E(\mathcal{G}) \subseteq V \times V$, be the set of nodes and edges respectively of $\mathcal{G}$. ▷ Using `V()` and `E()` from the `igraph` package [22]
for each node $v \in V$ **do**
    Append $\mathcal{G}_{\{v\}}$ to $\mathcal{S}$.
end for
for each subset $J \subset V$ of size $\geqslant 2$ **do**
    Extract the induced subgraph $\mathcal{G}_J$.
    if $\mathcal{G}_J$ is connected **then** ▷ See for instance the function `is.connected()` from the `R` package `igraph` [22].
        Append J to $\mathcal{S}$.
    end if
end for
return $\mathcal{S}$.

Algorithm 13 MatrixCoeffConstruction(MCS) - Construction of the coefficient matrix A

Input: tree $\mathcal{T} = (V, E)$, coefficient vector $\boldsymbol{a} = (a_{st} : (s,t) \in E) \in (0,1]^{|E|}$.
Output: coefficient matrix $A \in \mathcal{M}_{d,r}([0,1])$ as in Proposition 4.4.3.

Compute the index sets of all connected subgraphs $\mathcal{S} = \mathrm{GACS}(\mathcal{T})$ as defined in Algorithm 12.
Initialize an empty ordered list $\mathbf{c}$ to store nonzero coefficients.
Initialize an empty ordered list $\mathcal{J}$ to store corresponding extreme directions.
Let $k \leftarrow 1$. ▷ k represents index for nonzero coefficients.
for each $J \in \mathcal{S}$ **do**
    Compute $c \leftarrow \mathrm{CS}(\mathcal{T}, \boldsymbol{a}, J)$ as defined in Algorithm 11.
    if $c \neq 0$ **then**
        Append c to $\boldsymbol{c}$.
        Append J to $\mathcal{J}$.
        Update $k \leftarrow k + 1$.
    end if
end for
Let $r \leftarrow k - 1$ be the number of extreme directions.
Initialize a matrix $A \in \mathcal{M}_{r,d}([0,1])$ with zeros.
for each $k \in \{1, \ldots, r\}$ **do**
    for each $j \in \mathcal{J}_k$ **do**
        Set $A[j,k] \leftarrow c_k$.
    end for
end for
return A.

Algorithm 14 ExtremeDirectionsConstruction(EDC): Construction of extreme directions associated to a mixture Hüsler–Reiss forest model

Input: Tree $\mathcal{T} = (V, E)$ and coefficients $(\hat{a}_{st} : (s,t) \in E)$.

Output: Set of extreme directions $\mathcal{J}$

Define $E_1 \leftarrow \varnothing$ and $\mathcal{J} \leftarrow \varnothing$
for $(s,t) \in E$ **do**
    if $\hat{a}_{st} = 1$ **then**
        Append (s,t) to E_1
    end if
end for
Compute the list $\mathcal{S}$ of all subsets of V inducing connected subgraphs of $\mathcal{T}$:

$$\mathcal{S} \leftarrow \text{GACS}(\mathcal{T}) \text{ (see Algorithm 12)}$$

for $J \in \mathcal{S}$ **do**
    Test $\leftarrow$ True
    for $(s,t) \in E_1$ **do**
        if ($s \in J$ and $t \notin J$) or ($s \notin J$ and $t \in J$) **then**
            Test $\leftarrow$ False
            break ▷ Exit inner loop early
        end if
    end for
    if Test = True **then**
        Append J to $\mathcal{J}$
    end if
end for
return $\mathcal{J}$

Chapter 5

Comparison of Chapters 3 and 4

The goal here is to compare the two methods from Chapter 3 and Chapter 4 using river discharge data from $d = 5$ Stations $7, 18, 24, 27, 30$ chosen as described in Section 3.5.1. On the one hand, in Chapter 3, we fit a d-dimensional mixture logistic model. On the other hand, in Chapter 4, we fit a bivariate mixture Hüsler–Reiss model to every edge of the minimum spanning tree; see Step 1 from Algorithm 6. Proposition 4.4.3 shows then that this gives a d mixture Hüsler–Reiss graphical model with coefficient matrix and variogram matrix as given in the same Proposition.

Regarding the identification of extreme directions, both methods agree on three directions, which are also the ones assigned the highest weights by each method; see Table 5.1. However, the extreme directions identified by the penalized least-squares estimator are not compatible with any forest graphical structure $\mathcal{F}$ in dimension $d = 5$. According to Proposition 4.4.3, if a set J is an extreme direction, then the corresponding subgraph $\mathcal{F}_J$ must be connected. In this case, no forest structure satisfies this condition for the extreme directions found using the penalized least-squares estimator.

Next, we compare both methods based on their extremal correlations. On the one hand, we compute the empirical extremal correlation in (3.5.1), and on the other hand, for each method, we compute the estimated extremal correlation for each pair of stations s and t from the five stations. Recall that, for two stations s and t, the estimated extremal correlation resulting from the penalized least-squares estimator in Chapter 3 is given by (3.5.3), and the one resulting from the graphical structure-based estimator in Chapter 4 is given by (4.7.1). Results appear in Figure 5.1. We can observe that results are better for the penalized least-squares estimator, as points are closer to the diagonal line $x = y$. For the graphical structure-based estimator, we observe that the fitted model underestimates the extremal correlation for most station pairs. As explained in Section 4.7.1, this suggests that in the true underlying graph, those two sectors are actually joined

Penalized least-squares estimator	Extremal graphical model
$\{7, 18, 24, 27, 30\}$	$\{7, 18, 24, 27, 30\}$
$\{7, 18, 30\}$	$\{7, 18, 30\}$
$\{24, 27\}$	$\{24, 27\}$
$\{24\}$	$\{7, 18, 24, 27\}$
$\{7, 24\}$	$\{30\}$
—	$\{7, 24\}$

Table 5.1
Comparison of extreme directions obtained using the penalized least-squares estimator from Chapter 3 and the extremal graphical model from Chapter 4.

by a shorter path than the one recovered by the estimate. Another explanation of this underestimation would be that the model assumption of extremal conditional independence along a forest is not correct.

Finally, for each method, we compare the empirical and the estimated (fitted) probability of a failure set. A typical set used in applications is the one consisting of events where the sum of risks (river discharge in our case) exceeds a threshold u:

$$F_u = \{\boldsymbol{x} > \mathbf{0} : \|\boldsymbol{x}\|_1 > u\}. \tag{5.0.1}$$

We begin by computing the yearly componentwise maxima across the five stations, denoted by $\boldsymbol{M}_t = (M_{t1}, \ldots, M_{t5})$, for $t = 1960, \ldots, 2010$. We then define the sum of the componentwise annual maxima as $S_t = \sum_{j=1}^{5} M_{tj}$. Later, we will choose the threshold u in (5.0.1) as the quantile q_{1-p} for some exceedance probability $p \in [0, 1]$.

Probability of F_u from the fitted models. Recall that for each method, we obtain a fitted mixture model: a mixture logistic model for the penalized least-squares estimator in Chapter 3 and a mixture Hüsler–Reiss model for the graphical-structure based estimator in Chapter 4. From each method, we simulate N observations from the fitted model using Algorithm 5 (see also the R implementation in `https://github.com/AnasMourahib/Penalized_least-squares_estimator`). Note that the simulation code produces data with unit Fréchet margins. To match the margins of the observed data, we fit a generalized extreme value distribution to the componentwise maxima $\boldsymbol{M}_t$, for each year $t = 1960, \ldots, 2010$, using the `fgev()` function from the `R` package `evd`. This provides the shape, scale,

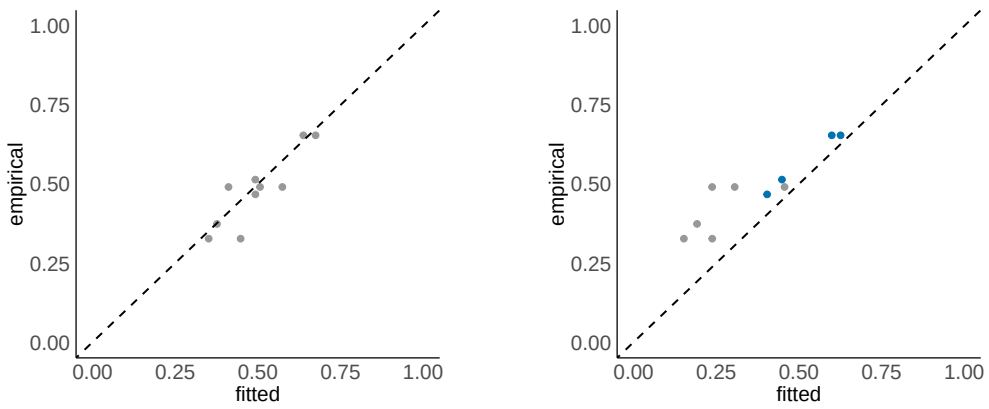


Figure 5.1
Empirical versus estimated extremal correlations based on the Danube river discharge data at the five stations mentioned before. Extremal. The estimated extremal correlations are calculated using the penalized-least squares estimator (left) and the graphical structure-based estimator (right). Blue points on the graph at right correspond to those of edges from the estimated forest.

and location parameters $(\gamma_j, \alpha_j, \mu_j)$ for each of the five margins $(j = 1, \ldots, 5)$. We then transform the simulated unit Fréchet data using the following formula:

$$Z \sim \text{Fréchet}(1) \quad \Rightarrow \quad \mu + \frac{\alpha}{\gamma}\left(Z^{\gamma} - 1\right) \sim \text{GEV}(\gamma, \alpha, \mu).$$

We finally compute the proportion of points in the failure set F_u. We repeat this n times and compute an average probability denoted by $\hat{p}_{F_u,\text{PLS}}$ and $\hat{p}_{F_u,\text{graph}}$ for the penalized least-squares estimator and the graphical-structure based method, respectively. Figure 5.2 shows the difference

$$D = \log(\hat{p}_{F_u}) - \log(p), \tag{5.0.2}$$

in orange for $\hat{p}_{F_u} = \hat{p}_{F_u,\text{PLS}}$ and in blue for $\hat{p}_{F_u} = \hat{p}_{F_u,\text{graph}}$. For both methods, the results are reasonably good. We also observe that the mixture logistic fit obtained from the penalized least-squares estimator performs slightly better overall.

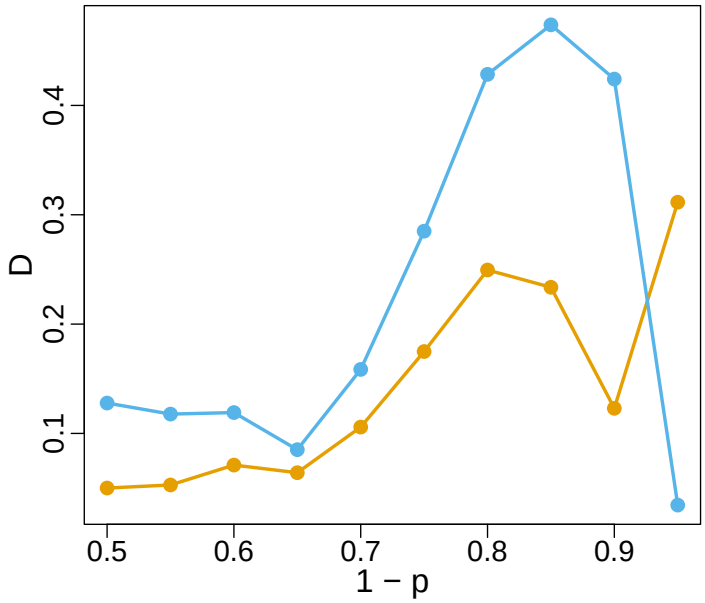


Figure 5.2
The distance D in (5.0.2) for the penalized least-squares estimator (orange) and the graphical structure-based estimator (blue), plotted as a function of the exceedance probability p used to construct the threshold $u = q_{1-p}$.

Chapter 6

Conclusion

6.1 Summary

Given a vector of risk factors—whether river discharges across some stations, returns on prices of stocks, or temperatures of some countries—we can always expect that a group of risk factors (not all factors) can be extreme without the remaining ones. In this thesis, we have called such a group a *non-standard extreme direction*. The main contribution of this thesis concerns the statistical practice of the threshold-exceedances approach in the presence of non-standard extreme directions.

In Chapter 2, we presented a construction method for threshold-exceedance parametric models that captures non-standard extreme directions. Such a model is called a max-linear model and can be seen as a smooth version of the max-linear model [100] in the sense that we do not assume complete dependence for the random factors (see Definition 2.4.1). We then show the generality of our construction in the sense that any generalized extreme value distribution can be constructed as the distribution function of the max-linear model with certain well-chosen parameters. Although the biggest part of this project is based on theory, precisely measure theory, as one can see from the proofs in the appendices of Chapter 2, we also worked on an algorithm to simulate from the mixture model; see Section 2.5. Moreover, we implemented this algorithm in `R` and developed a GitHub repository; see `https://github.com/AnasMourahib/MGPD-simulation`.

Once the mixture model from Chapter 2 is constructed, a natural question to consider is how to identify the parameters of this model. It turns out that this is a challenging question: some parameters being on the boundary of the parameter space make this estimation complicated. To solve this, we have adapted a penalized version of the weighted least-squares estimator proposed in Einmahl et al. [33]; see Section 3.3. This estimation also leads to an algorithm that

identifies groups of variables that can be large simultaneously, as presented in Section 3.3.2. To assess the performance of our estimation, we Use a simulation study and show that our estimation gives good results when choosing good tuning parameters. Finally, we apply our algorithm to two real-world datasets: first, to discharge measurements at stations along the Danube River, and second, to financial portfolio losses from stocks listed on the NYSE, AMEX, and NASDAQ. In both applications, we identify the extreme directions. To position this project within the available literature: on the one hand, even though we work with models for asymptotically dependent extremes, we can still handle situations when not all variables are extreme. In this way, we counter a common criticism in papers on conditional extremes [52] or geometric extremes [98]. On the other hand, we show that our extreme directions algorithm is more capable of identifying non-standard extreme direction structures compared to the DAMEX algorithm in Goix et al. [50]. Although theoretical results for the estimator of this project are still to be developed, the simulation part is complete and rich. In fact, the estimator is mostly implemented in `R` except the loss function (Equation (3.3.1)) which is implemented in `C` to reduce the cost of its evaluating during the minimization process. Techniques of parallel computing are also used in the coding part of this project. The code for this project is available at the GitHub repository `https://github.com/AnasMourahib/Penalized_least-squares_estimator`

Chapter 4 deals more with the notion of conditional independence in the extreme case using extremal graphical models in Engelke and Hitz [34]. Parametric extremal graphical models developed in the literature accommodate only the standard extreme direction $\{1, \ldots, d\}$. As discussed before, this is unrealistic since some risks may be extreme excluding others. The main contribution of this project is to develop extremal graphical models with non-standard extreme directions; see Theorem 4.3.1. As an example, we generalize the Hüsler–Reiss graphical model to accommodate multiple extreme directions and call our new model the mixture Hüsler–Reiss graphical model. Next to this construction, we propose a data-driven algorithm (Algorithm 6) to learn both the extreme directions and the forest dependence structure. The latter generalizes the tree structure learning algorithm proposed in Engelke et al. [38] to the case of extremal independence between some risks. The theoretical part of this project relies, on the one hand, on background from graph theory (see Section 4.A), and on the other hand, on background from measure theory. Some theoretical results are still in progress, for instance, the consistency of the forest structure learning algorithm. In terms of code, algorithms proposed in Sections 4.5 and 4.G are implemented in `R` and will soon be available in a repository on GitHub.

6.2 Future directions

Theoretical results for the PLS estimator. As presented in Section 3.4, we obtain good simulation results for our PLS estimator proposed in (3.3.1). However, no results on consistency or asymptotic normality are provided. Two main difficulties prevent us from establishing asymptotic theoretical results:

1. When the penalization exponent of the penalization term is such that $p < 1$, our PLS estimator is the solution of a non-convex minimization problem.

2. Some of the true parameters lie on the boundary of the parameter space. Under certain conditions, this is treated in Kiriliouk [64] for the unpenalized version of our estimator, i.e., the updated weighted least-squares estimator in Einmahl et al. [33]. However, these conditions seem difficult to verify for the parametric mixture models.

In the simplest case where the true parameters are in the interior of the parameter space and no penalization is needed, under certain conditions, asymptotic results are proved in Einmahl et al. [33]. However, as pointed out in the same paper, these conditions are either difficult to verify, for instance with the Hüsler–Reiss parametric model, or are not verified at all, for instance with the max-linear model.

General graph structure learning for extremes. Extremal graph structure learning has so far been treated in the case of tree graphs in Engelke and Volgushev [36] and forest graphs in Chapter 4. However, data structures may exhibit other general graph structures of extremal conditional independencies. For Gaussian graphical models, a partial-correlation testing approach is proposed; see Højsgaard et al. [55]. Consider a vector $(X_1, \ldots, X_d)$ of risk factors (river discharge at different stations from the Danube dataset, for example). For every two components s, t, this is based on the two steps:

1. Estimate the sample correlation $\hat{\rho}_{st\,|\,\{1,\ldots,d\}\setminus\{s,t\}}$;
2. Perform hypothesis testing $\hat{\rho} = 0$ vs. $\hat{\rho} \neq 0$.

In the context of extreme value theory, extremal correlations or the extremal variogram might be used instead; see Lee and Cooley [69].

Extremal structural causal models and extreme directions. Idea of Frank Röttger and Johan Segers. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a vector of risk factors. Suppose that $\boldsymbol{X}$ follows a structural causal model (SCM) with respect to a directed acyclic graph (DAG) $\mathcal{G}$; see Peters et al. [80]. Under certain conditions on the structure functions, Engelke et al. [41, Theorem 1] show that the extremal behavior of $\boldsymbol{X}$ can be described by a multivariate Pareto random vector $\boldsymbol{Y}$. Moreover, the random vector $\boldsymbol{Y}$ also follows a SCM with respect to a DAG $\tilde{\mathcal{G}} \subset \mathcal{G}$, meaning that

some causal links may disappear in the tail. Directed extremal graphical models are then defined using multivariate Pareto random vectors; see Engelke et al. [41, Section 4]. However, this definition is restricted to the case of a unique extreme direction $\{1, \ldots, d\}$. A possible extension is to define directed extremal causal models with non-standard extreme directions and investigate how the two notions of extreme directions and causality are related.

An `R` package for multivariate generalized Pareto distributions. Mourahib et al. [75] proposes a simulation algorithm (Algorithm 1) for a mixture multivariate generalized Pareto random vector $\boldsymbol{Y}$. This algorithm is implemented in `R` for the mixture Hüsler–Reiss and mixture logistic models; see `https://github.com/AnasMourahib/MGPD-simulation`. Other models can also be investigated, for instance mixture Dirichlet and Schlather models in Coles and Tawn [18], etc. An `R` package handling both estimation and simulation for multivariate generalized (Pareto) distributions would be an interesting idea.

Combinatorics and extreme directions. Recall Algorithm 6. One task of this algorithm is to identify the extreme directions. For a d-dimensional graph, one step of this algorithm visits every subset $J \subset \{1, \ldots, d\}$ and checks whether J satisfies a certain condition (P); see Theorem 4.3.1. This is as complex as enumerating all subsets of d elements. To our knowledge, there is no algorithm with less complexity than $\mathcal{O}(2^d)$ that performs this task. A question to investigate: given the nature of condition (P), is there a better way to solve the problem:

$$\text{Find all } J \subset \{1, \ldots, d\} \text{ such that } J \text{ satisfies (P).}$$

Extreme directions and reinsurance. Identifying extreme directions has practical relevance in risk theory, particularly in applications such as reinsurance. For instance, suppose we consider $d = 10$ risk components and identify two extreme directions $\{1, 2, 3\}$ and $\{4, \ldots, 10\}$. A reinsurance company seeking to diversify its exposure to extreme losses would prefer to reinsure one risk from each group, rather than two from the same group. This strategy reduces the likelihood of simultaneous large claims, thereby improving portfolio diversification and mitigating accumulation risk.

Extreme directions identification using a neural Bayes estimator. Suppose we want to estimate a parameter $\boldsymbol{\theta} \in \Theta$ of a certain model using data $\boldsymbol{x} \in \mathcal{X}$. This can be, for instance, the matrix A and the dependence parameter α for the mixture logistic model constructed in Section 2.4.2. Let $\pi(\cdot)$ be a prior of $\boldsymbol{\theta}$ and

$L : \Theta \times \Theta \to [0, \infty)$ be a loss function. The *Bayes risk* of an estimator $\hat{\boldsymbol{\theta}} : \mathcal{X} \to \Theta$ is given by

$$\int_{\Theta} \left\{ \int_{\mathcal{X}} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\boldsymbol{x})) \cdot p(\boldsymbol{x} \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{x} \right\} \cdot \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}, \tag{6.2.1}$$

where p is the likelihood function. Any minimizer of the Bayes risk is called a *Bayes estimator*. The idea of Bayes neural estimation is to assume a flexible parametric form for $\hat{\boldsymbol{\theta}}(\cdot)$ and optimize its parameters in order to approximate a Bayes estimator. More precisely, let $\hat{\boldsymbol{\theta}}_{\gamma} : \mathcal{X} \to \Theta$ denote a neural network parametrized by γ. Then a Bayes estimator can be approximated by $\hat{\boldsymbol{\theta}}_{\gamma^*}$ such that

$$\gamma^* = \arg\min_{\gamma} \frac{1}{N} \sum_{i=1}^{N} L\left(\boldsymbol{\theta}^{(i)}, \hat{\boldsymbol{\theta}}_{\gamma}(\boldsymbol{x}^{(i)})\right),$$

with $\boldsymbol{\theta}^{(i)} \sim \pi(\cdot)$ and independently $\boldsymbol{x}^{(i)} \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})$. We use neural Bayes estimation in order to estimate the dependence parameter α and one entry a of the coefficient matrix A of a 2-dimensional mixture logistic model; see Section 2.4.2. As an architecture for the neural network, we use a *DeepSets* architecture [101], as it ensures permutation invariance, meaning that the output of the neural network is not influenced by the order of the elements in the input set. Our model follows the DeepSets architecture, consisting of two core components: ψ and ϕ. The ψ network encodes each individual element in the set through three fully connected layers with *ReLU* activation, mapping the 2D input to a higher-dimensional latent space of size $w = 32$. These encoded representations are then aggregated via summation, ensuring permutation invariance, and passed to ϕ, which is another deep neural network composed of two hidden layers (also using ReLU) and a final `Parallel` layer. This final layer outputs two values, each in the interval $(0, 1)$ (since α and a are in $[0, 1]$), via sigmoid activations (σ), suitable for modeling parameters constrained to that range.

For training the neural network, we generate $K = 5000$ prior samples for the training set and $K/10$ for the validation set. For each prior, we simulate data from a mixture logistic model with a sample size of $m = 250$. During testing, we generate $K = 1000$ new prior samples. For each of these, we again simulate a dataset of size $m = 250$ from the mixture logistic model and estimate the corresponding prior parameters using the trained model. The results are presented in Figure 6.1. Results are quite good; however, when $a = 0$, the Bayes neural estimator fails to detect exactly 0, which is a key point to identify extreme directions; see Figure 6.2.

One possible idea would be to add an adequate penalty term to the loss function L, as in Chapter 3, in order to encourage the parameter a to take zero values.

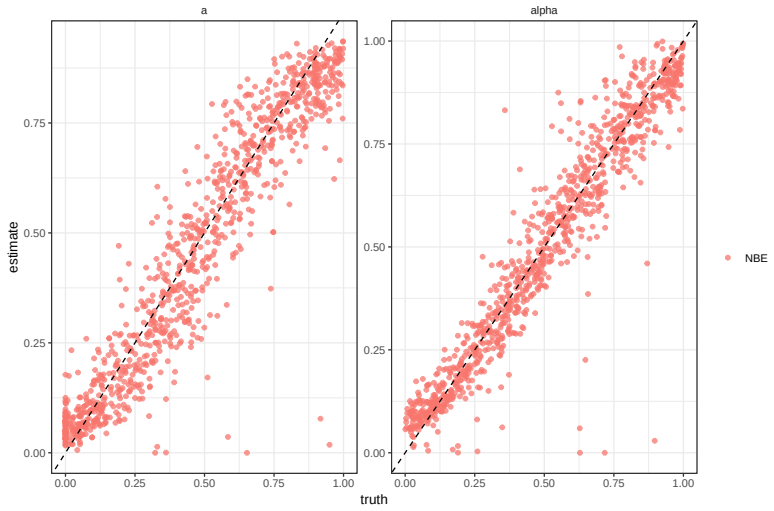


Figure 6.1
Bayes estimates of the mixture logistic model parameters using the trained DeepSets neural network on test priors. Each point corresponds to an estimate based on a simulated dataset of size $m = 250$ generated from a test prior.

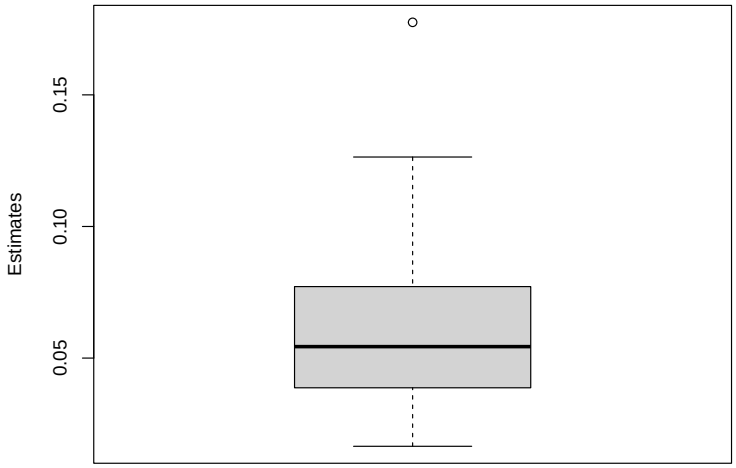


Figure 6.2
Boxplots of DeepSets neural network estimates for prior a centered at 0.

Bibliography

[1] Peiman Asadi, Anthony C Davison, and Sebastian Engelke. Extremes on river networks. 2015.

[2] Stefka Asenova and Johan Segers. Extremes of Markov random fields on block graphs: max-stable limits and structured Hüsler–Reiss distributions. *Extremes*, 26(3):433–468, 2023.

[3] Felix Ballani and Martin Schlather. A construction principle for multivariate extreme value distributions. *Biometrika*, 98(3):633–645, 2011.

[4] Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef L Teugels. *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons, 2004.

[5] Jan Beirlant, Mikael Escobar-Bach, Yuri Goegebeur, and Armelle Guillou. Bias-corrected estimation of stable tail dependence function. *Journal of Multivariate Analysis*, 143:453–466, 2016.

[6] Leo Belzile, Jennifer L Wadsworth, Paul J Northrop, Scott D Grimshaw, Jin Zhang, Michael A Stephens, and Art B Owen. Package 'mev'. Technical report, 2024.

[7] Léo R Belzile and Johanna G Nešlehová. Extremal attractors of liouville copulas. *Journal of Multivariate Analysis*, 160:68–92, 2017.

[8] Elsa Bernard, Philippe Naveau, Mathieu Vrac, and Olivier Mestre. Clustering of maxima: Spatial dependencies among heavy rainfall in France. *Journal of climate*, 26(20):7929–7937, 2013.

[9] Oliver Böhm and K-F Wetzel. Flood history of the Danube tributaries Lech and Isar in the Alpine foreland of Germany. *Hydrological Sciences Journal*, 51(5):784–798, 2006.

[10] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[11] Emilie Chautru. Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics*, 9:383–418, 2015.

[12] Maël Chiapino and Anne Sabourin. Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 132–147. Springer, 2016.

[13] Maël Chiapino, Anne Sabourin, and Johan Segers. Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22(2): 193–222, 2019.

[14] Maël Chiapino, Stéphan Clémençon, Vincent Feuillard, and Anne Sabourin. A multivariate extreme value theory approach to anomaly clustering and visualization. *Computational Statistics*, 35(2):607–628, 2020.

[15] Stuart Coles, Janet Heffernan, and Jonathan Tawn. Dependence measures for extreme value analyses. *Extremes*, 2:339–365, 1999.

[16] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[17] Stuart G Coles and Jonathan A Tawn. Modelling extreme multivariate events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2): 377–392, 1991.

[18] Stuart G. Coles and Jonathan A. Tawn. Modelling extreme multivariate events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):377–392, 1991.

[19] Daniel Cooley and Emeric Thibaud. Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604, 2019.

[20] Daniel Cooley, Richard A. Davis, and Philippe Naveau. The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117, 2010.

[21] Robert G Cowell, Philip Dawid, Steffen L Lauritzen, and David J Spiegelhalter. *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer Science & Business Media, 2007.

[22] Maintainer Gabor Csardi. Package 'igraph'. *Last accessed*, 3(09):2013, 2013.

[23] Anthony C Davison and Richard L Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 52(3):393–425, 1990.

[24] Laurens de Haan and Ana Ferreira. *Extreme value theory: an introduction*, volume 3. Springer, 2006.

[25] Laurens de Haan and Sidney I Resnick. Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 40(4):317–337, 1977.

[26] Gabriel Andrew Dirac. On rigid circuit graphs. In *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, volume 25, pages 71–76. Springer, 1961.

[27] Clément Dombry, Sebastian Engelke, and Marco Oesting. Asymptotic properties of the maximum likelihood estimator for multivariate extreme value distributions. *arXiv preprint arXiv:1612.05178*, 2016.

[28] Clément Dombry, Sebastian Engelke, and Marco Oesting. Exact simulation of max-stable processes. *Biometrika*, 103(2):303–317, 2016.

[29] Holger Drees and Xin Huang. Best attainable rates of convergence for estimators of the stable tail dependence function. *Journal of Multivariate Analysis*, 64(1):25–46, 1998.

[30] Holger Drees and Xin Huang. Best attainable rates of convergence for estimators of the stable tail dependence function. *Journal of Multivariate Analysis*, 64(1):25–46, 1998.

[31] Holger Drees and Anne Sabourin. Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, 15:908–943, 2021.

[32] John HJ Einmahl, Andrea Krajina, and Johan Segers. An M-estimator for tail dependence in arbitrary dimensions. *The Annals of Statistics*, 40(3): 1764–1793, 2012.

[33] John HJ Einmahl, Anna Kiriliouk, and Johan Segers. A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes*, 21:205–233, 2018.

[34] Sebastian Engelke and Adrien S Hitz. Graphical models for extremes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):871–932, 2020.

[35] Sebastian Engelke and Jevgenijs Ivanovs. Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, 8:241–270, 2021.

[36] Sebastian Engelke and Stanislav Volgushev. Structure learning for extremal tree models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2055–2087, 2022.

[37] Sebastian Engelke, Alexander Malinowski, Zakhar Kabluchko, and Martin Schlather. Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1):239–265, 2015.

[38] Sebastian Engelke, Jevgenijs Ivanovs, and Kirstin Strokorb. Graphical models for infinite measures with applications to extremes and Lévy processes. Available at `https://arxiv.org/abs/2211.15769`, 2022.

[39] Sebastian Engelke, Manuel Hentschel, Michaël Lalancette, and Frank Röttger. Graphical models for multivariate extremes. *arXiv preprint arXiv:2402.02187*, 2024.

[40] Sebastian Engelke, Adrien S Hitz, Nicola Gnecco, and Manuel Hentschel. Package 'graphicalExtremes'. 2024.

[41] Sebastian Engelke, Nicola Gnecco, and Frank Röttger. Extremes of structural causal models. *arXiv preprint arXiv:2503.06536*, 2025.

[42] Shimon Even. *Graph algorithms*. Cambridge University Press, 2011.

[43] Michael Falk. *Multivariate extreme value theory and D-norms*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2019.

[44] Ana Ferreira and Laurens De Haan. On the block maxima method in extreme value theory: Pwm estimators. *The Annals of Statistics*, 2015.

[45] Vladimir Fomichov and Jevgenijs. Ivanovs. Spherical clustering in detection of groups of concomitant extremes. *Biometrika*, 110(1):135–153, 2023.

[46] Anne-Laure Fougères, Cécile Mercadier, and John P Nolan. Dense classes of multivariate extreme value distributions. *Journal of Multivariate Analysis*, 116:109–129, 2013.

[47] Christian Francq and Jean-Michel Zakoian. *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons, 2019.

[48] Nicolas Goix, Anne Sabourin, and Stéphan Clémençon. Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial Intelligence and Statistics*, pages 75–83. PMLR, 2016.

[49] Nicolas Goix, Anne Sabourin, and Stéphan Clémençon. Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial intelligence and statistics*, pages 75–83. PMLR, 2016.

[50] Nicolas Goix, Anne Sabourin, and Stéphan Clémençon. Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis*, 161:12–31, 2017.

[51] Frank Harary. A characterization of block-graphs. *Canadian Mathematical Bulletin*, 6(1):1–6, 1963.

[52] Janet E Heffernan and Jonathan A Tawn. A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(3):497–546, 2004.

[53] Manuel Hentschel, Sebastian Engelke, and Johan Segers. Statistical inference for Hüsler–Reiss graphical models through matrix completions. *Journal of the American Statistical Association*, pages 1–13, 2024.

[54] Zhen Wai Olivier Ho and Clément Dombry. Simple models for multivariate regular variation and the Hüsler–Reiss Pareto distribution. *Journal of Multivariate Analysis*, 173:525–550, 2019.

[55] Søren Højsgaard, David Edwards, Steffen Lauritzen, Søren Højsgaard, David Edwards, and Steffen Lauritzen. Gaussian graphical models. *Graphical models with R*, pages 77–116, 2012.

[56] Shuang Hu, Zuoxiang Peng, and Johan Segers. Modeling multivariate extreme value distributions via Markov trees. *Scandinavian Journal of Statistics*, 51 (2):760–800, 2024.

[57] Raphaël Huser and Anthony C Davison. Composite likelihood estimation for the Brown–Resnick process. *Biometrika*, 100(2):511–518, 2013.

[58] Jürg Hüsler and Rolf-Dieter Reiss. Maxima of normal random vectors: Between independence and complete dependence. *Statistics and Probability Letters*, 7(4):283–286, 1989.

[59] Hamid Jalalzai and Rémi Leluc. Feature clustering for support identification in extreme regions. In *International Conference on Machine Learning*, pages 4733–4743. PMLR, 2021.

[60] Anja Janßen and Phyllis Wan. K-means clustering of extremes. *Electronic Journal of Statistics*, 14:1211–1233, 2020.

[61] Piotr Jaworski, Fabrizio Durante, Wolfgang Karl Hardle, and Tomasz Rychlik. *Copula theory and its applications*, volume 198. Springer, 2010.

[62] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[63] Anna Kiriliouk. Vignette for the tailDepFun package. Technical report, Technical report, 2016.

[64] Anna Kiriliouk. Hypothesis testing for tail dependence parameters on the boundary of the parameter space. *Econometrics and Statistics*, 16:121–135, 2020.

[65] Anna Kiriliouk and Philippe Naveau. Climate extreme event attribution using multivariate peaks-over-thresholds modeling and counterfactual theory. *The Annals of Applied Statistics*, 2020.

[66] Anna Kiriliouk, Holger Rootzén, Johan Segers, and Jennifer L Wadsworth. Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics*, 61(1):123–135, 2018.

[67] Anna Kiriliouk, Johan Segers, and Laleh Tafakori. An estimator of the stable tail dependence function based on the empirical beta copula. *Extremes*, 21: 581–600, 2018.

[68] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[69] Jeongjin Lee and Daniel Cooley. Partial tail correlation for extremes. *arXiv preprint arXiv:2210.02048*, 2022.

[70] Amanda Lenzi, Julie Bessac, Johann Rudi, and Michael L Stein. Neural networks for parameter estimation in intractable models. *Computational Statistics & Data Analysis*, 185:107762, 2023.

[71] Helmut Lütkepohl. Vector autoregressive models. In *Handbook of research methods and applications in empirical macroeconomics*, pages 139–164. Edward Elgar Publishing, 2013.

[72] Marco Avella Medina, Richard A Davis, and Gennady Samorodnitsky. Spectral learning of multivariate extremes. *Journal of Machine Learning Research*, 25(124):1–36, 2024.

[73] Nicolas Meyer and Olivier Wintenberger. Sparse regular variation. *Advances in Applied Probability*, 53(4):1115–1148, 2021.

[74] Nicolas Meyer and Olivier Wintenberger. Multivariate sparse clustering for extremes. *Journal of the American Statistical Association*, 119(forthcoming): 1–23, 2023.

[75] Anas Mourahib, Anna Kiriliouk, and Johan Segers. Multivariate generalized Pareto distributions along extreme directions. *Extremes*, pages 1–34, 2024.

[76] Anas Mourahib, Anna Kiriliouk, and Johan Segers. A penalized least squares estimator for extreme-value mixture models. *arXiv preprint arXiv:2506.15272*, 2025.

[77] Thomas Opitz. Extremal t processes: Elliptical domain of attraction and a spectral representation. *Journal of Multivariate Analysis*, 122:409–413, 2013.

[78] Simone A Padoan, Mathieu Ribatet, and Scott A Sisson. Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489):263–277, 2010.

[79] Ioannis Papastathopoulos and Kirstin Strokorb. Conditional independence among max-stable laws. *Statistics & Probability Letters*, 108:9–15, 2016.

[80] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[81] Sidney I Resnick. *Extreme values, regular variation, and point processes*, volume 4. Springer Science & Business Media, 2008.

[82] Jordan Richards, Matthew Sainsbury-Dale, Andrew Zammit-Mangion, and Raphaël Huser. Neural bayes estimators for censored inference with peaks-over-threshold models. *Journal of Machine Learning Research*, 25(390):1–49, 2024.

[83] Holger Rootzén and Nader Tajvidi. Multivariate generalized Pareto distributions. *Bernoulli*, 12(5):917–930, 2006.

[84] Holger Rootzén, Johan Segers, and Jennifer L Wadsworth. Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis*, 165:117–131, 2018.

[85] Holger Rootzén, Johan Segers, and Jennifer L Wadsworth. Multivariate peaks over thresholds models. *Extremes*, 21(1):115–145, 2018.

[86] Martin Schlather and Jonathan Tawn. Inequalities for the extremal coefficients of multivariate extreme value distributions. *Extremes*, 5:87–102, 2002.

[87] Johan Segers. Max-stable models for multivariate extremes, revstat stat. *Statistical Journal*, 10:61–82, 2012.

[88] Johan Segers. Discussion on 'Graphical models for extremes' by Sebastian Engelke and Adrien Hitz. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):926, 2020.

[89] Johan Segers. One-versus multi-component regular variation and extremes of Markov trees. *Advances in Applied Probability*, 52(3):855–878, 2020.

[90] Emma S Simpson, Jennifer L Wadsworth, and Jonathan A Tawn. Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532, 2020.

[91] Richard L Smith. The extremal index for a Markov chain. *Journal of applied probability*, 29(1):37–45, 1992.

[92] Richard L Smith, Jonathan A Tawn, and Stuart G Coles. Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268, 1997.

[93] RL Smith, JA Tawn, and HK Yuen. Statistics of multivariate extremes. *International Statistical Review/Revue Internationale de Statistique*, pages 47–58, 1990.

[94] Alec Stephenson. Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59, 2003.

[95] Jonathan A. Tawn. Bivariate extreme value theory: models and estimation. *Biometrika*, 75(3):397–415, 1988.

[96] Jonathan A Tawn. Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245–253, 1990.

[97] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[98] Jennifer L Wadsworth and Ryan Campbell. Statistical inference for multivariate extremes via a geometric approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1243–1265, 2024.

[99] Jennifer L Wadsworth and Jonathan A Tawn. Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika*, 101(1):1–15, 2014.

[100] Yizao Wang and Stilian A Stoev. Conditional sampling for spectrally discrete max-stable random fields. *Advances in Applied Probability*, 43(2):461–483, 2011.

[101] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.