Extreme Value Theory and Machine Learning
Théorie des Valeurs Extrêmes et Apprentissage

Habilitation à Diriger des Recherches de l'Institut Polytechnique de Paris

Soutenue publiquement à Palaiseau, le 27 Octobre 2021, par

ANNE SABOURIN

LTCI, Télécom Paris, Institut polytechnique de Paris

Composition du Jury :

Stéphane Boucheron Professeur, Université de Paris, Paris Diderot (LPSM UMR 8001)	Rapporteur
Richard Davis Professor, Columbia University (dept. of Statistics), USA	Rapporteur
Matthieu Lerasle Professeur et chargé de recherches, ENSAE (CREST UMR 9194)	Rapporteur interne
Clément Dombry Professeur, Université de Franche-Comté (LMB UMR 6623)	Examinateur
Holger Drees Professor, University of Hamburg (detp. of Mathematics), Germany	Examinateur
Gabor Lugosi Research Professor, Pompeu Fabra University (Dept. of Economics), Spain	Examinateur
Johan Segers Professor, UC Louvain (LIDAM ISBA), Belgium	Examinateur

Thanks

I thank all members of the S2A team, past and present, for the friendly, though stimulating ambience they create at work. In particular I thank Stephan Clémençon for many things, including a guided tour through statistical learning, saving my skin a couple times, and keeping morale up.

I feel lucky to have worked with so many people from so many different backgrounds. So thank you to all my co-authors, without distinction.

Finally I warmly thank the reviewers of this thesis Stéphane Boucheron, Richard Davis and Matthieu Lerasle for their time, their kind attention, their insightful comments and the interesting questions they raise in their reports. Thanks also in advance to them and the other jury members Clément Dombry, Holger Drees, Gabor Lugosi, Johan Segers for participating in the defense.

Contents

Chapter 1

Introduction

1.1 Foreword

Extreme value theory (EVT) is concerned with characterizing the probabilistic structure of tail events. Depending on the context, the focus is on the distribution of the maximum of a large sample or on the conditional distribution of excesses above large thresholds. In a multivariate or infinite dimensional setting (when stochastic processes are considered), the maximum is usually defined componentwise, and the large thresholds are relative to the norm of the random element under consideration. Since large-scale events are of particular concern for risk analysis, applications of extreme value analysis are numerous, ranging from insurance and finance to environmental sciences, including robustness of industrial installations. It is thus no wonder that extreme value statistics have aroused interest in the statistical community for decades. In contrast, until recently, extreme events have only carried little weight in the statistical learning and machine learning community. The largest values of a dataset are routinely treated as outliers and removed from the training test in most machine learning algorithms, if they are treated at all. Typical machine learning tasks are more related with mean behaviors than rare events. Also extreme value analysis makes extensive use of statistical models, while model-free approaches are preferred in the machine learning community. From a theoretical perspective, one common working assumption made in the statistical learning literature is that the random variables under considerations are sub-Gaussian, that is satisfy a concentration inequality of the same kind as a Gaussian variable, which is not compatible with regular variation assumptions typically made in extreme value analysis. Nonetheless, in some specific machine learning tasks such as anomaly detection, predictive maintenance, failure anticipation, extreme events and the distributional tail play a central role, while data scarcity suggests to use an extrapolation model such as those issued from extreme value theory. Also some aspects of the statistical learning literature such as normalized Vapnik-Chervonenkis inequalities are a hint that concentration results can be derived for the empirical measure of rare events, opening the road to a finite sample analysis of various estimators featured by extreme value statistics.

This thesis gathers my contribution to bridging the gap between extreme value theory and statistical learning from a theoretical perspective as well as in applications. This line of thoughts has recently generated interest, in particular the topic of dimensionality reduction and sparse pattern detection has drawn considerable attention in the past few years, as reviewed by Engelke and Ivanovs (2020). To my best knowledge however when I started working on this subject, that is, after completing my PhD in 2013, the only existing works in this direction was, first, a concentration study for extreme order statistics (Boucheron and Thomas (2012)) allowing for an adaptive choice of the number of extreme order statistics in tail index estimation (Boucheron and Thomas (2015)), see also Carpentier and Kim (2015) under additional regularity assumptions, and second, in a multivariate setting, a clustering algorithm aiming at identifying the support of the distribution of tail events (Chautru et al. (2015)).

1.2 Layout of the thesis

Chapter 2 starts off with a brief exposition of the necessary background on concentration inequalities for statistical learning (Section 2.1.1) and presents a specific concentration inequality adapted to rare classes which is proved in Goix et al. (2015) (see Section 2.1.2) and used on several occasions in the different contributions which are gathered in this thesis. Section 2.2 provides a first example of application of this inequality to statistical analysis of multivariate extremes after recalling basic facts pertaining to multivariate extreme value theory and regular variation (Section 2.2.1).

In Chapter 3 we consider the problem of classification in extreme regions of the predicting variable. The opening section 3.1 presents some background on the empirical risk minimization paradigm for classification from a statistical learning perspective, which is the viewpoint adopted by Jalalzai et al. (2018) (Section 3.2) where we derive the form of optimal tail classifiers and prove finite sample generalization bounds regarding empirical classifiers learnt in this framework. This general methodology is applied in Jalalzai et al. (2020) (Section 3.3) in a natural language processing framework, where a representation learning strategy with heavy tailed target is designed for improved classification of extreme sentence embeddings and dataset augmentation.

Chapter 4 gathers my contribution to the topic of dimensionality reduction for multivariate extremes. Two main directions are explored for this purpose: (i)

multiple subspace clustering, in other terms identification of the support of the limit distribution of extremes among a large number of possible unions of lower dimensional subsets of the original sample space (Goix et al. (2017), Section 4.1 and Chiapino and Sabourin (2016); Chiapino et al. (2019b), Section 4.2) ; and (ii) Principal component analysis of the limit distribution (Drees and Sabourin (2021), Section 4.3).

Chapter 5 is dedicated to the machine learning treatment of anomalies located in the tails of the dataset, that is anomaly detection and clustering of anomalies. In Goix et al. (2016) (Section 5.1) the dimension reduction device proposed in Goix et al. (2017) is exploited to detect anomalies deviating from the estimated tail support. In a somewhat different spirit, in Thomas et al. (2017) (Section 5.2) we focus on moderate dimensional problems and adopt a strategy based on minimum volume sets to perform anomaly detection on the angular component of the limit law of extremes, thus exploiting the pseudo-polar decomposition of the latter distribution. Finally, in Chiapino et al. (2019a) (Section 5.3) we consider the problem of clustering extremes of a large dimensional vector in the context of aviation safety management. In the latter framework, all extreme values (defined by the fact that the norm of the considered vector is comparatively large) are considered as potential anomalies, contrarily to the former sections of this chapter where the goal is to distinguish between normal and abnormal data among extremes.

Chapter 6 presents a piece of work (Sabourin and Segers (2017)) which is somewhat disconnected to the rest of the thesis. The focus is on semi-continuous processes, namely upper semi-continuous ones, which have been proposed in the literature of spatial extremes to model some meteorological extreme events such as rainstorms. A widely used pre-processing step in applications related to spatial extremes is to apply a preliminary standardization step to the marginal distributions of the considered object. The question we ask is under which conditions it is legitimate to do so with the above described processes while preserving max-stability properties.

Chapter 7 opens perspectives and sketches the main lines of ongoing works.

1.3 Selected list of publications

This thesis presents the material of the following main publications I contributed to since the end of my PhD thesis.

- Nicolas Goix, Anne Sabourin, and Stéphan Clémençon, "Learning the dependence structure of rare events: a non-asymptotic study." Conference on Learning Theory. PMLR, 2015.
  (Goix et al. (2015)) → Chapter 2.
- Jalalzai, Hamid, Stephan Clémençon, and Anne Sabourin. "On Binary Classification in Extreme Regions." NeurIPS, 2018.
  (Jalalzai et al. (2018)) → Section 3.2.
- Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin, "Heavy-tailed Representations, Text Polarity Classification and Data Augmentation", NeurIPS, 2020.
  (Jalalzai et al. (2020)) → Section 3.3.
- Nicolas Goix, Anne Sabourin, and Stephan Clémençon. "Sparse representation of multivariate extremes with applications to anomaly detection." Journal of Multivariate Analysis 161 (2017): 12-31. (Goix et al. (2017))
  and
  Nicolas Goix, Anne Sabourin, and Stéphan Clémençon. "Sparse representation of multivariate extremes with applications to anomaly ranking." Artificial Intelligence and Statistics. PMLR, 2016. (Goix et al. (2016))
  → Sections 4.1 and 5.1
- Maël Chiapino and Anne Sabourin. "Feature clustering for extreme events analysis, with application to extreme stream-flow data." International Workshop on New Frontiers in Mining Complex Patterns. Springer, Cham, 2016. (Chiapino and Sabourin (2016))
  and
  Maël Chiapino, Anne Sabourin, and Johan Segers. "Identifying groups of variables with the potential of being large simultaneously." Extremes 22.2 (2019): 193-222. (Chiapino et al. (2019b))
  → Section 4.2.
- Holger Drees and Anne Sabourin. "Principal component analysis for multivariate extremes." arXiv preprint arXiv:1906.11043 (2019), to appear in the Electronic Journal of Statistics.
  (Drees and Sabourin (2021)) → Section 4.3.
- Albert Thomas, Stephan Clémençon, Alexandre Gramfort and Anne Sabourin. "Anomaly Detection in Extreme Regions via Empirical MV-sets on the

Sphere." Artificial Intelligence and Statistics, 2017.
(Thomas et al. (2017)) → Section 5.2.

- Maël Chiapino, Stephan Clémençon, Vincent Feuillard, and Anne Sabourin, "A multivariate extreme value theory approach to anomaly clustering and visualization." Computational Statistics, 1-22. (2019)
(Chiapino et al. (2019a)) → Section 5.3.
- Anne Sabourin and Johan Segers. "Marginal standardization of upper semi-continuous processes. with application to max-stable processes." Journal of Applied Probability (2017): 773-796.
(Sabourin and Segers (2017)) → Chapter 6.

List of notations

μ	Exponent measure, page 15
Φ	Angular measure, page 16
$\mathcal{S}_{\mathcal{G}}(n)$	Shattering coefficient, page 12
P_n	Empirical distribution of an $i.i.d.$ sample following the distribution P, page 12
AD	Anomaly Detection, page 50
ERM	Empirical Risk Minimization, page 19
EVT	Extreme Value Theory, page 5
MDA	Maximum Domain of Attraction, page 16
PCA	Principal Component Analysis, page 33
r.v.	Random Variable, page 12
STDF	Stable tail dependence function, page 11

Chapter 2

Statistical learning guarantees for Extreme Value Analysis

In a statistical context, the goal is to learn some features of the distribution of the random element of interest, based on data. A crude partitioning of the field would be *(i)* the frequentist asymptotic approach, *(ii)* the Bayesian approach, *(iii)* the statistical learning approach. The first approach provides explicit asymptotic guarantees concerning estimated quantities depending on the true distribution. The second one allows to build confidence regions depending on the observed data, which are valid for any sample size. The third approach provides explicit universal error bounds which do not depend on the data and are also valid for any sample size. Until recently, the literature in extreme value statistics has mainly followed the first two approaches. This chapter gathers my contributions to the third one. It starts with a general concentration result for rare events (Section 2.1) which can be applied to extreme value analysis to obtain finite sample guarantees for estimators related to the empirical risk minimization paradigm. Section 2.2 provides a first example of application of this concentration inequality to the analysis of the empirical estimator of the *Stable Tail Dependence Function* (STDF), a classical summary of the dependence structure of multivariate extremes.

2.1 Concentration inequalities for rare events

The material gathered in this section and the next one relies on the publication Goix et al. (2015). After introducing some notation and recalling the minimum necessary background on statistical learning and concentration inequalities for VC classes (Section 2.1.1) we state and comment the main concentration result of the cited reference.

2.1.1 Statistical learning on VC classes

We recall some standard definitions and results borrowed from learning theory. For an in depth introduction, refer *e.g.* to Lugosi (2002) or Bousquet et al. (2003). If P is the distribution on the sample space $\mathcal{X}$ of a random variable (*r.v.*) X, then P_n denotes the empirical distribution on $\mathcal{X}$ of an independent and identically distributed ($i.i.d.$) sample $X_1, \dots, X_n \sim P$. The framework developed by Vapnik and Chervonenkis allows to control the deviations of the empirical measure uniformly over classes of sets of bounded complexity.

Definition 2.1 (Shattering coefficient and VC dimension)**.** *Let $\mathcal{G}$ be a class of subsets of a space $\mathcal{X}$. The shattering coefficient $\mathcal{S}_\mathcal{G}(n)$ of the class $\mathcal{G}$ is*

$$\begin{aligned}\mathcal{S}_\mathcal{G}(n) &= \max_{x_1,\dots,x_n \in \mathcal{X}} \left|\{A \cap (x_1, \dots, x_n) : A \in \mathcal{G}\}\right| \\ &= \max_{x_1,\dots,x_n \in \mathcal{X}} \left|\{z_A(x) = (\mathbb{1}_A(x_1), \dots, \mathbb{1}_A(x_n)) \in \{0,1\}^n : A \in \mathcal{G}\}\right|\end{aligned}$$

and the VC dimension of $\mathcal{G}$ is the integer

$$\mathcal{V}_\mathcal{G} = \sup_{n \in \mathbb{N}} \{n : \mathcal{S}_\mathcal{G}(n) = 2^n.\}$$

If $\mathcal{V}_\mathcal{G} < \infty$ we say that the class has finite VC dimension.

Many intuitive families of sets (half-spaces, finite classes, unions and intersections of such classes ...) have finite VC dimension. Sauer's lemma permits to bound the shattering coefficient in terms of VC dimension. It is proved *e.g.* in Lugosi (2002).

Lemma 2.2 (Sauer)**.** *The shattering coefficient satisfies $\mathcal{S}_\mathcal{G}(n) \le \sum_{k=0}^{\mathcal{V}_\mathcal{G}} \binom{n}{k}$. As a corollary, for all $n \in \mathbb{N}$, $\mathcal{S}_\mathcal{G}(n) \le \big(n+1\big)^{\mathcal{V}_\mathcal{G}}$ and for $n \ge \mathcal{V}_\mathcal{G}$, $\mathcal{S}_\mathcal{G}(n) \le \big(\frac{en}{\mathcal{V}_\mathcal{G}}\big)^{\mathcal{V}_\mathcal{G}}$.*

The celebrated VC inequality is a uniform bound in probability on the deviations of the empirical measure evaluated on a class $\mathcal{G}$. Notice that by Sauer's Lemma, the term $\ln(\mathcal{S}_\mathcal{G}(n))$ in the statement is bounded by $\mathcal{V}_\mathcal{G} \ln(en/\mathcal{V}_\mathcal{G})$ for all n, and by $\mathcal{V}_\mathcal{G} \ln(n+1)$ for n sufficiently large.

Theorem 2.3 (Vapnik-Chervonenkis inequality)**.** *Then for all $n \in \mathbb{N}$, with probability $(1-\delta)$,*

$$\sup_{A \in \mathcal{G}} |P - P_n|(A) \le B_n(\delta)$$

with B_n of order (as $n \to \infty$)

$$B_n(\delta) = O\left[\sqrt{\frac{\ln(1/\delta) + \ln(\mathcal{S}_\mathcal{G}(n))}{n}}\right]$$

2.1.2 Contributions

In extreme value analysis, one typically uses the k ($k \ll n$) largest order statistics of the sample at hand to estimate quantities pertaining to the tail distribution. In a multivariate context (or even in a general metric space endowed with a scalar multiplication) it is still possible to order the data according to their norm, or their distance to the origin. This amounts to evaluating the empirical measure P_n is on a class of sets of the kind $\mathcal{G} = \{tB, B \in \mathcal{G}_1\}$ where $\mathcal{G}_1$ is is any class of sets bounded away from the origin and $t > 0$ is chosen such that the class union $\mathbb{A} = \cup_{A \in \mathcal{G}} A$ has small probability $p = P[\mathbb{A}] = O(k/n)$. The classical empirical measure is then replaced with the tail empirical measure, $\nu_k(A) = \frac{n}{k} P_n(A) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_A(X_i)$, for $A \in \mathcal{G}$. It is thus reasonable to expect concentration inequalities for the tail empirical measure of the kind:

$$\textit{With probability } 1-\delta, \qquad \sup_{A \in \mathcal{G}} |\nu_k(A) - n/kP(A)| \le B_k(\delta),$$

where $B_k(\delta)$ is as in Theorem 2.3 with n replaced with k. Dividing both sides of the inequality by n/k and identifying p and k/n, the desired result becomes:

With probability $1 - \delta$,

$$\sup_{A \in \mathcal{G}} |P_n(A) - P(A)| \le O\left(\sqrt{\frac{p\left[\ln(1/\delta) + \ln(\mathcal{S}_{\mathcal{G}}(np))\right]}{n}} \right). \tag{2.1}$$

The following normalized VC-inequality (Vapnik and Chervonenkis (2015); Bousquet et al. (2003)) is a first step towards this end, stating that with probability $1 - 2\delta$,

$$\sup_{A \in \mathcal{G}} \left| \frac{P_n(A) - P(A)}{\sqrt{P(A)}} \right| \;\le\; 2\sqrt{\frac{\ln \mathcal{S}_{\mathcal{G}}(2n) + \ln \frac{4}{\delta}}{n}} \,,$$

which immediately yields

$$\sup_{A \in \mathcal{G}} |P_n(A) - P(A)| \;\le\; 2\sqrt{\frac{p\left[\ln \mathcal{S}_{\mathcal{G}}(2n) + \ln \frac{4}{\delta}\right]}{n}} \,. \tag{2.2}$$

Notice that the upper bound in the above display involves a logarithmic term $\ln \mathcal{S}_{\mathcal{G}}(2n)$ depending on the total sample size, not the effective sample size np as in (2.1). The VC-inequality stated below achieves the goal stated in (2.1) and is the cornerstone of the statistical learning contributions gathered in this thesis.

Theorem 2.4 (Concentration on low probability regions, Goix et al. (2015)). *Let $X_1, \dots, X_n$ be $i.i.d.$ realizations of a* r.v. *X with distribution P and let $\mathcal{G}$ be a VC-class of sets with VC-dimension $\mathcal{V}_{\mathcal{G}}$. Consider the class union $\mathbb{A} = \cup_{A \in \mathcal{G}} A$, and let $p = P(\mathbb{A})$. Then there is an absolute constant C such that for all $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\sup_{A \in \mathcal{G}} |P_n(A) - P(A)| \;\leq\; C \left[\sqrt{p} \sqrt{\frac{\mathcal{V}_{\mathcal{G}}}{n} \ln \frac{1}{\delta}} + \frac{1}{n} \ln \frac{1}{\delta} \right] . \tag{2.3}$$

To give an idea of concentration tools at play we provide in the appendix section an unpublished alternative statement (Theorem A.1) of Theorem 2.4 together with a complete proof. In this alternative statement, which is part of an ongoing work with Stéphane Lhaut and Johan Segers [1] about concentration for rare events, the unknown constant C is replaced with a logarithmic factor $\sqrt{\ln(pn)}$ which arises from using shattering coefficient to control the symmetrized deviations of the empirical measure conditionally to the number of points hitting the rare class. From a technical point of view it may be seen as a simplification of Goix et al. (2015)'s approach insofar as it does not require a call to the Bernstein-type inequality from McDiarmid (1998) applied to a maximum deviation functional $f(X_{1:n}) = \sup_{A \in \mathcal{G}} |P_n(A) - P(A)|$. Instead, the classical Bernstein inequality for binomial variables is used to control the number of points hitting the rare class. This conditioning trick is thus a central step in the argument, also present in Goix et al. (2015)'s proof.

2.2 Learning guarantees for the dependence structure of extremes

The concentration inequalities for rare classes obtained in Section 2.1 can be exploited to obtain finite sample guarantees for various estimators of tail quantities. In this section we focus on a classical functional summary of the tail dependence structure, the STDF defined in (2.8). We start off with a brief account of the probabilistic framework adopted in Goix et al. (2015) as well as in most of the contributions gathered in this thesis, that is multivariate regular variation (de Haan and Resnick (1977); Resnick (1987, 2007)) For an extensive account of Extreme Value Theory, in particular for relationships between the Max-domain of attraction and regular variation, we refer the reader to (Resnick, 1987; Beirlant et al., 1996; de Haan and Ferreira, 2006). One should keep in mind that modelling the

[1] Between the time the manuscript was sent to the reviewers and the defense, an arXiv version has been submitted (Lhaut et al. (2021))

upper tail of a random vector may require alternative and sometimes finer assumptions than regular variation, which underlie in particular the conditional extreme value models (Wadsworth et al. (2017)) and the theory of hidden regular variation (Das et al. (2013)). Here we focus on the original regular variation framework.

2.2.1 Background in Multivariate Extreme Value theory and regular variation

We place ourselves in $\mathcal{X} = \mathbb{R}^d$ endowed with its Borel σ-field $\mathcal{B}(\mathbb{R}^d)$ and we consider a random vector $X = (X_1, \ldots, X_d) \sim P$ and n $i.i.d.$ replications of it $X_i = (X_{i,1}, \ldots, X_{i,d}), i \leq n$. A traditional assumption in EVT is that after a suitable marginal standardization to unit Pareto margins, the conditional distribution of the standardized vector V (see (2.4) below) given that $\|V\| > t$ converges to a certain limit. Precisely, denoting by F the cumulative distribution (*c.d.f.*) of X and letting $F_j(x) = \mathbb{P}(X_j \leq x)$ define

$$\begin{aligned} T(x) &= \frac{1}{1 - F_j(x_j)}, j \in \{1, \ldots, d\} \\ V &= T(X) \end{aligned} \tag{2.4}$$

Then our key assumption is that there exists a Radon measure μ on $\mathbb{R}^d_+ \setminus \{0\}$, called the exponent measure, which is finite on sets bounded away from 0 (that is $0 \notin \overline{A}$ such that

$$t\mathbb{P}(V \in tA) \xrightarrow[t\to\infty]{} \mu(A)\,, \tag{2.5}$$

for all set $A \in \mathcal{B}(\mathbb{R}^d)$ which is bounded away from 0 such that $\mu(\partial A) = 0$. This is equivalent to vague convergence of the measures $\mu_t = t\mathbb{P}\,(V \in t\,\cdot\,)$ on the space $[0, \infty]^d \setminus \{0\}$ (see Resnick (1987, 2007)) and to M_0 convergence of the same measures on $\mathbb{R}^d_+$ as defined in Hult and Lindskog (2006) on a complete separable metric space. Notice that including or not the points at infinity does not matter in practice since μ assigns no mass to subspaces at infinity, that is to the subspaces $\{x \in [0, \infty] : x_j = \infty \text{ for } j \in J \subset \{1, \ldots, d\}\}$, $J \neq \varnothing$. Indeed an immediate consequence of (2.5) is that μ is homogeneous of order -1, $\mu(tA) = t^{-1}\mu(A)$ for $t > 0$ and $A \in \mathcal{B}(\mathbb{R}^d)$.

A few remarks are in order concerning Condition (2.5). First, notice that condition (2.5) is a special case of *regular variation*: a random vector Z is regularly varying if there exists a real function $b(t) > 0$ and a limit measure ν, such that

$$b(t)\mathbb{P}(Z \in tA) \xrightarrow[t\to\infty]{} \nu(A) \qquad (\nu(\delta A) = 0, 0 \notin \overline{A}) \tag{2.6}$$

where b is a positive function such that $b(tx)/b(t) \to x^{-\alpha}$ for all $x, t > 0$. The exponent α is called the *index of regular variation*. In the standard form (2.5) the normalizing function is $b(t) = t$ so that $\alpha = 1$. Thus condition (2.5) may seem overly stringent since it requires regular variation of V in a standard form. However it is in fact weaker than the starting point of multivariate extreme value theory. Indeed the latter framework relies on the maximum domain of attraction (MDA) condition stipulating that the componentwise maximum $M_n = \max_{i \le n} X_i$ converges in distribution, after affine normalization. Namely the MDA condition is that $(M_n - b_n)/a_n$ converges weakly to a non-degenerate limit, with $a_n \in (\mathbb{R}_+^*)^d$ and $b_n \in \mathbb{R}^d$ two sequences of vectors, and where all algebraic operations are understood componentwise (see *e.g.* de Haan and Ferreira (2006); Beirlant et al. (1996) and the references therein). This assumption is equivalent (De Haan and Resnick (1987), proposition 5.10) to a marginal MDA condition on the marginal distributions together with our standard regular variation assumption (2.5). In contrast our assumption (2.5) concerns only the dependence structure of X represented by V.

An additional assumption that we make throughout this chapter is that the marginal distributions F_j are continuous, so that the marginally ordered samples have no ties with probability one.

Assumption 2.1. *The margins of X have continuous c.d.f., namely $F_1, \dots, F_d$ are continuous.*

It should be noted that this assumption is present for convenience of the statistical analysis mainly and could be replaced with an assumption that the weight of the atoms δ_x decay sufficiently fast at infinity, which is the case anyway under marginal MDA conditions (see Leadbetter et al. (1983), Theorem 1.7.13).

The exponent measure can be characterized in many different ways. One such characterization relies on a transformation to polar coordinates: given $\| \cdot \|$ a norm on $\mathbb{R}^d$, for $v \in \mathbb{R}_+ \setminus \{0\}$, set $\mathcal{T}(v) = (r(v), \theta(v))$ where $r(v) = \|v\|$ and $\theta(v) = r(v)^{-1}v$. Let $\mathbb{S}_+$ denote the positive orthant of the unit sphere on $\mathbb{R}^d$. Then the homogeneity property of μ implies that $\mu \circ \mathcal{T}^{-1}$ is a product measure on $\mathbb{R}_+^* \times \mathbb{S}_+$, namely $\mathrm{d}(\mu \circ \mathcal{T}^{-1})(r, \theta) = \frac{\mathrm{d}r}{r^2} \otimes \mathrm{d}\Phi(\theta)$. The angular component Φ, usually called the *angular measure* has finite mass and the above definition may be rephrased as follows: for all $t > 0$ and $B \in \mathcal{B}(\mathbb{S}_+)$, where $\mathcal{B}(\mathbb{S}_+)$ is the trace σ-field of $\mathcal{B}(\mathbb{R}^d)$ on $\mathbb{S}_+$,

$$\mu\Big\{x \in \mathbb{R}_+^d : r(x) \ge t, \theta(x) \in B\Big\} = t^{-1}\Phi(B). \tag{2.7}$$

The fact that the angular measure characterizes the exponent measure suggests estimating Φ instead of μ using extreme angles $\theta(V_i)$'s such that $r(V_i)$ is large. This reduces the dimension of the problem by one. This may seem little, but it should be noticed that the removed radial dimension is the one along which the data points are likely to be the most spread out since the radial distribution behaves asymptotically as a power law, while the angular component is contained in the compact set $\mathbb{S}_+$. This line of thoughts is the one underlying the developments of the following Chapters 3, 4, 5.

In contrast in Goix et al. (2015) the focus is on the STDF denoted by l which is the evaluation of μ on L-shaped regions of the kind $[0,\infty]^d \setminus [0,y]$ (see (2.8) below). From a technical viewpoint, in a realistic setting where the marginal distributions F_j are unknown, working with such rectangular regions instead of angular regions makes it easier to control the error induced by marginal estimation. Indeed, the deviations $\widehat{F}_j(x) - F_j(x)$ may be analyzed separately and a mere union bound ensures a joint control of the marginal error. Here and throughout for $a, b \in [-\infty,\infty]^d$ such that $a_j \le b_j$ for all $j \in \{1,\dots,d\}$ the notation $[a,b]$ stands for the rectangle $\{x \in [-\infty,\infty]^d : \forall j \in \{1,\dots,d\}, a_j \le x_j \le b_j\}$. Also it is convenient to work on $[0,\infty]^d$, which is stable under the change of variable $x \mapsto 1/x$ and then for $a, b \in [0,\infty]^d$ as above, the complementary set of $[a,b]$ is understood as $[a,b]^c = [0,\infty]^d \setminus [a,b]$. Finally, when clear from the context, the notation ∞ (*resp.* 0) may indifferently denote the point at infinity (*resp.* the origin) in $\overline{\mathbb{R}}_+$ or $\overline{\mathbb{R}}_+^d$. Changing variable as described above amounts to considering $U = (U_1,\dots,U_d)$ with $U_j = 1 - F_j(X_j)$. Under the assumption that F_j is continuous, U_j is uniform on $[0,1]$, a convenient feature for using concentration theory and empirical processes. Equipped with these notations, for $x = (x_1,\dots,x_d) \in [0,\infty]^d \setminus \{\infty\}$, the STDF evaluated at x is

$$\begin{aligned} l(x) &= \mu\left([0,x^{-1}]^c\right) \\ &= \lim_{t\to 0} t^{-1}\mathbb{P}\left(U_1 \le t\,x_1 \text{ or } \dots \text{ or } U^d \le t\,x_d\right) \end{aligned} \tag{2.8}$$

The empirical estimator of l denoted by l_n below is routinely defined (see Huang (1992), Qi (1997), Drees and Huang (1998), Einmahl et al. (2006)) as

$$l_n(x) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}\{X_{i,1} \ge X_{(n-\lfloor kx_1\rfloor+1),1} \text{ or } \dots \text{ or } X_{i,d} \ge X_{(n-\lfloor kx_d\rfloor+1),d}\}, \tag{2.9}$$

which derives naturally from (2.8), up to replacing t with k/n, taking empirical counterparts of the F_j's to define rank transformed variables $\widehat{U}_{i,j}$'s and replacing the distribution Q of the *r.v.* U with $\widehat{Q}_n$, the empirical distribution of the $\widehat{U}_i$'s.

2.2.2 Contribution: finite sample guarantees on the STDF

Although extensive studies have proved consistency and asymptotic normality for the empirical version of the STDF (see Huang (1992), Drees and Huang (1998) and de Haan and Ferreira (2006) for the asymptotic normality in dimension 2, Qi (1997) for consistency in arbitrary dimension, and Einmahl et al. (2012)

for asymptotic normality in arbitrary dimension under differentiability conditions on l), Goix et al. (2015) is to my best knowledge the first contribution to a non-asymptotic analysis. In that paper, upper bounds are derived on the maximal deviation $\sup_{0\leq \mathbf{x}\leq T}|l_n(\mathbf{x})-l(\mathbf{x})|$ with expected rate of convergence in $O(k^{-1/2})$ without any smoothness condition on l.

The main idea is to adapt Theorem 2.4 to the particular setting in view. Consider Q and Q_n, respectively the distribution of the standardized vector U and the empirical measure relative to an $i.i.d.$ sample of size n and use the VC class of sets

$$\mathcal{A}=\left\{\left[\frac{k}{n}\mathbf{x},\infty\right]^c \;:\quad x\in\mathbb{R}^d_+,\quad 0\leq x_j\leq T\;(1\leq j\leq d)\right\}$$

so that $\mathcal{V}_{\mathcal{G}}=d$ (Devroye et al. (1996), Theorem 13.8). Then it is easy to show that $Q(\mathbb{A})=p\leq dT\frac{k}{n}$, so that a direct consequence of Theorem 2.4 is that for $\delta\geq e^{-k}$,

$$\sup_{x_j\in[0,T],j\in\{1,\dots,d\}}\frac{n}{k}\left|(Q_n-Q)(\frac{k}{n}[x,\infty]^c)\right|\;\leq\;Cd\sqrt{\frac{T}{k}\ln\frac{1}{\delta}}\,. \tag{2.10}$$

Inequality (2.10) is the cornerstone of the following theorem, which is the main result of Goix et al. (2015). The remaining steps of the proof aim at controlling the discrepancy between Q_n and $\widehat{Q}_n$, the latter being the actual observable statistic while the former is based on pseudo-observations U_i which are not observed since the margins F_j are unknown. This is done following the same general lines as in Qi (1997), replacing asymptotic arguments from empirical processes with non asymptotic upper bounds, again issued from empirical process theory.

Theorem 2.5. *Let T be a positive number such that $T\geq\frac{7}{2}(\frac{\ln d}{k}+1)$, and δ such that $\delta\geq e^{-k}$. If the marginal distributions are continuous (Assumption 2.1) then there is an absolute constant C such that for each $n>0$, with probability at least $1-\delta$:*

$$\sup_{0\leq\mathbf{x}\leq T}|l_n(\mathbf{x})-l(\mathbf{x})|\;\leq\;Cd\sqrt{\frac{T}{k}\ln\frac{d+3}{\delta}}\;+\;\sup_{0\leq\mathbf{x}\leq 2T}\left|\frac{n}{k}\tilde{F}(\frac{k}{n}\mathbf{x})-l(\boldsymbol{x})\right|$$

The second term of the upper bound of Theorem 2.5 is a bias term which depends on the discrepancy between the left hand side and the limit in (2.8) at level $t=k/n$. The value k can be interpreted as the effective number of observations used in the empirical estimate, *i.e.* the effective sample size for tail estimation.

Chapter 3

Classification of extreme events

The material gathered in this chapter relies on the published papers Jalalzai et al. (2018) and Jalalzai et al. (2020).

Classification is the flagship of supervised learning problems. It is also one of a most natural framework in which uniform concentration bounds such as those introduced as background in Chapter 2 reveal themselves fruitful for proving generalization guarantees of classifiers obtained *via* Empirical Risk Minimization (ERM). A natural question to ask is whether the concentration tools for rare events laid out in Section 2.1.2 can be used to propose classifiers dedicated to the tails of the explanatory variable with satisfactory generalization guarantees. To fix ideas, consider a classification problem where a random pair (X, Y) is observed, where X is an explanatory variable and $Y \in \{-1, +1\}$ is the label to be predicted. Suppose that the goal is to predict the labels associated to large explanatory variables say $\|X\| \geq t$ for some large threshold t. As detailed in the next background section, classification by ERM consists in selecting a classifier g_n from a class $\mathcal{G}$ such that the empirical risk of the classifier (the number of errors on the training set) is minimal among the class. If the focus is on the error made above threshold t, one should think of a specific strategy to avoid two pitfalls: (i) the classical ERM solution is not guaranteed to perform well in the tails because the relative weight of the training error made in this region is negligible and has thus negligible influence on the output g_n , (ii) if one restricts the training set to tail regions $\{\|x\| > t\}$, the size of the training set may be too small for large values of t to guarantee any generalization properties. After providing some background on generalization guarantees for classifiers issued from the ERM strategy (Section 3.1) we summarize in Section 3.2 the main findings of Jalalzai et al. (2018).

We conclude this chapter (Section 3.3) with an application (Jalalzai et al. (2020)) of the framework and the results from Section 3.2 in Natural Language Processing (NLP) which involves in addition a representation learning strategy. The aim of this paper is double: (i) improve classification of sentences which vectorial representation has a large norm, (ii) take advantage of the radial invariance of the classifiers dedicated to the tail to generate new data with prescribed label, which is a major challenge for dataset augmentation. In the present thesis I shall mainly focus on the methods and results related to the first goal, as the second one is more specialized to the NLP setting. A key ingredient of the proposed methodology is a representation learning device with heavy tailed target, *i.e.* such hat the obtained representation is multivariate regularly varying. Even though we have not tried to derive theoretical guarantees regarding the quality of the representation learning procedure, experimental results show that the obtained representation fulfills the regular variation requirements. This may open the road to a novel line of research aiming at broadening the impact of multivariate EVT in a regular variation setting by proposing data pre-processing strategies based on representation learning in order to ensure that the pre-processed data satisfies the regular variation conditions (2.5) or (2.6).

3.1 Classification in the ERM paradigm: background

We recall in this section the standard statistical learning framework for binary classification. All the facts stated below are proved in Lugosi (2002) or Bousquet et al. (2003). As sketched out at the beginning of this chapter, $(X, Y) \sim P$ is a random pair on a product space $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y}$ a set made of two elements, say $\mathcal{Y} = \{-1, +1\}$. As it is the case in Chapter 2, here and throughout we take $\mathcal{X} = \mathbb{R}^d$. The explanatory variable X is assumed to contain relevant information for predicting Y. Given a family $\mathcal{G}$ of classifiers $g : \mathcal{X} \to \{-1, 1\}$ the goal is to select g_n sufficiently close to the minimizer of the $0-1$ risk $R(g) = \mathbb{P}\left(g(X) \neq Y\right)$ over the class $\mathcal{G}$. If the latter class is the whole family $\mathcal{G}^*$ of measurable functions $g : \mathcal{X} \to \mathcal{Y}$ the solution of the risk minimization problem is the so called *Bayes classifier* $g^* : x \mapsto 2\mathbb{1}\{\eta(x) \geq 1/2\} - 1$ where η is the *regression function*, $\eta(x) = \mathbb{P}\left(Y = 1 \mid X = x\right)$. However P is unknown and in a supervised setting, one can only use a training set $(X_i, Y_i)_{i \leq n}$ made of n $i.i.d.$ copies of (X, Y). The ERM strategy consists in selecting g_n as the minimizer over $\mathcal{G}$ of the empirical risk

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(X_i) \neq Y_i\} = P_n(\mathbb{1}\{g(x) \neq y\})$$

The latter expression suggests to consider, for each classifier $g \in \mathcal{G}$, the set $A_g = \{z = (x, y) : g(x) \neq y\} \subset \mathcal{X} \times \mathcal{Y}$. The family of classifiers $\mathcal{G}$ is thus in one-to-one correspondence with the class of sets $\mathcal{G} = \{A_g, g \in \mathcal{G}\}$. With these notations,

$$P - P_n(A_g) = R(g) - R_n(g).$$

Thus, as soon as the class $\mathcal{G}$ is simple enough so that $\mathcal{G}$ has finite VC dimension, Vapnik's result (Theorem 2.3) provides a probabilistic upper bound on $\sup_{g\in\mathcal{G}} |R-R_n|(g) = \sup_{A\in\mathcal{G}} |P - P_n|(A)$. One may wonder at some point why such a uniform control is necessary. Here is an answer: in practice, a quantity of interest for which upper bounds are welcome is the deviation of the true risk $R(g_n)$ of the selected classifier from its empirical version $R_n(g_n)$ based on the training sample, and the excess risk $R(g_n) - R^*$, where $R^* = \inf_{g\in\mathcal{G}} R(g)$. The starting point for deriving such guarantees is that, on the one hand

$$R(g_n) \leq R_n(g_n) + \sup_g (R(g) - R_n(g)).$$

On the other hand, for $\epsilon > 0$, consider g_ϵ an epsilon minimizer of R, that is $R(g_\epsilon) \leq R^* + \epsilon$. Then

$$\begin{aligned} R(g_n) - R^* &\leq R(g_n) - R(g_\epsilon) + \epsilon \\ &\leq \big(R(g_n) - R_n(g_n)\big) + \underbrace{\big(R_n(g_n) - R_n(g_\epsilon)\big)}_{\leq 0} + \big(R_n(g_\epsilon) - R(g_\epsilon)\big) + \epsilon \end{aligned}$$

Letting $\epsilon \to 0$ yields a control of the excess risk,

$$R(g_n) - R^* \leq 2 \sup_{g\in\mathcal{G}} |R(g) - R_n(g)| .$$

3.2 Binary classification in extreme regions

Following the opening argument of this chapter, the main focus of Jalalzai et al. (2018) is the problem of producing a classifier minimizing the probability of an error conditional to a excess above a radial threshold,

$$R_t(g) := R_{P_t}(g) = \mathbb{P}\left(Y \neq g(X) \mid \|X\| > t\right), \tag{3.1}$$

as $t \to \infty$, where P_t denotes the conditional distribution of (X, Y) given that $\|X\| > t$, and where for any probability distribution on $\mathcal{X}\times\mathcal{Y}$, $R_Q(g) = Q\{(x,y) : y \neq g(x)\}$. Thus we introduce the risk at infinity,

$$R_\infty(g) = \limsup_{t\to\infty} R_t(g), \qquad g \text{ any classifier.} \tag{3.2}$$

Notice already that the Bayes classifier g^* defined in Section 3.1 relative to P is a minimizer of R_∞. Indeed for $\|x\| > t$ the regression functions relative to P and P_t coincide, and so do the Bayes classifiers, $g^*_P = g^*_{P_t}$, so that $R_t(g^*) \leq R_t(g)$ for any classifier g. Taking the limit superior as $t \to \infty$, the desired result follows. However their is still no guarantee that the ERM classifier g_n performs well in the tail, especially if $\mathcal{G}$ is a parametric class, because the number of potential errors made by g_n in the tail is by definition negligible compared to the number of potential errors in the bulk.

To avoid the second pitfall (data scarcity) mentioned above, we need to make assumptions about P_t as $t \to \infty$. In our context it is rather natural to assume that the class distributions $\mathbb{P}\left(X \in \cdot \mid Y = \sigma 1\right)$, $\sigma \in \{-1, +1\}$ are regularly varying (see (2.6)). Also for the problem to be meaningful one needs to ensure that the ratio $\mathbb{P}\left(Y = +1 \mid \|X\| > t\right) / \mathbb{P}\left(Y = -1 \mid \|X\| > t\right)$ has a limit in $(0, \infty)$, otherwise the problem is either trivial (one class has asymptotic weight equal to 1) or insoluble (the quantities $\mathbb{P}\left(Y = +1 \mid X \in tA\right)$ have no limit as $t \to \infty$). Equivalently, one must assume that the ratio of the normalizing functions $b_+(t)/b_-(t)$ converges to a finite, non zero limit. Thus necessarily the indices of regular variation are the same, $\alpha_+ = \alpha_-$. In Jalalzai et al. (2018) we make the simplifying assumption that the tail index is equal to 1 and that the normalizing functions $b_+(t), b_-(t)$ in (2.6) may both be chosen as $b_+(t) = b_-(t) = t$. This would indeed be the case if the explanatory variable had been marginally standardized as in (2.4), so that one would work with a the random pair (V, Y). This explains the notation μ instead of ν in Assumption 3.1 below and the term 'angular measure' referring to the angular component of the limit measure. Of course in practice the margins are unknown and taking into account the marginal error is the subject of ongoing work (see Section 7.1). Summarizing, the first assumption in Jalalzai et al. (2018) is

Assumption 3.1. *For all $\sigma \in \{-,\ +\}$, the conditional distribution of X given $Y = \sigma 1$ is regularly varying with limit measure μ_σ, angular measure $\Phi_\sigma(d\theta)$ (respectively, limit measure $\mu_\sigma(dx)$) and normalizing function $b(t) = t$: for $A \subset [0, \infty]^d \setminus \{0\}$ a measurable set such that $0 \notin \partial A$ and $\mu(\partial A) \neq 0$,*

$$t\mathbb{P}\left(t^{-1}X \in A \mid Y = \sigma\, 1\right) \xrightarrow[t\to\infty]{} \mu_\sigma(A), \qquad \sigma \in \{-, +\},$$

and for $B \subset S$ a measurable set,

$$\Phi_\sigma(B) = \mu_\sigma\{x \in \mathbb{R}^d_+ : R(x) > 1, \theta(x) \in B\}, \qquad \sigma \in \{-, +\},$$

Remark 3.1. An inspection of the proofs of Jalalzai et al. (2018) shows that the choice of the functions $b_+(t), b_-(t)$ plays no role since only conditional probabilities above t come into play, as long as $b_+(t)/b_-(t) \to \ell \in (0, \infty)$. Thus the results of the paper are unchanged when replacing the assumption that $b_\sigma(t) = t$ with the latter condition concerning the limit of their ratio, up to a minor modification of the definitions of the limiting pair (X_∞, Y_∞) introduced below which should take into account the limit ℓ.

In Jalalzai et al. (2018)'s framework, denoting by p the marginal probability $p = \mathbb{P}(Y = +1)$, it is easy to see that $\mathbb{P}(Y = +1 \mid \|X\| > t) \to p_\infty = p\Phi_+(\mathbb{S}_+)/\Phi_-(\mathbb{S}_+)$. It is quite natural to define a limiting pair (X_∞, Y_∞) on $\mathbb{R}^d_+ \cap \{x : \|x\| \geq 1\} \times \{-1, 1\}$ through its distribution :

$$\begin{aligned}
&\mathbb{P}(Y_\infty = 1) = p_\infty \\
&\mathbb{P}(X_\infty \in A \mid Y_\infty = y) = \lim_t \mathbb{P}(X \in tA, Y_\infty = y \mid \|X| > t) \\
&\qquad = \frac{\mu_{\mathrm{sign}(y)}(A)}{\Phi_{\mathrm{sign}(y)}(\mathbb{S}_+)}
\end{aligned} \tag{3.3}$$

Then by homogeneity of μ, it can be shown that the regression function $\eta_\infty(x) = \mathbb{P}(Y_\infty = 1 \mid X_\infty = x)$ relative to (X_∞, Y_∞) depends on the angle $\theta(x) = \|x\|^{-1}x$ only. A reasonable conjecture is that the Bayes classifier $g^*_\infty(x) = 2\mathbb{1}\{\eta_\infty(x) \geq 1/2\} - 1$ relative to the the distribution P_∞ of the pair (X_∞, Y_∞) is also optimal for the asymptotic risk R_∞. We prove that it is the case under the following regularity assumption

Assumption 3.2. (UNIFORM CONVERGENCE ON THE SPHERE OF $\eta(tx)$) *The limiting regression function η_∞ is continuous on S and*

$$\sup_{\theta \in \mathbb{S}_+} |\eta(t\theta) - \eta_\infty(\theta)| \xrightarrow[t \to \infty]{} 0$$

Assumption 3.2 is satisfied under the condition of uniform convergence of densities required in the framework of De Haan and Resnick (1987); Cai et al. (2011). We may now state the main result of Jalalzai et al. (2018) concerning optimal classification at extreme levels. To understand the statement, the reader should keep in mind that we already have $R^*_t = R_t(g^*)$ and $R^*_{P_\infty} = R_{P_\infty}(g^*_\infty)$ from the definitions and the above argument.

Theorem 3.2. *(Jalalzai et al. (2018)) Under Assumptions 3.1 and 3.2,*

$$R^*_t = R_t(g^*) \xrightarrow[t \to \infty]{} R^*_{P_\infty}. \tag{3.4}$$

Hence, we have: $R^*_\infty = R^*_{P_\infty}$. In addition, the classifier g^*_∞ minimizes the asymptotic risk in the extremes:

$$\inf_{g \text{ measurable}} R_\infty(g) = L_\infty(g^*_\infty) = \mathbb{E}\min(\eta_\infty(\Theta_\infty), 1 - \eta_\infty(\Theta_\infty)). \tag{3.5}$$

Equation (3.4) means that limit of the infimum R^*_t of the risk above level t coincide with the infimum of risk R_{P_∞} relative to the limit distribution. Equation (3.5) ensures that the minimizer g_∞ of the risk relative to the limit distribution P_∞ also minimizes the limit risk R_∞. A very useful consequence in practice is that there exists an optimal classifier for R_∞ which depends solely on the angle $\theta(x)$ of the explanatory variable. This suggests an ERM strategy based on restricting the attention to angular classifiers, *i.e.* classifiers depending on the angular component of x only. In the sequel, the notation $\mathcal{G}_{\mathbb{S}}$ stands for any family of such classifiers. As is customary in tail analysis, we consider an empirical version of R_t based on the k observations (X_i, Y_i) such that the norms $\|X_i\|$ rank among the k largest. Introducing the order statistics $(X_{(i)}, Y_{(i)})$ such that $\|X_{(1)}\| > \ldots > \|X_{(n)}\|$ we fix $\tau > 0$ a small probability and we let $k = \lfloor n\tau \rfloor$. If t_τ is the $1-\tau$ quantile of the *r.v.* $\|X\|$, then

$$\widehat{R}_k(g) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\{Y_{(i)} \neq g(X_{(i)})\} = R_{\widehat{P}_k}(g), \tag{3.6}$$

is the empirical version of R_{t_τ}. Then we suggest performing classification above large threshold t using the ERM classifier

$$\widehat{g}_k \in \operatorname{argmin}_{g\in\mathcal{G}_{\mathbb{S}}} \widehat{R}_k(g)$$

This strategy is guaranteed to be successful as stated below. The proof relies on the concentration inequalities for rare events stated in Section 2.1

Theorem 3.3 (ERM classification for extremes, Jalalzai et al. (2018))**.** *Suppose that the angular class $\mathcal{G}_S$ is of finite* VC *dimension $V_{\mathcal{G}_S} < +\infty$. Let $\widehat{g}_k$ be any minimizer of* (3.6)*. Then, for $\delta \in (0,1)$, $\forall n \geq 1$, we have with probability larger than $1-\delta$:*

$$R_{t_\tau}(\widehat{g}_k) - R^*_{t_\tau} \leq \frac{1}{\sqrt{k}}\left(\sqrt{2(1-\tau)\ln(2/\delta)} + C\sqrt{V_{\mathcal{G}_S}\ln(1/\delta)}\right)$$
$$+\frac{1}{k}\left(5 + 2\ln(1/\delta) + \sqrt{\ln(1/\delta)}(C\sqrt{V_{\mathcal{G}_S}} + \sqrt{2})\right) + \left\{\inf_{g\in\mathcal{G}_S} R_{t_\tau}(g) - R^*_{t_\tau}\right\},$$

where C is a constant independent from n, τ and δ.

The last term is a bias term relative to the richness of the class $\mathcal{G}_{\mathbb{S}}$. Under the assumption that the class is rich enough to discriminate the pairs in the tails, we obtain

Corollary 3.4. *Under the assumptions of Theorems 3.2 and 3.3, assume in addition that the model bias asymptotically vanishes as $\tau \to 0$, i.e.*

$$\inf_{g\in\mathcal{G}_S} R_{t_\tau}(g) - R^*_{t_\tau} \longrightarrow 0 \quad \text{as } \tau \to 0.$$

Then, as soon as $k \to +\infty$ as $n \to \infty$, the sequence of classifiers $(\widehat{g}_k)$ is asymptotically consistent,

$$R_\infty(\widehat{g}_k) \to R^*_\infty \text{ as } n \to \infty.$$

3.3 Heavy-tailed representations, classification and data augmentation in a NLP framework

The present section relies on the material published in Jalalzai et al. (2020).

Introduction Representing the meaning of natural language in a mathematically grounded way is a scientific challenge that has received increasing attention with the explosion of digital content and text data in the last decade. Relying on the richness of contents, several embeddings have been proposed Peters et al. (2018); Radford et al. (2018); Devlin et al. (2018) with demonstrated efficiency for the considered tasks when learnt on massive datasets. However, none of these embeddings take into account the fact that word frequency distributions are heavy tailed Baayen (2002); Church and Gale (1995); Mandelbrot (1953), so that extremes are naturally present in texts. Similarly, Babbar et al. (2014) shows that, contrary to image taxonomies, the underlying distributions for words and documents in large scale textual taxonomies are also heavy tailed. Exploiting this information, several studies, as Clinchant and Gaussier (2010); Madsen et al. (2005), were able to improve text mining applications by accurately modeling the tails of textual elements.

In this work we rely on the multivariate EVT framework for classification presented in Section 3.2. The tail region (where samples are considered as extreme) of the input variable $x \in \mathbb{R}^d$ is of the kind $\{\|x\| \geq t\}$, for a large threshold t. The latter is typically chosen such that a small but non negligible proportion of the data is considered as extreme, namely 25% in our experiments. A major advantage of this framework in the case of labeled data is that classification on the tail regions may be performed using the angle $\Theta(x) = \|x\|^{-1}x$ only. The main idea behind the present paper is to take advantage of the scale invariance for two tasks regarding sentiment analysis of text data: *(i)* Improved classification of extreme inputs, *(ii)* Label preserving data augmentation, as the most probable label of an input x is unchanged by multiplying x by $\lambda > 1$. Jalalzai et al. (2018) demonstrate the usefulness of their framework with simulated and some real world datasets. However, there is no reason to assume that the previously mentioned text embeddings satisfy the required regularity assumptions. The aim of the present work is to extend Jalalzai et al. (2018)'s methodology to datasets which do not satisfy their assumptions, in particular to text datasets embedded by state of the art techniques. This is achieved by the algorithm *Learning a Heavy Tailed Representation* (in short **LHTR**) which learns a transformation mapping the input data X onto a random vector Z which does satisfy the aforementioned assumptions. The transformation is learnt by an adversarial strategy Goodfellow et al. (2016). In the appendix section of the paper we propose an interpretation of the extreme nature of an input in both **LHTR** and BERT representations. In a word, these sequences are longer and are more difficult to handle (for next token prediction and classification tasks) than non extreme ones.

Our second contribution is a novel data augmentation mechanism **GENELIEX** which takes advantage of the scale invariance properties of Z to generate synthetic sequences that keep invariant the attribute of the original sequence. Label preserving data augmentation is an effective solution to the data scarcity problem and is an efficient pre-processing step for moderate dimensional datasets Wang and Perez (2017); Wei and Zou (2019). Adapting these methods to NLP problems remains a challenging issue.The problem consists in constructing a transformation h such that for any sample x with label $y(x)$, the generated sample $h(x)$ would remain label consistent: $y\big(h(x)\big) = y(x)$ Ratner et al. (2017). The dominant approaches for text data augmentation rely on word level transformations such as synonym replacement, slot filling, swap deletion Wei and Zou (2019) using external resources such as wordnet Miller (1995). Linguistic based approaches can also be combined with vectorial representations provided by language models Kobayashi (2018). However, to the best of our knowledge, building a vectorial transformation without using any external linguistic resources remains an open problem. In this work, as the label $y\big(h(x)\big)$ is unknown as soon as $h(x)$ does not belong to the training set, we address this issue by learning both an embedding φ and a classifier g satisfying a relaxed version of the problem above mentioned, namely $\forall \lambda \geq 1$

$$g\big(h_\lambda(\varphi(x))\big) = g\big(\varphi(x)\big). \tag{3.7}$$

In order to exploit the scale invariance of the regression function in the tails, h_λ is chosen as the homothety with scale factor λ, $h_\lambda(x) = \lambda x$. In this paper, we work with output vectors issued by BERT Devlin et al. (2018). BERT and its variants are currently the most widely used language model but we emphasize that the proposed methodology could equally be applied using any other representation as input. BERT embedding does not satisfy the regularity properties required by EVT, as demonstrated empirically in the appendix section of the paper. Besides, there is no reason why a classifier g trained on such embedding would be scale invariant, *i.e.* would satisfy for a given sequence u, embedded as x, $g(h_\lambda(x)) = g(x)$ $\forall \lambda \geq 1$. On the classification task, we demonstrate on two datasets of sentiment analysis that the embedding learnt by **LHTR** on top of BERT is indeed following a heavy-tailed distribution. Besides, a classifier trained on the embedding learnt by **LHTR** outperforms the same classifier trained on BERT. On the dataset augmentation task, quantitative and qualitative experiments demonstrate the ability of **GENELIEX** to generate new sequences while preserving labels.

Learning a Regularly varying representation We now introduce a novel algorithm *Learning a heavy-tailed representation* (**LHTR**) for text data from high dimensional vectors as issued by pre-trained embeddings such as BERT. The idea behind is to modify the output X of BERT so that classification in the tail regions enjoys the statistical guarantees presented in Section 3.2, while classification in the bulk (where many training points are available) can still be performed using standard models. Stated otherwise, **LHTR** increases the information carried by the resulting vector $Z = \varphi(X) \in \mathbb{R}^{d'}$ regarding the label Y in the tail regions of Z in order to improve the performance of a downstream classifier. In addition **LHTR** is a building block of the data augmentation algorithm **GENELIEX** (not detailed in the present thesis). **LHTR** proceeds by training an encoding function φ in such a way that *(i)* the marginal distribution $q(z)$ of the code Z be close to a user-specified heavy tailed target distribution p satisfying the regularity condition (2.6) with $b(t) = t$, and *(ii)* the classification loss of a multilayer perceptron trained on the code Z be small.

A major difference distinguishing **LHTR** from existing auto-encoding schemes is that the target distribution on the latent space is not chosen as a Gaussian distribution but as a heavy-tailed, regularly varying one. A workable example of such a target is provided in our experiments As the Bayes classifier (*i.e.* the optimal one among all possible classifiers) in the extreme region has a potentially different structure from the Bayes classifier on the bulk (recall from Section 3.2 that the optimal classifier at infinity depends on the angle $\Theta(x)$ only), **LHTR** trains two different classifiers, g^{ext} on the extreme region of the latent space on the one hand, and g^{bulk} on its complementary set on the other hand. Given a high threshold t, the extreme region of the latent space is defined as the set $\{z : \|z\| > t\}$. In practice, the threshold t is chosen as an empirical quantile of order $(1-\kappa)$ (for some small, fixed κ) of the norm of encoded data $\|Z_i\| = \|\varphi(X_i)\|$. The classifier trained by **LHTR** is thus of the kind $g(z) = g^{\text{ext}}(z)\mathbb{1}\{\|z\| > t\} + g^{\text{bulk}}(z)\mathbb{1}\{\|z\| \leq t\}$. If the downstream task is classification on the whole input space, in the end the bulk classifier g^{bulk} may be replaced with any other classifier g' trained on the original input data X restricted to the non-extreme samples (*i.e.* $\{X_i, \|\varphi(X_i)\| \leq t\}$). Indeed training g^{bulk} only serves as an intermediate step to learn an adequate representation φ.

Remark 3.5. Recall from Section 3.2 that the optimal classifier in the extreme region as $t \to \infty$ depends on the angular component $\theta(x)$ only, or in other words, is scale invariant. One can thus reasonably expect the trained classifier $g^{\text{ext}}(z)$ to enjoy the same property. This scale invariance is indeed verified in our experiments and is the starting point for our data augmentation algorithm **GENELIEX**. An alternative strategy would be to train an angular classifier, *i.e.* to impose scale invariance. However in preliminary experiments (not shown here), the resulting classifier was less efficient and we decided against this option in view of the scale invariance and better performance of the unconstrained classifier.

The goal of **`LHTR`** is to minimize the weighted risk

$$\begin{aligned} R(\varphi, g^{\text{ext}}, g^{\text{bulk}}) = & \rho_1 \mathbb{P}\left(Y \neq g^{\text{ext}}(Z), \|Z\| \geq t\right) + \\ & \rho_2 \mathbb{P}\left(Y \neq g^{\text{bulk}}(Z), \|Z\| < t\right) + \\ & \rho_3 \mathfrak{D}(q(z), p(z)) \end{aligned}$$

where $Z = \varphi(X)$, $\mathfrak{D}$ is the Jensen-Shannon distance between the heavy tailed target distribution p and the code distribution q, and ρ_1, ρ_2, ρ_3 are positive weights. Following common practice in the adversarial literature, the Jensen-Shannon distance is approached (up to a constant term) by the empirical proxy $\widehat{L}(q,p) = \sup_{D \in \Gamma} \widehat{L}(q,p,D)$, with $\widehat{L}(q,p,D) = \frac{1}{m}\sum_{i=1}^{m} \log D(Z_i) + \log\left(1 - D(\tilde{Z}_i)\right)$, where Γ is a wide class of discriminant functions valued in $[0,1]$, and where independent samples $Z_i, \tilde{Z}_i$ are respectively sampled from the target distribution and the code distribution q. The classifiers g^{ext}, g^{bulk} are of the form $g^{\text{ext}}(z) = 2\mathbb{1}\{C^{\text{ext}}(z) > 1/2) - 1$, $g^{\text{bulk}}(z) = 2\mathbb{1}\{C^{\text{bulk}}(z) > 1/2) - 1$ where $C^{\text{ext}}, C^{\text{bulk}}$ are also discriminant functions valued in $[0,1]$. Following common practice, we shall refer to $C^{\text{ext}}, C^{\text{bulk}}$ as classifiers as well. In the end, **`LHTR`** solves the following min-max problem $\inf_{C^{\text{ext}}, C^{\text{bulk}}, \varphi} \sup_D \widehat{R}(\varphi, C^{\text{ext}}, C^{\text{bulk}}, D)$ with

$$\begin{aligned} \widehat{R}(\varphi, C^{\text{ext}}, C^{\text{bulk}}, D) = & \frac{\rho_1}{k}\sum_{i=1}^{k} \ell(Y_{(i)}, C^{\text{ext}}(Z_{(i)})) \; + \cdots \\ & \frac{\rho_2}{n-k}\sum_{i=k+1}^{n-k} \ell(Y_{(i)}, C^{\text{bulk}}(Z_{(i)})) + \cdots \\ & \rho_3 \, \hat{L}(q,p,D), \end{aligned}$$

where $\{Z_{(i)} = \varphi(X_{(i)}), i = 1, \dots, n\}$ are the encoded observations with associated labels $Y_{(i)}$ sorted by decreasing magnitude of $\|Z\|$ (*i.e.* $\|Z_{(1)}\| \geq \cdots \geq \|Z_{(n)}\|$), $k = \lfloor \kappa n \rfloor$ is the number of extreme samples among the n encoded observations and $\ell(y, C(x)) = -(y \log C(x) + (1-y)\log(1 - C(x)), y \in \{0,1\}$ is the negative log-likelihood of the discriminant function $C(x) \in (0,1)$. A summary of **`LHTR`** and an illustration of its workflow are provided in the appendix sections of the paper.

Summary of experiments In our experiments we work with the infinity norm. The proportion of extreme samples in the training step of **`LHTR`** is chosen as $\kappa = 1/4$. The threshold t defining the extreme region $\{\|x\| > t\}$ in the test set is $t = \|\tilde{Z}_{(\lfloor \kappa n \rfloor)}\|$ as returned by **`LHTR`**.

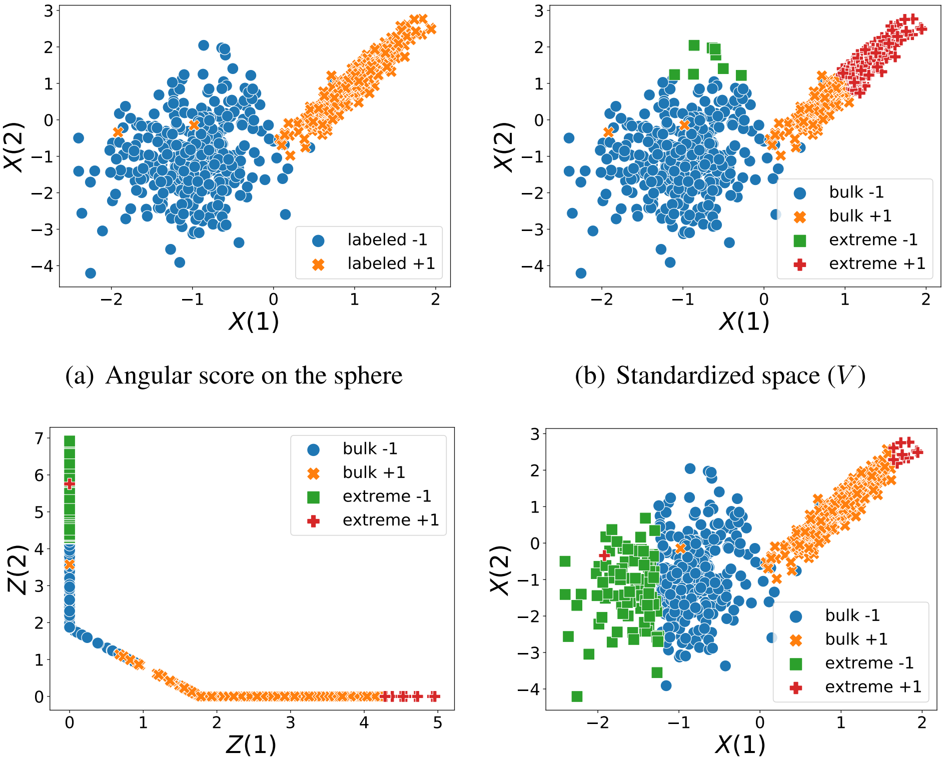
Classifiers $C^{\text{bulk}}, C^{\text{ext}}$ involved in **LHTR** are Multi Layer Perceptrons (MLP). The regularly varying target distribution is chosen as a multivariate logistic distribution $F(x) = \exp\left\{-(\sum_{j=1}^{d} {x_j}^{\frac{1}{\delta}})^{\delta}\right\}$ with parameter $\delta = 0.9$. This distribution is widely used in the context of extreme values analysis and differ from the classical logistic distribution.

We start with a simple bivariate illustration of the heavy tailed representation learnt by **LHTR**. Our goal is to provide insight on how the learnt mapping φ acts on the input space and how the transformation affects the definition of extremes (recall that extreme samples are defined as those samples which norm exceeds an empirical quantile).

Labeled samples are simulated from a Gaussian mixture distribution with two components of identical weight. The label indicates the component from which the point is generated. **LHTR** is trained on 2250 examples and a testing set of size 750 is shown in Figure 3.1. The testing samples in the input space (Figure 3.1(a)) are mapped onto the latent space *via* φ (Figure 3.1(c)) In Figure 3.1(b), the extreme raw observations are selected according to their norm after a component-wise standardisation of X_i. The extreme threshold t is chosen as the 75% empirical quantile of the norm on the training set in the input space. Notice in the latter figure the class imbalance among extremes. In Figure 3.1(c), extremes are selected as the 25% samples with the largest norm in the latent space. Figure 3.1(d) is similar to Figure 3.1(b) except for the selection of extremes which is performed in the latent space as in Figure 3.1(c). On this toy example, the adversarial strategy appears to succeed in learning a code which distribution is close to the logistic target, as illustrated by the similarity between Figure 3.1(c) and Figure 3.2.In addition, the heavy tailed representation allows a more balanced selection of extremes than the input representation.

We next compare the performance of three models on NLP data. The baseline **NN model** is a MLP trained on BERT. The second model **LHTR**$_1$ is a variant of **LHTR** where a single MLP (C) is trained on the output of the encoder φ, using all the available data, both extreme and non extreme ones. The third model (**LHTR**) trains two separate MLP classifiers C^{ext} and C^{bulk} respectively dedicated to the extreme and bulk regions of the learnt representation φ. All models take the same training inputs, use BERT embedding and their classifiers have identical structure.

Comparing **LHTR**$_1$ with **NN model** assesses the relevance of working with heavy-tailed embeddings. Since **LHTR**$_1$ is obtained by using **LHTR** with $C^{\text{ext}} = C^{\text{bulk}}$, comparing **LHTR**$_1$ with **LHTR** validates the use of two separate classifiers so that extremes are handled in a specific manner. As we make no claim concerning the usefulness of **LHTR** in the bulk, at the prediction step we suggest working with a combination of two models: **LHTR** with C^{ext} for extreme samples and any other off-the-shelf ML tool for the remaining samples (*e.g.* **NN model**).



(a) Angular score on the sphere

(b) Standardized space (V)

(c) Input space (X).

(d) Input space (X).

Figure 3.1 – 3.1(a): Bivariate samples X_i normally distributed and designed for binary classification, in the input space. 3.1(b): X_i's in the input space with extremes from each class selected in the input space. 3.1(c): Latent space representation $Z_i = \varphi(X_i)$. Extremes of each class are selected in the latent space. 3.1(d): X_i's in the input space with extremes from each class selected in the latent space.

In our experiments we rely on two large datasets from *Amazon* (231k reviews) McAuley and Leskovec (2013) and from *Yelp* (1,450k reviews) Yu et al. (2014); Liu et al. (2015). Reviews, (made of multiple sentences) with a rating greater than or equal to $\frac{4}{5}$ are labeled as $+1$, while those with a rating smaller or equal to $\frac{2}{5}$ are labeled as -1. The gap in reviews' ratings is designed to avoid any overlap between labels of different contents.

Results. To illustrate the generalization ability of the proposed classifier in the extreme regions we consider nested subsets of the extreme test set $\mathcal{T}_{\text{test}}$, $\mathcal{T}^{\lambda} = \{z \in \mathcal{T}_{\text{test}}, \|z\| \geq \lambda t\}$, $\lambda \geq 1$. For all factor $\lambda \geq 1$, $\mathcal{T}^{\lambda} \subseteq \mathcal{T}_{\text{test}}$. The greater λ, the fewer the samples retained for evaluation and the greater their norms. On both datasets,

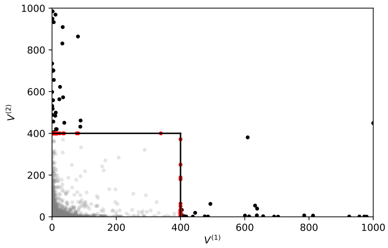


Figure 3.2 – Illustration of the distribution of the angle $\Theta(X)$ obtained with bivariate samples X generated from a logistic model with coefficient of dependence $\delta = 0.9$ Non extreme samples are plotted in gray, extreme samples are plotted in black and the angles $\Theta(X)$ (extreme samples projected on the sup norm sphere) are plotted in red. Note that not all extremes are shown since the plot was truncated for a better visualization. However all projections on the sphere are shown.

`LHTR`$_1$ outperforms the baseline **`NN model`**(see the cited paper for details), which shows the improvement offered by the heavy-tailed embedding on the extreme region. In addition, **`LHTR`**$_1$ is in turn largely outperformed by the classifier **`LHTR`**, which proves the importance of working with two separate classifiers. Finally the classification scores of the proposed model respectively on the bulk region, tail region and overall shows that using a specific classifier dedicated to extremes improves the overall performance.

Chapter 4

Dimensionality reduction

When monitoring a multivariate random vector $X = (X_1, \dots, X_d)$, the distributional structure of the tail is of particular importance due to the potentially disastrous impact of tail events involving several variables, for many applications ranging from insurance and finance to environmental risk management, network surveillance (Finkenstadt and Rootzén, 2003; Smith, 2003) or anomaly detection (Clifton et al., 2011; Lee and Roberts, 2008). As explained in Section 2.2.1, the angular measure Φ encapsulates key information regarding the structure of such extremes since after a suitable marginal standardization as in (2.4) and appropriate regular variation assumptions, the standardized vector V satisfies

$$\mathbb{P}\left(\theta(V) \in A \mid \|V\| > r\right) \approx c\,\Phi(A)$$

for some normalizing constant c and all measurable subset of the sphere $\mathbb{S}_+$ such that $\Phi(\partial A) = 0$. As it is usually the case in multivariate statistics, estimators of Φ or any other summary of the tail dependence structure such as the STDF are prone to suffer from the curse of dimensionality as d increases. In particular, most parametric models available for Φ or l (see *e.g.* Coles and Tawn (1991); Fougères et al. (2009); Cooley et al. (2010); Sabourin and Naveau (2014)) have been designed for the moderate dimensionality case, and existing proofs of asymptotic normality of the non parametric version of Φ are only available in dimension $d = 2$ (Einmahl et al. (2001),Einmahl and Segers (2009)) with techniques of proofs that do not allow for an easy extension to the general d-dimensional setting. This chapters gathers three lines of work aiming at reducing the dimensionality of the tail estimation problem.

Section 4.1 presents the main findings of the Goix et al. (2016) and Goix et al. (2017), the former being a short version of the latter which does not take into account marginal uncertainty. The idea behind these papers is that in high dimension (think *e.g.* of the discrete output of a climate model, where each X_j is the value of a physical field at location i), some subgroups of components, say

$\{X_j, j \in c_m\}$, $m = 1\dots, M$ are much likelier to exceed simultaneously a large threshold than others. In a regular variation setting, this qualitative feature can be formalized as the fact that the angular measure concentrates on a relatively small number M (compared to 2^d) of subcones of the positive orthant $\mathbb{R}^d_+$. In Goix et al. (2016, 2017) the family of such subcones is retrieved by empirical estimation of a thickened version of these sets. Non-asymptotic guarantees are derived using an extension of the concentration inequalities for the STDF presented in Section 2.2 to the empirical measure of the latter thickened sets.

This strategy is not always successful, in particular when the empirical angular measure spreads a small amount of mass onto a large number of subcones. The main purpose of Chiapino and Sabourin (2016); Chiapino et al. (2019b) (Section 4.2) is to overcome this issue by a clustering strategy which is similar in spirit to the apriori algorithm (Agrawal et al. (1994)). In Chiapino et al. (2019b) the asymptotic distribution of the stopping criterion for the algorithm proposed in Chiapino and Sabourin (2016) is derived by leveraging the results of Einmahl et al. (2012) on the asymptotic distribution of the empirical STDF. In addition alternative criteria are proposed which are based on a multivariate version of the coefficient of tail dependence (Ledford and Tawn (1996); Ramos and Ledford (2009); De Haan and Zhou (2011); Eastoe and Tawn (2012)) in order to provide asymptotic statistical guarantees in a Neyman-Pearson framework.

Finally Section 4.3 presents an alternative approach proposed in the accepted paper Drees and Sabourin (2021) which consists in applying Principal Component Analysis (PCA) to a suitably rescaled version of a regularly varying vector. We show in the cited reference that doing so, one recovers a good approximation of the support of μ, in the sense that the excess of reconstruction risk for the square error loss is bounded from above with high probability for finite sample size. Blanchard et al. (2007) provide such guarantees for standard PCA and our analysis follows their footsteps while using specific concentration tools for rare classes, namely a Bernstein-type concentration inequality from McDiarmid (1998) already mentioned in Section 2.1.

As hinted in the introduction, dimension reduction in multivariate extremes has drawn considerable attention in the past few years and various trails have been followed towards this end: Simpson et al. (2020) propose a an alternative to Goix et al. (2016, 2017)'s method based on modeling the regular variation indices of each subcone, Meyer and Wintenberger (2019) propose in an unpublished paper an alternative definition of sparsity with convenient algorithmic features. The PCA strategy has been investigated by Cooley and Thibaud (2019) with illustrations on financial and precipitation data. Several clustering approaches have been proposed (Chautru et al. (2015), Janßen and Wan (2020), Fomichov and Ivanovs (2020)). In a quite different spirit, graphical models can be used to seek sparsity patterns in the dependence graph at extreme levels, see *e.g.* Engelke and Hitz (2020). Finally

Engelke and Ivanovs (2020) propose an extensive review of the state-of-the art dimension reduction strategies available for multivariate extremes.

4.1 Sparse representation of multivariate extremes

The motivating assumption behind Goix et al. (2016, 2017) is that the dependence structure of extremes is such that
(i) Only a small number of groups of components may be concomitantly extreme, so that only a small number of hyper-cubes (those corresponding to these subsets of indexes precisely) have non zero mass (the adjective small is relative to the total number of groups 2^d).
(ii) Each of these groups contains a limited number of coordinates (compared to the original dimensionality), so that the corresponding hyper-cubes with non zero mass have small dimension compared to d.

These informal assumptions can be made rigorous as follows. In the remaining of this section the norm we consider is the infinity norm, $\|\cdot\| = \|\cdot\|_\infty$, and the positive orthant $\mathbb{S}_+$ of the sphere and the angular measure are defined accordingly. When clear from the context, the comparison operators $\leq, \geq, <, >$ are those of the partial ordering on $\mathbb{R}^d$: for vectors $x, y \in \mathbb{R}^d$ write $x \leq y$ if $x_j \leq y_j$ for all $j \in \{1, \dots, d\}$. We introduce the truncated subcones of $\mathbb{R}^d_+$

$$\mathcal{C}_a = \{v \geq 0,\ \|v\|_\infty \geq 1,\ v_j > 0 \text{ for } j \in a,\ v_j = 0 \text{ for } j \notin a\}. \tag{4.1}$$

One remarkable property of the $\mathcal{C}_a$'s is that they form a partition of the truncated positive orthant: for $\emptyset \neq a \neq b \subset \{1, \dots, d\}$, $\mathcal{C}_a \cap \mathcal{C}_b = \emptyset$ and $\bigcup_{\emptyset \neq a \subset \{1,\dots,d\}} \mathcal{C}_a = \{x \in \mathbb{R}^d_+ : \|x\| \geq 1\}$. Let Ω_a denote the intersection of $\mathcal{C}_a$ and $\mathbb{S}_+$. Then we clearly have $\mu(\mathcal{C}_a) = \Phi(\Omega_a)$ for any $\emptyset \neq a \subset \{1, \dots, d\}$ and the partitioning property of the $\mathcal{C}_a$'s passes on to the Ω_a's. Hence, one may naturally decompose the exponent measure as

$$\mu = \sum_{\emptyset \neq a \subset \{1,\dots,d\}} \mu_a, \tag{4.2}$$

where each component μ_a is concentrated on the untruncated cone generated by Ω_a. Similarly, we may write $\Phi = \sum_{\emptyset \neq a \subset \{1,\dots,d\}} \Phi_a$, where $\Phi_a = \Phi_{|\Omega_a}$. Then '$\mu_a \neq 0$' means that conditioned upon the event '$R(V)$ is large' (*i.e.* , an excess of a large radial threshold), the components $V_j (j \in a)$ may be simultaneously large while the other V_j's $(j \notin a)$ are small, with non negligible probability. This is an easy consequence of the definition (2.5), together with the fact that even though $\mathcal{C}_a$ may not be a continuity set of the exponent measure μ, it holds (Lemma 1 from Goix et al. (2017))that $\mu(\mathcal{C}_a) = \lim_{\epsilon \to 0} \mu(R_a^\epsilon)$ where R_a^ϵ is a thickened version of

$\mathcal{C}_a$, namely

$$R_a^\epsilon = \{v \geq 0,\ \|v\| \geq 1,\ v_j > \epsilon \ \text{ for } j \in a,\ v_j \leq \epsilon \ \text{ for } j \notin a\}. \tag{4.3}$$

Each index subset a thus defines a specific direction in the tail region. Note that the R_a^ϵ' form a partition of the truncated positive orthant, just as the $\mathcal{C}_a$'s do. Figures 4.1 and 4.2 below illustrate the different objects introduced thus far.

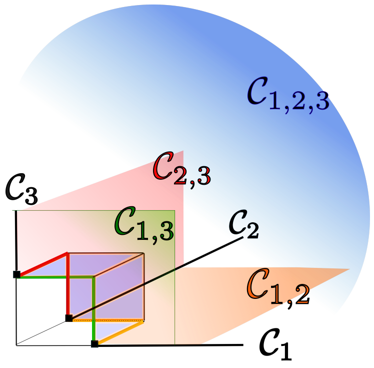


Figure 4.1 – Truncated cones in 3D

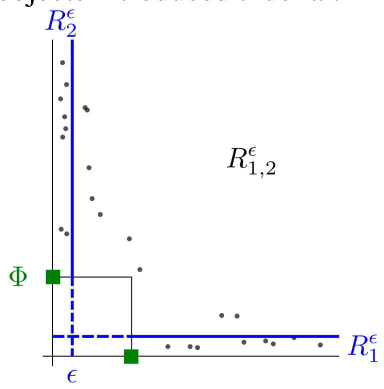


Figure 4.2 – Truncated ϵ-rectangles in 2D

The aim of Goix et al. (2016, 2017) is twofold. First, recover a rough approximation of the support of Φ based on the partition $\{\Omega_a, a \subset \{1, \ldots, d\}, a \neq \emptyset\}$, that is, determine which Ω_a's have nonzero mass, or equivalently, which $\mu_a' s$ (resp. Φ_a's) are nonzero. This support estimation is potentially sparse in the sense that only a small number of Ω_a may have non-zero mass and the latter may possibly be of low dimensionality (if the dimension of the sub-cones Ω_a with non-zero mass is low). The second objective is to investigate how the exponent measure μ spreads its mass on the $\mathcal{C}_a$'s, the theoretical quantity $\mu(\mathcal{C}_a)$ indicating to which extent extreme observations may occur in the 'direction' a for $\emptyset \neq a \subset \{1,\ \ldots,\ d\}$.

In a word, the goal is to recover the $(2^d - 1)$-dimensional unknown vector

$$\mathcal{M} = \{\mu(\mathcal{C}_a) :\ \emptyset \neq a \subset \{1,\ \ldots,\ d\}\} \tag{4.4}$$

from $X_1,\ \ldots,\ X_n \overset{i.i.d.}{\sim} P$ and to build an estimator $\widehat{\mathcal{M}}$ such that

$$\|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty = \sup_{\emptyset \neq a \subset \{1,\ \ldots,\ d\}} |\widehat{\mathcal{M}}(a) - \mu(\mathcal{C}_a)|$$

is small with large probability. These two goals are achieved using empirical versions of the angular measure evaluated on the ϵ-thickened rectangles R_a^ϵ. The following regularity conditions are required in addition to the continuity of margins (Assumption 2.1):

Assumption 4.1. *Each component μ_a of (4.2) is absolutely continuous with respect to Lebesgue measure $\mathrm{d}x_a$ on $\mathcal{C}_a$.*

Assumption 4.1 has a very convenient consequence regarding Φ (Lemma 2 from Goix et al. (2017)):

- Φ is concentrated on the (disjoint) faces

$$\Omega_{a,j_0} = \{x : \ \|x\| = 1, \ x_{j_0} = 1, \ \begin{array}{ll} 0 < x_j < 1 & \text{for } j \in a \setminus \{j_0\} \\ x_j = 0 & \text{for } j \notin a \end{array} \quad \}$$

  for $j_0 \in a, \emptyset \neq a \subset \{1, \ \dots, \ d\}$.

- The restriction Φ_{a,j_0} of Φ to Ω_{a,j_0} is absolutely continuous *w.r.t.* the Lebesgue measure $\mathrm{d}x_{a\setminus j_0}$ on the cube's faces, whenever $|a| \geq 2$.

Thus the angular measure Φ decomposes as $\Phi = \sum_a \sum_{i_0 \in a} \Phi_{a,i_0}$ and that there exist densities $\mathrm{d}\Phi_{a,i_0} / \mathrm{d}x_{a \smallsetminus i_0}, \ |a| \geq 2, \ i_0 \in a$, such that for all $B \subset \Omega_a, \ |a| \geq 2$,

$$\Phi(B) \ = \ \Phi_a(B) \ = \ \sum_{j_0 \in a} \int_{B \cap \Omega_{a,j_0}} \frac{\mathrm{d}\Phi_{a,j_0}}{\mathrm{d}x_{a \smallsetminus j_0}}(x) \, \mathrm{d}x_{a\setminus j_0}.$$

In order to formulate the next assumption, for $|a| \geq 2$, we set

$$M_a = \sup_{j \in a} \ \sup_{x \in \Omega_{a,j}} \ \frac{\mathrm{d}\Phi_{a,j}}{\mathrm{d}x_{a\setminus j}}(x).$$

Assumption 4.2. *(*SPARSE SUPPORT*) The angular density is uniformly bounded on $\mathbb{S}_+$ ($\forall |a| \geq 2, \ M_a < \infty$), and there exists a constant $M > 0$, such that we have $\sum_{a \subset \{1,\dots,d\}, |a| \geq 2} M_a < M$.*

We show that in the situation where $\mathcal{M}$ is most informative, *i.e.* when the angular density is constant on each subface Ω_a, the constant M is moderate, namely $M \leq d$. Assumptions 4.1 and 4.2 are not necessary to prove a preliminary result on a class of rectangles. However, they are required to bound the bias induced by the tolerance parameter ϵ, in particular in the main result of the paper.

Since the marginal distributions F_j are unknown, we classically consider the empirical counterparts of the V_i's, $\widehat{V}_i = (\widehat{V}_{i,1}, \dots, \widehat{V}_{i,d})$ for all $i \in \{1, \dots, n\}$, as standardized variables obtained from a rank transformation (instead of a probability integral transformation),

$$\widehat{V}_i = \left(\left(1 - \widehat{F}_1(X_{i,1})\right)^{-1}, \dots, \left(1 - \widehat{F}_d(X_{i,d})\right)^{-1} \right),$$

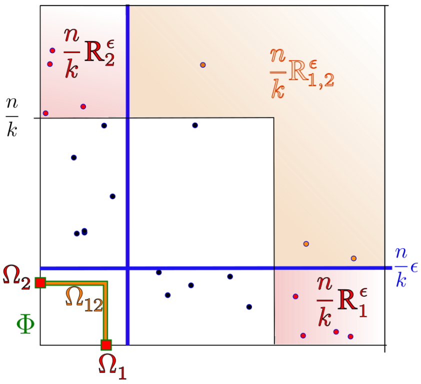
where $\widehat{F}_j(x) = (1/n) \sum_{i=1}^n \mathbf{1}_{\{X_{i,j} < x\}}$. The empirical probability distribution of the rank-transformed data is then given by $\widehat{P}_n = (1/n) \sum_{i=1}^n \delta_{\widehat{V}_i}$. A natural empirical version of μ is defined as

$$\mu_n(A) \ = \ (n/k)\widehat{P}_n((n/k)A) \ = \ \frac{1}{k} \sum_{i=1}^n \mathbb{1}\{\widehat{V}_i \in (n/k)A\} \, . \qquad (4.5)$$

Our non-parametric estimator $\widehat{\mathcal{M}}(a)$ of $\mathcal{M}(a) = \mu(\mathcal{C}_a)$ is then

$$\widehat{\mathcal{M}}(a) = \mu_n(R_a^\epsilon), \qquad \emptyset \neq a \subset \{1, \dots, d\}. \tag{4.6}$$

Figure 4.3 illustrates the estimation strategy in dimension 2.



$\frac{n}{k}\mathbf{R}_2^\epsilon$
$\frac{n}{k}\mathbb{R}_{1,2}^\epsilon$
$\frac{n}{k}$
$\frac{n}{k}\epsilon$
Ω_2
Ω_{12}
Φ
$\frac{n}{k}\mathbf{R}_1^\epsilon$
Ω_1

Figure 4.3 – Estimation procedure

We decompose the error as

$$\begin{aligned} \|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty &= \max_{\emptyset \neq a \subset \{1,\dots,d\}} |\mu_n(R_a^\epsilon) - \mu(\mathcal{C}_a)| \\ &\leq \max_{\emptyset \neq a \subset \{1,\dots,d\}} |\mu - \mu_n|(R_a^\epsilon) + \max_{\emptyset \neq a \subset \{1,\dots,d\}} |\mu(R_a^\epsilon) - \mu(\mathcal{C}_a)| \, . \end{aligned} \tag{4.7}$$

The second term on the right-hand side of the above display is a bias term stemming from the ϵ-thickening of the truncated cones. It is controlled using the regularity assumptions regularity assumptions 4.1 and 4.2. As for the first term, an inspection of the above definitions together with those of the STDF and its empirical counterpart shows that the latter may be defined as $l(x) = \mu([0, x^{-1}]^c)$ and its empirical counterpart is, up to negligible terms of order $\mathrm{O}(1/k)$, $l_n(x) = \mu_n[0, x^{-1}]^c$. Thus the guarantees obtained in Chapter 2 concern the maximal deviations $\sup_{1/T \leq x} |\mu_n - \mu| ([0, x]^c)$. In Goix et al. (2017) these guarantees are extended to a larger class of rectangles, the intersection of which with the sphere S includes the thickened cones R_a^ϵ. Here, the tolerance parameter ϵ plays the same role as $1/T$ in the analysis of the empirical STDF. Thus, the maximal deviations in the first term of the sum is controlled with an upper bound of order $\mathrm{O}(d\sqrt{\frac{\ln(d/\delta)}{\epsilon k}} + \mathrm{bias}(n/k, \epsilon))$ where $\mathrm{bias}(t, \epsilon)$ accounts for the difference between the distribution of V above level $t > 0$ and the measure μ evaluated on rectangles, namely

$$\mathrm{bias}(t, \epsilon) = \max_{x,z \geq \epsilon/2} \max_{\substack{a \,\subset\, \{1,\dots,d\} \\ a \neq \emptyset}} \mu(R_{a,x,z}) - t\mathbb{P}\left(V \in tR_{a,x,z}\right) \tag{4.8}$$

where for $\emptyset \neq a \subset \{1, \dots, d\}$ and $x, z \in \mathbb{R}_+^d$,

$$R_{a,x,z} = \{y \in \mathbb{R}_+^d : \ \forall j \in a, y_j \geq x_j \ \text{ and } \ \forall j \in \{1, \dots, d\} \setminus a, y_j < z_j\}.$$

We can now state the main result of the paper, revealing the accuracy of the estimate (4.6).

Theorem 4.1. *Suppose that Assumptions 2.1, 4.1 and 4.2 are satisfied. There is an universal constant $C > 0$ such that for every n, k, ϵ, δ verifying $\delta \geq e^{-k}$, $0 < \epsilon < 1/2$ and $\epsilon \leq 2/(7(1 + \ln(d)/k))$, the following inequality holds true with probability greater than $1 - \delta$:*

$$\|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty \ \leq \ Cd \left(\sqrt{\frac{1}{\epsilon k} \ln \frac{d}{\delta}} + Md\epsilon \right) + \ 4 \, \textit{bias} \, (n/k, \epsilon).$$

Notice that $7(1 + \ln(d)/k)/2$ is smaller than 4 as soon as $\ln(d)/k < 1/7$, so that a sufficient condition on ϵ is $\epsilon < 1/4$. The term $Md\epsilon$ is also a bias term, which stems from considering ϵ-thickened rectangles. It depends linearly on the sparsity constant M defined in Assumption 4.2.

For the purpose of dimensionality reduction and anomaly detection, the goal is to obtain a hopefully short list of subsets a such that $\mathcal{M}(a) \neq 0$. Even though Theorem 4.1 establishes satisfactory guarantees on the supremum norm $\|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty$ it does not imply that the estimate $\widehat{\mathcal{M}}$ should be sparse (*i.e.* with many null entries) even though $\mathcal{M}$ is so. A natural way around is to threshold the estimate $\widehat{\mathcal{M}}$ to a suitably chosen level $m > 0$ and to declare as zero any entry $\widehat{\mathcal{M}}(a) < m$. Doing so is equivalent (up to a bias term) to solving a risk minimization problem with an L^1 penalization term, see Remark 5 in Goix et al. (2017). Goix et al. (2016, 2017) propose an algorithm named DAMEX taking as input training data $X_1, \dots, X_n$, together with hyper-parameters ϵ, k, m, and returning a thresholded list $\{\widehat{\mathcal{M}}(a) : a \subset \{1, \dots, d\}, \widehat{\mathcal{M}}(a) > m\}$. The algorithm is particularly well suited to large datasets since its complexity is of order $\mathrm{O}(dn \ln(n))$. Experimental results on real and simulated data demonstrate the usefulness of the proposed approach. Chapter 5 details at length how DAMEX can be used for anomaly detection.

Finally, as it is the case in the previous chapters we have left the choice of hyper parameters (here, k, ϵ, m) outside our scope. Investigating the accuracy of validation strategies based on splitting the sample is part of my research perspectives.

4.2 Subspace clustering and hypothesis testing

Despite promising results on particular datasets described in Goix et al. (2016, 2017), DAMEX fails to recover a meaningful dependence structure in the tails when the dataset does not exhibit a clear-cut sparsity pattern. This observation is the starting point of Chiapino and Sabourin (2016), in which we propose a feature clustering strategy to circumvent this issue. In Chiapino and Sabourin (2016) the quantity of interest is river daily water-flow recorded at 92 locations of the French river system form 1969 to 2008, which results after preprocessing into $n = 14610$ records. For this dataset, the subsets of components $a \subset \{1, \dots, d\}$ impacted by extreme events vary from one event to another, DAMEX thus finds a very large number of subsets to be dependent, but not significantly so, (*i.e.* $0 < \widehat{\mathcal{M}}(a) \ll 1$), thus no sparsity pattern emerges. However one remarkable feature of this dataset is that many subsets of variables $a \subset \{1, \dots, d\}$ such that $\widehat{\mathcal{M}}(a) > 0$, form *clusters* in the sense that their symmetric difference has small cardinality. In practice, this means that several distinct extreme events have impacted 'almost' the same locations. Our paper proposes a methodology enabling to cluster together such 'close-by' subsets. This is done by relaxing the constraint in Goix et al. (2016, 2017)'s approach that 'features not in a take small values' when constructing the representation of the dependence structure. The output of the CLEF algorithm is an alternative representation of the dependence structure which remains usable in this 'weakly sparse' context. Namely, the aim of CLEF is to recover the *maximal* subsets $a \subset \{1, \dots, d\}$ (for the inclusion order), such that the probability of an extreme event impacting concomitantly all components $X_j, j \in a$ is non negligible, in the sense that the associated joint tail coefficient χ_a is non zero, where

$$\chi_a = \lim_{t \to \infty} t\mathbb{P}\left(\forall j \in a : V_j > t\right) = \mu(\{x \in [0, \infty)^d \mid \forall j \in a : x_j > 1\}). \tag{4.9}$$

In the bivariate case $\chi_{1,2}$ is the upper tail dependence coefficient denoted by χ in Coles et al. (1999). As shown in Chiapino and Sabourin (2016) (Lemma 1), the maximal subsets a such that $\chi_a > 0$ are the same as the maximal subsets a such that $\mathcal{M}(a) > 0$, thus the problem considered here coincides to some extent with the one considered in Goix et al. (2016, 2017). On the other hand it shares similarities with that of frequent itemsets mining and the Apriori algorithm introduced by Agrawal et al. (1994), see also Gunopulos et al. (2003). Indeed, encoding as '1' any value above a specified threshold and as '0' any value below this threshold, CLEF recovers the groups of items (= components) for which concomitant '1' values are frequent. The combinatorial issue that arises with possibly $2^d - 1$ subsets is circumvented in Apriori (see also a subset clustering method proposed in Agrawal et al. (2005)) by considering subsets of increasing sizes, letting a subset 'grow' until its frequency in the database is not significant anymore. CLEF proceeds in a similar fashion and has a natural interpretation in terms of multivariate EVT.

Since $\chi_a \leq \chi_b$ as soon as $a \supset b$, any positive tolerance level with which we would like to compare an estimate of χ_a should depend on a and in particular be decreasing as a function of the cardinality $|a|$. To circumvent this issue, Chiapino and Sabourin (2016) consider for a such that $|a| \geq 2$ the conditional tail dependence coefficient

$$\kappa_a = \lim_{t\to\infty} \mathbb{P}\left[\forall j \in a : V_j > t \;\middle|\; \textstyle\sum_{j\in a} \mathbb{1}\{V_j > t\} \geq |a| - 1\right], \tag{4.10}$$

which is the limiting conditional probability that all variables in a exceed a large threshold given that all but at most one already do. In contrast to χ_a, the coefficient κ_a has no particular reason to decrease as a function of $|a|$. Note that $\chi_a = \mu(\Gamma_a)$ while $\kappa_a = \mu(\Gamma_a)/\mu(\Delta_a) = \chi_a/\mu(\Delta_a)$ where $\Gamma_a = \{x \in [0,\infty)^d \mid \forall j \in a : x_j > 1\}$ is a subset of $\Delta_a = \{x \in [0,\infty)^d \mid \sum_{j\in a} \mathbb{1}_{\{x_j>1\}} \geq |a| - 1\}$, provided $|a| \geq 2$. In words, Δ_a is the set of vectors $x \in [0,\infty)^d$ such that $x_j \geq 1$ for all but at most one $j \in a$. Another way to see Δ_a is as the union of the sets $\Gamma_{a\setminus\{j\}} = \{x \mid \forall i \in a \setminus \{j\} : x_j \geq 1\}$ over all $j \in a$. If $\mu(\Delta_a) = 0$, then $\mu(\Gamma_a) = 0$ and also $\mu(\Gamma_{a\setminus\{j\}}) = 0$ for all $j \in a$, and in that case, we define $\kappa_a = 0$.

In CLEF (Chiapino and Sabourin, 2016) summarized in Algorithm 1 below, the criterion to decide whether $\chi_a > 0$ or not is that $\widehat{\kappa}_a \geq C$, where C is a user-defined tolerance level, $\widehat{\kappa}_a = \mu_n(\Gamma_a)/\mu_n(\Delta_a)$, and μ_n is the empirical exponent measure (4.5).

The level C can be chosen independently of a. Still, its choice is somewhat arbitrary, and in particular, the user has no control of false positives. Another popular summary of the tail dependence of components $X_j, j \in a$ is the extremal coefficient θ_a Smith (1990); Coles (1993); Schlather and Tawn (2002, 2003),

$$\theta_a = \lim_{t\to\infty} t\mathbb{P}\left(\exists j \in a : V_j > t\right) = \mu(\{v \in [0,\infty)^d \mid \exists j \in a : v_j > 1\}), \tag{4.11}$$

The joint tail coefficients χ_a and the extremal coefficients θ_a are related *via* the inclusion–exclusion formula, a property which is exploited in Chiapino et al. (2019b) to derive the asymptotic distribution of the stopping criterion in CLEF $\widehat{\kappa}_a$ provided that $\kappa_a \neq 0$. Indeed Einmahl et al. (2012, Theorem 4.6) find the weak limit of the STDF empirical process $\sqrt{k}(l_n - l)$ on $[0,T]^d$ for any $T > 0$ and the following conditions stem from the cited article.

Assumption 4.3 (Uniform tail convergence)**.** *There exists $\gamma > 0$ such that, uniformly in $x \in [0,1]^d$ with $\sum_{j=1}^d x_j = 1$, we have*

$$t^{-1}\mathbb{P}\left(\exists j = 1,\dots,d : \; F_j(X_j) > tx_j\right) - l(x) = \mathrm{O}(t^\gamma), \qquad t \to \infty.$$

Algorithm 1 CLEF (CLustering Extreme Features)

```
**Input**: Tolerance parameter $C > 0$.
**STAGE 1: constructing the collection $\widehat{\mathbb{M}}_{\max}$ of tail-dependent groups.**
**Step 1:** Put $\hat{\mathcal{A}}_1 = \{\{1\}, \ldots, \{d\}\}$ and $S = 1$.
**Step $s = 2, \ldots, d$**: If $\hat{\mathcal{A}}_{s-1} = \emptyset$, end **STAGE 1**. Otherwise:
    • Generate candidates of size $s$:
      $\mathcal{A}'_s = \{a \subset \{1, \ldots, d\} : |a| = s$ and $a \setminus j \in \hat{\mathcal{A}}_{s-1}$ for all $j \in a\}$.
    • Put $\hat{\mathcal{A}}_s = \{a \in \mathcal{A}'_s : \hat{\kappa}_a > C\}$.
    • If $\hat{\mathcal{A}}_s \neq \emptyset$, put $S = s$.
**Output**: $\widehat{\mathbb{M}} = \emptyset$ if $S = 1$ and $\widehat{\mathbb{M}} = \bigcup_{s=2}^{S} \hat{\mathcal{A}}_s$ if $S \geq 2$.
**STAGE 2: pruning, keeping maximal groups $a$ only.**
If $S = 1$, then $\widehat{\mathbb{M}}_{\max} = \emptyset$. Otherwise:
*Initialization:* $\widehat{\mathbb{M}}_{\max} \leftarrow \hat{\mathcal{A}}_S$.
for $s = (S - 1) : 2$,
    for $a \in \hat{\mathcal{A}}_s$,
        If there is no $b \in \widehat{\mathbb{M}}_{\max}$ such that $a \subset b$, then $\widehat{\mathbb{M}}_{\max} \leftarrow \widehat{\mathbb{M}}_{\max} \cup \{a\}$.
**Output**: $\widehat{\mathbb{M}}_{\max}$
```

Assumption 4.4 (Moderate k)**.** *The sequence $k = k(n)$ satisfies $k = \mathrm{o}(n^{2\gamma/(1+2\gamma)})$ as $n \to \infty$, with $\gamma > 0$ as in Condition 4.3.*

Assumption 4.5 (Smoothness)**.** *For all $j \in \{1, \ldots, d\}$, the partial derivative $\partial_j l = \partial l / \partial x_j$ exists and is continuous on the set $\{x \in [0, \infty)^d \mid x_j > 0\}$.*

Given the importance of the joint tail coefficient (4.9) for the problem at hand we introduce the joint tail dependence function $r_a : [0, \infty]^a \setminus \{\boldsymbol{\infty}\} \to [0, \infty)$, where $\boldsymbol{\infty} = (\infty, \ldots, \infty)$, given by

$$r_a(x) = \lim_{t \to 0} t^{-1} \mathbb{P}\left(\forall j \in a : V_j > 1/x_j\right) = \mu(\{y \mid \forall j \in a : y_j > 1/x_j\}) \quad (4.12)$$

The empirical counterpart of r_a is defined as $\widehat{r}_a(x) = \mu_n(\{y \mid \forall j \in a : y_j > 1/x_j\})$, where μ_n is the same as in (4.5). We consider Hoffmann-Jørgensen weak convergence in metric spaces as in van der Vaart (1998); van der Vaart and Wellner (1996); notation $\rightsquigarrow$. We work in the metric space $L^\infty(E)$ of bounded, real functions f on an arbitrary set E, the metric being the one induced by the supremum norm, $\|f\|_\infty = \sup_{x \in E} |f(x)|$

Einmahl (1997) and Einmahl et al. (2012) characterize the weak limit of the empirical process $\sqrt{k}(l_n - l)(x)$ in terms of a centered Gaussian process W indexed by the Borel sets of $[0, \infty]^d \setminus \{\boldsymbol{\infty}\}$ bounded away from $\boldsymbol{\infty}$ with covariance function

$$\mathbb{E}\left(W(A)\,W(B)\right) = \Lambda(A \cap B). \tag{4.13}$$

where the measure Λ is the image measure of the exponent measure under the mapping $i : x \mapsto 1/x$, the inverse operation being understood componentwise, $\Lambda = \mu \circ i^{-1}$. Note that $W(\emptyset) = 0$ almost surely. For $\emptyset \neq a \subset \{1, \ldots, d\}$ and $x \in [0, \infty)^a$, write

$$W_a(x) = W(\{y \in [0, \infty]^d \mid \forall j \in a : y_j < x_j\}).$$

Leveraging Einmahl et al. (2012)'s result through the inclusion-exclusion formula and the Delta method we obtain

Proposition 4.2. *Let $X_i = (X_{i,1}, \ldots, X_{i,d})$, for $i \in \{1, \ldots, n\}$, be an independent random sample from P, having continuous margins (Assumption 2.1) and satisfying (2.5). Let $k = k(n) \to \infty$ as $n \to \infty$, while $k(n) = \mathrm{o}(n)$. If Conditions 4.3, 4.4 and 4.5 hold, then, for $T > 0$, in the product space $\prod_{\varnothing \neq a \subset \{1,\ldots,d\}} L^\infty([0, T]^a)$, we have, as $n \to \infty$, the weak convergence*

$$\sqrt{k}\,\{\widehat{r}_a(x) - r_a(x)\} \rightsquigarrow W_a(x) - \sum_{j \in a} \partial_j r_a(x)\, W_{\{j\}}(x_j) = Z_a(x). \tag{4.14}$$

The asymptotic normality of the vector $\sqrt{k}\,(\widehat{\kappa}_a - \kappa_a)_{\emptyset \neq a \subset dd}$ follows (Proposition 2 in Chiapino et al. (2019b)), and the asymptotic variance can be consistently estimated provided it is non zero.

If $\chi_a = 0$ (or $\kappa_a = 0$), the limit distributions of the statistics $\sqrt{k}(\widehat{\chi}_a - \chi_a)$ and $\sqrt{k}(\widehat{\kappa}_a - \kappa_a)$ are degenerate at zero. We therefore have no control on the asymptotic type-I error rate of tests based on those statistics under $H_0 : \kappa_0 = 0$. Instead we define a CLEF stopping criterion in terms of a test of $H_0 : \kappa_a \geq \kappa_{\min}$ versus $H_1 : \kappa_a < \kappa_{\min}$, in terms of a user-defined level $\kappa_{\min} > 0$. Again the choice of $\kappa_{\min}$ is somewhat arbitrary; in our simulation experiments we choose $\kappa_{\min} = 0.08$.

To overcome the issue of the choice of $\kappa_{\min}$ we also consider alternative CLEF stopping criteria based on estimators of the coefficient of tail dependence $\eta_a \in (0, 1]$, defined in (4.15) below. For bivariate distributions, this coefficient has been introduced by Ledford and Tawn (1996) and extended by Ramos and Ledford (2009) in order to model a wide range of situations including asymptotic dependence ($\chi_{\{1,2\}} > 0$, $\eta_{\{1,2\}} = 1$), and asymptotic independence ($\eta_{\{1,2\}} < 1$) with positive or negative association depending on the sign of $\eta_{\{1,2\}} - 1/2$. De Haan and Zhou (2011) and Eastoe and Tawn (2012) proposed and studied a multivariate extension of η_a for $|a| \geq 3$. The model assumption is that there exist $\eta_a \in (0, 1]$ and a slowly varying function $\mathcal{L}_a$ such that

$$\mathbb{P}\left(\forall j \in a : V_j > t\right) = t^{-1/\eta_a} \mathcal{L}_a(t). \tag{4.15}$$

Suppose that the limit χ_a in (4.9) exists and that (4.15) holds. Then $\chi_a > 0$ implies $\eta_a = 1$. The converse is true as well, provided $\liminf_{t\to\infty} \mathcal{L}_a(t) > 0$. Modulo this side condition, which we take for granted, the null hypothesis $\chi_a > 0$ corresponds to the simple hypothesis $\eta_a = 1$. We test the null hypothesis $\eta_a = 1$ via multivariate extensions of nonparametric estimators of η_a in Peng (1999) and Draisma et al. (2004) which are respectively related to the Pickands estimator and the Hill estimator for the extreme value index of $T_a = \min_{j\in a} V_j$, namely

$$\widehat{\eta}_a^P = \log(2)/\log\{\widehat{r}_a(\mathbf{2}_a)/\widehat{r}_a(\mathbf{1}_a)\} \ , \tag{4.16}$$

$$\widehat{\eta}_a^H = \frac{1}{k}\sum_{i=1}^{k}\log\frac{\widehat{T}_{(n-i+1),a}}{\widehat{T}_{(n-k),a}}, \qquad \text{where } \widehat{T}_{i,a} = \min_{j\in a}\widehat{V}_{i,j}, \tag{4.17}$$

where for $t \in \mathbb{R}$ the notation $\boldsymbol{t}_a$ stands for the constant vector of size $|a|$ with entries equal to t. The maximum likelihood estimator, also considered in Draisma et al. (2004), is less suitable to our context due to its relative computational complexity, since the test is destined to be performed on a large number of subsets a of $\{1,\dots,d\}$. See also the review Bacro and Toulemonde (2013) and the references therein.

The asymptotic normality of $\sqrt{k}(\widehat{\eta}_a^P - 1)$ (Proposition 4 in Chiapino et al. (2019b)) follows from Proposition 4.2 and the delta method under the assumption hat $\chi_a > 0$. To prove the asymptotic normality of $\sqrt{k}(\widehat{\eta}_a^H - 1)$ (Proposition 6 in Chiapino et al. (2019b)) we extend to the multivariate setting the proof of Theorem 2.1 in Draisma et al. (2004) which covers the bivariate case only, and we provide a general expression for the asymptotic variance. A second-order regular variation condition is required. Again, the asymptotic variance can be estimated as soon as it is positive. Since the null limits of the test statistics $\widehat{\eta}_a^P, \widehat{\eta}_a^H$ are nondegenerate under the condition that $\chi_a > 0$, the asymptotic type-I error rate of the test can be controlled, with no need to introduce an additional tolerance parameter $\kappa_{\min}$.

We end up with three variants of the original CLEF algorithm, named CLEF-asymptotic, CLEF-Peng and CLEF-Hill, obtained by replacing in Algorithm 1 the condition '$\widehat{\kappa}_a > C$' (under which the components of subset a are considered as tail-dependent) respectively with $\widehat{\kappa}_a > \kappa_{\min} - z_{1-\alpha}\widehat{\sigma}_{\kappa,a}/\sqrt{k}$, $\quad \widehat{\eta}_{a,P} > 1 - z_{1-\alpha}\widehat{\sigma}_{a,P}/\sqrt{k}$, and $\widehat{\eta}_{a,H} > 1 - z_{1-\alpha}\widehat{\sigma}_{a,H}/\sqrt{k}$, where $\alpha > 0$ is the type-I error, $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal variable, and $\widehat{\sigma}_{\kappa,a}, \widehat{\sigma}_{\kappa,a}, \widehat{\sigma}_{\kappa,a}$ are the estimated limit standard deviations of the considered estimators. The performance of these three variants are compared together with CLEF and DAMEX on simulated and real data. For real data, the ground truth is unknown so an unsupervised cross-validation procedure is proposed to assess the quality of the output. Experimental results indicate that the choice of a particular algorithm and its tuning parameters should be made according to the particular dataset under consideration, which confirms the importance of cross-validation for model selection in this context.

4.3 Principal Component Analysis for Multivariate extremes

We end this chapter with a presentation of the paper Drees and Sabourin (2021) in which the unmissable PCA strategy for dimension reduction is adapted to multivariate extremes and an upper bound on the reconstruction error in the tail is obtained. We consider a (high dimensional) vector X satisfying the regular variation assumption (2.6), that is, no standardization is required for a limit measure ν to exist. We denote by $\alpha > 0$ the regular variation index. A reasonable assumption is that the support of the limit measure ν is concentrated on a lower dimensional subspace, meaning that certain linear combinations of the components are much likelier to be large than others. Identifying this subspace and thus reducing the dimension will facilitate a refined statistical analysis. In this work we apply Principal Component Analysis to a re-scaled version of radially thresholded observations.

In a classical setting, when $\|X\|$ has finite second moments, PCA (Anderson (1963)) is the method of choice to determine such supporting linear subspaces if $i.i.d.$ random vectors X_i, $1 \leq i \leq n$, with the same distribution as X are observed. Theoretical guarantees obtained so far concern the reconstruction error (Koltchinskii and Giné (2000); Shawe-Taylor et al. (2005); Blanchard et al. (2007); Koltchinskii and Lounici (2017); Reiß and Wahl (2020)) or the approximation error for the eigenspaces of the covariance matrix (Zwald and Blanchard (2006)), under the assumption that the sample space (or the feature space for Kernel-PCA) has finite diameter or that sufficiently high order moments exist.

For motivation of our version of PCA, it is useful to keep the following working hypothesis in mind, although it is not required for most results to hold.

Assumption 4.6. *The vector space $V_0 = \operatorname{span}(\operatorname{supp}\nu)$ generated by the support of ν has dimension $p < d$.*

Note that then the points $(X_i/t)\mathbb{1}\{\|X_i\| > t\}$ are more and more concentrated on a neighborhood of V_0 as t increases, but usually they will not lie on V_0. If the dimension p of V_0 is known, then it suggests itself to approximate V_0 by the subspace of dimension p that is 'closest' in expectation to these points.

In PCA one measures the closeness by the squared Euclidean distance which hugely alleviates the optimization problem as one may work with orthogonal projections in the Hilbert space L_2. However, this approach requires finite second moments which cannot be taken for granted in the above setting. Indeed, if $\alpha < 2$ then $\mathbb{E}(\|X_i\|^2) = \infty$. Hence, we instead consider a rescaling function θ and the re-scaled vectors Θ_i defined by

$$\begin{aligned} \theta(x) &= \omega(x)x, \qquad x \in \mathbb{R}^d \\ \Theta_i &= \theta(X_i) \quad 1 \leq i \leq n, \end{aligned} \tag{4.18}$$

where $\omega : \mathbb{R}^d \to (0, \infty)$ is a suitable scaling function. The most common choice is $\omega(x) = 1/\|x\|$, leading to Θ_i on the unit sphere which describes the direction of X_i, and we focus on this re-scaling when we derive finite sample bounds on the reconstruction error. Throughout this paper we use the Euclidean norm $\|\cdot\| = \|\cdot\|_2$ and $\mathbb{S} := \{x \in \mathbb{R}^d : \|x\| = 1\}$ denotes the associated unit sphere. Consistency results are proved for more general scaling functions than the projection onto the sphere, namely those satisfying the homogeneity condition:

$$\begin{aligned} \exists\, \beta \in \left(1 - \frac{\alpha}{2}, 1\right] \forall\, \lambda > 0, x \in \mathbb{R}^d : \quad \omega(\lambda x) = \lambda^{-\beta}\omega(x) \\ \text{and} \quad c_\omega := \sup_{x \in \mathbb{S}} \omega(x) < \infty, \end{aligned} \tag{4.19}$$

Before stating our main results we introduce some notation and the ERM setting adopted in our work. For V a subspace of $\mathbb{R}^d$, denote by $\mathbf{\Pi}_V$ (*resp.* $\mathbf{\Pi}_V^\perp$) the orthogonal projection onto V (*resp.* on the orthogonal $V^\perp$), or the associated projection matrix. Let P_t denote the conditional distribution $\mathbb{P}\left(X \in \cdot \mid \|X\| > t\right)$. A consequence of (2.6) is that P_t converges weakly to $P_\infty = \nu(\,\cdot\,)/\nu(\{x : \|x\| \geq 1\})$ on the complementary set of the unit ball $B_1(0)^c$ in $\mathbb{R}^d$. For $t > 0$ we consider the conditional risk

$$R_t(V) := P_t\big(\|\mathbf{\Pi}_V^\perp \theta)\|^2\big) = \mathbb{E}\left(\|\mathbf{\Pi}_V^\perp \Theta\|^2 \mid \|X\| > t\right) \tag{4.20}$$

and its conditional counterpart for t taken as the empirical quantile $\hat{t}_{n,k}$ of the norm $\|X\|$ at level $1 - k/n$,

$$\hat{R}_{n,k}(V) := \hat{R}_{\hat{t}_{n,k}}(V) = \frac{1}{k}\sum_{i=1}^{k} \|\mathbf{\Pi}_V^\perp \Theta_{(i)}\|^2 \tag{4.21}$$

where $\Theta_{(i)} = \theta(X_{(i)})$ and the observations are ranked by decreasing order relative to their norm, $\|X_{(1)}\| \geq \|X_{(2)}\| \geq \ldots \|X\|$. In the regular variation setting it is natural to consider also the reconstruction risk at infinity

$$R_\infty(V) := P_\infty\|\mathbf{\Pi}_V\theta - \theta\|^2 = P_\infty\|\mathbf{\Pi}_V^\perp\theta\|^2. \tag{4.22}$$

A direct consequence of standard facts from PCA is that the eigen vectors of the conditional convariance matrix

$$\Sigma_t = \mathbb{E}\left(XX^\top \mid \|X\| > t\right)$$

determine the best projection subspaces. Namely, considering $\mathcal{V}_p$ the set of all subspaces of $\mathbb{R}^d$ of dimension p, and if $u_1, \dots, u_d$ denote the eigen vectors of Σ_t ranked by decreasing order of the associated eigen values, then

$$V_t^* = \mathrm{span}(u_1, \dots, u_p) \in \mathrm{argmin}_{V \in \mathcal{V}_p} R_t(V).$$

Similarly, denoting by $(\hat{u}_1, \hat{u}_p)$ the eigen vectors of the empirical conditional covariance matrix

$$\widehat{\Sigma}_{n,k} = \frac{1}{k} \sum_{i=1}^{l} \Theta_{(i)} \Theta_{(i)}^\top ,$$

it holds that

$$\widehat{V}_{n,k} = \mathrm{span}(\hat{u}_1, \dots, \hat{u}_p) \in \mathrm{argmin}_{V \in \mathcal{V}_p} \hat{R}_{n,k}(V).$$

It is easy to show (Lemma 2.5 from Drees and Sabourin (2021)) that under Assumption 4.6, V_0 is the unique minimizer of R_∞ over $\mathcal{V}_p$, that $R_\infty(V_0) = 0$ and that any other subspace V such that $R_\infty(V) = 0$ contains V_0. Natural questions arise: (i) under which conditions and in which sense do the minimizers V_t^* of R_t converge to V_∞^*, a minimizer of R_∞? (ii) What can be said about the empirical minimizer $\widehat{V}_n$ of $\hat{R}_{n,k}$? (iii) Can one obtain uniform non asymptotic bounds of the deviations of $\hat{R}_{n,k}(V)$ in order to bound the excess risk $R_\infty(\widehat{V}_n) - R_\infty(V_\infty^*)$ (as sketched in Section 3.1 in the context of classification)? (iv) What is the practical relevance of this dimension reduction device, *e.g.* for non-parametric estimation of the probability of failure regions?

First it is shown (Proposition 2.2 in Drees and Sabourin (2021) that condition (4.19) on the scaling function combined with weak convergence of P_t towards P_∞ entails that $\lim_{t \to \infty} t^{2(\beta-1)} R_t(V) = R_\infty(V)$ for any fixed subspace $V \subset \mathbb{R}^d$. We then endow $\mathcal{V}_p$ with the metric induced by the operator norm of the orthogonal projection, $\rho(V, W) = |||\mathbf{\Pi}_V - \mathbf{\Pi}_W|||$. It can be shown that $\mathcal{V}_p$ is compact w.r.t. ρ and that the normalized conditional risk functions $t^{2(\beta-1)} R_t$ are uniformly Lipschitz continuous from which the convergence of the risk minimizers follows by standard arguments.

Theorem 4.3 (Theorem 2.5 in Drees and Sabourin (2021))**.** *Suppose that ω satisfies condition* (4.19) *and that R_∞ has a unique minimizer V_∞^* in $\mathcal{V}_p$. Then, for any minimizer V_t^* of R_t in $\mathcal{V}_p$, one has*

$$\lim_{t \to \infty} \rho(V_t^*, V_\infty^*) = 0.$$

The consistency of the empirical risk minimizers follows the same line, after replacing the uniform Lipschitz property of the risk functions with asymptotic equicontinuity in probability of a rescaled version of the empirical risks $\hat{R}_{n,k}$.

Theorem 4.4 (Theorem 2.7 in Drees and Sabourin (2021)). *If ω satisfies condition* (4.19) *and R_∞ has a unique minimizer V_∞^* in $\mathcal{V}_p$, then $\rho(\hat{V}_n, V_\infty^*) \to 0$ in probability for all minimizers $\hat{V}_n$ of $\hat{R}_{n,k}$ in $\mathcal{V}_p$.*

Under the stronger condition on ω that $\omega(x) \leq 1/\|x\|$ (thus $\|\Theta\|$ becomes a bounded random vector), we derive an upper bound in probability for the deviations of the empirical risk $\hat{R}_{n,k}$. To do so we adapt the arguments of Blanchard et al. (2007) to take into account the (small) variance of $\Theta \mathbb{1}\{\|X\| > t\}$ and we use a Bernstein-like concentration inequality from McDiarmid (1998) which is already mentioned in Section 2.1.2 as a key ingredient of the proof of the general concentration inequality for rare classes (Inequality (2.3)). Denote by $t_{n,k}$ the $1 - k/n$ quantile of the norm $\|X\|$. We state below a simplified version of Theorem 3.1 in Drees and Sabourin (2021) which holds in the most intuitive case $\omega(x) = 1/\|x\|$:

Theorem 4.5 (Uniform risk bound I). *If $\omega(x) = 1/\|x\|$, with probability $1 - \delta$,*

$$\sup_{V \in \mathcal{V}_p} |\hat{R}_{n,k}(V) - R_{t_{n,k}}(V)| \leq \Big[\frac{p \wedge (d-p)}{k}\big(1 - (k/n)\operatorname{tr}(\Sigma_{t_{n,k}}^2)\big)\Big]^{1/2} + \Big[\frac{8}{k}(1 + k/n)\log(4/\delta)\Big]^{1/2} + \frac{4\log(4/\delta)}{3k}.$$

Note that the upper bound in Theorem 4.5 involves a term Σ_t which cannot be calculated from the data and can thus not directly be used to construct confidence intervals for the true reconstruction error $R_{t_{n,k}}(\hat{V}_n)$ or the minimal reconstruction error $\inf_{V \in \mathcal{V}_p} R_{t_{n,k}}(V)$. Therefore, we derive data-dependent bounds directly from (a minor improvement of) the bound established by Blanchard et al. (2007). This result is be applied to the conditional distribution of Θ given $\|X\| > t$ and the resulting bound is to be interpreted conditionally on the number N_t of exceedances over the chosen threshold t.

Theorem 4.6 (Conditional data-dependent risk bounds, Theorem 3.3 in Drees and Sabourin (2021)). *If $\omega(x) \leq 1/\|x\|$ for all $\ell > 1, u, v > 0$,*

$$\mathbb{P}\Big(\sup_{V \in \mathcal{V}_p} |\hat{R}_t(V) - R_t(V)| \geq \Big[(p \wedge (d-p))\Big(\frac{\tilde{S}_t}{\ell - 1} + \frac{v}{\ell}\Big)\Big]^{1/2} + u \,\Big|\, N_t = \ell\Big) \leq 2\exp\big(-2\ell u^2\big) + \exp\big(-\lfloor \ell/2 \rfloor v^2/2\big)$$

with $\tilde{S}_t := N_t^{-1}\sum_{i=1}^n \|\Theta_{i,t}\|^4 - \operatorname{tr}\Big((N_t^{-1}\sum_{i=1}^n \Theta_{i,t}\Theta_{i,t}^\top)^2\Big)$ *and* $\lfloor x \rfloor := \max\{k \in \mathbb{Z} : k \leq x\}$.

In our simulation study, we we examine the impact of PCA on the standard non-parametric estimator of the angular probability measure related to the limit measure ν, *i.e.* $H(A) = \nu(\{t\theta, \theta \in A\})/\nu(B_0(1)^c)$, based on the k largest observations

$$\hat{H}_{n,k} := \frac{1}{k} \sum_{i=1}^{n} \delta_{\theta_{t_{n,k}}(X_i)}.$$

(with $\theta(x) = x/\|x\|$) We investigate how $\hat{H}_{n,k}$ is influenced if the data is first projected onto a lower dimensional subspace using PCA:

$$\hat{H}_{n,k}^{PCA} := \frac{1}{k} \sum_{i=1}^{n} \delta_{\theta_{t_{n,k}}(\mathbf{\Pi}_V^{\perp} X_i)}.$$

Here, V denotes the subspace picked by PCA based on the same number k of largest observations. It turns out that sometimes it is advisable to use a smaller number $\tilde{k}$ for the PCA procedure; the resulting estimator of the spectral measure is then denoted by $\hat{H}_{n,k,\tilde{k}}^{PCA}$.

We simulate from different models of d-dimensional regularly varying vectors for which the spectral measure is (approximately) concentrated on a p-dimensional subspace. Since PCA is equivariant under rotations, w.l.o.g. we assume that this subspace is spanned by the first p unit vectors. We consider moderate to high dimensions (respectively $(d = 5, p = 2)$ and $(d = 100, p = 5)$). The performance of the spectral estimators is measured in terms of the errors of the resulting estimators of the following probabilities in the limit model, which can be expressed in terms of the angular measure:

(i) $\lim_{u\to\infty} \mathbb{P}(p^{-1} \sum_{1\le j\le p} X_j/\|X\| > t_{(i)} \mid \|X\| > u)$
$= H\{x : p^{-1} \sum_{j=1}^{p} x_j > t_{(i)}\}$ for some $t_{(i)} \in (0, p^{-1/2})$

(ii) $\lim_{u\to\infty} \mathbb{P}(\min_{1\le j\le p} X_j > u, \max_{p+1\le j\le d} X_j \le u \mid \|X\| > u)$
$= \int \big((\min_{1\le j\le p} x_j)^\alpha - (\max_{p+1\le j\le d} x_j)^\alpha\big)^+ H(dx)$

(iii) $\lim_{u\to\infty} \mathbb{P}(X^1 > u \mid \max_{1\le j\le d} X_j > u)$
$= \int (x^1)^\alpha H(dx) / \int (\max_{1\le j\le p} x_j)^\alpha H(dx)$

(iv) $\lim_{u\to\infty} \mathbb{P}(\min_{1\le j\le d} X_j > u \mid \|X\| > u) = \int (\min_{1\le j\le d} x_j)^\alpha H(dx)$

The first probability is related to the *c.d.f.* of the mean contribution of the first p coordinates to the norm of the random vector, thus quantifying, in some sense, how strongly the norm is spread over the coordinates. Probability (ii) indicates how likely it is that the first p components are all large, while this is not true for any of the other components, given that the norm of the vector is large. Probability (iii) specifies how likely it is that the first component is extreme, given that any component is extreme. In a financial context, such probabilities are used to quantify how strongly a specific market participant is exposed to a failure of any market participant. Finally, probability (iv) specifies the minimal contribution of any coordinate to the norm. Note that under Hypothesis 1 this probability equals 0. The other true values are determined by Monte Carlo simulations.

The marginal distributions are chosen as Fréchet with *c.d.f.* $\exp(-x^{-\alpha})$, $\alpha \in \{1, 2\}$, and α is assumed to be known since we are interested in the effect of the PCA procedure on the estimator of the spectral measure, which should not be compounded with the estimation error of the tail index.

Without dwelling into the details of the experimental result, we jump to their conclusions here: while the PCA step does not always improve the estimator of the angular measure, for probability (i) the resulting estimators are superior to the standard estimator and in most other cases they seem competitive if $\tilde{p}$ (the retained dimension of the projecting subspace) is chosen appropriately. To this end, experimental results show that the plot of the empirical risk is a very useful tool, however we have not tried to derive theoretical guarantees concerning the choice of $\tilde{p}$. The added value of PCA is all the more visible for moderate dimensional problems. For higher dimensional data, there may be some ambiguity about the dimension of the subspace onto which the data should be projected. In case of doubt, it is advisable to choose a higher dimensional subspace, in particular for the PCA method that uses the same number of largest observations to estimate the support and to calculate the estimator of the spectral measure. The PCA estimators that determine the support based only on the largest 10 observations often exhibit a desirable insensitivity to the choice of largest observations used to estimate the spectral measure, which makes them easier to apply in practice.

Chapter 5

Anomaly detection, clustering and vizualization

What is usually referred to as an anomaly or an outlier in data analysis is an observation which has been generated by a mechanism which is distinct from the one having generated the vast majority of other points (the inliers). In other terms the situation considered is that of a mixture model with highly imbalanced classes. Depending on the context the goal may be, as in Anomaly Detection (AD) to separate the anomalies from the rest in a pre-processing step, or on the contrary, the focus may be on the anomalies themselves, the goal being then to characterize their distribution through appropriate summaries. The first two sections of this chapter relate to the former goal, while the third section is related to the latter and presents a clustering and visualization algorithm dedicated to the anomalies in the tail.

In AD, the underlying assumption is that anomalies lie in regions of the sample space where the density on the inliers is low. Most anomaly detection strategies consist in constructing a score function $s : \mathcal{X} \to \mathbb{R}_+$ representing the degree of abnormality of a new unlabeled data point x_{new}. The lower $s(x_{\text{new}})$, the likelier it is that x_{new} is an anomaly. The remaining ingredient to cook-up an AD algorithm is a user-defined threshold relative to $s(x)$ below which any new point is declared as abnormal. When analyzed in a Neyman-Pearson framework in the limiting case where the density of the outliers is uniform on $\mathcal{X}$, the optimal scoring functions are those which are non-decreasing transforms of the density function of the inliers, so that the scoring function and the density function share the same level sets. In *semi-supervised* AD the training data consists of only (or an overwhelming proportion of) normal instances which are used to select a scoring function within a class of controlled complexity, so that generalization bounds can be derived. Numerous algorithms have been proposed, be it in a parametric setting (Barnett and Lewis (1994); Eskin (2000) or in a non-parametric one, in which case popular approaches include estimation of level sets of the inlier density (Breunig et al. (2000); Schölkopf et al. (2001); Steinwart et al. (2005); Scott and Nowak (2006a); Vert and Vert (2006), dimensionality reduction (Shyu et al. (2003); Aggarwal and Yu (2001)), decision trees (Liu et al. (2008); Shi and Horvath (2012)), see also Chandola et al. (2009) and the references therein. In high dimensional setting, deep learning strategies have recently been proposed based on auto-encoding approaches (*e.g.* in Zhou and Paffenroth (2017)) or adversarial learning (Zenati et al. (2018)), see also the review Chalapathy and Chawla (2019).

Investigating AD from the view point of Extreme Value Analysis is a natural idea considering the fact that in many applications the inliers' distribution is unimodal and low density regions and tail regions are the same. In such a case, under appropriate regularity assumptions and if the problem at hand requires a very low false alarm rate, EVT can help choosing a scoring function which approximates well the lowest levels of the density. This simple consideration underlies several recent works using uni-variate EVT for AD (Roberts (1999, 2000); Lee and Roberts (2008); Clifton et al. (2008); Tressou (2008); Clifton et al. (2011); Siffer et al. (2017); Vignotto and Engelke (2020)). Until recently there has been no AD algorithm relying on multivariate EVT. In Goix et al. (2016, 2017) and Thomas et al. (2017) we take a step towards bridging the gap between the practice of AD in multivariate settings and multivariate EVT, as detailed respectively in sections 5.1 and 5.2 of the present thesis.

5.1 Anomaly detection *via* dimensionality reduction of the multivariate tail

In this section we explain how the dimensionality reduction device presented in Section 4.1 can be used to construct an AD algorithm named DAMEX in Goix et al. (2016, 2017). In the suggested framework, *extreme* data are observed values X such that the norm of their standardization $\|V\|$ is large, denoting by $\|\cdot\|$ the sup norm $\|\cdot\|_\infty$ on $\mathbb{R}^d$ throughout. *Anomalies* among extremes are those which *direction* $\theta(V) = V/\|V\|$ is unusual, which is an appropriate model for anomalies in many applications. Recall the summary of the dependence structure $\{\mathcal{M}(a) = \mu(\mathcal{C}_a), \emptyset \neq a \subset \{1, \dots, d\}\}$ defined through equations (4.1), (4.4) and its estimator $\{\widehat{\mathcal{M}}(a), \emptyset \neq a \subset \{1, \dots, d\}\}$ from (4.5), (4.6). For large $t \in \mathbb{R}_+$ and a fixed measurable set $A \subset [0, \infty)^d \setminus \{0\}$ ($\mu(\partial A) = 0$), the standardized regular variation condition (2.5) implies that $\mathbb{P}\left(V \in tA\right) \approx t^{-1}\mu(A) = \mu(tA)$. Considering $A = \mathcal{C}_a$ is tempting in view of the decomposition (4.2) however since the Lebesgue measure of $\mathcal{C}_a$ is 0 for $a \neq \{1, \dots, d\}$, the condition $\mu(\partial\mathcal{C}_a) = 0$ fails in general. Notice that this is the same argument which underlies the use of thickened cones R_a^ϵ for estimation purpose in (4.6). Thus $\widehat{\mathcal{M}}_a$ may as well be seen as an estimator of $\mu(R_a^\epsilon)$. In addition, doing so cancels out one of the bias terms in the error bound stated in Theorem 4.1. In this context it seems appropriate to propose a scoring function s_n on the tail region of the standardized variables $\{v : \|v\| \geq t\}$ such that for $v \in R_\alpha^\epsilon$,

$$s_n(v) = (\|v\|)^{-1}\widehat{\mathcal{M}}_a \approx \mathbb{P}\left(\|V\| > \|v\|, V \in R_a^\epsilon\right).$$

Algorithm 2 DAMEX (Detecting Anomalies in Multivariate EXtremes

```
Input: parameters $\epsilon > 0, \quad k = k(n), \quad p \geq 0$.
Compute the marginal standardization function based on ranks
    $$\widehat{T}(x) = \big(1/(1 - \widehat{F}_j(x_j)\big)_{j \in \{1,\dots,d\}}, \qquad x \in \mathbb{R}^d \qquad (5.1)$$
Standardize the training data $\widehat{V}_i := \widehat{T}(X_i)$
Assign to each $\widehat{V}_i$ the cone $R_a^\epsilon$ it belongs to.
Compute $\widehat{\mathcal{M}}(a)$ from (4.6).
(Optional) Set to $0$ any $\widehat{\mathcal{M}}(a)$ below some small threshold $m$ (e.g. a small fraction $p$ of the total empirical mass $\sum_a \widehat{\mathcal{M}}(a)$)
Output 1: (sparse) representation of the dependence structure
    $$\left\{\widehat{\mathcal{M}}(a) : \emptyset \neq a \subset \{1,\dots,d\}\right\}.$$
Output 2: Scoring function
    $$s_n(x) := (1/\|\widehat{T}(x)\|_\infty) \sum_a \widehat{\mathcal{M}}(a) \mathbf{1}\{\widehat{T}(x) \in R_a^\epsilon\}.$$
```

Notice that the scoring function issued by DAMEX is not intended to approximate a density function (or any non decreasing transformation of it) of the tail probability measure $\mu(\,\cdot\,)/\mu(B_1(0)^c)$, which in pseudo-polar coordinates is proportional to $\|v\|^{-2}\frac{\mathrm{d}\Phi}{\mathrm{d}\eta}(\theta(v))$, where Φ is the angular measure (2.7), η is any appropriate reference measure on $\mathbb{S}_+$ (*e.g.* Φ itself). Instead in DAMEX the radial contribution is proportional to $1/\|v\|$. This choice is in part motivated by the interpretability of such a score in terms of probability of a failure region as discussed above. Most importantly, in a high dimensional context where we assume that the support of μ is concentrated on lower dimensional subspaces, there is no universal (or at least consensual, as it is the case with the Lebesgue measure in moderate dimension) dominating reference measure according to which the anomalies could be assumed to be distributed.

Experiments have been carried on with five reference AD datasets: *shuttle*, *forestcover*, *http*, *SF* and *SA*. These datasets are available for instance on he UCI Machine learning repository. The experiments are performed in a semi-supervised framework (the training set consists of normal data) with various values of the hyper-parameters. DAMEX significantly improves the performance (both in term of precision and of ROC curves) in extreme regions for each considered dataset. One should notice that DAMEX may be combined with any standard AD algorithm to handle extreme *and* non-extreme data by splitting the input space between an extreme region and a non-extreme one, then using Algorithm 2 to treat new observations that appear in the extreme region, and the standard algorithm to deal with those which appear in the non-extreme region.

5.2 Anomaly detection in moderate dimension using spherical Mass-Volume sets

The purpose of the paper Thomas et al. (2017) is to promote an anomaly detection algorithm in moderate dimensional multivariate problems based on Multivariate EVT. Here, 'moderate' means that we may assume that the angular measure of extremes is concentrated on the interior of the positive orthant of the unit sphere. This would typically be the case on subspaces indexed by components $a \subset \{1, \dots, d\}$ issued from the DAMEX algorithm, such that $|a| \geq 2$ and for any $b \subsetneq a, \mu(\mathcal{C}_b) = 0$. In this work, as in Section 5.1, anomalies are sought among extreme observations, *i.e.* among observations which norm –after standardization– exceeds a large quantile. The main idea consists in applying a classical multivariate anomaly detection approach, that is based on minimum volume sets (MV-sets in short) estimation, to the angular component of the standardized variable. As in previous chapters for $v \in \mathbb{R}^d$ we use the pseudo polar decomposition $x = r(v)\theta(v)$ with $r(v) = \|v\|$ (*radius*) and $\theta(v) = \|v\|^{-1}v$ (*angle*) with $\|v\| = \|v\|_\infty$ throughout this section. The choice of the sup norm is mainly dictated by algorithmic reasons as it facilitates the construction of empirical MV-sets as detailed below. Another incentive for the use of such a norm is that it facilitates the connection with previous works Goix et al. (2016, 2017).

MV-sets Given a random vector Z taking its values in $\mathcal{Z} \subset \mathbb{R}^d$ with $d \geq 1$, MV-sets correspond to subsets of the feature space $\mathcal{Z} \subset \mathbb{R}^d$ where the probability distribution P of the random variable Z is most concentrated. More precisely, given a measure $\lambda(dz)$ of reference on the space $\mathcal{Z}$ equipped with its Borel σ-

algebra $\mathcal{B}(\mathcal{Z})$ and $\alpha \in (0,1)$, a MV-set of level α for Z is any solution Ω_α^* of the problem:

$$\min_{\Omega \in \mathcal{B}(\mathcal{Z})} \lambda(\Omega) \text{ subject to } \mathbb{P}\left(Z \in \Omega\right) \geq \alpha\,, \tag{5.2}$$

generalizing the well-known notion of quantile for 1-dimensional distributions, refer to Einmahl and Mason (1992); Polonik (1997) for details on minimum volume set theory and to Scott and Nowak (2006a); Vert and Vert (2006) for related statistical learning results. State-of-the-art methods for MV-sets estimation and anomaly detection (*e.g.* Scott and Nowak (2006a); Schölkopf et al. (2001); Liu et al. (2008)) are usually sensitive to scaling effects and consider a fixed level $\alpha \in (0,1)$ in their theoretical analysis (*e.g.* Scott and Nowak (2006a); Vert and Vert (2006)) whereas the approach we suggest is concerned with extreme regions (the level α tends to 1) and is insensitive to scaling effects.

Assume that Z's distribution P is absolutely continuous w.r.t. λ and denote by $f(z) = dP/d\lambda(z)$ the related density. For any $\alpha \in (0,1)$, under the assumption that the density f is bounded and $f(Z)$ has a continuous distribution F_f, one may show (Polonik, 1997) that the set $\Omega_\alpha^* = \{z \in \mathcal{Z} : \ f(z) \geq F_f^{\leftarrow}(1-\alpha)\}$ is the unique solution of the *minimum volume set* problem (5.2), where for any *c.d.f.* $K(t)$ on $\mathbb{R}$, $K^{\leftarrow}$ denotes the left-continuous inverse $K^{\leftarrow}(u) = \inf\{t \in \mathbb{R} : \ K(t) \geq u\}$. For high values of the mass level α, minimum volume sets are expected to contain the modes of the distribution, whereas their complementary sets correspond to *abnormal observations*.

Empirical MV-sets. A mass level $\alpha \in (0,1)$ being preliminarily fixed, estimating an empirical MV-set consists in building from training data $Z_1, \ \ldots, \ Z_n$ an estimate of a specific density level set Ω_α^* by solving a natural statistical counterpart of problem (5.2):

$$\min_{\Omega \in \mathcal{G}} \lambda(\Omega) \text{ subject to } P_n(\Omega) \geq \alpha - \psi_n\,, \tag{5.3}$$

where ψ_n plays the role of a *tolerance parameter*, and where optimization is restricted to a subset $\mathcal{G}$ of $\mathcal{B}(\mathcal{Z})$. The class $\mathcal{G}$ is supposed to be rich enough to include Ω_α^* or a reasonable approximation of it. It is ideally made of sets Ω whose volume $\lambda(\Omega)$ can be efficiently computed or estimated, *e.g.* by Monte-Carlo simulation. The empirical distribution based on the training sample (or a smoothed version of the latter) $P_n = (1/n)\sum_{i=1}^n \delta_{z_i}$ replaces P and the tolerance parameter Ψ_n is chosen of the same order of magnitude as the supremum $\sup_{\Omega \in \mathcal{G}} |P_n(\Omega) - P(\Omega)|$. Under usual complexity assumptions on the class $\mathcal{G}$ combined with an appropriate choice of ψ_n, non-asymptotic statistical guarantees for solutions $\widehat{\Omega}_\alpha$ of (5.3) are given in Scott and Nowak (2006a), together with algorithmic approaches to compute such solutions.

MV-sets and multivariate EVT Our approach may be summarized as follows: Since the angular measure Φ encapsulates the dependence structure of the tail, recovering MV-sets on the sphere of high mass (*i.e.* corresponding to high values of α) for the angular measure Φ gives access to the most probable directions of extremes. In the case where the angular component alone should be considered for anomaly detection, those angular MV-sets would allow to pin the complementary sets as abnormal. In practice, the radial part does play a role (see equation (2.7)) and we define an anomaly score which is a product of a radial score and an angular score based on a family of nested MV-sets.

Our approach shares similarity with (e.g. Cai et al., 2011), where estimation of low levels of the density function using multivariate EVT is also considered in a somewhat different context, that is assuming joint regular variation with a single regular variation index as in (2.6). In the cited reference, consistency of the extreme level sets is established. Here we take a different approach by assuming regular variation of the standardized vector V and working with preliminary standardized data. Also we obtain non-asymptotic upper bounds concerning the estimated level sets. Despite this seemingly stronger result our work may not be seen as an improvement of Cai et al. (2011) in that we do not take into account the impact of marginal standardization (which amounts to assuming that the marginal distributions are known). Of course, the feature variables $\widehat{V}_i = \widehat{T}(X_i)$ obtained through (5.1) are not independent anymore and analyzing the accuracy of an estimate of the angular distribution Φ is far from straightforward. However, it has been shown in Einmahl et al. (2001); Einmahl and Segers (2009) that using the rank transformed variables $\widehat{V}_i$'s instead of the probability integral transformed ones V_i does not damage the asymptotic properties of the empirical estimator of the angular measure (in dimension 2, under suitable regularity assumptions). In arbitrary dimension, as presented in Section 2.2, Goix et al. (2015) have obtained a similar result for the finite sample case, concerning an alternative characterization of the angular measure, which is an integrated version of Φ. Relaxing the (unrealistic) assumption of known margins without substantial worsening the upper bound amounts to establishing concentration results on the empirical angular measure, which is the subject of ongoing work, see Section 7.1.

We now rigorously formulate the MV-sets statistical problem on the sphere. Denoting by λ_d the Lebesgue measure on $\mathbb{S}_+$ equipped with its Borel σ-algebra $\mathcal{B}(\mathbb{S}_+)$, the generic goal in a MV-set context is to recover from training observations $X_1, \ldots, X_n$ which are independent copies of the generic heavy-tailed *r.v.* X, a solution of the problem $\min_{\Omega \in \mathcal{B}(\mathbb{S}_+)} \lambda_d(\Omega)$ subject to $\Phi(\Omega) \geq \alpha$. Note that $\alpha \in (0, \Phi(\mathbb{S}_+))$, instead of $\alpha \in (0,1)$, as Φ is not a probability distribution.

In practice, the angular measure Φ is an asymptotic object, whereas the data at hand is non asymptotic. Also, it may be argued from a practical perspective that our interest lies in large, but non asymptotic regions $\{x : \|T(x))\| > t\}$ where T

is the transform (2.4). In this work we thus consider the *sub-asymptotic* angular measure at finite level t, $\Phi_t(\Omega) = t\,\mathbb{P}(\|V\| > t, V/\|V\| \in \Omega)$ and notice from (2.5) and (2.7) that $\Phi_t(\Omega) \to \Phi(\Omega)$ as $t \to \infty$ as soon as $\Phi(\partial\Omega) = 0$. In the sequel we shall thus consider the modified, non asymptotic optimization problem

$$\min_{\Omega \in \mathcal{B}(\mathbb{S}_+)} \lambda_d(\Omega) \quad \text{subject to} \quad \Phi_t(\Omega) \geq \alpha\,. \tag{5.4}$$

In order to ensure existence and uniqueness of the solution of this optimization problem, we consider the following assumptions, which are commonly used in the MV-set literature to ensure the existence and uniqueness of the MV-set optimization problem Polonik (1997).

$\boldsymbol{A_1}$ For any $t > 1$, the distribution $\Phi_t(\,\cdot\,)$ is absolutely continuous w.r.t. the Lebesgue measure λ_d on $\mathbb{S}_+$ with density ϕ_t. In addition, the *r.v.* $\phi_t(\theta(V))$ has no flat parts: $\forall c > 0$, $\mathbb{P}\{\phi_t(\theta(V)) = c\} = 0$.

$\boldsymbol{A_2}$ The density $\phi_t(\theta)$ of $\Phi_t(\,\cdot\,)$ is uniformly bounded:

$$\sup_{t>1,\theta\in\mathbb{S}_+} \phi_t(\theta) < \infty.$$

Given assumptions $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ one can show that (5.4) has a unique solution, given by the density level set $B^*_{\alpha,t} = \{\theta \in \mathbb{S}_+ :\ \phi_t(\theta) \geq K^{-1}_{\Phi_t}(\Phi(\mathbb{S}_+) - \alpha)\}$, where $K_{\Phi_t}(y) = \Phi_t(\{\theta \in \mathbb{S}_+ :\ \phi_t(\theta) \leq y\})$.

The general method described next consists in replacing in (5.4) the angular measure Φ_t by a sharp estimate, involving a fraction of the original observations (*i.e.* the most extreme observations).

Empirical estimation The algorithm we propose to estimate an MV-set of the distribution of extreme data directions is implemented in three main steps described in Algorithm 3. The output is meant to approach a MV-set of the angular measure Φ_t for $t = n/k$, where $k \in \{1,\ \ldots,\ n\}$ is the number of extreme observations to be retained along each axis. The choice of k should depend on n, in the sense that $k = o(n)$ and $k \to \infty$ as $n \to \infty$. The practical choice of k results from a bias/variance trade-off which is a recurrent issue in extreme values analysis, that we shall not investigate. In practice, k is chosen in a stability region of the output, and $k = O(\sqrt{n})$ appears to be a reasonable default choice.

Statistical guarantees Statistical guarantees for the general algorithm 3 and a practical method for solving the optimization problem (5.5) it involves are detailed next. from a practical perspective, a crucial advantage of the approach we promote lies in the compactness of the feature space $\mathbb{S}_+$ used to detect abnormal directions. Our analysis proceeds as if the marginal distributions were known, *i.e.* as if

Algorithm 3 Empirical estimation of an angular MV-set

Inputs: Training data set $\{X_1, \ \ldots, \ X_n\}$, $k \in \{1, \ \ldots, \ n\}$, mass level α, tolerance $\psi_k(\delta)$, confidence level $1-\delta$, collection $\mathcal{G}$ of subsets of $\mathbb{S}_+$

Standardization: Apply the rank-transformation (5.1) to the X_i's, yielding the empirically marginally standardized vectors $\widehat{V}_i = \widehat{T}(X_i)$, $i = 1, \ \ldots, \ n$.

Thresholding: Retain the indexes

$$\begin{aligned}\mathcal{I} &= \left\{ i \in \{1, \ldots, n\} : r(\widehat{V}_i) \geq \frac{n}{k} \right\} \\ &= \left\{ i \in \{1, \ldots, n\} : \exists j \leq d, \widehat{F}_j(X_i^{(j)}) \geq 1 - k/n \right\}\end{aligned}$$

and consider the angles $\theta_i = \theta(\widehat{V}_i)$ for $i \in \mathcal{I}$.

Empirical MV-set estimation: Form the empirical angular measure $\widehat{\Phi} = (1/k) \sum_{i \in \mathcal{I}} \delta_{\theta_i}$ and solve the constrained minimization problem.

$$\min_{\Omega \in \mathcal{G}} \lambda_d(\Omega) \text{ subject to } \widehat{\Phi}(\Omega) \geq \alpha - \psi_k(\delta)\,. \tag{5.5}$$

Output: Estimated MV-set $\widehat{\Omega}_\alpha \in \mathcal{G}$ of the angular measure $\Phi_{n/k}$.

the true transformed variables V_i's were observables. Controlling the additional sample error induced by the discrepancy $\widehat{V}_i - V_i$ is reserved for future work.

The result stated below shows that with high probability over the data set the empirical MV-set estimated on the extremes is an approximation of the true MV-set.

Theorem 5.1. *Assume that assumptions $\boldsymbol{A_1} - \boldsymbol{A_2}$ are fulfilled by the finite distance angular measure $\Phi_t, t \geq 1$ related to X's heavy-tailed distribution with λ_d as reference measure. Let $\mathcal{G}$ be a finite class of sets with cardinality $|\mathcal{G}|$.*

Fix a mass level α and $\delta \in (0, 1)$ and consider the empirical MV*-set $\widehat{\Omega}_\alpha$ solution of* (5.5) *where the empirical $\widehat{V}_i$'s are replaced with V_i's in the definition of $\widehat{\Phi}$, and where the tolerance is set to*

$$\psi_k(\delta) = \sqrt{\frac{d}{k}} \Big[2\sqrt{2 \ln(|\mathcal{G}|)} + 3\sqrt{\ln(1/\delta)} \Big].$$

Then, with probability at least $1 - \delta$, we simultaneously have:

$$\left\{ \Phi_{n/k}(\widehat{\Omega}_\alpha) \geq \alpha - 2\psi_k(\delta) \right\} \textit{ and } \left\{ \lambda_d(\widehat{\Omega}_\alpha) \leq \inf_{\Omega \in \mathcal{G}_\alpha} \lambda_d(\Omega) \right\},$$

where $\mathcal{G}_\alpha = \{\Omega \in \mathcal{G}, \Phi(\Omega) \geq \alpha\}$.

As expected, the rate of statistical recovery of the solution $B^*_{\alpha,n/k}$ of (5.4) when $t = n/k$ is of order $O_{\mathbb{P}}(\sqrt{1/k})$, the learning procedure involving the $|\mathcal{I}| \in [k, dk]$ most extreme standardized observations only.

Remark 5.2 (On the finite class assumption). The argument originally developed in Goix et al. (2015) for controlling the accuracy of an empirical estimation of the STDF is crucially exploited to cope with the dependence structure of the transformed variable V. The finite class assumption fits our purposes in the present paper, since we consider unions of rectangles paving the sphere as described below A minor modification of the proof would allow to replace $\log(|\mathcal{G}|)$ with $\mathcal{V}_{\mathcal{G}} \ln(dke/V_{\mathcal{G}})$, where $V_{\mathcal{G}}$ is the VC-dimension of the class $\mathcal{G}$, and dk is an upper bound for the average number of points hitting the extreme regions (see the proof of Lemma 1 in the Supplementary Material of Thomas et al. (2017)). Then the learning rate bound given by the result above is of order $\mathrm{O}(\sqrt{(\ln k)/k})$, as expected, since $\mathrm{O}(k)$ observations are actually involved in the learning procedure, due to the thresholding stage.

Computational aspects We build empirical MV-sets on the sphere by binding together elementary subsets S of $\mathbb{S}_+$ with same volume (*i.e.* same Lebesgue measure $\lambda_d(S)$). Again, empirical estimation $\widehat{\Phi}$ of the angular measure is based on the fraction $\{\theta(\widehat{V}_i)_i : \ i \in \mathcal{I}\}$ of the transformed data (see Algorithm 3) and we consider the partition of $\mathbb{S}_+$ in dJ^{d-1} hypercubes S_j with same volume as shown in Figure 5.1.

We therefore consider the class $\mathcal{G}$ that corresponds to the class $\mathcal{G}_J$ of subsets obtained as a union of cubes S_j. In this case, $|\mathcal{G}| = \exp(dJ^{d-1}\ln 2)$. Figure 5.1 shows an example of such a partition for $d = 3$ and $J = 5$. Sorting the elements by decreasing order with respect to the number of samples they contain and binding them together until reaching a mass greater than $\alpha - \psi_k(\delta)$ yields $\widehat{\Omega}_\alpha$ (see Scott and Nowak (2006a)).

The number of hypercubes of the partition increases exponentially with the dimension d. Therefore as d increases, most hypercubes will be empty and there is no need to take them into account when sorting the elements of the partition. The solution is to rather loop over the samples $\theta(V_i)$, $i \in \mathcal{I}$ and apply a geometric hash function assigning a signature to each sample. The signature of a sample $\theta(V)$ characterizes the hypercube it belongs to. Such a signature can be defined as the sign of $\langle \boldsymbol{e}_p, \theta(V)\rangle - j/J$ for $p \in \{1, \dots, d\}$, $j \in \{1, \dots, J\}$, where $\boldsymbol{e}_p$ denotes the vector of $\mathbb{R}^d$ such that $\boldsymbol{e}_p^{(\ell)} = \delta_{i\ell}$ for all $\ell \in \{1, \dots, d\}$. The hash function thus takes its values in $\{-1, 1\}^{dJ}$. Its computation for one $\theta(V_i)$ requires a single loop over the dimensions $\ell \in \{1, \dots, d\}$ and examination of the integer part of $Jx^{(\ell)}$. The complexity for m samples is thus $O(dm)$.

The number of unique signatures is equal to the number of non empty hypercubes of the partition and the number of identical signatures is equal to the number of samples in the corresponding hypercube. We have therefore identified all the non empty hypercubes and the number of samples in each of them. Using Algorithm 4 we then obtain an estimated MV-set with level mass α, *i.e.* , the solution of (5.5).

Algorithm 4 Solution of (5.5) when $\mathcal{G}$ is the regular grid on $\mathbb{S}_+$

Sorting: Sort the elementary subsets S_j so that: $\widehat{\Phi}(S_{(1)}) \geq \ldots \geq \widehat{\Phi}(S_{(J)})$.
Concatenation: Bind together the elementary subsets sequentially, until the empirical angular measure of the resulting set exceeds $\alpha - \psi_k(\delta)$, yielding the region

$$\widehat{\Omega}_{J,\alpha} = \bigcup_{j=1}^{J(\alpha)} S_{(j)} \,, \tag{5.6}$$

where $J(\alpha) = \min\{j \geq 1 : \sum_{j=1}^{J} \widehat{\Phi}(S_{(j)}) \geq \alpha - \psi_k(\delta)\}$

Remark 5.3. While the complexity of the algorithm is linear in the dimension d, this approach suffers from the curse of dimensionality. Indeed, as the number of hypercubes increases exponentially with d, only a small proportion of hypercubes will be non-empty and the solution will tend to overfit.

Remark 5.4. When implementing the hash function we have to carefully deal with the samples $\theta(V)$ that are located on the edges of $\mathbb{S}_+$, *i.e.* , such that at least two of their components are equal to 1. Under assumption $\mathbf{A}_1$, the probability of a sample $\theta(V)$ to be located on an edge of $\mathbb{S}_+$ is equal to 0. However it is not always the case in practice, especially if we use the empirical marginals for the standardization step. The hash function defined above assigns a signature to an edge sample $\theta(V)$ that is equal to none of the signatures of the adjacent hypercubes of $\theta(V)$. Therefore we arbitrarily assign such samples to one of their adjacent hypercubes.

Remark 5.5 (Bias induced by the finite grid.). Looking for the MV-set in the class $\mathcal{G}$ instead of all the measurable subsets of the sphere induces a bias which can be controlled with mild assumptions on the angular distribution, such as the box counting class introduced in Scott and Nowak (2006b).

Remark 5.6 (Model Selection). The resolution level J should be chosen with care as it can impact significantly the MV-set estimation procedure. This issue can be addressed through *complexity penalization* (see Supplementary Material). However for the numerical experiments we resort to cross validation selecting the resolution level giving an empirical angular mass close to α on a test set. Indeed if the

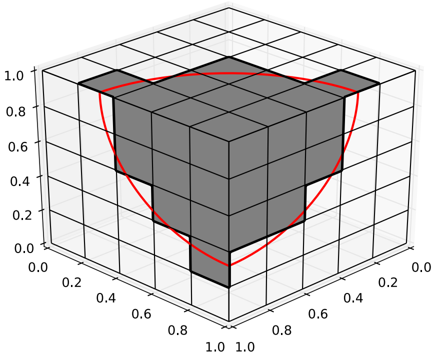


Figure 5.1 – Estimated angular MV-set on the sphere based on Gaussian data. In red, the border of the true MV-set with mass at least 0.9. In gray the estimated MV-set with relative level mass 0.9 and $J = 5$.

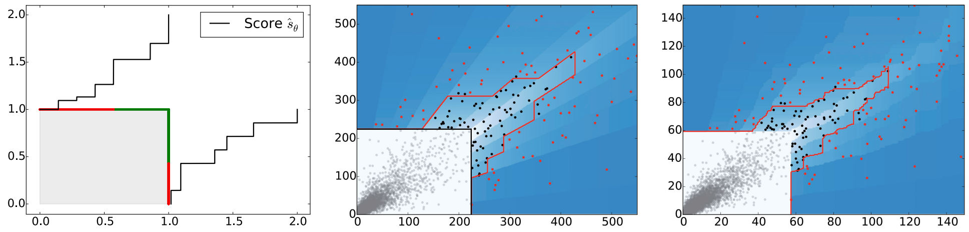
grid is too coarse, the estimated MV-set should have an empirical measure much greater than α on a test set. Similarly, if the grid is too fine, the estimated MV-set will have an empirical measure much smaller than α on a test set.

Application to Anomaly Detection As already mentioned above considering angular MV-sets only does not yield an optimal decision function, since the density of the largest observations includes a radial part. the density (with respect to $dr \otimes d\theta$) on the most extreme regions is proportional to $\frac{1}{r^2}\phi(\theta)$. A standard approach in anomaly detection is to define a scoring function $\hat{s}$, which should be ideally proportional to the density, and then to declare as abnormal regions of the kind $\{x : \hat{s}(x) \leq s_0\}$, where s_0 can be tuned so that a given proportion of the samples are pinned as abnormal. It turns out that as a byproduct of our algorithm, we can also estimate a scoring function $\hat{s}_\theta$ on $\mathbb{S}_+$, such that the smaller $\hat{s}_\theta(\boldsymbol{\theta})$ is, the more abnormal the direction $\boldsymbol{\theta}$. We define $\hat{s}_\theta$ as the piecewise constant function defined on each hypercube of the partition of $\mathbb{S}_+$ by the number of samples it contains (see Figure 5.2(a)). One can then consider the scoring function on the whole space defined by

$$\hat{s}(r(\boldsymbol{V}), \theta(\boldsymbol{V})) = 1/r(\boldsymbol{V})^2 \cdot \hat{s}_\theta(\theta(\boldsymbol{V})). \tag{5.7}$$

Again, the smaller $\hat{s}(r(\boldsymbol{V}), \theta(\boldsymbol{V}))$ is, the more abnormal $(r(\boldsymbol{V}), \theta(\boldsymbol{V}))$, i.e. $\boldsymbol{V}$. Using such a scoring function, observations with very large sup norm but with high angular score have a chance to be considered as anomalies, which would not be the case if the MV-set estimates on $\mathbb{S}_+$ only were considered. x

Numerical Experiments We compare our approach to two state-of-the-art unsupervised anomaly detection algorithms, Isolation Forest (Liu et al., 2008) and



(a) Angular score on the sphere

(b) Standardized space (V)

(c) Input space (X).

Figure 5.2 – Illustration of our approach on a sample generated from a logistic model. Figure (a) shows the angular score obtained with our algorithm. In (b) and (c) the red contour shows the frontier between abnormal and normal regions. Non extreme samples are in gray and extreme anomalies are in red.

One-Class SVM (OCSVM) (Schölkopf et al., 2001), on five real data sets (shuttle, SF, http, ann, forestcover) available on the UCI ML repository. We set $k = \sqrt{n}$ in all experiments. As we do not know the normalization constant $\Phi(\mathbb{S})$ of the angular measure, we use $|\mathcal{I}|$ to normalize the empirical angular measure and consider relative mass levels in $(0, 1)$. The penalty $\psi_k(\delta)$ in (5.5) would require k to be too large to allow us to consider it in practice. We therefore solve the optimization problem (5.5) setting $\psi_k(\delta) = 0$. One may think that the connection with the theoretical result would be lost, however by corollary 12 in Scott and Nowak (2006a), one can solve the empirical minimum volume set optimization problem without the tolerance parameter in the mass constraint and obtain a theoretical result similar to the one of Theorem 5.1. Finally, we use the implementation of Isolation Forest and OCSVM provided by Scikit-Learn Pedregosa et al. (2011). In all experiments, the suggested algorithm, Isolation Forest and OCSVM are trained on half of the normal instances, chosen at random. The test set for both algorithms consists in all instances (normal and abnormal) not used in the training set. This test set is then restricted to the extreme region in accordance with the thresholding step of Algorithm 1 and performance is assessed with the available labels. Areas under the Receiver Operating Characteristic curve (ROC-AUC) obtained on all data sets show that our approach outperforms Isolation Forest and OCSVM in the extreme region on three out of five data sets and is never the worst one.

5.3 Clustering and Visualization of tail events

This section presents the work published in Chiapino et al. (2019a) aiming at clustering and visualizing tail events. The initial motivation of this work is an aeronautic application. Here the monitored quantities are various flight data aiming at building system health indicators. Since any unusual high value is a potential indicator of a preset or future anomaly, in this particular context all tail events (*i.e.* points with a unusually high norm) may be considered as anomalies. Thus, contrarily to the previous sections of this chapter, the aim is not to differentiate anomalies from normal data in the tail, but rather to identify and visualize tail patterns in a fully unsupervised manner.

In the unsupervised setting, several extensions of the basic linear Principal Component Analysis for dimensionality reduction and visualization techniques have been proposed in the statistics and data-mining literature, accounting for non linearities or increasing robustness for instance, *cf* Gorban et al. (2008) and Kriegel et al. (2008). These approaches intend to describe parsimoniously the 'center' of a massive data distribution, see *e.g.* Naik (2017) and the references therein. Similarly, for clustering purposes, several multivariate heavy-tailed distributions have been proposed that are robust to the presence of outliers, see *e.g.* Forbes and Wraith (2014), Punzo and Tortora (2018). However the issue of clustering extremes or outliers is only recently receiving attention, at the instigation of industrial applications such as those mentioned above and because of the increasing availability of extreme observations in databases: generally out-of-sample in the past, extreme values are becoming observable in the Big Data era. In Chiapino et al. (2019a) we propose a novel mixture model-based approach for clustering extremes in the multivariate setup. It relies on a dimensionality reduction technique of the tail distribution summarized by the DAMEX algorithm fromGoix et al. (2016, 2017) described in Sections 4.1 and 5.1. In practice, a sparse representation of the extremal dependence structure is obtained with DAMEX when only a few such groups of variables can be exhibited (compared to $2^d - 1$) and/or when these groups involve a small number of variables (with respect to d). Here we develop this framework further, in order to propose a (soft) clustering technique in the region of extremes and derive effective 2-d visual displays, shedding light on the structure of anomalies/extremes in sparse situations. This is achieved by modelling the distribution of extremes as a specific *mixture model*, where each component generates a different type α of extremes. In this respect, the present paper may be seen as an extension of Boldi and Davison (2007); Sabourin and Naveau (2014), where a Bayesian inference framework is designed for moderate dimensions ($d \leq 10$ say) and situations where the sole group of variables with the potential of being simultaneously large is $\{1, \dots, d\}$ itself. In the context of mixture modelling (see *e.g.* Fruhwirth-Schnatter et al. (2018)), the Expectation-Maximization algorithm (EM) permits to partition/cluster the set of extremal data through the statistical recovery of *latent observations*, as well as posterior probability distributions (inducing a soft clustering of the data in a straightforward manner) and, as a by-product, a similarity measure on the set of extremes: the higher the probability that their latent variables are equal, the more similar two extreme observations X and X' are considered. The similarity matrix thus obtained naturally defines a *weighted graph*, whose vertices are the anomalies/extremes observed, paving the way for the use of powerful graph-mining techniques for community detection and visualization, see *e.g.* Schaeffer (2007), Hu and Shi (2015) and the references therein. Beyond its detailed description, the methodology proposed is applied to a real fleet monitoring dataset in the aeronautics domain and shown to provide useful tools for analyzing and interpreting abnormal data.

A mixture model for high dimensional extremes We place ourselves in the probabilistic framework presented in Section 4.1, up to a change of norm: to facilitate probabilistic modeling on the unit sphere – namely the use of Dirichlet distribution – we use the L^1 norm, $\|v\| = \sum_1^d |v_j|$. The angular and radial components of the transformed variable $R = r(V) = \|V\|$ and $\Theta = \theta(V) = V/\|V\|$ are defined accordingly. The positive orthant of the unit sphere is then the unit simplex with our choice of norm and it is a well known fact that in multivariate EVT that the unit-Pareto standardization combined with the regular variation assumption (2.5) implies

$$\int_{\mathbb{S}_+} \theta_i \, \mathrm{d}\Phi(\theta) = 1, \text{ for } i = 1, \dots, d. \tag{5.8}$$

In addition, the normalizing constant is explicit:

$$\Phi(\mathbb{S}_+) = \int_{\mathbb{S}_+} (\theta_1 + \ldots + \theta_d) \, \mathrm{d}\Phi(\theta) = d. \tag{5.9}$$

Recall from Section 4.1 the notation $\mathcal{M} = (\mu(\mathcal{C}_a), \emptyset \neq a \subset \{1, \dots, d\})$. The change of norm does not define the set of non-zero entries in $\mathcal{M}$, which we denote by $\mathbb{M}$. As a first step we develop a novel mixture model for the angular distribution Φ of the largest instances of the dataset, indexed by $a \in \mathbb{M}$. Each component $a \in \mathbb{M}$ of the mixture generates instances V such that V_j is likely to be large for $j \in a$ and the latent variables of the model take their values in $\mathbb{M}$. In practice, we adopt a *plug-in* approach and identify $\mathbb{M}$ with $\{\widehat{\mathcal{M}}(a) : \widehat{\mathcal{M}}(a) \neq 0\}$, the output of DAMEX. As the distribution of extremes may be entirely characterized by the distribution of their angular component $\Theta = \theta(V) \in \mathbb{S}_+$, a natural model is that of Dirichlet mixtures. We next show how to design a 'noisy' version of the model for subasymptotic observations and how to infer it by means of an EM procedure based on a truncated version of the original dataset, surmounting difficulties related to the geometry of Φ's support. Let K denote the number of subsets $a \in \mathbb{M}$ of cardinality at least 2 and let $d_1 \in \{0, \dots, d\}$ be the number of singletons $\{j\} \in \mathbb{M}$. Without loss of generality we assume that these singletons correspond to the first d_1 coordinates, so that $\mathbb{M} = \{a_1, \dots, a_K, \{1\}, \dots, \{d_1\}\}$. For simplicity, we also suppose that the sets $a \in \mathbb{M}$ are not nested, an hypothesis which can be relaxed at the price of additional notational complexity. In view of (4.2), the angular measure then admits the decomposition

$$d^{-1}\Phi(\,\cdot\,) = \sum_{\ell=1}^{K} \pi_\ell \Phi_{a_\ell}(\,\cdot\,) + \sum_{j \le d_1} \pi_{K+j} \delta_{\boldsymbol{e}_j}(\,\cdot\,),$$

where Φ_{a_ℓ} is a probability measure on $\mathbb{S}_{+,a_\ell} = \mathbb{S}_+ \cap \{\theta \in \mathbb{R}_+^d : \theta_j = 0 \text{ for } j \notin a_\ell\}$, the weights π_ℓ satisfy $\sum_{\ell \le K+d_1} \pi_\ell = 1$ and $\boldsymbol{e}_j = (0, \dots, 1, \dots, 0)$ is the j^{th} canonical basis vector of $\mathbb{R}^d$.

The singletons weights derive immediately from the moment constraint (5.8): for $j \le d_1$, it is easily shown that our assumptions imply $\pi_{K+j} = d^{-1}$ so that

$$\Phi(\,\cdot\,) = d \sum_{k=1}^{K} \pi_\ell \Phi_{a_\ell}(\,\cdot\,) + \sum_{j \le d_1} \delta_{\boldsymbol{e}_j}(\,\cdot\,), \tag{5.10}$$

where the vector $\boldsymbol{\pi} \in [0,1]^{K+d_1}$ must satisfy

$$\sum_{\ell=1}^{K} \pi_\ell = 1 - d_1/d. \tag{5.11}$$

For likelihood-based inference, a parametric model for each component Φ_{a_ℓ} of the angular measure must be specified. One natural model for probability distributions on a simplex is the Dirichlet family, which provides a widely used prior in Bayesian statistics for data clustering purposes in particular. We recall that the Dirichlet distribution on a simplex $\mathbb{S}_{+,a}$ admits a density φ_a with respect to the $(|a|-1)$-dimensional Lebesgue measure which is denoted by $\mathrm{d}\boldsymbol{w}$ for simplicity. It can be parameterized by a mean vector $\boldsymbol{m}_a \in \mathbb{S}_{+,a}$ and a concentration parameter $\nu_a > 0$, so that for $\theta \in \mathcal{S}_a$,

$$\varphi_a(\theta | \boldsymbol{m}_a, \nu_a) = \frac{\Gamma(\nu_a)}{\prod_{j \in a} \Gamma(\nu_a m_{a,i})} \prod_{j \in a} \theta_j^{\nu_a m_{a,j} - 1}.$$

Refer to *e.g.* Müller and Quintana (2004) for an account of Dirichlet processes and mixtures of Dirichlet Processes applied to Bayesian nonparametrics. We emphasize that our context is quite different: a Dirichlet Mixture is used here as a model for the angular component of the largest observations, not as a prior on parameters. This modeling strategy for extreme values was first proposed in Boldi and Davison (2007) and revisited in Sabourin and Naveau (2014) to handle the moment constraint (5.8) via a model re-parametrization. In both cases, the focus was on moderate dimensions. In particular, both cited references worked under the assumption that the angular measure concentrates on the central simplex $\Omega_{\{1,\dots,d\}}$ only. In this low dimensional context, the main purpose of the cited authors was to derive the posterior predictive angular distribution in a Bayesian framework, using a variable number of mixture components concentrating on $\Omega_{\{1,\dots,d\}}$. Since the set of Dirichlet mixture distributions with an arbitrary number of components is dense among all probability densities on the simplex, this model permits in theory to approach any angular measure for extremes. The scope of the present paper is different. Indeed we are concerned with high dimensional data (say $d \simeq 100$) and consequently we do not attempt to model the finest details of the angular measure. Instead we intend to design a model accounting only for information which is relevant for clustering. Since an intuitive summary of an extreme event in a high dimensional context is the subset a of features it involves, we assign one mixture component per sub-simplex Ω_a such that $a \in \mathbb{M}$. Thus we model each Φ_a by a single Dirichlet distribution with unknown parameters $\boldsymbol{m}_a, \nu_a$. Using the standard fact that for such a distribution, $\int_{\mathbb{S}_a} \theta\varphi_a(\theta|\boldsymbol{m}_a, \nu_a)\,\mathrm{d}\theta = \boldsymbol{m}_a$, the moment constraint (5.8) becomes:

$$\frac{1}{d} = \sum_{\ell=1}^{K} \pi_\ell \boldsymbol{m}_{\ell,j}, \quad j \in \{d_1+1,\dots,d\}, \tag{5.12}$$

where $\boldsymbol{m}_\ell = \boldsymbol{m}_{a_\ell}$ for $\ell \leq K$.

Statistical model for large but sub-asymptotic data Recall from (2.7) that Φ is the *limiting* distribution of $\Theta = \theta(V)$ for large $R = r(V)$. In practice, we dispose of no realization of this limit probability measure and the observed angles corresponding to radii $R > r_0$ follow a sub-asymptotic version of Φ. In particular, if the margins V_j have a continuous distribution, we have $\mathbb{P}(V_j \neq 0) = 1$, $j \in \{1,\dots,d\}$ so that all the observations $V_i = (V_{i,1},\dots,V_{i,d})$, $1 \leq i \leq n$, lie in the central cone $\mathcal{C}_{\{1,\dots,d\}}$, as already emphasized in Section 4.1. This is also true using the empirical versions $\hat{V}_i = \widehat{T}(V_i)$ defined in (5.1). In the approach we propose, the deviation of V from its asymptotic support, which is $\bigcup_{a\in\mathbb{M}} \mathcal{C}_a$, is accounted for by a noise $\boldsymbol{\varepsilon}$ with light tailed distribution, namely an exponential distribution. That is, we assume that $V = R\Theta + \boldsymbol{\varepsilon}$, see Model 1 below. As is usual for mixture modeling purposes, we introduce a multinomial latent variable $\boldsymbol{Z} = (Z_1, \ \dots, \ Z_{K+d_1})$ such that $\sum_\ell Z_\ell = 1$ and $Z_\ell = 1$ if W has been generated by the ℓ^{th} component of the angular mixture (5.10). In a nutshell, the type of anomaly/extreme is encoded by the latent vector $\boldsymbol{Z}$. Then, for $\ell \leq K$, $\mathbb{P}(Z_\ell = 1) = \pi_\ell$, while, for $K < \ell \leq K + d_1$, $\mathbb{P}(Z_\ell = 1) = d^{-1}$. The unknown parameters of the model are $\boldsymbol{\eta} = (\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\nu})$, where $\nu_\ell > 0$ and

$\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$, $\boldsymbol{m} = (\boldsymbol{m}_1, \dots, \boldsymbol{m}_K)$ must satisfy the constraints (5.11) and (5.12), as well as the exponential rates $\boldsymbol{\lambda} = (\lambda_1, \ \dots, \ \lambda_{K+d_1})$, where $\lambda_\ell > 0$. Figure 5.3 illustrates Model 1 in dimension $d = 3$.

Model 1 (Sub-asymptotic mixture model).

- Consider a standard regularly varying random vector V satisfying (2.5) with tail index 1 which margins satisfy $\lim_{t\to\infty} t\mathbb{P}(V_j > t) = 1$, $j = 1, \dots, d$ (typically $V_j = (1 - \hat{F}_j(X_j))$ for $\hat{F}_j$ an estimate of the marginal distribution F_j of X_j
- Let $R = \|V\|$. Fix some high radial threshold r_0, typically a large quantile of the observed radii. Let $\boldsymbol{Z}$ be a hidden variable indicating the mixture component in (5.10). Conditionally to $\{R > r_0, Z_\ell = 1\}$, V decomposes as

$$V \ = \ V_\ell + \boldsymbol{\varepsilon}_\ell \ = \ R_\ell \Theta_\ell + \boldsymbol{\varepsilon}_\ell, \tag{5.13}$$

where $V_\ell \in \mathcal{C}_{a_\ell}$, $\boldsymbol{\varepsilon}_\ell \in \mathcal{C}_{a_\ell}^\perp$, $R_\ell = \|V_\ell\|$, $\Theta_\ell = R_\ell^{-1} V_\ell \in \mathcal{S}_{a_\ell}$. The components $R_\ell, W_\ell, \boldsymbol{\varepsilon}_\ell$ are independent from each other. The noise's components are *i.i.d.* according to a translated exponential distribution with rate λ_ℓ, R_ℓ is Pareto distributed above r_0 and W_ℓ is distributed as Φ_ℓ, that is

$$\begin{cases} \mathbb{P}\left(R_\ell > r\right) = r_0 r^{-1}, r > r_0\,, \\ W_\ell \sim \Phi_\ell\,, \\ \varepsilon_j \sim 1 + \mathcal{E}xp(\lambda_\ell), j \in \{1, \dots, d\} \setminus a_\ell\,, \end{cases}$$

with $\Phi_\ell = \varphi_\ell(\,\cdot\,|\boldsymbol{m}_\ell, \nu_\ell)$ if $\ell \le K$, and $\Phi_\ell = \delta_{\mathbf{e}_{\ell-K}}$ if $K < \ell \le K + d_1$.

Statistical inference in Model 1 is carried out using an EM algorithm which is described at length in Chiapino et al. (2019a). The major issue here is the fact that the parameters to be optimized are subject to several linear constraints, which jeopardizes the convergence properties of the algorithm. Indeed the constraints are

$$\nu_\ell > 0 \; (1 \le \ell \le K)\;, \qquad \lambda_\ell > 0 \; (1 \le \ell \le K + d_1), \tag{5.14}$$

and that $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ and $\boldsymbol{m} = (\boldsymbol{m}_1, \dots, \boldsymbol{m}_K)$ satisfy (5.11) and (5.12). The latter linear constraint on $(\boldsymbol{\pi}, \boldsymbol{m})$ implies that $\boldsymbol{m}$ and $\boldsymbol{\pi}$ cannot be optimized independently, which complicates the M-step of an EM-algorithm. This major drawback has been discussed in Boldi and Davison (2007) and Sabourin and Naveau (2014) in a a lower dimensional context. In the latter work, which was part of my PhD thesis, we propose a re-parametrization of the Dirichlet Mixture model

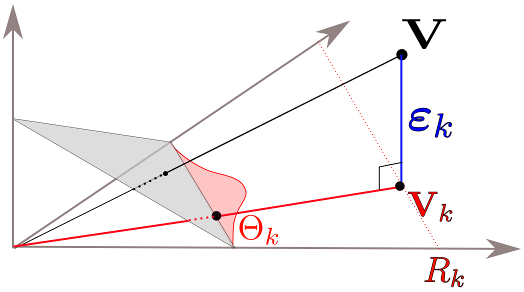


Figure 5.3 – Trivariate illustration of the sub-asymptotic model 1:
the observed point $\boldsymbol{V}$ has been generated by component $a_\ell = \{1, 2\}$. The grey triangle is the unit simplex, the shaded red area stands for the Dirichlet density φ_ℓ.

relying on barycenters of decreasing subsets of the mixture components. In the present work we propose a novel and much simpler re-parametrization which alleviates the linear constraints and hugely facilitates numerical optimization. Also, how to adapt the construction of Sabourin and Naveau (2014) in a high dimensional context where several simplices are involved remains an open question. The re-parametrization that we propose here consists in working with the product parameter $\rho_{\ell,j} = \pi_\ell m_{\ell,j}$ instead of the pair $(\pi_\ell, m_{\ell,j})$. Namely, consider a $K \times (d - d_1)$ matrix $\boldsymbol{\rho} = (\boldsymbol{\rho}_1{}^\top, \dots, \boldsymbol{\rho}_K{}^\top)$ where $\rho_{\ell,j} > 0$ for $j \in a_\ell$ and $\rho_{\ell,j} = 0$ otherwise. Then, for all $\ell \in \{1 \dots, K\}$, set

$$\pi_\ell := \sum_{j \in a_\ell} \rho_{\ell,j} \text{ and } m_{\ell,j} := \frac{\rho_{\ell,j}}{\pi_\ell}, \forall j \in a_\ell. \tag{5.15}$$

Then (5.11) and (5.12) together are equivalent to

$$\sum_{\{\ell : j \in a_\ell\}} \rho_{\ell,j} = \frac{1}{d}, \quad \forall j \in \{d_1 + 1, \dots, d\}. \tag{5.16}$$

Graph clustering and visualization The output of the EM algorithm is a pair $(\gamma, \boldsymbol{\eta})$ where $\boldsymbol{\eta}$ gathers all the estimated parameters in the model and γ is a posterior weights matrix,

$$\gamma_{i,\ell} = \mathbb{P}\left(Z_{i,\ell} = 1 \mid V_i, \boldsymbol{\eta}\right), \qquad i \leq n_0, \ell \leq K + d_1$$

where n_0 denotes the number of extreme observations.

Beyond the hard clustering that may be straightforwardly deduced from the computation of the likeliest values $z_1, \ \dots, \ z_{n_0}$ for the hidden variables given the V_i's and the parameter estimates produced by the EM algorithm, the statistical model previously introduced defines a natural structure of undirected weighted graph on the set of observed extremes, which interpretable layouts (graph drawing) can be directly derived using classical solutions. Indeed, a partition (hard clustering) of the set of (standardized) anomalies/extremes $V_1, \ldots, V_{n_0}$ is obtained by assigning membership of each V_i in a cluster (or cone/sub-simplex) determined by the component of the estimated mixture model from which it arises with highest probability: precisely, one then considers that the abnormal observation V_i is in the cluster indexed by

$$\ell_i = \text{argmax}_{\ell \in \{1, \ldots, K+d_1\}} \gamma_{i,\ell}$$

and is of type a_{ℓ_i}. However, our model-based approach brings much more information and the vector of posterior probabilities $(\gamma_{i,1}, \ldots, \gamma_{i,K+d_1})$ output by the algorithm actually defines soft membership and represent the uncertainty in whether anomaly V_i is in a certain cluster. It additionally induces a similarity measure between the anomalies: the higher the probability that two extreme values arise from the same component of the mixture model, the more similar they are considered. Hence, consider the undirected graph whose vertices, indexed by $i = 1, \ldots, n_0$, correspond to the extremal observations $V_1, \ldots, V_{n_0}$ and whose edgeweights are $w_{\boldsymbol{\eta}}(V_i, V_j)$, $1 \leq i \neq j \leq n_0$, where

$$w_{\boldsymbol{\eta}}(V_i, V_j) = \mathbb{P}\left(\boldsymbol{Z}_i = \boldsymbol{Z}_j \mid V_i\, V_j,\ \boldsymbol{\eta}\right) = \sum_{\ell=1}^{K+d_1} \gamma_{i,\ell}\gamma_{j,\ell}.$$

Based on this original graph description of the set of extremes, it is now possible to rank all anomalies (*i.e.* extreme points) by degree of similarity to a given anomaly V_i

$$w_{\boldsymbol{\eta}}(V_i, V_{(i,1)}) \geq w_{\boldsymbol{\eta}}(V_i, V_{(i,2)}) \geq \ldots \geq w_{\boldsymbol{\eta}}(V_i, V_{(i,n_0)})$$

and extract neighborhoods $\{V_{(i,1)}, \ldots, V_{(i,l)}\}$, $l \leq n_0$.

Remark 5.7 (Graph-theoretic clustering). We point out that many alternative methods to that consisting in assigning to each any anomaly/extreme its likeliest component (*i.e.* model-based clustering) can be implemented in order to partition the similarity graph thus defined into subgraphs whose vertices correspond to similar anomalies, ranging from tree-based clustering procedures to techniques based on local connectivity properties through spectral clustering. One may refer to *e.g.* Schaeffer (2007) for an account of graph-theoretic clustering methods.

In possible combination with clustering, graph visualization techniques (see *e.g.* Hu and Shi (2015)), when the number n_0 of anomalies to be analyzed is large, can also be used to produce informative layouts. Discussing the merits and limitations of the wide variety of approaches documented in the literature in this purpose is beyond the scope of this paper. The usefulness of the weighted graph representation proposed above combined with state-of-the-art graph-mining tools is simply illustrated in the experimental sections of Chiapino et al. (2019a). We point out however that alternatives to the (force-based) graph drawing method used therein can be naturally considered, re-using for instance the eigenvectors of the graph Laplacian computed through a preliminary spectral clustering procedure (see *e.g.* Athreya et al. (2017) and the references therein for more details on spectral layout methods).

Chapter 6

Miscellanea: standardization of semi-continuous max-stable processes

This chapter presents a published paper (Sabourin and Segers (2017)) which is somewhat disconnected in spirit from the rest of this thesis. This publication is devoted to the study of max-stable upper semi-continuous processes, examples of which have been exploited in the literature of spatial extremes.

A common way to describe multivariate max-stable distributions is as follows: their margins are univariate max-stable distributions; after standardization of the marginal distributions to a common one, the joint distribution has a specific representation, describing the dependence structure. The separation into margins and dependence is in line with Sklar's theorem Sklar (1959), which provides a decomposition of a multivariate distribution into its margins and a copula, that is, a multivariate distribution with standard uniform margins. If the margins of the original distribution are continuous, the copula is unique and can be found by applying the probability integral transform to each variable. Conversely, to recover the original distribution, it suffices to apply the quantile transformation to each copula variable. Although it is more common in extreme-value theory to standardize to the Gumbel or the unit-Fréchet distribution rather than to the uniform distribution, the principle is the same. The advantage of breaking up a distribution into its margins and a copula is that both components can be modelled separately.

Applications in spatial statistics have spurred the development of extreme-value theory for stochastic processes. If the trajectories of the process are continuous almost surely, then the process can be reduced to a process with standardized margins and continuous trajectories via the probability integral transform applied to each individual variable. Conversely, the original process can be recovered from the standardized one by applying the quantile transform to each standardized variable. Moreover, these maps, sending one continuous function to another one, are measurable with respect to the sigma-field on the space of continuous functions that is generated by the finite-dimensional cylinders. These results hinge on the following two properties. First, the marginal distributions of a stochastic process with continuous trajectories depend continuously on the index variable. Second, the distribution of the random continuous function associated to the process is determined by the finite-dimensional distributions.

For stochastic processes with upper semicontinuous (usc) trajectories, however, the two properties mentioned above do not hold: the marginal distributions need not depend in a continuous way on the index variable, and the distribution of path functionals such as the supremum of the process is not determined by the finite-dimensional distributions of the process. Still, max-stable processes with usc trajectories have been proposed as models for spatial extremes of environmental variables Davison and Gholamrezaee (2012); Huser and Davison (2014); Schlather (2002). As in the continuous case, construction of and inference on such models is carried out by a separation of concerns regarding the margins and the dependence structure. However, up to date, there is no theoretical foundation for such an approach. Another possible application of max-stable usc process is random utility maximization when the alternatives range over a compact metric space rather than a finite set McFadden (1981, 1989); Resnick and Roy (1991).

The present paper aims to fill the gap in theory and develop a framework for marginal standardization for stochastic processes with usc trajectories. For the mathematical framework, we follow Norberg (1987) and Resnick and Roy (1991) and we work within the space $\mathrm{USC}(\mathbb{D})$ of usc functions on a locally compact subset $\mathbb{D}$ of some Euclidean space. The space $\mathrm{USC}(\mathbb{D})$ is equipped with the hypo-topology: a usc function is identified with its hypograph, a closed subset of $\mathbb{D} \times \mathbb{R}$; the hypo-topology on $\mathrm{USC}(\mathbb{D})$ is then defined as the trace topology inherited from the Fell hit-and-miss topology on the space $\mathcal{F}$ of closed subsets of $\mathbb{D} \times \mathbb{R}$ Salinetti and Wets (1986); Vervaat (1986).

The theory is specialized to max-stable usc processes. Our definition of max-stability allows the shape parameter of the marginal distributions to vary with the index variable of the stochastic process. As a consequence, the stabilizing affine transformations in the definition of max-stability may depend on the index variable too. However, the coordinatewise affine transformation of a usc function does not necessarily produce a usc function, so that care is needed in the formulation of the definition and the results. In de Haan (1984); Resnick and Roy (1991), in contrast, the marginal distributions are assumed to be Fréchet with unit shape parameter, so that the stabilizing sequences in the definition of max-stability are the same for all margins. It is often taken for granted that the general case may be reduced to this simpler case, but, as argued in the paper, for usc processes, this is not guaranteed.

A general class of measurable transformations on the space of usc functions is first introduced. Under regularity conditions on the marginal distributions, this class includes the pointwise probability integral transform and its inverse. This property allows us to state a partial generalization of Sklar's theorem for usc processes in general and for max-stable ones in particular

6.1 usc processes

We first review essential definitions and introduce some notation. The material we need on the theory of usc process may for instance be found in Salinetti and Wets (1986), Beer (1993) (Chapter 5), Vervaat (1988a), Vervaat and Holwerda (1997) and Molchanov (2005)(Chapter 1.1 and Appendix B). Let $\mathbb{D}$ be a non-empty, locally compact subset of some finite-dimensional Euclidean space. A function $x : \mathbb{D} \to [-\infty, \infty]$ is *upper semicontinuous* (usc) if and only $\limsup_{n\to\infty} x(s_n) \leq x(s)$ whenever $s_n \to s \in \mathbb{D}$. The *hypograph* of x is a commonly defined by

$$\operatorname{hypo} x = \{(s, \alpha) \in \mathbb{D} \times \mathbb{R} : \alpha \leq x(s)\}.$$

Another characterization of upper semicontinuity for a function x as above is that the hypograph of x is closed. Let $\mathrm{USC}(\mathbb{D})$ be the collection of all upper semicontinuous functions from $\mathbb{D}$ into $[-\infty, \infty]$. By identifying the function $x \in \mathrm{USC}(\mathbb{D})$ with the set $\operatorname{hypo} x \subset \mathbb{D} \times \mathbb{R}$, any topology on the space $\mathcal{F} = \mathcal{F}(\mathbb{D} \times \mathbb{R})$ of closed subsets of $\mathbb{D} \times \mathbb{R}$ results in a trace topology on the space of usc functions. For usc processes, the theory is built on the use of the The Fell *hit-and-miss* topology on $\mathcal{F}$ A base for the latter topology is the family of sets of the form

$$\mathcal{F}^{K}_{G_1,\ldots,G_n} = \{F \in \mathcal{F} \ : \ F \cap K = \varnothing,\, F \cap G_1 \neq \varnothing,\, \ldots,\, F \cap G_n \neq \varnothing\}$$

for $K \in \mathcal{K}$ and $G_1, \ldots, G_n \in \mathcal{G}$. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a complete probability space. In this work, a usc process is always a function $\xi : \mathbb{D} \times \Omega \to [-\infty, \infty]$ such that the map $\Omega \to \mathcal{F} : \omega \mapsto \operatorname{hypo} \xi(\,\cdot\,, \omega)$ is Borel measurable.

6.2 Sklar's theorem for usc processes

A d-variate copula is the cumulative distribution function of a d-dimensional random vector with standard uniform margins. Sklar's Sklar (1959) celebrated theorem states two things:

1. For every copula C and every vector $F_1, \ldots, F_d$ of univariate distribution functions, the function $(x_1, \ldots, x_d) \mapsto C(F_1(x_1), \ldots, F_d(x_d))$ is a d-variate distribution function with margins $F_1, \ldots, F_d$.

2. Every d-variate distribution function F can be represented in this way.

Reformulated in terms of random vectors, the two statements read as follows:

1. For every random vector $(U_1, \ldots, U_d)$ with uniform components and for every vector $F_1, \ldots, F_d$ of univariate distribution functions, the random vector $(Q_1(U_1), \ldots, Q_d(U_d))$ has marginal distributions $F_1, \ldots, F_d$, where Q_j is the (right- or left-continuous) quantile function corresponding to F_j.
2. Every random vector $(X_1, \ldots, X_d)$ can be represented in this way.

The next results stated and proved in Sabourin and Segers (2017) specify up to what extent these facts hold for usc processes too.

Proposition 6.1 (à la Sklar I for usc processes)**.** *Let Z be a usc process having standard uniform margins. Let $(F_s : s \in \mathbb{D})$ be a family of (right-continuous) distribution functions and let $Q_s(p) = \sup\{x \in \mathbb{R} : F_s(x) \le p\}$ for all $(s, p) \in \mathbb{D} \times [0, 1]$. Define a stochastic process ξ by $\xi(s) = Q_s((Z(s) \vee 0) \wedge 1)$ for $s \in \mathbb{D}$. Then the following two statements are equivalent:*

(i) ξ is a usc process with marginal distributions given by F_s.

(ii) For every $p \in [0, 1]$, the function $s \mapsto Q_s(p)$ is usc.

Proposition 6.2 (à la Sklar II for usc processes)**.** *Let ξ be a usc process. Let $F_s(x) = \Pr[\xi(s) \le x]$ for $x \in [-\infty, \infty]$ and let $Q_s(p) = \sup\{x \in \mathbb{R} : F_s(x) \le p\}$ for $p \in [0, 1]$. Suppose the following two conditions hold:*

(a) For every $s \in \mathbb{D}$, the distribution of $\xi(s)$ has no atoms in $[-\infty, \infty]$.

(b) For every $x \in \mathbb{R} \cup \{+\infty\}$, the function $s \mapsto F_s(x)$ is usc.

Then the following statements hold:

(i) The process Z defined by $Z(s) = F_s(\xi(s))$ is a usc process with standard uniform margins.

(ii) The process $\tilde{\xi}$ defined by $\tilde{\xi}(s) = Q_s(Z(s)) = Q_s(F_s(\xi(s))$ is a usc process such that $\Pr[\tilde{\xi}(s) = \xi(s)] = 1$ for every $s \in \mathbb{D}$. In particular, the finite-dimensional distributions of $\tilde{\xi}$ and ξ are identical.

6.3 Max-stable usc processes and their standardization

In dimension 1, the only possible non degenerate limits in distribution of affine normalized maxima $\bigvee_{i=1}^{n}(X_i - b_n)/a_n$, with $b_n \in \mathbb{R}$, $a_n > 0$ and $(X_i)_{i \ge 1}$ an $i.i.d.$ sequence, are the generalized extreme-value (GEV) distributions with parameter vector $\boldsymbol{\eta} = (\gamma, \mu, \sigma) \in E = \mathbb{R} \times \mathbb{R} \times (0, \infty)$ given by

$$F(x; \boldsymbol{\eta}) = \begin{cases} \exp[-\{1 + \gamma(x-\mu)/\sigma\}^{-1/\gamma}] & \text{if } \gamma \neq 0 \text{ and } \sigma + \gamma(x-\mu) > 0, \\ \exp[-\exp\{-(x-\mu)/\sigma\}] & \text{if } \gamma = 0 \text{ and } x \in \mathbb{R}. \end{cases} \tag{6.1}$$

(see *e.g.* Beirlant et al. (1996), Chap. 1 for univariate EVT). In fact a distribution is a GEV if and only if it is max-stable, *i.e.* for every n, there exist unique scalars $a_{n,\boldsymbol{\eta}} \in (0, \infty)$ and $b_{n,\boldsymbol{\eta}} \in \mathbb{R}$ such that the following max-stability relation holds:

$$F^n(a_{n,\boldsymbol{\eta}} x + b_{n,\boldsymbol{\eta}}; \boldsymbol{\eta}) = F(x; \boldsymbol{\eta}), \qquad x \in \mathbb{R}. \tag{6.2}$$

This property motivates the use of such distributions for modeling maxima over many variables. The location and scale sequences are given by

$$a_{n,\boldsymbol{\eta}} = n^{\gamma}, \qquad\qquad b_{n,\boldsymbol{\eta}} = \begin{cases} (\sigma - \gamma\mu)(n^{\gamma} - 1)/\gamma & \text{if } \gamma \neq 0, \\ \sigma \log n & \text{if } \gamma = 0. \end{cases} \tag{6.3}$$

Max-stable usc processes have been defined in the literature Molchanov (2005); Vervaat (1986, 1988b) as usc processes whose distribution is invariant under the componentwise maximum operation, up to an affine rescaling involving *constants* which are not allowed to depend upon the index variable $s \in \mathbb{D}$. Likewise, the processes in de Haan (1984) and Resnick and Roy (1991) and have marginal distributions which are Fréchet with shape parameter $\gamma \equiv 1$ and lower endpoint $\mu - \gamma/\sigma \equiv 0$, so that the scaling sequences are $a_n \equiv n$ and $b_n \equiv 0$. Moreover, max-stability is defined in de Haan (1984) in terms of finite-dimensional distributions only. Definition 6.3 below extends these previous approaches by allowing for an index-dependent rescaling, through scaling *functions* a_n and b_n, and by viewing the random objects as random elements in $\mathrm{USC}(\mathbb{D})$.

Definition 6.3. *A usc process ξ is* max-stable *if, for all integer $n \geq 1$, there exist functions $a_n : \mathbb{D} \to (0, \infty)$ and $b_n : \mathbb{D} \to \mathbb{R}$ such that, for each vector of n independent and identically distributed (iid) usc processes $\xi_1, \ldots, \xi_n$ with the same law as ξ, we have*

$$\bigvee_{i=1}^{n} \xi_i \overset{d}{=} a_n \xi + b_n \quad \textit{in } \mathrm{USC}(\mathbb{D}). \tag{6.4}$$

A max-stable usc process ξ^* is said to be simple *if, in addition, its marginal distributions are unit-Fréchet, $\mathbb{P}\left(\xi^*(s) \leq x = e^{-1/x}\right), x > 0$. In that case, the norming functions are given by $a_n(s) = n$ and $b_n(s) = 0$ for all $n \geq 1$ and $s \in \mathbb{D}$, i.e., for iid usc processes $\xi_1^*, \ldots, \xi_n^*$ with the same law as ξ^*, we have*

$$\bigvee_{i=1}^{n} \xi_i^* \overset{d}{=} n\xi^* \quad \textit{in } \mathrm{USC}(\mathbb{D}). \tag{6.5}$$

In (6.4) and (6.5), the meaning is that the induced probability distributions on the space $\mathrm{USC}(\mathbb{D})$ equipped with the sigma-field of hypo-measurable sets are equal. In Definition 6.3, it is implicitly understood that the functions a_n and b_n are such that the right-hand side of (6.4) still defines a usc process. If a_n is continuous and b_n is usc, then this is automatically the case, see Lemma 3.1 and Example 3.1 (iii) in the paper.

The evaluation map $\mathrm{USC}(\mathbb{D}) \to [-\infty, \infty] : z \mapsto z(s)$ is hypo-measurable for all $s \in \mathbb{D}$. Equation (6.4) then implies the following distributional equality between random variables:

$$\bigvee_{i=1}^{n} \xi_i(s) \stackrel{d}{=} a_n(s)\,\xi(s) + b_n(s), \qquad s \in \mathbb{D}.$$

As a consequence, the marginal distribution of $\xi(s)$ is max-stable and therefore it is a GEV (Generalized Extreme Value) distribution with some parameter vector $\boldsymbol{\eta}(s) = (\gamma(s), \mu(s), \sigma(s))$. The normalizing functions a_n and b_n of a max-stable usc process must then be of the form $a_n(s) = a_{n,\boldsymbol{\eta}(s)}$ and $b_n(s) = b_{n,\boldsymbol{\eta}(s)}$ as in (6.3).

We now investigate the relation between general and simple max-stable usc processes via the pointwise probability integral transform and its inverse. Max-stability of usc processes is defined in (6.4) via an equality of distributions on $\mathrm{USC}(\mathbb{D})$ rather than of finite-dimensional distributions. It is therefore not clear from the outset that max-stability is preserved by pointwise transformations.

Proposition 6.4 gives a necessary and sufficient condition on the GEV margins to be able to construct a general max-stable usc process starting from a simple one. Proposition 6.5 treats the converse question, that is, when can a max-stable usc process be first reduced to a simple one and then be reconstructed from it.

Proposition 6.4 (à la Sklar I for max-stable usc processes)**.** *Let ξ^* be a simple max-stable usc process. Let $\boldsymbol{\eta} : \mathbb{D} \to E$. Define a stochastic process ξ by $\xi(s) = Q(\Phi(\xi^*(s)); \boldsymbol{\eta}(s))$ for $s \in \mathbb{D}$. Then the following two statements are equivalent:*

(i) ξ is a usc process with marginal distributions $\mathrm{GEV}(\boldsymbol{\eta}(s))$.

(ii) For every $p \in [0, 1]$, the function $s \mapsto Q(p; \boldsymbol{\eta}(s))$ is usc.

If these conditions hold, then ξ is a max-stable usc process with normalizing functions $a_n(s) = a_{n,\boldsymbol{\eta}(s)}$ and $b_n(s) = b_{n,\boldsymbol{\eta}(s)}$.

Proposition 6.5 (à la Sklar II for max-stable processes)**.** *Let ξ be a usc process with $\mathrm{GEV}(\boldsymbol{\eta}(s))$ margins for $s \in \mathbb{D}$. Assume that for every compact $K \subset \mathbb{D}$, we have $\sup_{s \in K} F(\xi(s); \boldsymbol{\eta}(s)) < 1$ with probability one. Define two usc processes ξ^* and $\tilde{\xi}$ by $\xi^*(s) = -1/\log F(\xi(s); \boldsymbol{\eta}(s))$ and $\tilde{\xi}(s) = Q(\Phi(\xi^*(s)); \boldsymbol{\eta}(s))$, for $s \in \mathbb{D}$. Then, almost surely, $\xi = \tilde{\xi}$. Furthermore, the following two statements are equivalent:*

(i) The usc process ξ is max-stable.
(ii) The usc process ξ^* is simple max-stable.

As a conclusion, the aim of the paper has been to extend Sklar's theorem from random vectors to usc processes. We have stated necessary and sufficient conditions to be able to construct a usc process with general margins by applying the pointwise quantile transformation to a usc process with standard uniform margins (Propositions 6.1 and 6.4). Furthermore, we have stated sufficient conditions for the pointwise probability integral transform to be possible for usc processes (Propositions 6.2, 6.5). These conditions imply in particular that the marginal distribution functions are continuous with respect to the space variable (Lemma 4.1 in the paper). We have also provided several examples of things that can go wrong when these conditions are not satisfied, which are not presented in the present thesis for the sake of concision. However, finding *necessary* and sufficient conditions remains an open problem.

The motivation has been to extend the margins-versus-dependence paradigm used in multivariate extreme-value theory to max-stable usc processes. The next step is to show that marginal standardization is possible in max-domains of attraction too. One question, for instance, is whether the standardized weak limit of the pointwise maxima of a sequence of usc processes is equal to the weak limit of the pointwise maxima of the sequence of standardized usc processes (Resnick (1987), Proposition 5.10). Interesting difficulties arise: weak convergence of finite-dimensional distributions does not imply and is not implied by weak hypoconvergence; Khinchin's convergence-of-types lemma does not apply in its full generality to unions of random closed sets (Molchanov (2005), p. 254, 'Affine normalization').

Chapter 7

Ongoing work and perspectives

The works exposed in the present thesis leave many questions unanswered and open the road to several research perspectives. This closing chapter lays out the main directions in which I intend to pursue research in the near future.

First of all, in a multivariate setting, although a standardization step allowing to work with nearly identical margins is often required, the theoretical guarantees obtained so far do not always take into account the uncertainty induced by the use of estimated marginal distributions. More precisely, when the pseudo-polar decomposition of the exponent measure comes into play in the analysis, *e.g.* for classification in Chapter 3 or estimation of MV-sets in Section 5.2, the theoretical tools presented in this thesis do not allow to take into consideration the standardization step. The missing piece to complete the analysis is a concentration study of the empirical angular measure. This is the subject of an ongoing work which main lines are presented in Section 7.1.

Secondly, a central aspect of this thesis is the need for dimensionality reduction devices for analyzing multivariate extremes. So far we have only considered the unsupervised setting. Section 7.2 sketches the main ideas of an ongoing work inspired by inverse regression techniques for dimension reduction, when the focus is on the tails of the explained variable.

Section 7.3 describes two research projects at a prospective stage. The aim of the first one (Section 7.3.1) is to investigate in what extent the use of re-sampling techniques such as the bootstrap or cross-validation can help in choosing hyperparameters in various multivariate extreme values context. In the univariate case, re-sampling has been proved useful to select the number of extreme order statistics. The question of how to use the bootstrap or cross-validation to choose, say, the tolerance parameter ϵ in dimensionality reduction algorithms such as DAMEX (Section 5.1), or the number of order statistics to perform PCA (Section 4.3) remains open. Finally in Section 7.3.2 we are dipping the toe in the time series and functional data water, with applications to anomaly detection in view.

7.1 Concentration of the empirical angular measure

Joint work with Stephan Clémençon, Hamid Jalalzai, and Johan Segers, working paper Clémençon et al. (2021) available `arXiv:2104.03966`

Throughout this thesis the marginal standardization of the random quantity under consideration (a random vector or process) is paramount to describe the dependence structure in the upper tail. The impact of empirical estimation of margins (which is how one goes from probability integral transform T in (2.4) to the rank transform (5.1)) is fully taken into account when analyzing the supremum deviations of the STDF (Section 2.2) and of summaries of the exponent measure, namely $\mathcal{M}_a, a \in \{1, \dots, d\}$ in Section 4.1 and the joint tail coefficients χ_a in Section 4.2.

However when we consider classification in extreme regions (Chapter 4), Principal Component Analysis (Section 4.3), anomaly detection using MV-sets (Section 5.2 and clustering of anomalies (Section 5.3) we either rely on the assumption that the margin are known, or that the tail index is the same for all components, which in many practical situations means that some preliminary standardization has been performed. One main reason why we make such a strong assumption is the following: the analysis in the latter papers relies heavily on empirical estimation of quantities directly related to the angular measure Φ defined in (2.7). In other words the various estimation steps involve the empirical measure related to a sample of components $\theta(V_i)$, the angles of the considered vectors V_i which are assumed to satisfy the regular variation condition (2.6). Thus deriving concentration guarantees similar to the established ones while taking into account the impact of the standardization (*i.e.* using $\widehat{V}_i$'s instead of V_i's) would amount to establishing concentration inequalities for the empirical angular measure,

$$\widehat{\Phi}(A) = \mu_n(\mathcal{C}_A) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}\{\|\widehat{V}_i > n/k\} \delta_{\theta(\widehat{V}_i)}(A), \qquad A \in \mathcal{B}(\mathbb{S}_+).$$

To this date, the only existing results concerning the empirical angular are stated and proved in an asymptotic framework for the bivariate case only, with techniques of proof which do not allow for a straightforward generalization to the general multivariate case. The main difficulty in analyzing the empirical measure comes from the fact that the errors $\widehat{F}_j - F_j$ propagate in a non linear fashion onto the angular error of the rank transformed samples $\theta(\widehat{V}_i) - \theta(V_i)$. The proof of asymptotic normality in the bivariate case relies heavily on rewriting the empirical angular measure in terms of a random set $\widehat{A}$ accounting from marginal randomness, $\delta_{\theta(\widehat{V}_i)}(A) = \delta_{\theta(V_i)}(\widehat{A})$. The next step is to construct two deterministic framing sets A_-, A_+ such that $A_- \subset \widehat{A} \subset A_+$ with high probability. Due to non-linearities the expression for these framing sets is somewhat complicated, whence the difficulty to extend the proof to the multivariate case. It is the purpose of ongoing work to establish concentration inequalities for the empirical angular measure in the multivariate case. Our point of departure is the general concentration inequality for rare classes in Goix et al. (2015) together with the construction of appropriate framing sets in dimension greater than two. From a technical viewpoint, a major advantage of the non-asymptotic approach is that the framing sets are not required to be 'tight', in the sense that the approximation error arising from such a framing in the error decomposition can be of the same order of magnitude as the other variance terms (*i.e.* , with a leading term of order $\mathrm{O}(1/\sqrt{k})$), instead of being negligible compared to the other terms as it is required in the asymptotic analysis to obtain weak convergence of the empirical process. From a broader perspective, such concentration results would unblock several bottlenecks in the statistical learning approach of multivariate extremes, starting with those mentioned above, namely the assumption of known margins in a classification setting (Jalalzai et al. (2018)) and for anomaly detection based on angular MV-set estimation (Thomas et al. (2017)).

7.2 Sliced Inverse Regression with extreme target

Joint work with Anass Aghbalou, François Portier, and Chen Zhou, working paper Aghbalou et al. (2021) available `arXiv:2108.01432`

Dimensionality reduction plays an important role in this thesis. It is the main subject developed in Chapter 4. So far we have only investigated unsupervised dimensionality reduction techniques, leaving aside supervised problems for such matter.

In statistical regression, the aim is to predict the response random variable Y valued in, say, $\mathbb{R}$ for simplicity, using a set of predictors or features, represented by a random vector X. When the dimensionality of X is high, estimating quantities defined through the conditional distribution of Y given X is subject to the curse of dimensionality. A way around is to rely on the sparsity assumption that only certain linear combinations of the feature variables are useful to predict Y, *i.e.* that there is an orthogonal projection P onto a subspace E (the *central space*) of dimension $p < d$ such that given PX, Y is independent from $P^{\perp}X$, so that only the dependence between Y and PX should be examined. This conditional independence assumption (Constantinou and Dawid (2017)) is at the heart of several dimensionality reduction techniques (see *e.g.* Fukumizu et al. (2004) and the references therein), including Sliced Inverse Regression (SIR) and Sliced Average Variance Estimation (SAVE) (Li (1991); Cook and Ni (2005); Zhu et al. (2010);

Cook and Weisberg (1991); Cook et al. (2002)). The principle behind SIR is that under a *linearity condition*,which is satisfied when the distribution of X is elliptical, $\mathbb{E}(X|Y)$ belongs to the central space. The SAVE method is based on a variant of this principle and uses second moments.

To my best knowledge, tail dimension reduction for regression has only been considered in Gardes (2018) in the specific context of extreme quantiles estimation. In that paper, a notion of tail conditional independence is introduced, which may be rephrased as the fact that the survival function of Y above large thresholds, conditionally on PX is asymptotically equivalent to the same survival function conditionally to X. Here P is the projector on a tail dimension reduction subspace. The main result of the paper is that the estimated quantile conditional to the reduced variable PX is asymptotically consistent, with a quantifiable rate of convergence. However asymptotic results are only proved under the assumption that the projector P is known. An estimation procedure is proposed together with a heuristic justification.

It is the purpose of ongoing work to investigate the consequence of a tail conditional independence assumption, to derive workable examples where this condition is met, and to obtain asymptptotic guarantees regarding the estimation of the tail dimension reduction subspace using a SIR/SAVE-like method. For such purposes one major technical tool is the framework of classes of functions changing with n (van der Vaart and Wellner (1996)).

7.3 Perspectives

7.3.1 Re-sampling multivariate extremes

(Part of the ongoing PhD Thesis of Anass Aghbalou, co-supervision with François Portier and Patrice Bertail)
Bootstrap, subsampling and cross-validation methods are widely used tools in Machine Learning and statistics. The objective may be depending on the context, to quantify the uncertainty attached to a statistical procedure (*e.g.* the width of a confidence interval) or to select optimal hyper-parameters. The theoretical properties of the methods based on the bootstrap and cross-validation are well known when the whole dataset is considered. However when the focus is on the distribution tails, most analyses conducted so far concern univariate problems (Danielsson et al. (2001); Kyselỳ (2008); Bertail et al. (2004); Gomes et al. (2016)). The validity of resampling methods for multivariate problems is a largely unexplored question, to the exception of the works on the bootstrap for the STDF Peng and Qi (2008) and on the multiplier bootstrap for extreme value copulas (the counterpart of the STDF using uniform margins for the standardization step) by Bücher et al.

(2013). Concerning the use of cross-validation for extreme value analyses, to my best knowledge there is to this date no theoretical analysis of the optimality properties of the selected hyper-parameters, *e.g.* the number of components in PCA, the number of clusters in k-means, the tolerance parameters for support identification in Goix et al. (2017) or in Chiapino and Sabourin (2016), the number of order statistics to be used for tail estimation,. . . . The starting point of the PhD project is to use existing asymptotic results from Peng and Qi (2008); Bücher et al. (2013) to propose alternative tests of extremal dependence to the ones proposed in Chiapino et al. (2019b), based on bootstrap statistics. The next step would be to investigate the properties of the bootstrap for the angular measure of extremes. Another direction concerns the use of cross-validation for model and hyper parameter selection by extending the approaches of Györfi et al. (2006) (chap. 8) or Arlot et al. (2010) to an extreme value context.

7.3.2 Extremes of time series and functional data, with application to anomaly detection

(Joint work with Stephan Clémençon and master student Nathan Huet)
In many industrial contexts the temporal dependence structure cannot be ignored when identifying the 'normal' (=not abnormal) behavior. The data then takes the form of time series (Malhotra et al. (2015); Wei et al. (2005)) or functional data (Dai and Genton (2018); Staerman et al. (2019)). Our focus is on the shape of the normal region associated to a very low false alarm rate $\alpha \ll 1$, i.e. quantile regions in a functional space at level $1-\alpha$ with $\alpha \to 0$. For such purpose the tail behavior (in an adequate representation) of the data is paramount. EVT and the theory of regular variation (De Haan and Resnick (1987); Hult and Lindskog (2006)) provide a convenient framework for characterizing such probabilistic behaviors, together with statistical practice allowing estimation. EVT has been successfully applied to the task of AD in the univariate (Siffer et al. (2017)) and multivariate cases (Goix et al. (2016, 2017),Thomas et al. (2017), Cai et al. (2011)). To our best knowledge, using EVT for AD with time series and functional data has not been addressed in the literature. Extremes of functional data has been the subject of a recent PhD thesis (Xiong (2018)), two contributions of which relate to our topic: Kokoszka and Xiong (2018) and Kokoszka et al. (2018), consider the Karhunen Loeve (KL) expansion of $i.i.d.$ strictly stationary zero mean time series $X_i \in L^2([0,T])$, $i \leq N$. They provide sufficient conditions under which it is legitimate to use the estimated covariance matrix to handle extreme values. In the context of time series analysis, a series of papers following Basrak and Segers (2009) characterize multivariate regular variation of strictly stationary time series. The serial dependence at extreme level is characterized in terms of a tail process Y and

a spectral tail process Θ, which are respectively the limit distributions of a rescaled version of the process X, conditionally to $\|X_0\| > u$, where the scaling function is respectively u and the value of $\|X_0\|$. One major result is that $Y = \|Y_0\|\Theta$ where $\|Y_0\|$ and Θ are independent. Thus (Y_0, Θ) may be seen as a pseudo-polar decomposition of the tail process Y.

For practical purposes such as AD, one needs a finite dimensional approximation of the infinite dimensional tail objects Y or Θ. However the viewpoints developed in the PhD thesis Xiong (2018) and the papers therein on the one hand, and in the series of papers following Basrak and Segers (2009) on the other hand, although complementary, have not been connected to each other. In this research project we shall attempt to bridge this gap between probabilistic analysis of extremes and machine learning applications. One promising point of departure is the finite dimensional counterpart of this matter, that is PCA for multivariate extremes which has been recently considered in the upcoming paper Drees and Sabourin (2021). The main idea consists in working with the angular component of the limit distribution of extremes, so that existence of moments is not an issue anymore. We may thus consider KL expansions of the spectral process Θ or alternatively the spectral process related to the KL expansion of the original series.

Appendix A

Alternative concentration inequality for rare classes

Theorem A.1 (Concentration in rare regions: explicit bound. Ongoing joint work with Johan Segers)**.** *Let* $X_1, \ldots, X_n$ *be an independent random sample from* P*. Let* $\mathcal{G}$*,* $\mathbb{A}$ *and* p *be as in Theorem 2.4. For* $\delta \in (0,1)$*, if* $np \geq 8\ln(8/\delta)$*, then with probability at least* $1-\delta$*,*

$$\sup_{A\in\mathcal{G}} |P_n(A) - P(A)| \leq 4\sqrt{\frac{2p}{n}\big(\ln(8/\delta) + \ln(\mathcal{S}_{\mathcal{G}}(8np))\big)}.$$

Proof. Let $X_1, \ldots, X_n, X'_1, \ldots, X'_n$ be an independent random sample of size $2n$ from P. Let P_n and P'_n be the empirical distributions of $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$, respectively. By symmetrization (Lemma A.2), we have, for $t > 0$ such that $nt^2 \geq 2\min(1, 4p)$,

$$\mathbb{P}\left[\sup_{A\in\mathcal{G}} |P_n(A) - P(A)| \geq t\right] \leq 4\,\mathbb{P}\left[\sup_{A\in\mathcal{G}} \big(P_n(A) - P'_n(A)\big) \geq t/2\right].$$

We seek t such that the probability on the right-hand side is bounded by $\delta/4$. By Lemma A.4, this is the case as soon as

$$\frac{t}{2} \geq 2\sqrt{\frac{2p}{n}\big(\ln(2/(\delta/4)) + \ln\mathcal{S}_{\mathcal{G}}(8np)\big)}$$

provided $np \geq 8\ln(2/(\delta/4))$. Simplifying, we find that t must satisfy

$$t \geq 4\sqrt{\frac{2p}{n}\big(\ln(8/\delta) + \ln(\mathcal{S}_{\mathcal{G}}(8np))\big)} =: t(\delta).$$

The threshold $t(\delta)$ is the bound stated in the theorem. We only need to check that $n(t(\delta))^2 \geq 2\min(1, 4p)$. But this easily holds, since

$$n(t(\delta))^2 \geq n \cdot 16 \cdot \frac{2p}{n}\ln(8/\delta) \geq 32\ln(8)p. \qquad \square$$

The following symmetrization lemma is almost identical to Lemma 2 in Bousquet et al. (2003). The difference is that the lower bound on admissible values for t features $p = \mathbb{P}(\mathbb{A})$. This is needed in view of the application of the inequality to rare event probabilities in the proof of Theorem A.1.

Lemma A.2 (Symmetrization)**.** *Let $X_1, \ldots, X_n, X'_1, \ldots, X'_n$ be an independent random sample of size $2n$ from P. Let P_n and P'_n be the empirical distributions of $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$, respectively. Let $\mathcal{G}$, $\mathbb{A}$ and p be as in Theorem. If $t \geq 0$ is such that $nt^2 \geq 2\min(1, 4p)$, then*

$$\left.\begin{array}{l} \mathbb{P}\left[\sup_{A\in\mathcal{G}}\big(P_n(A) - P(A)\big) \geq t\right] \\ \mathbb{P}\left[\sup_{A\in\mathcal{G}}\big(P(A) - P_n(A)\big) \geq t\right] \end{array}\right\} \leq 2\,\mathbb{P}\left[\sup_{A\in\mathcal{G}}\left(P_n(A) - P'_n(A)\right) \geq t/2\right].$$

Proof. Only the range of t needs to be widened compared to Lemma 2 in Bousquet et al. (2003). Following the proof of this result, we see that t needs to be such that

$$\forall A \in \mathcal{G}, \qquad \frac{4\,\mathbb{V}\mathrm{ar}(\mathbb{1}_A(X))}{nt^2} \leq \frac{1}{2}.$$

Now

$$4\,\mathbb{V}\mathrm{ar}(\mathbb{1}_A(X)) = 4P(A)(1 - P(A)) \leq \min\{1, 4P(A)\}.$$

Since $P(A) \leq P(\mathbb{A}) = p$, a sufficient condition is thus that $nt^2 \geq 2\min(1, 4p)$. □

Lemma A.3 (concentration after symmetrization and conditioning)**.** *Let $X_1, \ldots, X_n$, $X'_1, \ldots, X'_n$ be an independent random sample of size $2n$ from P. Let P_n and P'_n be the empirical distributions of $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$, respectively. Let $\mathcal{G}$, $\mathbb{A}$ and p be as in Theorem 2.4. Let K be the number of times the pair X_i, X'_i hits the union class $\mathbb{A}$,* i.e. *the random number of indices $i \leq n$ such that $X_i \in A$ or $X'_i \in A$. For $t > 0$ and $\kappa \in \{0, \ldots, n\}$ we have*

$$\mathbb{P}\left(\sup_{1\in\mathcal{G}} P'_n(A) - P_n(A) \geq t \mid K = \kappa\right) \leq \mathcal{S}_{\mathcal{G}}(2\kappa)e^{\frac{-n^2t^2}{2\kappa}}$$

Proof. Notice that for $\kappa = 0$ both sides of the inequality are zero. For $\kappa \geq 1$ we follow the classical argument yielding concentration of the symmetrized sample as *e.g.* in Bousquet et al. (2003), up to a conditioning step upon the number of pairs (X_i, X'_i) hitting the class. Let $\sigma_1, \ldots, \sigma_n$ be an independent random sample of Rademacher random variables, independent also from the $2n$-sample

$X_1, \dots, X_n, X'_1, \dots, X'_n$. By symmetry,

$$\begin{aligned}\sup_{A \in \mathcal{G}} \bigl(P_n(A) - P'_n(A)\bigr) &= \sup_{A \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \bigl(\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}\bigr) \\ &\overset{d}{=} \sup_{A \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bigl(\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}\bigr).\end{aligned}$$

We now condition on the counting variable

$$\begin{aligned}K &= \sum_{i=1}^{n} \mathbb{1}\{X_i \in A \text{ or } X'_i \in A\} \\ &= \sum_{i=1}^{n} \mathbb{1}\{(X_i, X'_i) \in \tilde{\mathbb{A}}\} \text{ where } \tilde{\mathbb{A}} = (\mathbb{A} \times \mathcal{X}) \cup (\mathcal{X} \times \mathbb{A}).\end{aligned}$$

The law of K is Binomial$(n, \tilde{p})$ with

$$\tilde{p} = (P \otimes P)(\tilde{\mathbb{A}}) = p(2 - p).$$

Let $(Y_1, Y'_1), \dots, (Y_n, Y'_n)$ be an independent random sample from the conditional distribution of (X_1, X'_1) given that $(X_1, X'_1) \in \tilde{\mathbb{A}}$. For a fixed $i = 1, \dots, n$, the variables Y_i and Y'_i are not independent, however they are exchangeable: $(Y_i, Y'_i) \overset{d}{=} (Y'_i, Y_i)$. Further, let $\sigma_1, \dots, \sigma_n$ be an independent sample of Rademacher variables, independent of $(Y_1, Y'_1), \dots, (Y_n, Y'_n)$. Working conditionally to $K = \kappa$ we may write

$$\begin{aligned}&\mathbb{P}\left(\sup_{A \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bigl(\mathbb{1}\{X_i \in A\} - \mathbb{1}\{X'_i \in A\}\bigr) \geq t \mid K = \kappa\right) \\ &= \mathbb{P}\left(\sup_{A \in \mathcal{G}} \frac{1}{\kappa} \sum_{i=1}^{\kappa} \sigma_i \bigl(\mathbb{1}\{Y_i \in A\} - \mathbb{1}\{Y'_i \in A\}\bigr) \geq \frac{n}{\kappa} t\right).\end{aligned}$$

The remaining of the proof follows the traditional argument: Fix κ pairs of points $(y_1, y'_1), \dots, (y_\kappa, y'_\kappa) \in \mathcal{X} \times \mathcal{X}$. The number of different vectors $\bigl(\mathbb{1}\{y_i \in A\} - \mathbb{1}\{y'_i \in A\}\bigr)_{i=1}^{k}$ that can arise as A ranges over $\mathcal{G}$ is bounded by $\mathcal{S}_{\mathcal{G}}(2\kappa)$. To each such vector, apply Hoeffding's inequality. Since $\mathbb{1}\{y_i \in A\} - \mathbb{1}\{y'_i \in A\} \in \{-1, 0, 1\} \subset [-1, 1]$, we find, for $u \geq 0$,

$$\mathbb{P}\left[\sup_{A \in \mathcal{G}} \frac{1}{\kappa} \sum_{i=1}^{\kappa} \sigma_i \bigl(\mathbb{1}\{y_i \in A\} - \mathbb{1}\{y'_i \in A\}\bigr) \geq u\right] \leq \mathcal{S}_{\mathcal{G}}(2\kappa) \exp\left(-\frac{\kappa u^2}{2}\right).$$

Setting $u = nt/\kappa$, we obtain for $t > 0$,

$$\mathbb{P}\left(\sup_{A\in\mathcal{G}}\big(P_n(A) - P'_n(A)\big) \geq t \mid K = \kappa\right) \leq \mathcal{S}_{\mathcal{G}}(2\kappa)\exp\left(-\frac{n^2t^2}{2\kappa}\right)$$

□

Lemma A.4 (Concentration after symmetrization and deconditioning)**.** *With the notations from Lemma A.3, For $\delta \in (0,1)$ and if $np \geq 8\ln(2/\delta)$, we have, with probability at least $1-\delta$,*

$$\sup_{A\in\mathcal{G}}\big(P_n(A) - P'_n(A)\big) \leq 2\sqrt{\frac{2p}{n}\big(\ln(2/\delta) + \ln\mathcal{S}_{\mathcal{G}}(8np)\big)}. \tag{A.1}$$

Remark A.5. The bound (A.1) can be improved to

$$2\sqrt{\frac{(1+s(\delta))p}{n}\big(\ln(2/\delta) + \ln\mathcal{S}_{\mathcal{G}}(4np(1+s(\delta)))\big)}$$

with $s(\delta) = \mathrm{O}\big(\sqrt{\ln(2/\delta)/(np)}\big)$, see (A.5).

Proof. We integrate with respect to K the upper bound from Lemma A.3 by considering whether K is less than $2np(1+s)$ or not, with $s > 0$ to be determined. Since the integrand is a non-decreasing function of κ, we find

$$\begin{aligned}\mathbb{P}&\left[\sup_{A\in\mathcal{G}}\big(P_n(A) - P'_n(A)\big) \geq t\right]\\ &\leq \mathbb{P}\left[K > 2np(1+s)\right] + \mathcal{S}_{\mathcal{G}}(4np(1+s))\exp\left(-\frac{nt^2}{4p(1+s)}\right).\end{aligned} \tag{A.2}$$

First we choose $s > 0$ such that the first term on the right-hand side of (A.2) is bounded by $\delta/2$. Second we choose t such that the second term on the right-hand side is bounded by $\delta/2$ too.

1. We use the standard Bernstein inequality for a binomial (n,p) random variable: for all $t \geq 0$,

$$\mathbb{P}\left(K - np > t\right) \leq \exp\left(-\frac{t^2}{2np(1-p) + \frac{2}{3}t}\right), \tag{A.3}$$

   equivalently, inverting the upper bound *w.r.t.* t, with probability greater than $1-\delta$,

$$K \leq np + \frac{2}{3}\ln(1/\delta) + \sqrt{2np(1-p)\ln(1/\delta)}. \tag{A.4}$$

Since K is Binomial$(n, \tilde{p})$ and $\tilde{p} = p(2-p) \leq 2p$, by Bernstein's inequality (A.4), with probability greater than $1 - \delta/2$, we have

$$\begin{aligned} K &\leq n\tilde{p} + \frac{2}{3}\ln(2/\delta) + \sqrt{2n\tilde{p}(1-\tilde{p})\ln(2/\delta)} \\ &\leq 2np + 2\sqrt{np\ln(2/\delta)} + \frac{2}{3}\ln(2/\delta) \\ &= 2np\left(1 + 2\sqrt{\frac{\ln(2/\delta)}{np}} + \frac{2}{3np}\ln(2/\delta)\right) \end{aligned}$$

Setting

$$s(\delta) = 2\sqrt{\frac{\ln(2/\delta)}{np}} + \frac{2}{3np}\ln(2/\delta). \tag{A.5}$$

The first term on the right-hand side of (A.2) is bounded by $\delta/2$.

2. We determine $t(\delta) > 0$ such that the second term on the right-hand side of (A.2) with $s = s(\delta)$ is equal to $\delta/2$. We find

$$t(\delta) = 2\sqrt{\frac{p(1+s(\delta))}{n}\big(\ln(2/\delta) + \ln \mathcal{S}_{\mathcal{G}}(4np(1+s(\delta)))\big)}.$$

If $np \geq 8\ln(2/\delta)$, then

$$s(\delta) \leq \tfrac{1}{\sqrt{2}} + \tfrac{1}{12} < 1$$

and thus

$$t(\delta) \leq 2\sqrt{\frac{2p}{n}\big(\ln(2/\delta) + \ln \mathcal{S}_{\mathcal{G}}(8np)\big)} \tag{A.6}$$

In view of (A.2) and our choice of $s(\delta)$ and $t(\delta)$, we find that, with probability $1-\delta$, the supremum of interest is bounded by the right-hand side of (A.6), as required. □

Bibliography

Aggarwal, C. and Yu, P. S. (2001). Outlier Detection for High Dimensional Data. In *SIGMOD REC*, volume 30, pages 37–46.

Aghbalou, A., Portier, F., Sabourin, A., and Zhou, C. (2021). Tail inverse regression for dimension reduction with extreme response. *arXiv preprint arXiv:2108.01432*.

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (2005). Automatic subspace clustering of high dimensional data. *DMKD*, 11(1):5–33.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148.

Arlot, S., Celisse, A., et al. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

Athreya, A., Fishkind, D., Tang, M., Priebe, C., Park, Y., Vogelstein, J., Levin, K., Lyzinski, V., and Qin, Y. (2017). Statistical Inference on Random Dot Product Graphs: A Survey. *Journal of Machine Learning Research*, 18(1):8393–8484.

Baayen, R. H. (2002). *Word frequency distributions*, volume 18. Springer Science & Business Media.

Babbar, R., Metzig, C., Partalas, I., Gaussier, E., and Amini, M.-R. (2014). On power law distributions in large-scale taxonomies. *ACM SIGKDD Explorations Newsletter*, 16(1):47–56.

Bacro, J.-N. and Toulemonde, G. (2013). Measuring and modelling multivariate and spatial dependence of extremes. *Journal de la Société Française de Statistique*, 154(2):139–155.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data*. Wiley New York.

Basrak, B. and Segers, J. (2009). Regularly varying multivariate time series. *Stochastic processes and their applications*, 119(4):1055–1080.

Beer, G. (1993). *Topologies on Closed and Closed Convex Sets*, volume 268 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht.

Beirlant, J., Vynckier, P., and Teugels, J. L. (1996). Tail index estimation, pareto quantile plots regression diagnostics. *Journal of the American Statistical Association*, 91(436):1659–1667.

Bertail, P., Haefke, C., Politis, D. N., and White, H. (2004). Subsampling the distribution of diverging statistics with applications to finance. *Journal of Econometrics*, 120(2):295–326.

Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3):259–294.

Boldi, M.-O. and Davison, A. (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):217–229.

Boucheron, S. and Thomas, M. (2012). Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17.

Boucheron, S. and Thomas, M. (2015). Tail index estimation, concentration and adaptivity. *Electronic Journal of Statistics*, 9(2):2751–2792.

Bousquet, O., Boucheron, S., and Lugosi, G. (2003). Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer.

Breunig, M., Kriegel, H., Ng, R., and Sander, J. (2000). LOF: identifying density-based local outliers. In *SIGMOD REC*, volume 29, pages 93–104.

Bücher, A., Dette, H., et al. (2013). Multiplier bootstrap of tail copulas with applications. *Bernoulli*, 19(5A):1655–1687.

Cai, J., Einmahl, J., and De Haan, L. (2011). Estimation of extreme risk regions under multivariate regular variation. *The Annals of Statistics*, pages 1803–1826.

Carpentier, A. and Kim, A. K. (2015). Adaptive and minimax optimal estimation of the tail coefficient. *Statistica Sinica*, pages 1133–1144.

Chalapathy, R. and Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*

Chautru, E. et al. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic journal of statistics*, 9(1):383–418.

Chiapino, M., Clémençon, S., Feuillard, V., and Sabourin, A. (2019a). A multivariate extreme value theory approach to anomaly clustering and visualization. *Computational Statistics*, pages 1–22.

Chiapino, M. and Sabourin, A. (2016). 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.

Chiapino, M., Sabourin, A., and Segers, J. (2019b). Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22(2):193–222.

Church, K. W. and Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, 1(2):163–190.

Clémençon, S., Jalalzai, H., Sabourin, A., and Segers, J. (2021). Concentration bounds for the empirical angular measure with statistical learning applications. *arXiv preprint arXiv:2104.03966*.

Clifton, D., Tarassenko, L., McGrogan, N., King, D., King, S., and Anuzis, P. (2008). Bayesian extreme value statistics for novelty detection in gas-turbine engines. In *AEROSP CONF PROC*, pages 1–11.

Clifton, D. A., Hugueny, S., and Tarassenko, L. (2011). Novelty detection with multivariate extreme value statistics. *J Signal Process Syst.*, 65:371–389.

Clinchant, S. and Gaussier, E. (2010). Information-based models for ad hoc ir. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 234–241.

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Coles, S. and Tawn, J. (1991). Modeling extreme multivariate events. *JR Statist. Soc. B*, 53:377–392.

Coles, S. G. (1993). Regional modelling of extreme storms via max-stable processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):797–816.

Constantinou, P. and Dawid, A. P. (2017). Extended conditional independence and applications in causal inference. *The Annals of Statistics*, 45(6):2618–2653.

Cook, R. D., Li, B., et al. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100(470):410–428.

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332.

Cooley, D., Davis, R., and Naveau, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *JMVA*, 101:2103–2117.

Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604.

Dai, W. and Genton, M. G. (2018). Multivariate functional data visualization and outlier detection. *Journal of Computational and Graphical Statistics*, 27(4):923–934.

Danielsson, J., de Haan, L., Peng, L., and de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate analysis*, 76(2):226–248.

Das, B., Mitra, A., and Resnick, S. (2013). Living on the multidimensional edge: seeking hidden risks using regular variation. *Advances in Applied Probability*, 45(1):139–163.

Davison, A. C. and Gholamrezaee, M. M. (2012). Geostatistics of extremes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 468(2138):581–608.

de Haan, L. (1984). A spectral representation for max-stable processes. *The Annals of Probability*, 12(4):1194–1204.

de Haan, L. and Ferreira, A. (2006). *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York. An introduction.

de Haan, L. and Resnick, S. (1977). Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 40(4):317–337.

De Haan, L. and Resnick, S. (1987). On regular variation of probability densities. *Stochastic processes and their applications*, 25:83–93.

De Haan, L. and Zhou, C. (2011). Extreme residual dependence for random vectors and processes. *Advances in Applied Probability*, 43(01):217–242.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Applications of mathematics : stochastic modelling and applied probability. U.S. Government Printing Office.

Draisma, G., Drees, H., Ferreira, A., and de Haan, L. (2004). Bivariate tail estimation: dependence in asymptotic independence. *Bernoulli*, pages 251–280.

Drees, H. and Huang, X. (1998). Best attainable rates of convergence for estimators of the stable tail dependence function. *J. Multivar. Anal.*, 64(1):25–47.

Drees, H. and Sabourin, A. (2021). Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, 15(1):908–943.

Eastoe, E. F. and Tawn, J. A. (2012). Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. *Biometrika*, 99(1).

Einmahl, J. H. (1997). Poisson and Gaussian approximation of weighted local empirical processes. *Stochastic Processes and Their Applications*, 70(1):31–58.

Einmahl, J. H., de Haan, L., and Piterbarg, V. I. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *Ann. Stat.*, 29:1401–1423.

Einmahl, J. H. and Mason, D. M. (1992). Generalized quantile processes. *The Annals of Statistics*, pages 1062–1078.

Einmahl, J. H. J., de Haan, L., and Li, D. (2006). Weighted approximations of tail copula processes with application to testing the bivariate extreme value condition. *Ann. Statist.*, 34(4):1987–2014.

Einmahl, J. H. J., Krajina, A., and Segers, J. (2012). An m-estimator for tail dependence in arbitrary dimensions. *Ann. Statist.*, 40(3):1764–1793.

Einmahl, J. H. J. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *Ann. Stat.*, 37:2953–2989.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):871–932.

Engelke, S. and Ivanovs, J. (2020). Sparse structures for multivariate extremes. *arXiv preprint arXiv:2004.12182*.

Eskin, E. (2000). Anomaly detection over noisy data using learned probability distributions. In *Proc. ICML*, pages 255–262.

Finkenstadt, B. and Rootzén, H. (2003). *Extreme values in finance, telecommunications, and the environment*. CRC Press.

Fomichov, V. and Ivanovs, J. (2020). Detection of groups of concomitant extremes using clustering. *arXiv preprint arXiv:2010.12372*.

Forbes, F. and Wraith, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing*, 24(6):971–984.

Fougères, A.-L., Nolan, J. P., and Rootzén, H. (2009). Models for dependent extremes using stable mixtures. *Scandinavian Journal of Statistics*, 36:42–59.

Fruhwirth-Schnatter, S., Celeux, G., and Robert, C. (2018). *Handbook of Mixture Analysis*. Chapman & Hall, CRC.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99.

Gardes, L. (2018). Tail dimension reduction for extreme quantile estimation. *Extremes*, 21(1):57–95.

Goix, N., Sabourin, A., and Clémençon, S. (2015). Learning the dependence structure of rare events: a non-asymptotic study. In *Conference on Learning Theory*, pages 843–860.

Goix, N., Sabourin, A., and Clémençon, S. (2016). Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial Intelligence and Statistics*, pages 75–83.

Goix, N., Sabourin, A., and Clémençon, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis*, 161:12–31.

Gomes, M. I., Caeiro, F., Henriques-Rodrigues, L., and Manjunath, B. (2016). Bootstrap methods in statistics of extremes. *Handbook of Extreme Value Theory and Its Applications to Finance and Insurance. Handbook Series in Financial Engineering and Econometrics (Ruey Tsay Adv. Ed.). John Wiley and Sons*, pages 117–138.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Gorban, A., Kégl, B., C. Wunsch, D., and Zinovyev, A. (2008). *Principal Manifolds for Data Visualisation and Dimension Reduction*. LNCSE 58. Springer.

Gunopulos, D., Khardon, R., Mannila, H., Saluja, S., Toivonen, H., and Sharma, R. S. (2003). Discovering all most specific sentences. *ACM Trans. Database Syst.*, 28(2):140–174.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hu, Y. and Shi, L. (2015). Visualizing large graphs. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(2):115–136.

Huang, X. (1992). Statistics of bivariate extreme values.

Hult, H. and Lindskog, F. (2006). Regular variation for measures on metric spaces. *Publications de l'Institut Mathematique*, 80(94):121–140.

Huser, R. and Davison, A. C. (2014). Space–time modelling of extreme events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):439–461.

Jalalzai, H., Clémençon, S., and Sabourin, A. (2018). On binary classification in extreme regions. In *Advances in Neural Information Processing Systems*, pages 3092–3100.

Jalalzai, H., Colombo, P., Clavel, C., Gaussier, E., Varni, G., Vignon, E., and Sabourin, A. (2020). Heavy-tailed representations, text polarity classification &amp; data augmentation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4295–4307. Curran Associates, Inc.

Janßen, A. and Wan, P. (2020). k-means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211–1233.

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Kokoszka, P., Stoev, S., and Xiong, Q. (2018). Principal components analysis of regularly varying functions. *arXiv preprint arXiv:1812.03108*.

Kokoszka, P. and Xiong, Q. (2018). Extremes of projections of functional time series on data–driven basis systems. *Extremes*, 21(2):177–204.

Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167.

Koltchinskii, V. and Lounici, K. (2017). New asymptotic results in principal component analysis. *Sankhya A*, 79(2):254–297.

Kriegel, H., Kröger, P., Schubert, E., and Zimek, A. (2008). A general framework for increasing the robustness of pca-based correlation clustering algorithms. In Ludäscher, B. and Mamoulis, N., editors, *Scientific and Statistical Database Management*, pages 418–435, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kyselỳ, J. (2008). A cautionary note on the use of nonparametric bootstrap for estimating uncertainties in extreme-value models. *Journal of Applied Meteorology and Climatology*, 47(12):3236–3251.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). Extremes and related properties of random sequences and processes. *Springer Series in Statistics*.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Lee, H. and Roberts, S. (2008). On-line novelty detection using the Kalman filter and extreme value theory. In *ICPR*, pages 1–4.

Lhaut, S., Sabourin, A., and Segers, J. (2021). Uniform concentration bounds for frequencies of rare events. *arXiv preprint arXiv:2110.05826*.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Liu, F., Ting, K., and Zhou, Z. (2008). Isolation Forest. In *ICDM*, pages 413–422.

Liu, J., Shang, J., Wang, C., Ren, X., and Han, J. (2015). Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1729–1744. ACM.

Lugosi, G. (2002). Pattern classification and learning theory. In *Principles of nonparametric learning*, pages 1–56. Springer.

Madsen, R. E., Kauchak, D., and Elkan, C. (2005). Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning*, pages 545–552.

Malhotra, P., Vig, L., Shroff, G., and Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series. In *Proceedings*, volume 89, pages 89–94. Presses universitaires de Louvain.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication theory*, 84:486–502.

McAuley, J. and Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM.

McDiarmid, C. (1998). Concentration. In Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., and Reed, B., editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, volume 16 of *Algorithms and Combinatorics*, pages 195–248. Springer Berlin Heidelberg.

McFadden, D. (1981). Econometric models of probabilistic choice. In Manski, C. and McFadden, D., editors, *Structural Analysis of Discrete Data with Econometric Applications*, pages 198–272. MIT Press, Cambridge, MA.

McFadden, D. (1989). Econometric modeling of locational behavior. *Annals of Operations Research*, 18:3–16.

Meyer, N. and Wintenberger, O. (2019). Sparse regular variation. *arXiv preprint arXiv:1907.00686*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Molchanov, I. (2005). *Theory of Random Sets*. Probability and its Applications (New York). Springer-Verlag London, Ltd., London.

Müller, P. and Quintana, F. (2004). Nonparametric bayesian data analysis. *Statistical science*, pages 95–110.

Naik, G., editor (2017). *Advances in Principal Component Analysis*. Research and Development. Springer.

Norberg, T. (1987). Semicontinuous processes in multi-dimensional extreme value theory. *Stochastic Processes and Their Applications*, 25:27–55.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dufour, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, L. (1999). Estimation of the coefficient of tail dependence in bivariate extremes. *Statistics & Probability Letters*, 43(4):399–409.

Peng, L. and Qi, Y. (2008). Bootstrap approximation of tail dependence function. *Journal of Multivariate Analysis*, 99(8):1807–1824.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69(1):1–24.

Punzo, A. and Tortora, C. (2018). Multiple scaled contaminated normal distribution and its application in clustering. *arXiv preprint arXiv:1810.08918*.

Qi, Y. (1997). Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics. *Acta Mathematicae Applicatae Sinica*, 13(2):167–175.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *unpublished manuscript*.

Ramos, A. and Ledford, A. (2009). A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society: Series B*, 71(1):219–241.

Ratner, A. J., Ehrenberg, H., Hussain, Z., Dunnmon, J., and Ré, C. (2017). Learning to compose domain-specific transformations for data augmentation. In *Advances in neural information processing systems*, pages 3236–3246.

Reiß, M. and Wahl, M. (2020). Nonasymptotic upper bounds for the reconstruction error of pca. *Annals of Statistics*, 48(2):1098–1123.

Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer Series in Operations Research and Financial Engineering.

Resnick, S. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.

Resnick, S. I. and Roy, R. (1991). Random usc functions, max-stable processes and continuous choice. *The Annals of Applied Probability*, pages 267–292.

Roberts, S. (1999). Novelty detection using extreme value statistics. *IEE P-VIS IMAGE SIGN*, 146:124–129.

Roberts, S. (2000). Extreme value statistics for novelty detection in biomedical data processing. *IEE P-SCI MEAS TECH*, 147:363–367.

Sabourin, A. and Naveau, P. (2014). Bayesian dirichlet mixture model for multivariate extremes: A re-parametrization. *Comput. Stat. Data Anal.*, 71:542–567.

Sabourin, A. and Segers, J. (2017). Marginal standardization of upper semicontinuous processes. with application to max-stable processes. *Journal of Applied Probability*, 54(3):773–796.

Salinetti, G. and Wets, R. J.-B. (1986). On the convergence in distribution of measurable multifunctions (random sets) normal integrands, stochastic processes and stochastic infima. *Mathematics of Operations Research*, 11(3):385–419.

Schaeffer, S. (2007). Graph clustering. *Computer Science Review*, 1(1):27 – 64.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, 5(1):33–44.

Schlather, M. and Tawn, J. A. (2002). Inequalities for the extremal coefficients of multivariate extreme value distributions. *Extremes*, 5(1):87–102.

Schlather, M. and Tawn, J. A. (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, 90(1):139–156.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001). Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13:1443–1471.

Scott, C. and Nowak, R. (2006a). Learning minimum volume sets. *JMLR*, 7:665–704.

Scott, C. and Nowak, R. D. (2006b). Minimax-optimal classification with dyadic decision trees. *IEEE transactions on information theory*, 52(4):1335–1353.

Shawe-Taylor, J., Williams, C. K., Cristianini, N., and Kandola, J. (2005). On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *Information Theory, IEEE Transactions on*, 51(7):2510–2522.

Shi, T. and Horvath, S. (2012). Unsupervised learning with random forest predictors. *J. Comp. Graph. Stat.*, 15.

Shyu, M., Chen, S., Sarinnapakorn, K., and Chang, L. (2003). A novel anomaly detection scheme based on principal component classifier. Technical report, DTIC Document.

Siffer, A., Fouque, P.-A., Termier, A., and Largouet, C. (2017). Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1067–1075.

Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231.

Smith, R. (1990). Max-stable processes and spatial extremes. Unpublished manuscript.

Smith, R. (2003). Statistics of extremes, with applications in environment, insurance and finance. *Extreme values in finance, telecommunications and the environment*, pages 1–78.

Staerman, G., Mozharovskyi, P., Clémençon, S., and d'Alché Buc, F. (2019). Functional isolation forest. In *Asian Conference on Machine Learning*, pages 332–347. PMLR.

Steinwart, I., Hush, D., and Scovel, C. (2005). A classification framework for anomaly detection. *JMLR*, 6:211–232.

Thomas, A., Clemencon, S., Gramfort, A., and Sabourin, A. (2017). Anomaly detection in extreme regions via empirical mv-sets on the sphere. In *AISTATS*, pages 1011–1019.

Tressou, J. (2008). Bayesian nonparametrics for heavy tailed distribution. application to food risk assessment. *Bayesian Analysis*, 3(2):367–391.

van der Vaart, A. W. (1998). *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Vapnik, V. N. and Chervonenkis, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer.

Vert, J.-P. and Vert, R. (2006). Consistency and convergence rates of one-class svms and related algorithms. *JMLR*, 6:828–835.

Vervaat, W. (1986). Stationary self-similar extremal processes and random semicontinuous functions. In *Dependence in probability and statistics (Oberwolfach, 1985)*, volume 11 of *Progr. Probab. Statist.*, pages 457–473. Birkhäuser Boston, Boston, MA.

Vervaat, W. (1988a). Narrow and vague convergence of set functions. *Statistics & Probability Letters*, 6(5):295–298.

Vervaat, W. (1988b). Random upper semicontinuous functions and extremal processes. *Department of Mathematical Statistics*, R 8801:1–43.

Vervaat, W. and Holwerda, H., editors (1997). *Probability and Lattices*, volume 110 of *CWI Tract*. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam.

Vignotto, E. and Engelke, S. (2020). Extreme value theory for anomaly detection– the gpd classifier. *Extremes*, 23(4):501–520.

Wadsworth, J. L., Tawn, J. A., Davison, A. C., and Elton, D. M. (2017). Modelling across extremal dependence classes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(1):149–175.

Wang, J. and Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit.*

Wei, J. and Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6383–6389.

Wei, L., Kumar, N., Lolla, V. N., Keogh, E. J., Lonardi, S., and Ratanamahatana, C. A. (2005). Assumption-free anomaly detection in time series. In *SSDBM*, volume 5, pages 237–242.

Xiong, Q. (2018). *Methods for extremes of functional data*. PhD thesis, Colorado State University.

Yu, X., Ren, X., Sun, Y., Gu, Q., Sturt, B., Khandelwal, U., Norick, B., and Han, J. (2014). Personalized entity recommendation: A heterogeneous information network approach. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 283–292. ACM.

Zenati, H., Romain, M., Foo, C.-S., Lecouat, B., and Chandrasekhar, V. (2018). Adversarially learned anomaly detection. In *2018 IEEE International conference on data mining (ICDM)*, pages 727–736. IEEE.

Zhou, C. and Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674.

Zhu, L.-P., Zhu, L.-X., and Feng, Z.-H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105(492):1455–1466.

Zwald, L. and Blanchard, G. (2006). On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems*, pages 1649–1656.